INVESTIGACIÓN DEL COMPORTAMIENTO

Cuarta edición

FRED N. KERLINGER (†)

HOWARD B. LEE
California State University

Traducción
Leticia Esther Pineda Ayala
Ignacio Mora Magaña
Traductores profesionales

Revisión técnica
Mtra. Cecilia Balbás Diez Barroso
Coordinadora del Área de Psicología Educativa
Escuela de Psicología
Universidad Anáhuac

Dra. Guadalupe Vadillo Bueno
Psicóloga y Master en Educación
Universidad de las Américas
Doctora en Educación
Universidad La Salle
Directora de Educación Continua y de Comunicación Humana
Universidad de las Américas

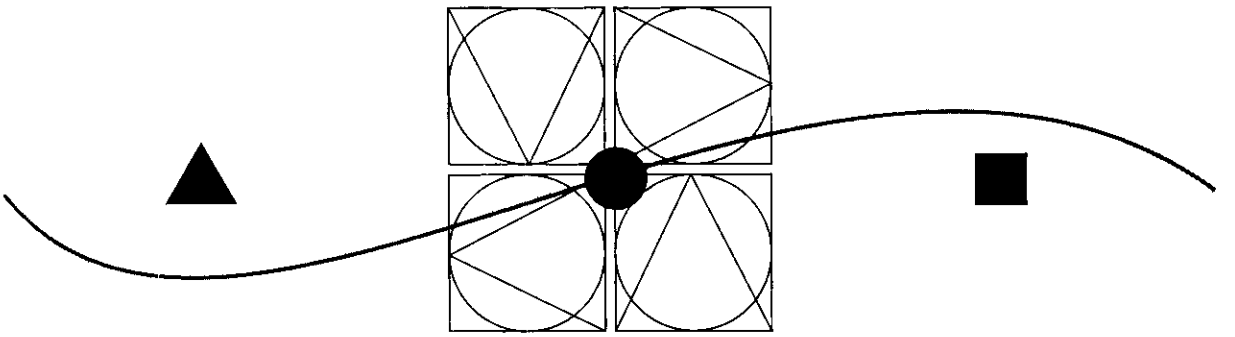
McGRAW-HILL

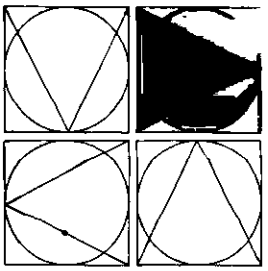
MÉXICO • BUENOS AIRES • CARACAS • GUATEMALA • LISBOA • MADRID
NUEVA YORK • SAN JUAN • SANTAFÉ DE BOGOTÁ • SANTIAGO • SÃO PAULO
AUCKLAND • LONDRES • MILÁN • MONTREAL • NUEVA DELHI
SAN FRANCISCO • SINGAPUR • ST. LOUIS • SIDNEY • TORONTO

Contenido

Parte Uno

El lenguaje y enfoque de la ciencia





Capítulo 1

La ciencia y el enfoque científico

- Ciencia y sentido común
- Cuatro métodos del conocimiento
- La ciencia y sus funciones
- Los objetivos de la ciencia, explicación científica y teoría
- La investigación científica: definición
- El enfoque científico
  - **Problema-obstáculo-idea**
  - **Hipótesis**
  - **Razonamiento-deducción**
  - **Observación-prueba-experimento**

Para entender cualquier actividad humana compleja es necesario comprender el lenguaje y el enfoque de quienes la realizan. Así sucede con la ciencia y la investigación científica. Se debe conocer y entender, al menos en parte, el lenguaje científico y el enfoque científico en la solución de problemas.

Una de las cosas que más confunde al estudiante de ciencia es la forma particular en que los científicos utilizan palabras ordinarias y, para complicar el asunto, inventan incluso nuevas palabras. Hay buenas razones para este uso tan especializado del lenguaje que serán evidentes más adelante. Por ahora basta decir que es necesario entender y aprender el lenguaje que usan los científicos sociales. Cuando los investigadores hablan de sus variables dependientes e independientes, se debe saber a qué se refieren. Cuando dicen que han aleatorizado sus procedimientos experimentales no sólo es necesario saber a qué se refieren, sino entender por qué lo hacen.

De igual forma, la manera en que el científico se aproxima a los problemas debe ser entendido con claridad. No porque su enfoque sea muy diferente al de cualquier persona. Por supuesto que sí *es* distinto, pero no es extraño ni esotérico. Por el contrario, cuando se

comprende la labor del científico parece natural y casi inevitable. De hecho, es probable que nos preguntemos por qué una gran parte del pensamiento humano y de la solución de problemas no están tan conscientemente estructurados de la misma manera.

El propósito de los capítulos 1 y 2 de este libro es ayudar al estudiante a aprender y entender tanto el lenguaje como el enfoque de la ciencia y de la investigación. En estos capítulos se estudiarán muchos de los constructos básicos del investigador social, conductual y educativo. En algunos casos no será posible dar definiciones completas y satisfactorias debido a la falta de antecedentes en este punto tan temprano del desarrollo del lector. En tales casos se intentará formular y usar un primer enfoque razonablemente preciso, para de ahí progresar a definiciones más satisfactorias. Comencemos nuestro estudio considerando cómo aborda el científico los problemas y cómo este enfoque difiere de lo que puede llamarse un enfoque por sentido común.

Ciencia y sentido común

Al inicio del siglo XX, Whitehead (1911/1992, p. 157) señaló que en el pensamiento creativo el sentido común es un mal instructor. "Su único criterio para juzgar es que las ideas nuevas deben parecerse a las viejas." Tiene razón: el sentido común puede a menudo ser un mal consejero para evaluar el conocimiento. ¿Pero, en qué se parecen y en qué difieren la ciencia y el sentido común? Desde un punto de vista ambos se asemejan: sus defensores dirían que la ciencia es una extensión sistemática y controlada del sentido común. James Bryant Conant (1951) establece que el sentido común es una serie de conceptos y esquemas conceptuales[1] satisfactorios para los usos prácticos de la humanidad. Sin embargo, estos conceptos y esquemas conceptuales pueden ser engañosos en la ciencia moderna —en particular en psicología y educación—. Muchos educadores del siglo XIX, por sentido común usaban el castigo como herramienta básica de la pedagogía. Sin embargo, a mediados del siglo XX la evidencia mostró que esta perspectiva de la motivación basada en el sentido común podía ser bastante errónea. La recompensa parece ser más efectiva que el castigo para apoyar el aprendizaje. Sin embargo, recientes hallazgos sugieren que diferentes formas de castigo son útiles en el aprendizaje en el salón de clases (Marlow *et al.*, 1997; Tingstrom *et al.*, 1997). La ciencia y el sentido común difieren marcadamente en cinco aspectos. Estos desacuerdos giran alrededor de las palabras *sistemático* y *controlado*.

Primero, los usos de los esquemas conceptuales y de las estructuras teóricas son notablemente diferentes. La persona común puede usar "teorías" y conceptos, pero en general lo hace de una forma vaga y a menudo acepta sin reparo explicaciones fantásticas de los fenómenos humanos y naturales. Por ejemplo, puede creerse que una enfermedad es un castigo por haber pecado (Klonoff y Landrine, 1994); o que se es alegre porque se tiene sobrepeso. Los científicos, por otro lado, construyen estructuras teóricas de forma sistemática, luego evalúan su consistencia interna, y someten algunos de sus aspectos a una

[1] Un *concepto* es una palabra que expresa una abstracción formada por la generalización de elementos particulares: "agresión" es un concepto, una abstracción que expresa un número de acciones particulares que tienen la característica común de dañar personas u objetos. Un *esquema conceptual* es un conjunto de conceptos interrelacionados por proposiciones hipotéticas y teóricas. Un *constructo* es un concepto con el significado adicional de haber sido creado o adaptado para propósitos científicos especiales. "Masa", "energía", "hostilidad", "introversión" y "rendimiento" son constructos que pueden ser llamados con más precisión "tipos construidos"; o "clases construidas", las clases o grupos de objetos o eventos se agrupan porque poseen una característica común definida por el científico. En un capítulo posterior se definirá el término "variable". Por ahora es suficiente con saber que representa un símbolo o nombre de una característica que puede adoptar diferentes valores numéricos.

prueba empírica. Además, se percatan de que los conceptos que emplean son términos acuñados por el hombre que pueden o no mostrar una relación estrecha con la realidad.

Segundo, los científicos prueban sus teorías e hipótesis de forma sistemática y empírica. Quienes no son científicos también prueban "hipótesis", pero lo hacen de manera selectiva: con frecuencia "seleccionan" evidencia sólo porque es consistente con las hipótesis. Veamos un estereotipo: los asiáticos tienen vocación científica y matemática. Si la gente lo cree podrán "verificar" con facilidad esta creencia al notar que muchos asiáticos son ingenieros y científicos (véase Tang, 1993). No se perciben las excepciones al estereotipo: los asiáticos que no son científicos o no son matemáticos. Los científicos sociales y conductuales identifican estas "tendencias de selección" como fenómeno psicológico común, y se cuidan mucho de no contaminar su investigación con sus propias preconcepciones o predilecciones y con el apoyo selectivo de hipótesis. Por principio, no se sienten satisfechos con realizar una exploración somera de las relaciones; los científicos deben probar estas relaciones en el laboratorio o en el campo. Ellos no se contentan por ejemplo, con las supuestas relaciones entre métodos de enseñanza y rendimiento, entre inteligencia y creatividad, entre valores y decisiones administrativas. Insisten en probar estas relaciones de manera sistemática, controlada y empírica.

Una tercera diferencia yace en la noción de control. En la investigación científica control significa diversas cosas. Por ahora considere que el científico trata sistemáticamente de descartar las variables que son posibles "causas" de los efectos bajo estudio, de otras variables que se ha hipotetizado son las "causas". La gente común rara vez se preocupa por controlar sus explicaciones de los fenómenos observados sistemáticamente. Por lo general, poco se esmeran por controlar las fuentes extrañas de influencia, y tienden a aceptar más aquellas explicaciones que concuerdan con sus preconcepciones y sesgos. Si creen que los barrios bajos producen delincuencia, tienden a hacer caso omiso de la delincuencia en otras zonas. Los científicos, por otro lado, buscan y "controlan" la incidencia de la delincuencia en diferentes tipos de barrios. La diferencia, por supuesto, es profunda.

Otra distinción entre ciencia y sentido común quizás no es tan clara. Antes se dijo que el científico está preocupado de manera constante por las relaciones entre fenómenos. La persona común también, pero usa su sentido común para explicar los fenómenos. El científico, sin embargo, persigue las relaciones de forma sistemática y concienzuda. La manera en que la persona común pretende conocer estas relaciones es relajada, no sistemática, y sin control: por ejemplo, con frecuencia observa la ocurrencia fortuita de dos fenómenos y los liga de inmediato de forma indisoluble como causa y efecto.

Tomemos la relación probada en un estudio clásico realizado hace muchos años por Hurlock (1925). Usando terminología más reciente, esta relación se puede expresar como: el reforzamiento positivo (recompensa) produce un mayor incremento de aprendizaje que el castigo. La relación se encuentra entre el reforzamiento (o recompensa y castigo) y el aprendizaje. Los educadores y padres de familia del siglo XIX con frecuencia suponían que el castigo era el agente más efectivo en el aprendizaje. Los educadores y padres de hoy asumen constantemente que el reforzamiento positivo (recompensa) es más efectivo. En ambos casos podemos decir que el punto de vista se basa sólo en el "sentido común". Es obvio, podrían decir, que si se premia o castiga a un niño aprenderá mejor. Por otro lado, el científico, ya sea que defienda en lo personal uno, otro o ninguno de estos puntos de vista, probablemente insistirá en una prueba sistemática y controlada de ambas relaciones (y de otras), como lo hizo Hurlock. Al usar el método científico Hurlock encontró que los incentivos estaban estrechamente relacionados con el rendimiento en aritmética. El grupo que fue elogiado obtuvo mayores puntuaciones que los que fueron reprobados o ignorados.

Una última diferencia entre el sentido común y la ciencia estriba en las diferentes explicaciones de los fenómenos observados. Cuando el científico intenta explicar las rela-

ciones entre los fenómenos observados descarta cuidadosamente lo que se ha llamado "explicaciones metafísicas". Una explicación metafísica es simplemente una proposición que no puede ser probada. Decir por ejemplo que la gente es pobre y padece hambre porque Dios así lo dispuso, o decir que es malo ser autoritario, es hablar metafísicamente.

Ninguna de estas proposiciones se puede probar; por lo tanto, son metafísicas y no son del interés de la ciencia. Esto no significa que los científicos necesariamente desprecien tales afirmaciones, digan que no son ciertas o mantengan que carecen de sentido. Sólo implica que *como científicos* no les interesan. En pocas palabras, la ciencia está involucrada con asuntos que pueden observarse y probarse. Si las propuestas o preguntas no implican esta observación y prueba públicas, no constituyen propuestas o preguntas científicas.

Cuatro métodos del conocimiento

Charles Sanders Peirce, como indica Buchler (1955), señaló cuatro formas generales de conocer o, como lo indicó, de establecer creencias. En la siguiente discusión los autores se tomaron algunas libertades con la propuesta original de Peirce en un intento de aclarar las ideas y hacerlas más apropiadas a esta presentación. El primero es el *método de la tenacidad*. En él la gente sostiene firmemente la verdad, la cual asumen como cierta debido a su apego a ella y a que siempre la han considerado como verdadera y real. La frecuente repetición de tales "verdades" parece aumentar su validez. A menudo la gente se aferra a sus creencias aun frente a hechos que claramente están en conflicto con ellas. Además infieren "nuevo" conocimiento a partir de proposiciones que pueden ser falsas.

Un segundo método de conocimiento o de establecer creencias es el *método de la autoridad*, o de creencias establecidas. Si la Biblia lo dice así es. Si un notable físico dice que hay un Dios, lo hay. Si una idea cuenta con el peso de la tradición y la sanción pública para apoyarla, entonces así. Peirce señaló que este método es superior al de la tenacidad porque es posible lograr progreso humano, aunque de manera lenta. De hecho, la vida no podría funcionar sin el método de la autoridad. Dawes (1994) establece que, como individuos, no podemos saberlo todo. Los norteamericanos reconocen la autoridad de la Oficina de Administración de Drogas y Alimentos de Estados Unidos (FDA) para determinar que cuanto comen y beben es seguro. Dawes estableció que no existe la mente completamente abierta que cuestione toda autoridad. Debemos asumir una gran cantidad de hechos e información con base en la autoridad. Por lo tanto, no podemos concluir que el método de la autoridad sea defectuoso; lo es sólo bajo ciertas circunstancias.

El *método a priori* es la tercera forma de conocimiento o de establecer creencias. Graziano y Raulin (1993) lo llamaron el *método de la intuición*. Basa su superioridad en el supuesto de que las proposiciones aceptadas por el "a priorista" son por sí mismas evidentes. Observe que la proposición *a priori* "concuerda con la razón" y no necesariamente con la experiencia. La idea parece ser que la gente, a través de la comunicación y trato libres pueda alcanzar la verdad porque sus inclinaciones naturales tienden hacia ella. La dificultad con esta postura subyace en la expresión "concuerda con la razón". ¿La razón de quién? Imagine a dos sujetos honestos y bien intencionados que usando el proceso racional alcanzan diferentes conclusiones. ¿Quién está en lo correcto? ¿Es cuestión de gustos, como lo señaló Peirce? Si algo es patente para muchas personas —por ejemplo que el aprender materias difíciles entrena la mente y fortalece el carácter moral, que la educación estadounidense es inferior a la asiática y europea— ¿significa que en realidad lo sea? De acuerdo con el método *a priori* lo es, justamente porque se mantiene frente a la razón.

El cuarto método es el *método de la ciencia*. Dice Peirce:

> Para satisfacer nuestras dudas..., por lo tanto, es necesario encontrar un método por el que nuestras creencias se determinen no a partir de algo humano, sino por algo con permanencia externa, por algo que nuestro pensamiento no pudiera afectar ... El método debe ser tal que la conclusión última de todo hombre fuera la misma. Éste es el método de la ciencia. Su hipótesis fundamental es ésta: "hay cosas reales cuyas características son totalmente independientes de nuestra opinión acerca de ellas..." (Buchler, 1995, p. 18).

El enfoque científico tiene una característica de la que carecen los otros métodos de obtención del conocimiento: autocorrección. Hay puntos de verificación intrínsecos a lo largo de todo el camino del conocimiento científico. Estos controles están concebidos y se utilizan de manera tal que dirigen y verifican las actividades científicas y las conclusiones con el fin de obtener conocimiento del que se pueda depender. Incluso si una hipótesis parece sustentarse en un experimento, el científico probará hipótesis alternas posibles que, si también reciben apoyo, pueden generar dudas sobre la primera hipótesis. Los científicos no aceptan declaraciones como verdaderas aunque en principio la evidencia pueda parecer prometedora. Insisten en probarlas. También subrayan la necesidad de que cualquier procedimiento de prueba esté abierto al escrutinio *público*. Una interpretación del método científico es que no hay método científico específico. Más bien existe una variedad de métodos que el científico puede emplear y, de hecho, usa, pero probablemente pueda decirse que hay un solo enfoque científico.

Como dijo Peirce, los controles usados en la investigación científica están anclados tanto como es posible en la realidad, más allá de las creencias personales del científico, y de sus percepciones, sesgos, valores, actitudes y emociones. Tal vez la mejor palabra para expresar esto es *objetividad*. La objetividad constituye el acuerdo entre jueces "expertos" sobre lo observado o sobre lo que se hizo o hará en la investigación (véase Kerlinger, 1979, para un análisis de objetividad, su significado y su carácter controvertido). De acuerdo con Sampson (1991, p. 12) la objetividad "se refiere a aquellas declaraciones acerca del mundo que se pueden justificar y defender en el presente usando los estándares de argumento y prueba empleados en la comunidad a la que pertenecemos —por ejemplo, la de científicos". Pero, como se verá más adelante, el enfoque científico involucra más que estas dos declaraciones. El hecho es que alcanzamos un conocimiento del que podemos depender más debido a que la ciencia apela finalmente a la evidencia: las proposiciones se someten a la prueba empírica. Puede surgir otra objeción: la teoría, que los científicos usan y exaltan, proviene de gente, de los científicos mismos. Pero, como Polanyi (1958/1974 p. 4) señala, "una teoría es algo distinto a mí". Así, una teoría ayuda al científico a lograr una mayor objetividad. En pocas palabras, los científicos sistemática y conscientemente usan los aspectos autocorrectivos del enfoque científico.

La ciencia y sus funciones

¿Qué es la ciencia? Esta pregunta no es fácil de contestar; de hecho, no intentaremos presentar definición alguna de ciencia de manera directa. En cambio hablaremos de nociones y perspectivas de la ciencia para después intentar explicar sus funciones.

Ciencia es una palabra malinterpretada. Parece ser que hay tres estereotipos populares que dificultan el entendimiento de la actividad científica. Uno es el de bata blanca-estetoscopio-laboratorio. Se percibe a los científicos como individuos que trabajan con hechos en laboratorios; usan equipo complicado, hacen muchos experimentos y amontonan hechos con el propósito final de perfeccionar a la humanidad. Así, aunque sean exploradores

poco imaginativos en busca de hechos, se les redime por sus nobles motivos. Puede creérseles cuando, por ejemplo, dicen que tal o cuál dentífrico es bueno para usted, o que no debería fumar cigarrillos.

El segundo estereotipo de los científicos consiste en que son individuos brillantes que piensan, elaboran teorías complejas y pasan el tiempo en torres de marfil alejados del mundo y sus problemas. Son teóricos poco prácticos, aun cuando su pensamiento y teorías ocasionalmente tengan resultados de significación práctica, como la energía atómica.

El tercer estereotipo equipara erróneamente a la ciencia con la ingeniería y la tecnología: la construcción de puentes, el mejoramiento de automóviles y misiles, la automatización de la industria, la invención de máquinas para enseñar. El trabajo del científico, según este estereotipo, está dedicado a optimizar inventos y artefactos. Se concibe al científico como una clase de ingeniero altamente especializado que trabaja para hacer la vida más cómoda y eficiente.

Estos estereotipos limitan al estudiante para entender la ciencia, las actividades y el pensamiento del científico, y la investigación científica en general. En pocas palabras, hacen que la tarea del alumno sea más difícil de lo que podría resultar. Por ello, se deben eliminar para hacer espacio a nociones más apropiadas.

Hay dos amplias visiones de la ciencia: la estática y la dinámica. De acuerdo con Conant (1951, pp. 23-27) la *visión estática*, aquella que parece influir en la mayoría de la gente común y en los estudiantes, consiste en que la ciencia es una actividad que aporta al mundo información sistematizada. El trabajo del científico es descubrir nuevos hechos y agregarlos al cuerpo ya existente de información. Se concibe incluso a la ciencia como un conjunto de hechos. Desde esta perspectiva la ciencia es también una forma de explicar los fenómenos observados. El énfasis está entonces en el estado *actual del conocimiento* y en *la adición* que se le hace, así como en el conjunto de leyes, teorías, hipótesis y principios actuales.

La *visión dinámica*, por otro lado, considera a la ciencia más como una actividad que como aquello que realizan los científicos. El estado actual del conocimiento es importante, por supuesto, pero lo es en tanto que constituye la base para futuras teorías e investigaciones científicas. A esto se le ha llamado *visión heurística*. La palabra *heurística* significa "que sirve para descubrir o revelar", y ahora tiene la connotación de autodescubrimiento. Un método heurístico de enseñanza, por ejemplo, subraya la importancia de que los estudiantes descubran las cosas por sí mismos. La visión heurística en la ciencia se centra en la teoría y esquemas conceptuales interconectados que resultan fructíferos para investigaciones posteriores. Un énfasis heurístico implica un énfasis en el descubrimiento.

Es esta visión heurística de la ciencia lo que la distingue en buena medida de la ingeniería y la tecnología. Con base en esta corazonada heurística el científico da un salto riesgoso. Como dice Polanyi (1958/1974, p. 123), "Es el impulso por el cual ganamos pie en la otra orilla de la realidad. En tales casos el científico tiene que apostar, pieza a pieza, toda su vida profesional". Michel (1991, p. 23) agrega: "quien teme ser malinterpretado, y por esta razón estudia un método científico 'seguro' o 'cierto', no debe involucrarse en investigación científica alguna". La visión heurística también puede llamarse solución de problemas, pero el énfasis está en lo imaginativo y no en la solución rutinaria de problemas. La visión heurística en la ciencia enfatiza la resolución de problemas, más allá de los hechos y conjuntos de información. Éstos resultan importantes para el científico heurístico porque le ayudan a encaminarse hacia teorías, descubrimientos e investigaciones futuras.

Al evitar aún una definición directa de la ciencia —pero ciertamente implicándola— ahora se abordará la función de la ciencia. Aquí se tienen dos visiones distintas. La persona práctica, generalmente quien no es científico, considera a la ciencia como una disciplina o actividad encaminada a mejorar las cosas, a generar progresos. También algunos científi-

cos asumen esta postura. La función de la ciencia, desde esta perspectiva, consiste en hacer descubrimientos, conocer hechos y avanzar el conocimiento con el fin de mejorar las cosas. Las ramas de la ciencia que claramente pertenecen a este género reciben un apoyo amplio y fuerte, como el caso de la investigación médica y meteorológica. El criterio de utilidad práctica y "resultado" son relevantes en esta perspectiva, en especial en la investigación educativa (véase Kerlinger, 1977; Bruno, 1972).

Un punto de vista muy diferente sobre la función de la ciencia está bien expresado por Braithwaite (1953/1996, p. 1):

> La función de la ciencia... consiste en establecer leyes generales sobre el comportamiento de eventos empíricos u objetos en los que la ciencia en cuestión está interesada, para así permitirnos conectar nuestro conocimiento de eventos conocidos por separado y hacer predicciones confiables de eventos aún desconocidos.

La conexión entre esta perspectiva de la función de la ciencia y la dinámica-heurística antes discutida es obvia, excepto que se ha agregado un elemento muy importante: el establecimiento de leyes generales —o teoría si se quiere—. Si hemos de comprender la investigación conductual moderna junto con sus fortalezas y debilidades debemos explorar los elementos de la declaración de Braithwaite. Lo hacemos al considerar el fin de la ciencia, la explicación científica y el papel e importancia de la teoría.

Sampson (1991) analiza dos puntos de vista opuestos de la ciencia. Existe una perspectiva convencional o tradicional y una perspectiva sociohistórica. La convencional percibe a la ciencia como un espejo de la naturaleza o como una vitrina de cristal transparente que presenta la naturaleza sin sesgo ni distorsión. El objetivo en este caso es describir con el máximo grado de exactitud cómo es el mundo en realidad. Sampson establece que la ciencia constituye un árbitro objetivo. Su trabajo es "resolver los desacuerdos y distinguir qué es cierto y correcto, de aquello que no lo es". Cuando la visión convencional de la ciencia es incapaz de resolver la disputa, esto sólo significa que hay datos o información insuficientes para hacerlo. Los convencionalistas, sin embargo, consideran que sólo es cuestión de tiempo para que la verdad salte a la vista.

La visión sociohistórica concibe a la ciencia como una historia. Los científicos son narradores. La idea es que la realidad sólo puede ser descubierta a través de las historias que se cuentan acerca de ella. Este enfoque es diferente de la visión tradicional-convencional en tanto que no hay un árbitro neutral. Cada historia está sazonada por la orientación del narrador. Como resultado, no hay una historia verdadera única. La interpretación del autor sobre la presentación de Sampson que compara estos dos aspectos se muestra en la tabla 1.1.

Aun cuando Sampson proporciona estas dos visiones de la ciencia a la luz de la psicología social, su presentación es aplicable en todas las áreas de las ciencias del comportamiento.

Los objetivos de la ciencia, explicación científica y teoría

El objetivo básico de la ciencia es la teoría. Quizás, dicho de forma menos críptica, el fin básico de la ciencia es explicar los fenómenos naturales. Tales explicaciones se llaman "teorías". En lugar de tratar de explicar cada una de las conductas de los niños por separado, el psicólogo científico busca explicaciones generales que abarquen y conjunten muchas conductas diferentes. En vez de intentar explicar los métodos que usan los niños para resolver los problemas aritméticos, por ejemplo, el científico busca explicaciones genera-

TABLA 1.1 *Los dos puntos de vista de Sampson acerca de la ciencia y la psicología social*

	Tradicional (Cuantatativo)	No tradicional (Sociohistórico) (Cuantatativo)
Objetivo primario	Describir la realidad de las interacciones y funciones humanas y sociales.	Describir la variedad de experiencias y actividades humanas y sociales a través de la información histórica y social y de los papeles que desempeñan en la vida humana.
Posición filosófica	La realidad puede ser descubierta de forma independiente por observadores neutros. La realidad puede ser apreciada sin ocupar ninguna posición particular que genere sesgos.	La realidad puede ser descubierta sólo desde algún punto de vista: el observador, entonces, siempre está posicionado.
Enunciado metafórico	Se puede percibir a la ciencia como un espejo. Está diseñada para reflejar las cosas tal como son en realidad.	La ciencia se percibe como un contador de historias que proporciona versiones diferentes o personales de la realidad.
Consideraciones metodológicas	Los métodos se crean y utilizan para controlar o eliminar factores que debilitarían la habilidad del investigador para descubrir la verdadera forma de la realidad.	Amplios factores históricos y sociales moldean la comprensión que el investigador tiene de la realidad. Los métodos pueden generar un entendimiento más rico y profundo de la realidad con base en el encuentro de diferentes versiones que la gente utiliza para comprender sus vidas.

les de todos los tipos de solución de problemas. Esto podría llamarse una teoría general de solución de problemas.

Este análisis sobre el objetivo básico de la ciencia como teoría puede resultar extraño al estudiante al que quizá se le ha inculcado la idea de que las actividades humanas han de producir resultados prácticos. Si dijéramos que el objetivo de la ciencia es el progreso de la humanidad, la mayoría leería las palabras con rapidez y las aceptaría. Pero el objetivo *básico* de la ciencia *no* es el progreso de la humanidad: es la teoría. Por desgracia, este enunciado vasto y realmente complejo no es fácil de entender. Aún así, debemos tratar de comprenderlo porque es importante. Hay una explicación más amplia de este punto en el capítulo 16 de Kerlinger (1979).

Otros objetivos de la ciencia que se han mencionado son: la explicación, comprensión, predicción y el control. Sin embargo, si aceptamos la teoría como el fin supremo de la ciencia, la explicación y el entendimiento se convierten en subobjetivos del objetivo fundamental debido a la definición y naturaleza de la teoría: ***una teoría es un conjunto de constructos (conceptos) interrelacionados, definiciones y proposiciones que presentan una visión sistemática de los fenómenos al especificar las relaciones entre variables con el propósito de explicar y predecir los fenómenos.***

Esta definición indica tres cosas: 1) una teoría es un conjunto de proposiciones constituidas por constructos definidos e interrelacionados, 2) una teoría establece las interre-

laciones entre un conjunto de variables (constructos) y, al hacerlo, presenta una visión sistemática del fenómeno descrito por las variables, y 3) una teoría explica fenómenos al especificar qué variables están relacionadas con cuáles otras y de qué forma están relacionadas. De esta manera permiten al investigador hacer predicciones de ciertas variables a partir de otras. Uno podría, por ejemplo, contar con una teoría sobre el fracaso escolar. Nuestras variables podrían ser inteligencia, aptitudes numérica y verbal, ansiedad, clase social, estado nutricional y motivación de logro. El fenómeno a ser explicado es, por supuesto, el fracaso escolar, o más precisamente, el rendimiento escolar. El fracaso escolar puede concebirse como un extremo del continuo de rendimiento escolar, mientras en el otro estaría el éxito escolar. El fracaso escolar se explica a partir de las relaciones especificadas entre cada una de las siete variables y el fracaso escolar, o por la combinación de las siete variables y el fracaso escolar. El científico, al utilizar con éxito este conjunto de constructos puede "entender" el fracaso escolar, es capaz de "explicarlo" y, al menos en alguna medida, "predecirlo".

Resulta evidente que la explicación y predicción pueden ser incluidas en una teoría. La misma naturaleza de una teoría consiste en la explicación de los fenómenos observados. Tomemos, por ejemplo, la teoría del reforzamiento. Una proposición sencilla que se deriva de ella es: si una respuesta es premiada (reforzada) cuando ocurre, tenderá a repetirse. El primer psicólogo científico que formuló esta proposición lo hizo como una forma de explicar la ocurrencia repetida de respuestas que había observado. ¿Por qué ocurrieron y volvieron a presentarse con una regularidad confiable? Porque fueron recompensadas. Aunque esto constituye una explicación puede no resultar satisfactoria para mucha gente. Algunos pueden preguntar por qué el premio aumenta la probabilidad de que ocurra una respuesta. Una teoría completa contendría la explicación. Sin embargo, hoy en día no existe una respuesta realmente satisfactoria. Todo lo que podemos decir es que con un alto grado de probabilidad, el reforzamiento de una respuesta hace que sea más probable que ocurra una y otra vez. (véase Nisbett y Ross, 1980). En otras palabras, las proposiciones de una teoría, las declaraciones de relaciones, constituyen la explicación, en cuanto a la teoría de fenómenos naturales observados.

En cuanto a la predicción y el control puede decirse que los científicos no tienen que estar realmente involucrados en la explicación y la comprensión. Sólo la predicción y el control son necesarios. Quienes proponen esta postura dirían que el poder predictivo marca qué tan adecuada resulta. Si al utilizar una teoría somos capaces de predecir con éxito, entonces la teoría se confirma y eso es suficiente, no es necesario buscar más explicaciones subyacentes. En tanto podemos predecir con confiabilidad podemos controlar, ya que el control se deriva de la predicción.

La perspectiva de la predicción en la ciencia tiene validez. Pero por lo que a este libro compete, la predicción se considera un aspecto de la teoría. Por su propia naturaleza, una teoría predice; cuando de las proposiciones originales de una teoría deducimos otras más complejas, en esencia estamos "prediciendo". Cuando explicamos fenómenos observados siempre establecemos una relación entre, por ejemplo, la clase *A* y la clase *B*. La explicación científica reside en especificar las relaciones entre una clase de eventos empíricos y otra, bajo ciertas circunstancias. Decimos: si *A*, entonces *B*; en donde *A* y *B* se refieren a clases de objetos o eventos.[2] Pero esto constituye una predicción, la predicción de *A* acer-

[2] Enunciados de la forma: "si *p*, entonces *q*" se llaman *enunciados condicionales* en lógica y son la base del cuestionamiento científico. Ellos y los conceptos o variables que incluyen son el ingrediente central de las teorías. El fundamento lógico del cuestionamiento científico que subyace a gran parte del razonamiento de este libro está resumido en Kerlinger (1977).

ca de *B*. Así, una explicación teórica implica una predicción, lo que nos lleva de nuevo a la idea de que la teoría es el objetivo final de la ciencia. Todo lo demás se deriva de la teoría.

Nuestra intención no es desacreditar o denigrar la investigación que no está específica y conscientemente orientada a la teoría. Muchas investigaciones valiosas en ciencias sociales y educativas se interesan por el objetivo más constreñido de encontrar relaciones específicas; es decir, el solo hecho de descubrir una relación forma parte de la ciencia. Pero las relaciones más útiles y satisfactorias son, en última instancia, aquéllas que tienen el máximo grado de generalización, las que están ligadas a otras relaciones en una teoría.

El concepto de generalidad es importante. En tanto que son generales las teorías se aplican a una variedad de fenómenos y a mucha gente de diversos lugares. Una relación específica, por supuesto, tiene un espectro de aplicación más reducido. Si por ejemplo, uno encuentra que la ansiedad al responder pruebas está relacionada con el desempeño, por interesante e importante que sea este hallazgo, tiene menor aplicabilidad y es menos comprendido que el descubrimiento de una relación en una red de variables interrelacionadas que forman parte de una teoría. Por lo tanto, los objetivos de investigación modestos, limitados y específicos son buenos, pero los de la investigación teórica son mejores, entre otras razones porque son más generales y aplicables a un amplio margen de situaciones. Además, cuando existe tanto una teoría simple como una compleja, y ambas dan cuenta de los hechos de forma igualmente efectiva, se prefiere la explicación sencilla (Navaja de Occam).* De aquí que en la discusión sobre la posibilidad de generalizar, una buena teoría también es parsimoniosa. Sin embargo, una cantidad de teorías incorrectas sobre la enfermedad mental persiste a causa de este atributo: algunos aún creen que los individuos están poseídos por demonios. Tal explicación es simple si se compara con las médicas y/o psicológicas.

Las teorías son explicaciones tentativas. Se evalúa cada teoría empíricamente para determinar qué tan bien predice los nuevos hallazgos. Las teorías pueden usarse para guiar un plan de investigación al generar hipótesis susceptibles de ser probadas, y para organizar hechos obtenidos al probar estas hipótesis. Una buena teoría es aquella que no se ajusta a todas las observaciones. Uno debería ser capaz de encontrar una ocurrencia que la contradiga. La teoría de Blondlot de los rayos N es un ejemplo de una teoría pobre. Blondlot expresó que toda la materia emitía rayos N (Weber, 1973). Aunque se demostró más tarde que los rayos N no existían, Barber (1976) indicó que casi 100 artículos se publicaron sobre estos rayos en un año, en Francia. Blondlot incluso desarrolló equipo complicado para observar rayos N. Los científicos que declararon haber observado rayos N sólo fortalecieron la teoría y hallazgos de Blondlot. Sin embargo, cuando alguien dijo no haber visto los rayos N, Blondlot declaró que sus ojos no eran lo suficientemente sensibles, o que no había ajustado el instrumento de forma apropiada. Ningún resultado se consideró evidencia en contra de la teoría. En tiempos más recientes otra teoría falsa que tomó más de 75 años derrocar fue la relativa al origen de la úlcera péptica. En 1910 Schwartz (citado en Blaser, 1996) declaró haber establecido firmemente la causa de las úlceras: los ácidos gástricos. En años posteriores investigadores médicos dedicaron su tiempo y energía al tratamiento de las úlceras a través del desarrollo de medicamentos para neutralizar o bloquear los ácidos, los cuales nunca tuvieron mucho éxito y resultaban costosos. Sin embargo, en 1985 J. Robin Warren y Barry Marshall (citados en Blaser, 1996) descubrieron que el *helioc bacter pylori* era la causa real de las úlceras gástricas. Casi todos los casos de este

---

* N. del T. La *navaja de Ocamm* (Ocamm's razor) procede de William de Occam (1285-1349), filósofo inglés que formuló la máxima: *entia non sunt multiplicanda praeter necessitatem*, esto es: *las suposiciones que explican un fenómeno no deberán ser multiplicadas más allá de lo necesario*. Sir William Hamilton en 1853 la denominó *ley de la parsimonia*.

tipo de úlcera fueron tratados con éxito por medio de antibióticos y a un precio considerablemente más bajo. Durante 75 años ningún resultado se tomó como evidencia contra la teoría del ácido-estrés de las úlceras.

La investigación científica: definición

Es más fácil definir la investigación científica que la ciencia. Sin embargo, no sería sencillo lograr un acuerdo de científicos e investigadores en relación con una definición. Aun así intentaremos una: *la investigación científica es una investigación sistemática, controlada, empírica, amoral, pública y crítica de fenómenos naturales. Se guía por la teoría y las hipótesis sobre las presuntas relaciones entre esos fenómenos.* Esta definición requiere poca explicación en tanto que es una declaración condensada y formalizada de muchos aspectos que ya se han presentado o que abordaremos pronto. Sin embargo, es necesario enfatizar dos puntos.

Primero, cuando decimos que la investigación científica es sistemática y controlada queremos decir, de hecho, que la investigación es tan ordenada que los investigadores pueden tener una confianza crítica en los resultados. Como se verá más adelante, las observaciones de la investigación científica son estrictamente disciplinadas. Más aún, entre las muchas explicaciones alternativas de un fenómeno, todas menos una se rechazan de forma sistemática. Así, uno puede tener mayor confianza en que una relación sometida a prueba es tal como es, que si no se hubieran controlado las observaciones y desechado las posibilidades alternativas. En algunos casos es posible establecer una relación de causa-efecto.

Segundo, la investigación científica es empírica. Si el científico cree que algo se da de cierta forma debe demostrarlo de un modo u otro por medio de una prueba independiente externa. En otras palabras, las consideraciones subjetivas deben ser verificadas contra una realidad objetiva. Los científicos siempre deben presentar sus nociones ante el tribunal del cuestionamiento y prueba empíricos. Los científicos son sumamente críticos de sus propios resultados y de los de las investigaciones de los demás. Cada científico que redacta un informe de investigación cuenta con otros científicos que leen lo que él escribe a lo largo de todo el proceso. Aunque es fácil errar, exagerar, generalizar de más al redactar uno mismo su propio trabajo, no es fácil escapar al escrutinio de otros científicos que vigilan la tarea.

En ciencia existe la revisión de pares. Esto significa que otros con igual capacidad y conocimiento son llamados para evaluar el trabajo del científico antes de que sea publicado en revistas especializadas. En este punto hay tantos aspectos positivos como negativos. Es a través de esta revisión de pares que se han descubierto estudios fraudulentos. El ensayo escrito por R.W. Wood (1973) acerca de sus experiencias con el profesor francés Blondlot, sobre la inexistencia de los rayos *N*, aporta una clara demostración de estas revisiones, que resultan buenas para la ciencia y promueven la calidad en la investigación. El sistema, sin embargo, no es perfecto. Hay ocasiones en que la evaluación de pares se ha volcado contra la ciencia. Esto está documentado a través de la historia con personas como Kepler, Galileo, Copérnico, Jenner y Semelweiss. Las ideas de estos individuos no fueron populares entre colegas. Más recientemente, en psicología, el trabajo de John García sobre las restricciones biológicas del aprendizaje fue contrario a los de sus colegas. García consiguió publicar sus hallazgos en una revista (*Bulletin of the Psychonomic Society*) que no tenía evaluación de pares. Algunos investigadores que leyeron y replicaron su trabajo lo encontraron valioso. En la gran mayoría de los casos la revisión de colegas resulta benéfica para la ciencia.

Tercero, el conocimiento obtenido científicamente no está sujeto a una evaluación moral. Los resultados no se juzgan por "malos" ni "buenos", sino en términos de validez y confiabilidad. Sin embargo, el método científico está sujeto a principios de moralidad; es decir, que el científico es responsable de los métodos utilizados para obtener el conocimiento científico. En psicología, los códigos de ética se establecen para proteger a quienes están bajo estudio. La ciencia es una aventura compartida. La formación científica está disponible para todos y el método científico es bien conocido y está a la mano de quienes eligen usarlo.

El enfoque científico

El enfoque científico es una forma especial y sistematizada del pensamiento y del cuestionamiento reflexivos. Dewey (1933/1991), en su influyente trabajo *How We Think (Cómo Pensamos)*, delineó un paradigma general del cuestionamiento. La presente discusión del enfoque científico se basa en gran parte en el análisis de Dewey.

Problema-obstáculo-idea

El científico puede experimentar dificultades para entender, una vaga inquietud acerca de los fenómenos observados y no observados, una curiosidad sobre por qué algo es de la forma en que se presenta. El primer paso —y el más importante— es tener que sacar a la luz una idea, expresar el problema de alguna forma razonablemente manejable. Nunca o rara vez el problema surgirá por completo en esta etapa. El científico debe esforzarse con él, luchar con él y vivir con él. Dewey (1933/1991, p. 108) dice: "Hay una situación problemática, perpleja, irritante, donde la dificultad se encuentra a todo lo largo y ancho de ella, afectándola como un todo." Más tarde o más temprano, explícita o implícitamente, el científico definirá el problema, incluso si su expresión resulta incipiente y tentativa. El científico intelectualiza, como Dewey (p. 109) señala "aquello que al principio es meramente una cualidad *emocional* de toda la situación" (las itálicas son agregadas). En algunos aspectos ésta es la parte más difícil e importante de todo el proceso. Sin algún tipo de definición del problema, el científico difícilmente podrá seguir adelante y esperar que su trabajo sea fructífero. Para algunos investigadores la idea puede provenir de la conversación con un colega, o de la observación de un fenómeno curioso. La idea es que el problema por lo general se inicia con un pensamiento vago o no científico, o con presentimientos no sistemáticos. Después seguirán pasos más refinados.

Hipótesis

Después de intelectualizar el problema, de referirse a experiencias pasadas para posibles soluciones, de observar fenómenos relevantes, el científico puede formular una hipótesis. Una hipótesis es una declaración conjetural, una proposición tentativa acerca de la relación entre dos o más fenómenos o variables. Nuestro científico dirá, "si ocurre tal y tal, entonces resultará tal y tal".

Razonamiento-deducción

Este paso o actividad con frecuencia pasa inadvertido o es poco enfatizado. Quizás es la parte más importante del análisis de Dewey sobre el pensamiento reflexivo. El científico

deduce las consecuencias de la hipótesis que él mismo ha formulado. Conant (1951), al hablar acerca del surgimiento de la ciencia moderna, indica que el nuevo elemento que se aportó en el siglo XVII fue el uso del razonamiento deductivo. Aquí es donde la experiencia, el conocimiento y la perspicacia son importantes.

Con frecuencia el científico, al deducir las consecuencias de una hipótesis formulada, llegará a un problema muy diferente del original. Por otro lado, las deducciones pueden hacer creer que el problema no puede resolverse con las herramientas técnicas actuales. Por ejemplo, antes que la estadística moderna se desarrollara, algunos problemas de investigación del comportamiento eran irresolubles. Era difícil, si no es que imposible, probar dos o tres hipótesis interdependientes de forma simultánea, y también era casi imposible probar el efecto interactivo de variables. Ahora hay razón para pensar que ciertos problemas no pueden resolverse a menos que se les aborde de forma multivariada. Un ejemplo de esto es la relación entre los métodos de enseñanza con el aprovechamiento escolar y otras variables. Es probable que los métodos de enseñanza, *per se*, no difieran mucho si sólo se estudian sus efectos simples. Los métodos de enseñanza trabajan en forma diferente bajo distintas condiciones, con diversos maestros y con alumnos variados. Se dice que los métodos "interactúan" con las condiciones y características de los docentes y de los estudiantes. Simon (1987) presentó otro ejemplo: un estudio de investigación sobre entrenamiento de pilotos propuesto por Williams y Adelson en 1954 no se podía realizar usando los métodos tradicionales de experimentación. El estudio proponía examinar 34 variables y su influencia en el entrenamiento de los pilotos. Con el uso de métodos tradicionales de investigación el número de variables bajo estudio era abrumador. Alrededor de 20 años después, Simon (1976) y Simon y Roscoe (1984) demostraron la forma en que se podía abordar con efectividad tales estudios usando diseños económicos de megafactor. Un ejemplo puede ayudarnos a entender este paso de razonamiento-deducción.

Suponga que un investigador está intrigado por la conducta agresiva. El investigador se pregunta por qué las personas son con frecuencia agresivas en situaciones donde este comportamiento puede ser inapropiado. La observación personal nos lleva a suponer que la conducta agresiva parece ocurrir cuando las personas han experimentado dificultades de uno u otro tipo. (Nótese la vaguedad del problema en este punto). Después de pensar por un tiempo, revisar la literatura para obtener algunas claves y llevar a cabo más observaciones, se formula la hipótesis: la frustración conduce a la agresividad. *La frustración* se define como el obstáculo para alcanzar una meta y la *agresividad* como la conducta caracterizada por ataque físico o verbal a otras personas u objetos.

Lo que sigue es una declaración como ésta: si la frustración conduce a la agresividad, entonces deberíamos encontrar un alto grado de agresividad entre los niños de escuelas restrictivas, que no les permiten mucha libertad ni posibilidad de expresión. De forma similar, en situaciones sociales difíciles, suponiendo que son frustrantes, esperaríamos más agresividad de lo "común". Si seguimos el razonamiento, si les diéramos a sujetos experimentales problemas interesantes para resolver, y después evitáramos que los solucionaran, podemos predecir algún tipo de conducta agresiva. En pocas palabras, el proceso de trasladarnos de un contexto amplio a una situación más específica se llama *razonamiento deductivo*.

El razonamiento puede, como se indicó antes, cambiar el problema. Podemos comprender que el problema inicial era sólo un caso especial de un problema más amplio, fundamental e importante. Podemos por ejemplo, iniciar con una hipótesis más limitada: las situaciones escolares restrictivas conducen a negativismo en los niños. Después podemos generalizar el problema a: la frustración induce la agresividad. Aun cuando es una forma diferente de pensamiento de lo arriba discutido, es importante, por lo que casi podríamos llamar su calidad heurística. El razonamiento puede ayudar a dirigirnos hacia

problemas más amplios, más básicos y por tanto más significativos, así como a aportar implicaciones operacionales (susceptibles de ser probadas) de la hipótesis original. Este tipo se llama *razonamiento inductivo*. Parte de un hecho particular hacia un enunciado general o hipótesis. Si uno no es cuidadoso este método puede inducir un razonamiento deficiente debido a su tendencia natural de excluir datos que no se ajustan a la hipótesis. El método de razonamiento inductivo tiende a buscar más datos de apoyo que a refutar evidencias.

Considere el estudio clásico de Peter Wason (Wason y Johnson-Laird, 1972) que ha suscitado un gran interés (Hoch 1986; Klayman y Ha, 1987). En este estudio se le pidió a estudiantes descubrir la regla que el experimentador tenía en mente al generar una secuencia de números. Un ejemplo fue generar una regla de la siguiente serie: "3, 5, 7". Se les dijo a los alumnos que podrían preguntar acerca de otras secuencias y que recibirían retroalimentación en cada serie sobre si se ajustaron o no a la regla pensada por el experimentador. Cuando los estudiantes se sintieran seguros podrían externar la regla. Algunos estudiantes indicaron: "9, 11, 13", y les dijeron que esta secuencia se ajustaba a la regla. Después siguieron con "15, 17, 19" y otra vez les respondieron que la serie correspondía. Los estudiantes entonces presentaron su respuesta: "la regla es tres números nones consecutivos", pero se les dijo que ésta *no* era la regla. Después de varias secuencias más propusieron contestaciones tales como: "números con incrementos de dos en dos", o bien "números nones con incrementos de valor dos". En cada uno de los casos se les indicó que ésa no era la regla que el experimentador había pensado. La regla que el experimentador tenía en mente era: "tres números positivos crecientes cualesquiera". Si los estudiantes hubieran propuesto las secuencias "8, 9, 10" o "1, 15, 4500" les habrían dicho que estos números también se ajustaban a la regla. Donde los estudiantes cometieron el error fue en probar sólo los casos que se ajustaban a la primera secuencia propuesta y que confirmaba su hipótesis.

Aunque simplificado en exceso, el estudio de Wason demostró lo que puede pasar en la investigación científica real. Un científico puede fácilmente condenarse a repetir el mismo tipo de experimento que siempre apoye la hipótesis.

Observación-prueba-experimento

A estas alturas debe quedar claro que la fase de observación-prueba-experimento forma sólo parte de todo el proceso científico. Si el problema ha sido bien planteado, la o las hipótesis se han formulado de manera adecuada y las implicaciones de las hipótesis se han deducido con cuidado, se puede presumir en este paso que el investigador es competente desde el punto de vista técnico.

La esencia de la prueba de hipótesis consiste en probar la *relación* expresada por la hipótesis. No se prueban las variables como tales, sino la relación entre ellas. La observación, la prueba y la experimentación tienen un propósito fundamental: probar empíricamente la relación del problema. Probar sin saber —al menos en alguna medida— *qué* y *por qué* uno evalúa implica un disparate. El hecho de sólo plantear un problema vago, por ejemplo, "¿cómo afecta al aprendizaje la educación abierta?" para después evaluar alumnos en las escuelas que dicen ser diferentes en cuanto a su "nivel de apertura", o preguntarnos: "¿cuáles son los efectos de la disonancia cognitiva?" para luego crear disonancia a través de manipulaciones experimentales y buscar los supuestos efectos, todo esto sólo podría llevarnos a información cuestionable. De forma similar, decir que se estudiarán los procesos de atribución sin realmente saber lo que se hace, o sin establecer relaciones entre las variables, constituye investigación sin sentido.

Otro aspecto importante es que por lo general no se prueba la hipótesis de manera directa. Como se indicó en el paso previo sobre el razonamiento, probamos las implicaciones que deducimos de las hipótesis. Nuestra hipótesis a probar puede ser: "los sujetos a quienes se les instruye para que eviten pensamientos indeseados estarán más preocupados con ellos que aquellos a los que se da una distracción". Ésta se dedujo de una hipótesis más amplia y general: "Cuantos mayores sean los esfuerzos para suprimir una idea, mayor será la preocupación sobre esta idea." No probamos "la supresión de ideas" o "la preocupación" sino la relación entre ellas, en este caso, la relación entre supresión de pensamientos indeseados y el nivel de preocupación (véase Wegner, Schneider, Carter y White, 1987; Wegner, 1989).

Dewey enfatizó que la secuencia temporal del pensamiento o cuestionamiento reflexivos no está fija. Podemos repetir y enfatizar lo que dijo en nuestro propio marco: los pasos del enfoque científico no están ordenados de manera impecable. El primer paso no se completa perfectamente antes de iniciar el segundo. Más aún, podemos hacer pruebas antes de deducir de forma adecuada las implicaciones de la hipótesis. La hipótesis, en sí misma, puede necesitar una mayor elaboración o refinamiento como resultado de las implicaciones deducidas de ella. Con frecuencia veremos que las hipótesis y su expresión parecen inadecuadas una vez que se deducen implicaciones a partir de ellas. Cuando una hipótesis es muy vaga se presenta la dificultad común de observar que una deducción es tan buena como otra; esto es, la hipótesis puede no conducir a pruebas precisas.

Retroalimentar el problema, la hipótesis y finalmente, la teoría de los resultados de la investigación es de la mayor importancia. Los teóricos e investigadores del aprendizaje, por ejemplo, con frecuencia han modificado sus teorías e investigaciones como resultado de hallazgos experimentales (véase Malone, 1991; Schunk, 1996; Hergenhahn, 1996). Teóricos e investigadores han estudiado los efectos del entorno y del entrenamiento tempranos en el desarrollo posterior. Kagan y Zentner (1996) revisaron los resultados de 70 estudios sobre las relaciones entre las experiencias de la edad temprana y la psicopatología en la vida adulta. Encontraron que es posible predecir la delincuencia juvenil a partir de la cantidad de impulsividad detectada en la etapa preescolar. Lynch, Short y Chua (1995) hallaron que el procesamiento musical estaba influido por la estimulación perceptual experimentada por el niño entre los seis meses y el año. Éstas y otras investigaciones han generado evidencia variada que converge en este problema de extrema importancia teórica y práctica. Una parte esencial de la investigación científica es el esfuerzo constante por replicar y verificar los hallazgos, por corregir la teoría con base en la evidencia empírica y por encontrar mejores explicaciones para los fenómenos naturales. Se puede incluso aseverar que la ciencia tiene un aspecto cíclico. Un investigador encuentra, por ejemplo, que *A* está relacionada con *B* de tal o cual forma. Entonces se conduce más investigación para determinar bajo qué otras condiciones *A* está relacionada de forma similar a *B*. Otros investigadores desafían esta teoría e investigación y ofrecen sus propias evidencias y explicaciones. El investigador original, se espera, modificará su trabajo a la luz de los nuevos datos: el proceso nunca termina.

Resumamos el llamado enfoque científico de la investigación. En primera instancia existe una duda, una barrera, una situación indeterminada que debe determinarse. El científico experimenta dudas vagas, disturbios emocionales e ideas incipientes. Hay un esfuerzo por formular el problema aunque sea de forma inadecuada. El científico entonces revisa la literatura y busca en su propia experiencia y en la de otros. Es frecuente que el investigador tenga que esperar un momento de inventiva: puede ser que ocurra, puede ser que no. Una vez que se cuenta con el problema formulado, con la o las preguntas básicas expresadas de manera apropiada, el resto es más fácil. Entonces, la hipótesis se construye, después de lo cual se deducen las implicaciones empíricas. Durante este proceso el proble-

ma original, y por supuesto la hipótesis original, pueden cambiarse, hacerse más generales o reducirse. Incluso pueden abandonarse. Más tarde se prueba la relación expresada por la hipótesis por medio de la observación y la experimentación. Con base en la evidencia procedente de la investigación, la hipótesis se apoya o rechaza. Esta información después retroalimenta al problema original, que se conserva o modifica de acuerdo con la evidencia. Dewey señaló que una fase del proceso puede expandirse y ser de gran importancia, otra puede reducirse, y puede haber más o menos pasos involucrados. La investigación rara vez es un asunto ordenado. De hecho, resulta mucho más desordenado que lo que la discusión anterior pueda sugerir. El orden y el desorden, sin embargo, no son de primera importancia. Lo que *sí* es importante es la racionalidad controlada de la investigación científica como un proceso de indagación reflexiva, la naturaleza interdependiente de las partes del proceso y la suprema importancia del problema y su enunciado.

RESUMEN DEL CAPÍTULO

1. Para comprender la compleja conducta humana es necesario entender el lenguaje y enfoque científicos.
2. La ciencia es una extensión sistemática y controlada del sentido común. Hay cinco diferencias entre el sentido común y la ciencia:
   a) La ciencia utiliza esquemas conceptuales y estructuras teóricas.
   b) La ciencia prueba de forma sistemática y empírica las teorías e hipótesis.
   c) La ciencia intenta controlar posibles causas extrañas.
   d) La ciencia busca relaciones de manera consciente y sistemática.
   e) La ciencia excluye explicaciones metafísicas (no demostrables).
3. Los cuatro métodos del conocimiento de Peirce son:
   a) Método de la tenacidad —influenciado por las creencias pasadas ya establecidas—.
   b) Método de autoridad —determinado por el peso de la tradición o la sanción pública—.
   c) Método *a priori* (también conocido como método de la intuición) —una natural inclinación hacia la verdad—.
   d) Método de la ciencia —autocorrectivo; los conceptos son objetivos y susceptibles de ser probados—.
4. Los estereotipos de la ciencia han limitado la comprensión de esta actividad por parte del público.
5. Visión y función de la ciencia
   a) Una visión estática considera a la ciencia como proveedora de información para el mundo; la ciencia aporta al cuerpo de información y al estado actual del conocimiento.
   b) La visión dinámica está interesada en la actividad de la ciencia (lo que hacen los científicos). Con ello aparece la visión heurística de la ciencia, que tiene un carácter de autodescubrimiento. La ciencia asume riesgos y resuelve problemas.
6. Los objetivos de la ciencia son:
   a) Generar teoría y explicar los fenómenos naturales.
   b) Promover la comprensión y desarrollar predicciones.
7. Una teoría tiene tres características:
   a) Posee un conjunto de propiedades con base en constructos definidos e interrelacionados.

b) Determina de forma sistemática las interrelaciones entre un grupo de variables.
c) Explica fenómenos.

8. La investigación científica es una búsqueda sistemática, controlada, empírica y crítica de fenómenos naturales. La teoría e hipótesis acerca de relaciones que se presume existen entre tales fenómenos la orienta. Es pública y amoral.
9. El enfoque científico, de acuerdo con Dewey, está integrado por:
   a) Problema-obstáculo-idea —formular el problema o pregunta de investigación a resolver—.
   b) Hipótesis —formular un enunciado conjetural acerca de la relación entre fenómenos o variables—.
   c) Razonamiento-deducción —el científico deduce las consecuencias de la hipótesis, lo cual puede conducir a un problema más significativo y generar ideas sobre cómo las hipótesis pueden ser probadas en términos observables—.
   d) Observación-prueba-experimento —constituyen la fase de recolección y análisis de datos—. Los resultados de la investigación se relacionan una vez más con el problema.

SUGERENCIAS DE ESTUDIO

Una parte de este capítulo se presta a gran controversia. Algunos pensadores aceptan esos puntos de vista, mientras otros los rechazan. La revisión de literatura selecta contribuye a mejorar la comprensión de la ciencia y de su propósito, de la relación entre ciencia y tecnología, y las diferencias entre investigación básica y aplicada. Esas lecturas pueden constituir la base de las discusiones en clase. En el libro del primer autor, *Behavioral Research: A Conceptual Approach* (Nueva York: Holt, Rinehart y Winston, 1979, capítulos 1, 15 y 16) es posible encontrar una amplia presentación de aspectos controvertidos de la ciencia, en especial, de la ciencia del comportamiento. Se ha publicado una gran cantidad de artículos de buena calidad en ciencia e investigación, en revistas científicas y en libros de filosofía de la ciencia. Aquí presentamos algunos. También se incluye un informe especial en *Scientific American*. Todos son relevantes para el contenido de este capítulo.

Barinaga, M. (1993). Philosophy of science: Feminists find gender everywhere in science. *Science, 260*, 392-393. Analiza la dificultad de separar puntos de vista culturales de las mujeres y la ciencia. Señala a la ciencia como un campo predominantemente masculino.

Hausheer, J., Harris, J. (1994). In search of a brief definition of science. *The Physics Teacher, 32(5)*, 318. Menciona que cualquier definición de ciencia debe incluir guías para evaluar teorías e hipótesis como científicas o no científicas.

Holton, G. (1996). The controversy over the end of science. *Scientific American, 273(10)*, 191. Este artículo versa sobre dos campos del pensamiento: los linealistas y los cíclicos. Los linealistas tienen una perspectiva más convencional de la ciencia; los cíclicos conciben a la ciencia como un proceso que se degenera de forma interna.

Horgan, J. (1994). Anti-omniscience: An eclectic gang of thinkers pushes at knowledge's limits. *Scientific American, 271*, 20-22. Analiza los límites de la ciencia.

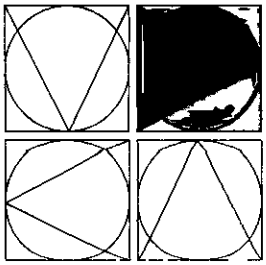
Horgan, J. (1997). *The end of science*. Nueva York: Broadway Books.

Miller, J.A. (1994). Postmodern attitude toward science. *Bioscience, 41(6)*, 395. Analiza las razones de algunos educadores y académicos en el área de humanidades que han adoptado una actitud hostil hacia la ciencia.

Scientific American. Science versus antiscience. (Special report). Enero 1997, 96-101. Presenta tres diferentes movimientos anticiencia: creacionista, feminista y medios de comunicación.

Smith, B. (1995). Formal ontology, common sense and cognitive science. *International Journal of Human-Computer Studies, 43(5-6),* 641-667. Un artículo que examina el sentido común y la ciencia cognitiva.

Timpane, J. (1995). How to convince a reluctant scientist. *Scientific American, 272,* 104. Este artículo advierte que demasiada originalidad en la ciencia puede conducir a la falta de aceptación y a la dificultad en su comprensión. También analiza cómo la aceptación científica está gobernada tanto por información previa como nueva, y por la reputación del científico.



CAPÍTULO 2

PROBLEMAS E HIPÓTESIS

- PROBLEMAS
- CRITERIOS DE LOS PROBLEMAS Y ENUNCIADOS DE PROBLEMAS
- HIPÓTESIS
- IMPORTANCIA DE LOS PROBLEMAS E HIPÓTESIS
- VIRTUDES DE LOS PROBLEMAS E HIPÓTESIS
- PROBLEMAS, VALORES Y DEFINICIONES
- GENERALIDAD Y ESPECIFICIDAD DE LOS PROBLEMAS E HIPÓTESIS
- LA NATURALEZA MULTIVARIABLE DE LA INVESTIGACIÓN Y PROBLEMAS DEL COMPORTAMIENTO
- COMENTARIOS FINALES: EL PODER ESPECIAL DE LAS HIPÓTESIS

Mucha gente cree que la ciencia es en lo fundamental una actividad de recolección de hechos. M.R. Cohen (1956/1997, p. 148) lo planteó de otra forma:

> No hay un progreso genuino en el discernimiento científico a través del método baconiano de acumular hechos empíricos sin una hipótesis o anticipación de la naturaleza. Sin alguna idea que nos guíe no sabemos qué hechos recolectar... no podemos determinar qué es y qué no es relevante.

Las personas sin formación científica con frecuencia tienen la idea que el científico es un individuo objetivo en extremo que recolecta datos sin tener ideas preconcebidas. Poincare (1952/1996, p. 143) señaló lo equivocado de esta idea: "Se dice a menudo que los experimentos deben realizarse sin ideas preconcebidas. Eso es imposible. No sólo haría que todo experimento fuera improductivo, sino que, aun deseándolo, no podríamos hacerlos."

Problemas

No siempre le es posible al investigador definir el problema de una manera simple, clara y completa. A menudo, puede tener una noción general, difusa e, inclusive, confusa del pro-

blema. Esto se debe a la naturaleza compleja de la investigación científica. Es posible incluso que, pueda tomarle años de exploración, reflexión e investigación el poder definir una pregunta de forma clara. Sin embargo, enunciar de manera adecuada el problema de investigación es una de las partes fundamentales del proceso. La dificultad para enunciar un problema de investigación de forma satisfactoria en un momento dado no debe hacernos perder de vista lo necesario y deseable que resulta.

Con esta dificultad en mente, podemos establecer un principio fundamental: si queremos resolver un problema, en general debemos conocerlo. Se puede decir que gran parte de la solución estriba en conocer lo que se trata de hacer. Otra parte está en entender qué es un problema y, en especial, un problema científico.

¿Qué constituye un buen enunciado del problema? Aunque los problemas de investigación difieren en gran medida y no existe una fórmula "correcta" para enunciar problemas, es posible aprender y utilizar para nuestro beneficio ciertas características de los problemas y de los enunciados de problemas. Para empezar, consideremos dos o tres ejemplos de problemas de investigación publicados y estudiemos sus características. Primero, tomemos el problema del estudio realizado por Hurlock (1925),[1] mencionado en el capítulo 1: ¿Cuáles son los efectos de diferentes tipos de incentivos en el rendimiento del alumno? Observe que el problema está enunciado en forma de pregunta. En este campo, la forma más simple es la mejor. También conviene señalar que el problema establece una relación entre variables, en este caso, entre las variables *incentivos* y *rendimiento del alumno* (logro). (El término *variable* será definido de manera formal en el capítulo 3. Por ahora, se usará para nombrar un fenómeno o un constructo, que asume un conjunto de diferentes valores numéricos.)

Un *problema*, entonces, es un enunciado u oración interrogativa que pregunta: ¿qué relación existe entre dos o más variables? La respuesta constituye aquello que se busca en la investigación. Un problema, en la mayoría de los casos, tendrá dos o más variables. En el ejemplo de Hurlock, el enunciado del problema relaciona incentivos con rendimiento del alumno. Otro problema, estudiado en el experimento clásico de Bahrick (1984, 1992) está asociado con preguntas de edad y vejez: ¿Cuánto de lo que ahora estás estudiando recordarás dentro de diez años? ¿Cuánto de esto mismo podrás recordar dentro de cincuenta años? ¿Cuánto recordarás después, si nunca lo utilizas? La pregunta formal de Bahrick es: ¿la memoria semántica involucra procesos separados? Una variable es la cantidad de tiempo que transcurre desde que el material se aprendió por primera vez; la segunda podría ser la calidad del aprendizaje original; y la otra variable es el recuerdo (u olvido). Veamos otro problema de Little, Sterling y Tingstrom (1996), que es muy diferente: ¿Influyen las claves geográficas y las características raciales en la atribución (culpa percibida)? Una variable son las claves geográficas; la segunda sería la información racial, y la tercera, la atribución.

No todos los problemas de investigación contienen dos o más variables claras. Por ejemplo, en psicología experimental, el foco de la investigación con frecuencia está en procesos psicológicos como la memoria y la categorización. Rosch (1973) en su influyente y justificadamente bien conocido estudio de categorías perceptuales hizo la siguiente pregunta: ¿Existen categorías no arbitrarias ("naturales") de color y forma? Aunque la relación entre dos o más variables no es aparente en este enunciado del problema, en la investigación real las categorías estaban relacionadas con el aprendizaje. Hacia el final de este libro se verá que los problemas de investigación factorial analítica también carecen

[1] Cuando referimos problemas e hipótesis de la literatura, no siempre usamos las palabras de los autores. De hecho, los enunciados de muchos de los problemas son nuestros y no de los autores citados. Algunos autores sólo usan enunciados de problemas, algunos sólo hipótesis, y otros usan ambos.

de la forma de las relaciones antes planteada. Sin embargo, en la mayoría de los problemas de investigación del comportamiento, se estudian las relaciones entre dos o más variables; por ello, enfatizaremos ese tipo de enunciados de relación.

Criterios de los problemas y enunciados de problemas

Existen tres criterios de buenos problemas y enunciados de problemas. El primero: el problema debe expresar una relación entre dos o más variables. En efecto plantea preguntas como la siguiente: ¿*A* está relacionada con *B*? ¿Cómo están relacionadas *A* y *B* con *C*? ¿Cómo está *A* relacionada con *B* bajo las condiciones *C* y *D*? La excepción a esta consideración ocurre casi siempre en investigación metodológica o taxonómica.

Segundo: el problema debe ser enunciado de manera clara y sin ambigüedades en forma de pregunta. En lugar de decir, por ejemplo "el problema es..." o "El propósito de este estudio es..." resulta necesario plantear una pregunta. Las preguntas tienen la virtud de presentar los problemas directamente. El propósito de un estudio no es por fuerza el mismo que el problema de un estudio. El propósito del estudio de Hurlock, por ejemplo, fue arrojar luz sobre el uso de incentivos en las situaciones escolares. El problema consistió en la pregunta acerca de la relación entre incentivos y rendimiento. Otra vez, mientras más simple, mejor: formule una pregunta.

El tercer criterio con frecuencia es difícil de satisfacer. Demanda que el problema y su enunciado *impliquen* la posibilidad de ser sometidos a una prueba empírica. Un problema que no contenga implicaciones para probar las relaciones que enuncia, no constituye un problema científico. Esto significa no sólo que se enuncie una relación real, sino también que las variables de la relación puedan ser medidas de alguna forma. Hay muchas preguntas interesantes e importantes que no constituyen preguntas científicas tan sólo porque no son susceptibles de prueba. Ciertas preguntas filosóficas y teológicas, aunque importantes para quienes las consideran, no pueden ser probadas empíricamente, por lo que no generan interés para el científico como tal. La pregunta epistemológica "¿Cómo conocemos?" es una pregunta de ese tipo. La educación plantea muchas preguntas interesantes pero no científicas, por ejemplo "¿Mejora la educación democrática el aprendizaje de los jóvenes?" "¿Son buenos los procesos grupales para los niños?" Estas preguntas pueden ser etiquetadas como metafísicas en el sentido en que están, al menos así enunciadas, fuera de la posibilidad de una prueba empírica. Las principales dificultades estriban en que algunas no constituyen relaciones, y es muy difícil o imposible definir la mayoría de sus constructos de forma que puedan ser medidos.

Hipótesis

Una *hipótesis* es un enunciado conjetural de la relación entre dos o más variables. Las hipótesis siempre se presentan en forma de enunciados declarativos y relacionan, de manera general o específica, variables con variables. Hay dos criterios que definen a las "buenas" hipótesis y a sus enunciados. Son los mismos que mencionamos para los problemas y sus enunciados. 1) Las hipótesis son enunciados acerca de las relaciones entre variables. 2) Las hipótesis contienen implicaciones claras para probar las relaciones enunciadas. Estos criterios significan que los enunciados de hipótesis contienen dos o más variables, que son medibles o pueden serlo, y que especifican cómo están relacionadas las variables.

Permítanos mencionar tres hipótesis de la literatura y aplicarles estos criterios. La primera hipótesis procede de un estudio de Wegner y colaboradores. (1987) que parece desafiar el sentido común: a mayor supresión de pensamientos indeseados, mayor preocu-

pación por ellos (represión ahora; obsesión más tarde). Aquí se establece una relación entre una variable, supresión de una idea o pensamiento, y otra variable, preocupación u obsesión. Dado que ambas se definen y miden con facilidad, las implicaciones para probar la hipótesis también se conciben sin esfuerzo. Los criterios están satisfechos. En el estudio de Wegner y colaboradores, se les pidió a los sujetos que *no* pensaran en un "oso blanco". Cada vez que pensaran en él, debían de tocar una campana. El número de campanadas indicaba el nivel de preocupación. Una segunda hipótesis, que resulta inusual y corresponde al estudio de Ayres y Hughes (1986), enuncia la relación de una forma que llamamos nula: el nivel de ruido o música no tiene efecto en el funcionamiento visual. La relación se establece con claridad: una variable, intensidad del sonido (por ejemplo, música), se relaciona con otra, funcionamiento visual, a través de las palabras "no tiene efecto en". En el criterio de potencia de ser probada, sin embargo, encontramos dificultades. Nos enfrentamos con el problema de definir "funcionamiento visual" e "intensidad" de forma que puedan medirse. Si podemos resolver este problema de manera satisfactoria, entonces, tenemos en definitiva una hipótesis. Ayres y Hughes lo hicieron al definir intensidad como 107 decibeles y funcionamiento visual en términos de una puntuación en una tarea de agudeza visual. Esta hipótesis permitió contestar una pregunta que la gente con mucha frecuencia se hace: "¿por qué bajamos el volumen del estéreo del auto cuando *buscamos* una dirección?". Ayres y Hughes encontraron una caída marcada en el funcionamiento perceptual cuando el nivel de música llegaba a 107 decibeles.

La tercera hipótesis representa una categoría numerosa e importante. En ella la relación es indirecta, oculta. En general enuncia que los grupos *A* y *B* diferirán en alguna característica. Por ejemplo: las mujeres creen, con mayor frecuencia que los hombres, que deben perder peso aun cuando éste se encuentre dentro de los límites normales (Fallon y Rozin, 1985). Esto es, que las mujeres difieren de los hombres en cuanto a la percepción de su figura corporal. Observe que este enunciado está un paso más allá de la hipótesis real que puede plantearse como: la percepción de la figura corporal es, en parte, una función del género. Si los enunciados posteriores constituyeran la hipótesis enunciada, entonces la primera podría llamarse una subhipótesis o una predicción específica basada en la hipótesis original.

Consideremos otra hipótesis de este tipo pero dando un paso más adelante. Los individuos que tienen características iguales o similares tendrán actitudes similares hacia objetos cognitivos significativamente relacionados con su papel ocupacional (Saal y Moore, 1993). (*Los objetos cognitivos* se definen como algo concreto o abstracto, percibido y "conocido" por los individuos. Personas, grupos, ascenso en el trabajo o en las calificaciones, el gobierno y la educación son algunos ejemplos.) La relación en este caso es, por supuesto, entre características personales y actitudes (hacia un objeto cognitivo relacionado con la característica personal, por ejemplo, género y actitudes hacia otros que reciben una promoción). Para probar esta hipótesis, sería necesario tener al menos dos grupos, cada uno con una característica diferente, y después comparar las actitudes de ambos grupos. Por ejemplo, como en el caso del estudio de Saal y Moore, la comparación sería entre hombres y mujeres. Serían comparados en relación a su evaluación hacia el ascenso dado a un compañero de trabajo del mismo sexo o del opuesto. En este ejemplo, se satisfacen los criterios.

Importancia de los problemas e hipótesis

Hay poca duda de que las hipótesis son herramientas importantes e indispensables de la investigación científica. Existen tres razones principales para esta creencia. La primera es que son, digamos, los instrumentos de trabajo de la teoría. Las hipótesis pueden deducirse

a partir de la teoría y de otras hipótesis. Si por ejemplo, trabajamos en una teoría sobre la agresividad, se presume que buscamos causas y efectos del comportamiento agresivo. Es posible que hayamos observado casos de comportamiento agresivo ocurrido después de circunstancias frustrantes. La teoría, entonces, puede incluir la proposición: la frustración produce agresividad (Berkowitz, 1983; Dill y Anderson, 1995; Dollard, Doob, Miller, Mowrer, y Sears, 1939). A partir de esta amplia hipótesis, podemos deducir hipótesis más específicas, como por ejemplo: impedir que los niños alcancen sus metas deseadas (frustración) generará pleitos entre ellos (agresión); si los niños son privados del amor paterno (frustración), reaccionarán en parte, con un comportamiento agresivo.

La segunda razón es que es posible someter a prueba las hipótesis y demostrar que son probablemente verdaderas o probablemente falsas. No se prueban hechos aislados, como se dijo antes, sólo relaciones. Es probable que la principal razón de usar hipótesis en la investigación científica sea que constituyen proposiciones relacionales. En esencia, son predicciones del tipo: "si *A*, entonces *B*", que utilizamos para probar la relación entre *A* y *B*. Dejamos que los hechos tengan la oportunidad de establecer la probable veracidad o falsedad de la hipótesis.

La tercera razón es que las hipótesis son herramientas poderosas para el avance del conocimiento porque permiten al científico ir más allá de sí mismo. Aunque desarrolladas por humanos, las hipótesis existen, pueden ser probadas y puede demostrarse que son probablemente correctas o incorrectas de manera independiente a los valores y opiniones de una persona (sesgos). Esto resulta crítico: no habría ciencia, en sentido completo alguno, sin las hipótesis.

Tan importantes como las hipótesis son los problemas tras ellas. Como Dewey (1938/1982, pp. 105-107) ha señalado, la investigación por lo general empieza con un problema. Indica que primero hay una situación indeterminada en la que las ideas son vagas, aparecen dudas, y el pensador queda perplejo. Agrega que el problema no se enuncia; de hecho, no puede ser enunciado hasta que uno ha experimentado una situación tan indeterminante.

La indeterminación, sin embargo, deberá, en última instancia ser eliminada. Aunque es cierto, como se señaló antes, que el investigador con frecuencia puede tener sólo una noción general y difusa del problema, tarde o temprano habrá de definir una idea clara de lo que el problema es. Aunque este enunciado parezca obvio, una de las cosas más difíciles de lograr, es enunciar de una manera clara y completa el problema de investigación. En otras palabras, uno debe saber qué es lo que trata encontrar. Cuando por fin se identifica, el problema ya está en camino a la solución.

Virtudes de los problemas e hipótesis

Los problemas y las hipótesis tienen virtudes importantes: 1) dirigen la investigación (las relaciones expresadas en las hipótesis indican al investigador lo que debe hacer); 2) los problemas e hipótesis, dado que son de ordinario enunciados relacionales generalizados, permiten al investigador deducir manifestaciones empíricas específicas implicadas en ellos. Podemos decir, de acuerdo con Guida y Ludlow (1989): si es un hecho verdadero que los niños de un tipo de cultura (Chile) tienen un mayor grado de ansiedad que los niños de otro tipo de cultura (blancos estadounidenses), entonces los niños en la cultura chilena deben rendir menos en lo académico que los niños en la cultura estadounidense. Los niños chilenos quizá también debieran presentar una menor autoestima o un *locus* de control más externo en lo que se refiere a la escuela y a la labor académica.

Hay diferencias importantes entre problemas e hipótesis. Las hipótesis, si están enunciadas de manera apropiada, pueden ser probadas. Una hipótesis dada puede ser demasia-

do amplia para ser probada de forma directa; sin embargo, si es una "buena" hipótesis, es posible deducir a partir de ella otras que sí lo sean. Los hechos o las variables no se prueban como tales. Se prueban las relaciones enunciadas por las hipótesis. Un problema no puede ser resuelto de manera científica a menos que se reduzca a su forma de hipótesis, ya que un problema es una pregunta, generalmente de naturaleza amplia, que no puede probarse en forma directa. No se someten a prueba preguntas como: ¿la presencia o ausencia de otra persona en un sanitario público afecta la higiene personal? (Pedersen, Keithly y Brady, 1986). ¿Las sesiones de consejería grupal disminuyen el nivel de morbilidad psiquiátrica en oficiales de policía? (Doctor, Cutris e Issacs, 1994). Quizás uno probaría una o más hipótesis deducidas de estas preguntas. Por ejemplo, para estudiar el último problema, uno puede hipotetizar que los oficiales de policía que asisten a sesiones de consejería para reducir el estrés requerirán menos días de incapacidad por enfermedad que aquéllos que no asisten. La hipótesis para el primer problema podría indicar que la presencia de una persona en un sanitario público hará que otras se laven las manos.

Los problemas e hipótesis permiten que avance el conocimiento científico al ayudar al investigador a confirmar o refutar una teoría. Suponga que un investigador en el campo de la psicología aplica a unos sujetos tres o cuatro pruebas, entre las cuales hay una para evaluar la ansiedad relacionada con una prueba aritmética. Al calcular de manera rutinaria las correlaciones entre las tres o cuatro pruebas, uno encuentra que la correlación entre ansiedad y aritmética es negativa. De lo anterior se deduce que a mayor ansiedad, menor puntuación en la prueba de aritmética. Sin embargo, es muy probable que esta relación sea fortuita e incluso espuria, pero si el investigador hubiera hipotetizado la relación con base en una teoría, tendría mayor confianza en sus resultados. El investigador que no hipotetiza relaciones en forma previa, no permite que los hechos prueben o rechacen nada. Las palabras *probar* y *rechazar* no deben tomarse en su sentido literal: una hipótesis nunca se prueba o refuta realmente. Para ser más precisos, deberíamos decir algo del tipo de: el peso de la evidencia está del lado de la hipótesis o el peso de la evidencia arroja dudas sobre la hipótesis. Braithwaite (1953/1996, p. 14) dice:

> De este modo, la evidencia empírica nunca prueba la hipótesis: en casos apropiados podemos decir que *se establece* (itálicas agregadas) la hipótesis, lo que significa que la evidencia hace que sea razonable aceptar la hipótesis; pero ésta nunca *prueba* la hipótesis en el sentido de que la hipótesis sea una consecuencia lógica de la evidencia.

Este uso de la hipótesis es similar a participar en un juego de azar. Se establecen las reglas del juego y se definen las apuestas por adelantado. Uno no puede cambiar las reglas después de un resultado, como tampoco se pueden cambiar las apuestas una vez hechas. Uno no tira los dados primero y luego apuesta. No sería "justo". De igual forma, si en primera instancia se recolectan datos y después se toma uno de ellos y se llega a una conclusión con base en él, se han violado las reglas del juego científico. El juego no es "justo" porque el investigador puede capitalizar fácilmente, digamos, dos relaciones importantes de las cinco a prueba. Las otras tres, por lo general, se olvidan. En un juego "justo", se cuenta cada tiro del dado, en el sentido de que se gana o no con base en el resultado de cada tirada.

Las hipótesis dirigen la investigación. Como Darwin señaló hace más de 100 años, todas las observaciones han de ser a favor o en contra de algún punto de vista para tener alguna utilidad. Las hipótesis incorporan aspectos de la teoría bajo prueba de forma susceptible o casi susceptible de ser probada. Antes se dio un ejemplo de la teoría del reforzamiento en el que se dedujeron hipótesis demostrables a partir del problema general. Podemos demostrar la importancia del reconocimiento de esta función de las hipótesis al introducirnos por la puerta trasera y usar una teoría muy difícil o quizás imposible de probar. La teoría de Freud de la ansiedad incluye el constructo de la represión. Con este

término Freud se refería a la introducción forzada de ideas inaceptables en lo profundo del inconsciente. Para probar la teoría freudiana de la ansiedad es necesario deducir relaciones sugeridas por la teoría. Estas deducciones por fuerza deberán incluir el concepto de represión que implica el constructo del inconsciente. Es posible formular hipótesis que utilizan estos constructos; para probar la teoría han de ser formuladas de esta manera. Pero probarlos resulta más difícil debido a la extrema dificultad para definir términos como "represión" e "inconsciente" de manera que puedan medirse. Hasta hoy nadie ha tenido éxito al definir estos dos constructos sin apartarse en gran medida del significado y uso freudianos originales. Las hipótesis constituyen, entonces, puentes importantes entre la teoría y la investigación empírica.

Problemas, valores y definiciones

Para establecer con claridad la naturaleza de los problemas y de las hipótesis, analizaremos ahora dos o tres errores comunes. En primer término, los problemas científicos no son preguntas morales ni éticas: ¿Son las medidas disciplinarias de tipo punitivo perjudiciales para los niños? ¿Debiera ser el liderazgo de una organización de tipo democrático? ¿Cuál es la mejor forma de enseñar a los estudiantes universitarios? Formular estas preguntas equivale a presentar cuestionamientos de valor y juicio que la ciencia no puede contestar. Muchas de las que se han llamado hipótesis no lo son en absoluto. Por ejemplo: el método de enseñanza a pequeños grupos es mejor que el método expositivo. Éste es un enunciado de valor; constituye un acto de fe, no una hipótesis. Si fuera posible establecer una relación entre las variables, y definirlas de manera que se pudiera probar esa relación, entonces podríamos contar con una hipótesis. Pero no hay una forma científica de someter a prueba preguntas de valor.

Una forma rápida y relativamente fácil de detectar preguntas y enunciados de valor consiste en buscar palabras como *debe, debería, mejor que* (en lugar de *mayor que*), así como palabras similares que indiquen juicios culturales o personales, o preferencias (sesgos). Sin embargo, los enunciados de valor son engañosos. Aunque resulta obvio que un enunciado que incluye la palabra "debería" es un enunciado de valor, otros tipos no son tan evidentes. Consideremos el enunciado: los métodos autoritarios de enseñanza conducen a un pobre aprendizaje. En este caso sí hay una relación. Pero el enunciado falla como una hipótesis científica en tanto que incorpora dos expresiones de valor: "métodos autoritarios de enseñanza" y "pobre aprendizaje", ninguna de las cuales puede definirse con propósitos de medición sin borrar las palabras *autoritario* y *pobre*.[2]

Con frecuencia se formula otra clase de enunciados que no constituyen hipótesis o que son hipótesis pobres, en especial en el campo de la educación. Consideremos, por ejemplo, el siguiente: los cursos de tronco común representan una experiencia enriquecedora. Otro tipo de enunciado que se usa con frecuencia, es la generalización vaga: es posible identificar las habilidades de lectura en el segundo grado; La meta del individuo auténtico es la autorrealización; El prejuicio se relaciona con ciertos rasgos de personalidad.

Otro defecto común de los enunciados de problema aparece a menudo en las tesis doctorales: enlistar aspectos metodológicos o "problemas" como subproblemas. Estos aspectos metodológicos poseen dos características que hacen fácil detectarlos: 1) No son

[2] Un caso ya casi clásico del uso de la palabra *autoritario* es la frase que a veces se escucha entre educadores: el método expositivo es autoritario. Esto parece indicar que quien lo dice no gusta del método expositivo y lo considera negativo. De forma similar, una de las formas más efectivas de criticar a un maestro es decir que es autoritario.

problemas sustantivos que surgen del problema básico; y 2) Se relacionan con técnicas o métodos de muestreo, medición o análisis. En general, no aparecen en forma de pregunta y contienen palabras tales como *probar, determinar, medir*: "para *determinar* la confiabilidad de los instrumentos usados en esta investigación;" "para *probar* la significación de las diferencias entre las medias" o "para *asignar* alumnos de manera aleatorizada a los grupos experimentales", son ejemplos de esta noción equivocada de problemas y subproblemas.

Generalidad y especificidad de los problemas e hipótesis

Una dificultad que el investigador por lo general encuentra y que casi todos los estudiantes que trabajan en una tesis hallan molesta, es la generalidad y la especificidad de los problemas e hipótesis. Si el problema es muy general, es demasiado vago para ser sometido a prueba. Así, desde el punto de vista científico resulta inútil, aunque puede ser interesante para leer. Los problemas e hipótesis demasiado generales o vagos son comunes. Por ejemplo: La creatividad es una función de la autorrealización del individuo; La educación democrática potencia el aprendizaje social y la forma cívica; El autoritarismo en el salón de clases universitario inhibe la imaginación creativa del estudiante. Todos resultan problemas interesantes, pero en su forma actual son por completo inútiles en el terreno científico, en tanto que no pueden ser sometidos a prueba y porque parecen sugerir una seguridad espuria de que constituyen hipótesis que "algún día" pueden ser probadas.

Términos tales como "creatividad", "autorrealización", "democracia" y "autoritarismo" no tienen, al menos hasta ahora, referentes empíricos adecuados.[3] Es cierto que podemos definir *creatividad*, en una forma limitada al especificar una o dos pruebas de creatividad. Éste puede ser un procedimiento legítimo; sin embargo, al emplearlo, corremos el riesgo de alejarnos del término original y de su significado. Esto es en particular cierto cuando hablamos de creatividad artística. Desde luego, con frecuencia aceptamos el riesgo con tal de investigar problemas importantes. Aun así, términos como "democracia" son casi imposibles de definir. Incluso cuando lo hacemos, a menudo descubrimos que hemos destruido su significado original. Una excepción sobresaliente es la definición y la medición de "democracia" de Bollen (1980). Examinaremos ambas en otros capítulos.

El otro extremo es caer en demasiada especificidad. Todo estudiante ha escuchado que es necesario reducir los problemas a una dimensión manejable. Esto es cierto, pero por desgracia, podemos reducirlo tanto hasta hacerlo desaparecer. En general, mientras más específicos son el problema o la hipótesis, más claras resultan sus implicaciones a probar. Sin embargo, el precio que podemos pagar es la trivialidad. Los investigadores no pueden manejar problemas demasiado amplios por su tendencia a ser demasiado vagos en cuanto a las operaciones adecuadas de investigación. Por otro lado, en su entusiasmo por reducir el problema a un tamaño manejable o por encontrar un problema manipulable, pueden terminar con su vida, y convertirlo en trivial o carente de importancia. Por ejemplo, una tesis sobre la simple relación entre velocidad de lectura y tamaño de la letra, por muy interesante e importante que pudiera parecer, resulta débil para un estudio doctoral. El estudiante de ese nivel necesitará ampliar el tema al recomendar una comparación entre géneros y considerar variables como cultura y antecedentes familiares. El investigador podría también expandir el estudio para concentrarse en los niveles de iluminación y

[3] Aunque se ha conducido con éxito una variedad de estudios sobre autoritarismo, no es claro que entendamos lo que significa autoritarismo en el salón de clases. Por ejemplo, una acción de un maestro autoritario en un salón de clases puede no serlo en otra aula. La mencionada conducta democrática exhibida por un maestro puede ser etiquetada como autoritarismo si la muestra otro docente. Tal elasticidad no pertenece a la ciencia.

el tipo de letra. Demasiada especificidad quizás sea más dañina que demasiada generalidad. El investigador puede estar en posición de contestar una pregunta específica pero no podrá generalizar los hallazgos a otras situaciones o grupos de personas. A cualquier precio, alguna clase de compromiso debe establecerse entre generalidad y especificidad. La capacidad para definir tal compromiso de manera efectiva es, en parte, función de la experiencia, y en parte, del estudio crítico de los problemas de investigación.

He aquí algunos ejemplos de problemas de investigación contrastantes, ya sea muy generales o muy específicos:

1)	Demasiado general:	Existen diferencias de género al jugar.
	Demasiado específico:	La puntuación de Carlos será 10 puntos mayor que la de Carol en el juego profesional Tetris.
	Cerca del ideal:	Habrá mayor transferencia de aprendizaje al practicar videojuegos con niños que con niñas.
2)	Demasiado general:	Las personas pueden leer letras de mayor tamaño más rápido que las letras más pequeñas.
	Demasiado específica:	Los alumnos de último año de la escuela Duarte pueden leer tipos de letra de 24 puntos más rápido que tipos de letra de 12 puntos.
	Cerca del ideal:	Una comparación de tres diferentes tamaños de letra y agudeza visual en la velocidad de lectura y de comprensión.

La naturaleza multivariable de la investigación y problemas del comportamiento

Hasta este punto, la discusión de problemas e hipótesis se ha limitado a dos variables, x y y. Debemos corregir cualquier impresión de que tales problemas e hipótesis son la norma en la investigación del comportamiento. Los investigadores en psicología, sociología, educación y otras ciencias del comportamiento se han concientizado de la naturaleza multivariable de este tipo de estudios. En lugar de decir: si p, entonces q, es frecuente y más apropiado decir: si $p_1, p_2 \ldots, p_k$, entonces q; o bien: si p, entonces q, bajo las condiciones r, s y t.

A continuación, un ejemplo que puede aclarar este punto. En lugar de simplemente formular la hipótesis: si hay frustración, entonces hay agresividad, es más realista reconocer la naturaleza multivariable de los determinantes e influencias de la agresividad. Esto se logra al decir, por ejemplo: si se es muy inteligente, de clase media, varón y frustrado, entonces hay agresividad; o bien: si hay frustración, entonces hay agresividad bajo las condiciones de alta inteligencia, clase media y sexo masculino. En lugar de tener una x, nosotros ahora tenemos cuatro x. Aunque un fenómeno puede ser el más importante para determinar o influir en otro fenómeno, es poco probable que la mayoría de los fenómenos de interés para los científicos del comportamiento sean determinados de forma simple. Es mucho más probable que lo sean de manera múltiple. Es mucho más probable que la agresividad sea el resultado de diversas influencias que actúan de forma compleja. Más aún, la agresividad en sí misma contiene múltiples aspectos. Después de todo, hay diferentes clases de agresividad.

Los problemas y las hipótesis, entonces deben reflejar la complejidad multivariable de la realidad psicológica, sociológica y educativa. Hablaremos de una x y una y, en particular en la parte inicial de este libro. Sin embargo, *es preciso* entender que la investigación del

comportamiento, que tuvo un enfoque casi exclusivamente univariado, se ha tornado cada vez más multivariable. Nos hemos propuesto usar la palabra "multivariable" en lugar de "multivariada" por una razón importante. De forma tradicional los estudios "multivariados" son aquéllos que tienen más de una variable *y* y una o más variables *x*. Cuando hablamos de una variable *y* y más de una variable *x* se usa el término "multivariable" que resulta más apropiado para hacer la distinción. Por ahora usaremos "univariado" para indicar una *x* y una *y*. De manera estricta, el término "univariado" también se aplica a *y*. Pronto encontraremos conceptos y problemas de naturaleza multivariada. Secciones posteriores del libro estarán enfocadas en especial a un enfoque y énfasis de este tipo. Para más explicaciones sobre las diferencias entre *multivariable* y *multivariado* (véase Kleinbaum, Kupper, Muller y Nizam, 1997).

Comentarios finales: el poder especial de las hipótesis

A veces se oye decir que las hipótesis son innecesarias en la investigación. Algunos sienten que las hipótesis restringen innecesariamente su imaginación investigadora y que el trabajo de la ciencia y de la investigación científica es descubrir cosas nuevas y no elaborar lo obvio. Algunos piensan que las hipótesis son obsoletas. Tales afirmaciones resultan engañosas y malinterpretan el propósito de las hipótesis.

Casi puede decirse que las hipótesis son uno de los instrumentos más poderosos que ha inventado el hombre para alcanzar un conocimiento confiable. Observamos un fenómeno; especulamos sobre sus causas posibles. Naturalmente, nuestra cultura tiene respuestas para dar cuenta de la mayoría de los fenómenos —muchas correctas, muchas incorrectas, muchas una mezcla de hechos y supersticiones, y muchas pura superstición—. Es obligación del científico dudar de la mayor parte de las explicaciones sobre los fenómenos. Esas dudas son sistémicas. Los científicos insisten en someter las explicaciones sobre los fenómenos a una prueba empírica controlada. Para lograrlo, formulan las explicaciones en términos de teorías e hipótesis. De hecho, las explicaciones constituyen hipótesis. Los científicos sólo disciplinan la cuestión al escribirla en forma de hipótesis sistemáticas y comprobables. Si una explicación no puede formularse en términos de una hipótesis comprobable, deberá considerarse como una explicación metafísica y, por lo tanto, no susceptible de investigación científica. Como tal, los científicos la rechazan como carente de interés.

El poder de las hipótesis va más allá. La hipótesis constituye una predicción: indica que si ocurre *x*, también ocurrirá *y*; esto es, *y* se predice a partir de *x*. Entonces si se hace que *x* ocurra (es decir, que varíe) y se observa que *y* también ocurre (o sea, varía de forma concomitante), entonces la hipótesis se confirma. Resulta una evidencia más poderosa que la simple observación, sin predicción, la covariación de *x* y *y*. Es más poderosa en el sentido de apuesta-juego antes discutido. El científico apuesta a que *x* conduce hacia *y*. Si en un experimento, *x* conduce en efecto a *y*, entonces habrá ganado la apuesta. Una persona no puede sólo entrar en el juego en cualquier momento y observar una ocurrencia común quizá fortuita de *x* y *y*. No se juega de esta forma (al menos no en nuestra cultura). La persona debe jugar de acuerdo con las reglas, y las reglas en la ciencia están hechas para minimizar el error y la falibilidad. Las hipótesis forman parte de las reglas del juego científico.

Aun cuando no se confirmen las hipótesis, tienen poder. Aun cuando *y* no covaríe con *x*, el conocimiento avanza. Los hallazgos negativos en ocasiones resultan tan importantes como los positivos, puesto que reducen el universo total de la ignorancia, y algunas veces señalan hacia otras hipótesis y líneas de investigación. *Pero el científico no puede distinguir la evidencia positiva de la negativa hasta usar una hipótesis.* Por supuesto, es posible conducir

una investigación sin hipótesis, en particular en el caso de estudios exploratorios, pero es difícil concebir a la ciencia moderna en toda su rigurosa y disciplinada fertilidad sin la guía y poder de las hipótesis.

RESUMEN DEL CAPÍTULO

1. Formular un problema de investigación no es una tarea fácil. El investigador empieza con una noción general difusa y vaga que gradualmente se refina. Los problemas de investigación varían en gran medida y no existe un único camino correcto para enunciar el problema.
2. Tres criterios de problemas y enunciados de problema adecuados son:
   a) El problema debe expresarse como una relación entre dos o más variables.
   b) El problema debe ser redactado en forma de pregunta.
   c) El enunciado del problema debe implicar la posibilidad de ser sometido a una prueba empírica.
3. Una hipótesis es un enunciado conjetural de la relación entre dos o más variables. Ésta se redacta en forma de enunciado declarativo. Los criterios para una hipótesis apropiada son los mismos que los usados para los problemas, que se señalan en el punto anterior.
4. La importancia de los problemas e hipótesis es que:
   a) Constituyen un instrumento de trabajo de la ciencia y un enunciado de trabajo específico de la teoría.
   b) Las hipótesis pueden ser sometidas a prueba y ser predictivas.
   c) Contribuyen al avance del conocimiento.
5. Las virtudes de los problemas y de las hipótesis son:
   a) Dirigen la investigación.
   b) Permiten al investigador deducir manifestaciones empíricas específicas.
   c) Sirven como puente entre teoría e investigación empírica.
6. Los problemas científicos no constituyen preguntas éticas y morales. La ciencia no puede contestar preguntas de valor o de juicio.
7. Para detectar preguntas de valor es necesario buscar palabras tales como *mejor que*, *debería*, o *habría que*.
8. Otro defecto común de los enunciados de problema es enlistar aspectos metodológicos como subproblemas. La falla consiste en que:
   a) No son problemas sustantivos que provengan del problema básico en forma directa.
   b) Están relacionados con técnicas o métodos de muestreo, medición, o análisis; no se presentan en forma de pregunta.
9. En cuanto a los problemas, es necesario establecer un equilibrio para que no sea ni demasiado general ni demasiado específico. Esta habilidad se desarrolla con la experiencia.
10. Los problemas e hipótesis deben reflejar la complejidad multivariada de la realidad del ámbito de las ciencias del comportamiento.
11. La hipótesis representa uno de los más poderosos instrumentos inventados para obtener conocimiento confiable. Tiene la capacidad de ser predictiva. Un hallazgo negativo para una hipótesis puede servir para eliminar una posible explicación y generar otras hipótesis y líneas de investigación.

SUGERENCIAS DE ESTUDIO

1. Utilice los siguientes nombres de variables para redactar problemas de investigación e hipótesis: frustración, logro académico, inteligencia, habilidad verbal, raza, clase social (estatus socioeconómico), sexo, reforzamiento, métodos de enseñanza, elección ocupacional, conservadurismo, educación, ingresos, autoridad, necesidad de logro, cohesión de grupo, obediencia, prestigio social, permisividad.
2. A continuación se presentan diez problemas de investigación tomados de la literatura. Estúdielos con cuidado, elija dos o tres y construya hipótesis con base en ellos.
   a) ¿Tienen diferentes puntuaciones en una prueba de ansiedad los niños de diferentes grupos étnicos? (Guida y Ludlow, 1989)
   b) ¿Las situaciones de cooperación social conducen a mayores niveles de motivación intrínseca? (Hom, Berger, Duncan, Miller y Belvin, 1994)
   c) ¿Las expresiones faciales de las personas influyen en las respuestas afectivas? (Strack, Martin y Stepper, 1988)
   d) ¿Respetarán los jurados las instrucciones e indicaciones judiciales prohibitivas? (Shaw y Skolnick, 1995)
   e) ¿Cuáles son los efectos positivos del uso de cojines de presión alternante para prevenir llagas en pacientes terminales atendidos en casa? (Stoneberg, Pitcock y Myton, 1986)
   f) ¿Cuáles son los efectos del condicionamiento pavloviano temprano en el condicionamiento pavloviano tardío? (Lariviere y Spear, 1996)
   g) ¿Depende la eficacia de la codificación de información en la memoria de largo plazo de lo novedoso que ésta sea? (Tulving y Kroll, 1995)
   h) ¿Cuál es el efecto del consumo de alcohol en la probabilidad de uso del condón durante el sexo ocasional? (MacDonald, Zanna y Fong, 1996)
   i) ¿Hay diferencias por género para predecir las decisiones relativas al retiro? (Talaga y Beehr, 1995)
   j) ¿Es el Juego de Buena Conducta una estrategia de intervención viable para niños que requieren procedimientos de cambio de comportamiento en el aula? (Tingstrom, 1994)
3. A continuación se presentan diez hipótesis. Discuta las posibilidades de someterlas a prueba. Después, lea dos o tres de los estudios para entender cómo lo hicieron los autores.
   a) Los solicitantes de trabajo que expresan una gran experiencia en tareas no existentes sobreestiman sus habilidades en tareas reales (Anderson, Warner y Spencer, 1984).
   b) En situaciones sociales, los hombres malinterpretan las expresiones amistosas de las mujeres como un signo de interés sexual (Saal, Johnson y Weber, 1989).
   c) A mayor éxito de un equipo, mayor será la atribución que cada miembro haga a su habilidad y suerte personales (Chambers y Abrami, 1991).
   d) El incremento de interés en una tarea aumentará la conformidad (Rind, 1997).
   e) Extractos de la sudoración del hombre pueden afectar el ciclo menstrual de la mujer (Cutler, Preti, Kreiger y Huggins, 1986).
   f) Las personas atractivas físicamente se consideran más inteligentes que las personas no atractivas (Moran y McCullers, 1984).
   g) Uno puede recibir ayuda de un extraño si éste es similar a uno mismo, o si la petición se hace a una cierta distancia (Glick, DeMorest, y Hotze, 1988).
   h) Fumar cigarrillos (nicotina) mejora el desempeño mental (Spilich, June y Remer, 1992).

i) Quienes guardan objetos valiosos en lugares extraños tendrán un mayor recuerdo del sitio que si colocaran estos artículos en lugares comunes (Winograd y Soloway, 1986).

j) Los hombres homosexuales con VIH sintomático presentan significativamente más angustia que aquéllos que desconocen su estatus de VIH (Cochran y Mays, 1994).

4. Los problemas e hipótesis multivariadas (por ahora, más de dos variables dependientes) son ahora comunes en la investigación del comportamiento. Para familiarizar al estudiante con tales problemas, hemos anexado algunos. Trate de imaginar cómo desarrollaría usted la investigación para estudiarlos.

a) ¿Difieren hombres y mujeres en cuanto a sus percepciones acerca de sus genitales, gozo sexual, sexo oral y masturbación? (Reinholtz y Muehlenhard, 1995)

b) ¿Son los fumadores jóvenes más extrovertidos mientras que los fumadores de más edad son más depresivos y aislados? (Stein, Newcomb y Bentler, 1996)

c) ¿Cuánto difiere la apreciación que los maestros tienen de las habilidades sociales de los estudiantes populares y los rechazados? (Frentz, Gresham y Elliot, 1991; Stuart, Gresham y Elliot 1991)

d) ¿Influye el grado de semejanza del consejero y del cliente en cuanto a grupo étnico, género y lenguaje en los resultados del tratamiento para niños de edad escolar? (Hall, Kaplan y Lee, 1994)

e) ¿Existen diferencias en las habilidades cognitivas y funcionales de pacientes con Alzheimer que residen en una unidad de cuidado especial en relación con aquellos que viven en una unidad de cuidados tradicional? (Swanson, Maas y Buckwalter, 1994)

f) ¿Difieren los niños hiperactivos con déficit de atención de los niños no hiperactivos con déficit de atención en cuanto a rendimiento en lectura, ortografía y lenguaje escrito? (Elbert, 1993)

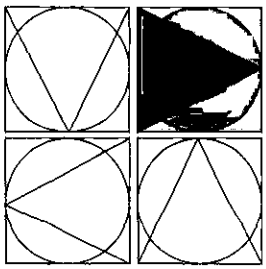
g) ¿La gente ve a las mujeres que prefieren el título de cortesía de señorita como poseedoras de mayores cualidades instrumentales y de menores cualidades de expresividad que las mujeres que prefieren los títulos de cortesía tradicionales? (Dion y Cota, 1991)

h) ¿Aumentará el estilo de liderazgo autoritario la satisfacción de los miembros del grupo? ¿Aumentará la percepción sobre la eficacia del grupo de trabajo su efectividad? (Kumpfer, Turner, Hopkins y Librett, 1993)

i) ¿Cómo influyen el grupo étnico, el género y los antecedentes socioeconómicos en la propensión a la psicosis: aberración perceptiva, ideación mágica y personalidad esquizoide? (Porch, Ross, Hanks y Whitman, 1995)

j) ¿Tendrá la exposición a los estímulos dos efectos, uno cognitivo y otro afectivo, que a su vez afecten la predilección, la familiaridad, certeza y precisión en el reconocimiento? (Zajonc, 1980)

Los últimos dos problemas y estudios resultan muy complejos en tanto que las relaciones establecidas son complejas. Los otros problemas y estudios, aunque complejos, poseen tan sólo un fenómeno presumiblemente afectado por otro, mientras que los últimos dos contienen varios fenómenos que afectan a dos o más fenómenos. El lector no deberá desalentarse si encuentra en ellos algo de dificultad: hacia el final del libro parecerán interesantes y naturales.



CAPÍTULO 3

CONSTRUCTOS, VARIABLES Y DEFINICIONES

- CONCEPTOS Y CONSTRUCTOS
- VARIABLES
- DEFINICIONES CONSTITUTIVAS Y OPERACIONALES DE CONSTRUCTOS Y VARIABLES
- TIPOS DE VARIABLES
  - **Variables independientes y dependientes**
  - **Variables activas y variables atributo**
  - **Variables continuas y categóricas**
- CONSTRUCTOS OBSERVABLES Y VARIABLES LATENTES
- EJEMPLOS DE VARIABLES Y DEFINICIONES OPERACIONALES

Los científicos operan en dos niveles: teoría-hipótesis-constructo y observación. Para ser más exactos, oscilan entre uno y el otro de forma continua. Un psicólogo científico podría decir: "La deprivación temprana produce deficiencia en el aprendizaje." Este enunciado es una hipótesis integrada por dos conceptos, "deprivación temprana" y "deficiencia en el aprendizaje", unidos por una palabra de relación, *produce*. Este enunciado se encuentra en el nivel teoría-hipótesis-constructo. Cuando los científicos formulan enunciados relacionales y utilizan conceptos, o constructos, como los llamaremos, están operando en este nivel.

Los científicos también deben operar en el nivel de observación. Deben reunir datos para probar sus hipótesis. Para hacerlo, es necesario que pasen del nivel de constructo al de observación. No pueden simplemente hacer observaciones de "deprivación temprana" y "deficiencia en el aprendizaje". Es necesario que definan estos constructos de modo que sea posible realizar las observaciones. El problema que se estudia en este capítulo es cómo examinar y aclarar la naturaleza de los conceptos científicos o constructos. Este capítulo también examinará y aclarará la forma en que los científicos del comportamiento pasan del nivel de constructo al de observación, cómo van de uno a otro.

Conceptos y constructos

Los términos "concepto" y "constructo" tienen significados similares, aunque existe una diferencia importante. Un "concepto" expresa una abstracción creada por una generalización a partir de instancias particulares. "Peso" es un concepto que expresa numerosas observaciones de cosas que son "más o menos" y "pesadas o ligeras". "Masa", "energía" y "fuerza" son conceptos usados por científicos de la física. Por supuesto, son mucho más abstractos que conceptos como "peso", "alto" y "longitud".

Un concepto de mayor interés para los lectores es el "aprovechamiento". Es una abstracción que se genera a partir de la observación de ciertos comportamientos de los niños, que se asocian con el dominio del "aprendizaje" en tareas escolares —lectura de palabras, resolución de problemas aritméticos, elaboración de dibujos, etcétera—. Los diversos comportamientos observados se reúnen y expresan en una palabra. "Aprovechamiento", "inteligencia", "agresividad", "conformidad" y "honestidad" son conceptos usados para expresar la variedad del comportamiento humano.

Un *constructo* es un concepto, que tiene el significado agregado de haber sido inventado o adoptado para un propósito científico especial, de forma deliberada y consciente. **"Inteligencia" es un concepto, una abstracción de la observación de comportamientos** presumiblemente inteligentes y no inteligentes. Sin embargo, como todo constructo científico, "inteligencia" implica tanto más como menos de lo que pueda significar como concepto. Esto quiere decir que los científicos de manera consciente y sistemática la usan en las dos formas: 1) se incorpora en los esquemas teóricos y se relaciona en diversas formas con otros constructos (podemos decir, por ejemplo, que el aprovechamiento escolar es en parte función de la inteligencia y motivación) y 2) "inteligencia" se define y especifica de tal forma que pueda ser observada y medida (podemos hacer observaciones de la inteligencia de los niños al aplicar una prueba de inteligencia, o al solicitar a los maestros que señalen el grado relativo de inteligencia de sus alumnos).

Variables

Los científicos de forma algo vaga, llaman a los constructos o propiedades que estudian, "variables". Algunos ejemplos de variables importantes en sociología, psicología, ciencia política y educación son: género, ingreso, educación, clase social, productividad organizacional, movilidad ocupacional, nivel de aspiración, aptitud verbal, ansiedad, afiliación religiosa, preferencias políticas, desarrollo político (de las naciones), orientación ocupacional, prejuicios raciales y étnicos, conformidad, recuerdo, memoria de reconocimiento y aprovechamiento. Puede decirse que una variable es una propiedad que asume diversos valores. Siendo redundantes, una variable es algo que varía. Aunque esta forma de expresarlo nos aporta una noción intuitiva de lo que son, necesitamos una definición al mismo tiempo más general y precisa.

Una *variable* es un símbolo al que se le asignan valores o números. Por ejemplo, x es una variable: es un *símbolo* al que se le asignan valores numéricos. La variable x puede tomar cualquier conjunto justificable de valores, por ejemplo, puntajes en una prueba de inteligencia o en una escala de actitudes. En el caso de la inteligencia, asignamos a x un conjunto de valores numéricos proporcionados por el procedimiento especificado en una determinada prueba de inteligencia. Este grupo de valores varía de bajo a alto, por ejemplo de 50 a 150.

Una variable, x, sin embargo, puede tener sólo dos valores. Si el género es el constructo bajo estudio, entonces a x se le pueden asignar 1 y 0, donde 1 representa uno de los géne-

ros y 0 el otro. Aún así, es una variable. Otros ejemplos de variables con dos valores son: dentro–fuera, correcto–incorrecto, viejo-joven, ciudadano-no ciudadano, clase media-clase trabajadora, maestro-no maestro, republicano-demócrata, etcétera. Tales variables se llaman *dicotomías*, variables dicotómicas o binarias.

Algunas de las variables usadas en la investigación conductual son verdaderas dicotomías, es decir, se caracterizan por la presencia o ausencia de una propiedad: masculino-femenino, con hogar-indigente, empleado-desempleado. Otras variables son *politomías*. Un buen ejemplo es la preferencia religiosa: protestante, católico, musulmán, judío, budista, otra. Tales dicotomías y politomías se denominan "variables cualitativas". La naturaleza de este calificativo se analizará más adelante. En teoría, muchas variables, sin embargo, pueden asumir valores continuos. Ha sido una práctica común en la investigación del comportamiento convertir las variables continuas en dicotómicas o politómicas. Por ejemplo, la inteligencia, una variable continua, se ha dividido en inteligencia alta, media y bajas, o en alta y baja. Variables como ansiedad, introversión y autoritarismo han recibido un trato similar. Aunque no es posible convertir una variable que de manera natural es dicotómica, como género, en una variable continua, siempre podemos convertir una variable continua en una dicotómica o politómica. Más adelante se verá que tal conversión puede tener un propósito conceptual útil, pero para el análisis de datos constituye una práctica negativa en tanto que descarta información.

Definiciones constitutivas y operacionales de constructos y variables

La diferencia que hicimos antes entre "concepto" y "constructo" conduce de manera natural a otra distinción importante entre los tipos de definiciones de constructos y variables. Se puede definir a las palabras o constructos de dos formas generales: primero, podemos definir una palabra con el uso de otras palabras, que es lo que hace un diccionario. Podemos definir *inteligencia* como "un intelecto operante", "agudeza mental" o "la habilidad para pensar de forma abstracta". Tales definiciones utilizan otros conceptos o expresiones conceptuales en lugar de la expresión o palabra que se define. Segundo, podemos definir una palabra por las acciones o comportamientos que expresa o implica. Definir *inteligencia* de esta forma requiere que especifiquemos qué comportamientos de los niños son "inteligentes" y cuáles son "no inteligentes". Podemos decir que un niño de siete años de edad que lee una historia con éxito es "inteligente", si el niño no puede leer la historia podemos asumir que el chico "no es inteligente". En otras palabras, esta clase de definición puede llamarse *definición observacional* o *conductual*. Se usan definiciones a partir de "otras palabras" y definiciones "observacionales" de manera cotidiana.

En esta discusión hay una imprecisión perturbadora. Aunque los científicos utilizan los tipos de definiciones que acabamos de describir, lo hacen en una forma más precisa. Expresamos este uso al definir y explicar la diferencia que Margenau (1950/1977) plantea para las definiciones constitutivas y operacionales. Una definición *constitutiva* define un constructo usando otros constructos. Por ejemplo, podemos definir *peso* diciendo que es la "pesadez" de los objetos, o *ansiedad* como "temor subjetivo". En ambos casos hemos sustituido un concepto por otro. Algunos de los constructos de una teoría científica pueden definirse de manera constitutiva. Torgerson (1958/1985), a partir de las ideas de Margenau, indica que para ser útiles desde el punto de vista científico, todos los constructos deben poseer significado constitutivo, es decir, poder ser usados en teorías.

Una definición *operacional* asigna significado a un constructo o variable al especificar las actividades u "operaciones" necesarias para medirlo y evaluar la medición. De manera

alternativa, una definición operacional constituye una especificación de las actividades del investigador para medir una variable o para manipularla. Implica algo así como un manual de instrucciones para el investigador. En efecto, dice, "haga tal y cual, de la forma tal y tal". En síntesis, define o aporta significado a una variable al delinear paso a paso lo que el investigador debe hacer para medirla y para evaluar dicha medición.

Michel (1990) presenta una excelente revisión histórica de cómo las definiciones operacionales se hicieron populares en las ciencias sociales y del comportamiento. Michel cita a P. W. Bridgeman, premio Nobel, como creador de la definición operacional en 1927. Bridgeman, como lo relata Michel (1990, p. 15), indica: "en general, con cualquier concepto dado queremos significar tan sólo un conjunto de operaciones; *el concepto es sinónimo del correspondiente conjunto de operaciones*". Cada operación diferente definiría un concepto distinto.

Un ejemplo bien conocido, aunque extremo, de una definición operacional es: inteligencia (ansiedad, aprovechamiento, etcétera) es la puntuación en una prueba *X* de inteligencia, o inteligencia es lo que la prueba *X* de inteligencia mide. Las puntuaciones altas indican un mayor nivel de inteligencia que las bajas. Esta definición indica qué hacer para medir la inteligencia y no precisa qué tan bien es medida la inteligencia por el instrumento especificado. (Se presume que se indagó sobre la adecuación de esta prueba antes de que el investigador la usara.) En este tipo de uso, una definición operacional equivale a una ecuación en la que decimos: "si la inteligencia es igual a la puntuación en la prueba *X* de inteligencia, las puntuaciones altas indican un mayor grado de inteligencia que los bajos". Parecería también que estamos diciendo "el significado de inteligencia (en este estudio) se expresa por el puntaje en la prueba *X* de inteligencia".

Existen, en general, dos clases de definiciones operacionales: 1) las *medidas* y 2) las *experimentales*. La definición de arriba está más estrechamente ligada con las definiciones medidas que con las experimentales. Una definición operacional *medida* describe cómo será medida una variable. Por ejemplo, aprovechamiento puede definirse por una prueba estandarizada de aprovechamiento, por un examen desarrollado por el maestro, o por las calificaciones. Doctor, Cutris e Isaacs (1994), al estudiar el efecto de la consejería para el estrés en oficiales de policía, definieron de manera operacional morbilidad psiquiátrica como las puntuaciones en el Cuestionario General de Salud y el número de días que habían tomado por incapacidad. Altas puntuaciones y un gran número de días indicaban niveles elevados de morbilidad. Little, Sterling y Tingstrom (1996) estudiaron los efectos de la raza y origen geográfico en la atribución. La atribución se definió operacionalmente como la puntuación en el cuestionario de estilo atribucional. Un estudio puede incluir la variable *consideración* que puede definirse operacionalmente a través de una lista de comportamientos de niños que se presume son comportamientos considerados. Después, se puede pedir a los maestros que evalúen a los alumnos en una escala de cinco puntos. Tales comportamientos pueden ejemplificarse como instancias en que un niño le dice a otro "lo siento" o "discúlpame", o cuando un chico le presta un juguete a otro o cuando se lo pide (sin la amenaza de agresión), o cuando un pequeño ayuda a otro con una tarea escolar. También puede definirse por la suma de comportamientos considerados: a mayor cantidad, mayor nivel de consideración.

Una definición operacional *experimental* señala los detalles (operaciones) de la manipulación de una variable por parte del investigador. El reforzamiento puede definirse operacionalmente al precisar los detalles de cómo los sujetos serán reforzados (premiados) y no reforzados (no premiados) por comportamientos específicos. Hom, Berger, Duncan, Miller y Belvin (1994) definieron operacionalmente el reforzamiento en forma experimental. En este estudio, los niños fueron asignados a uno de cuatro grupos. Dos de los grupos estuvieron sujetos a una condición de reforzamiento con base en cooperación,

mientras que los otros dos trabajaron en un esquema en que se reforzaba la postura individualista. Bahrick (1984) definió la memoria a largo plazo en términos de al menos dos procesos en lo referente a la retención de información de tipo académico. Un proceso llamado "almacenaje permanente" (*permastore*), elige de manera selectiva alguna información para ser almacenada de forma permanente y resulta muy resistente al decaimiento (olvido). El otro proceso al parecer elige información menos significativa por lo que es menos resistente al olvido. Esta definición contiene implicaciones claras para la manipulación experimental. Strack, Martin y Stepper (1988) definieron operacionalmente la sonrisa como la activación de los músculos asociados con la sonrisa humana. Lo lograron al solicitar que una persona sostuviera una pluma en su boca de cierta manera. Se trata de un procedimiento no intrusivo en tanto que no se les pidió a los participantes que posaran con una sonrisa. Se presentarán otros ejemplos de ambos tipos de definiciones más adelante.

Los investigadores científicos eventualmente enfrentan la necesidad de medir las variables de las relaciones que estudian. Algunas mediciones son fáciles, otras difíciles. Medir el género o la clase social es fácil; pero evaluar creatividad, conservadurismo o efectividad organizacional resulta difícil. La importancia de las definiciones operacionales no puede dejar de enfatizarse. Ellas son ingredientes indispensables de la investigación científica porque permiten al investigador medir variables y porque representan puentes entre el nivel de la teoría-hipótesis-constructo y el nivel de observación. No hay investigación científica sin observaciones, y éstas no son posibles sin instrucciones claras y específicas de qué y cómo observar. Las definiciones operacionales son tales instrucciones.

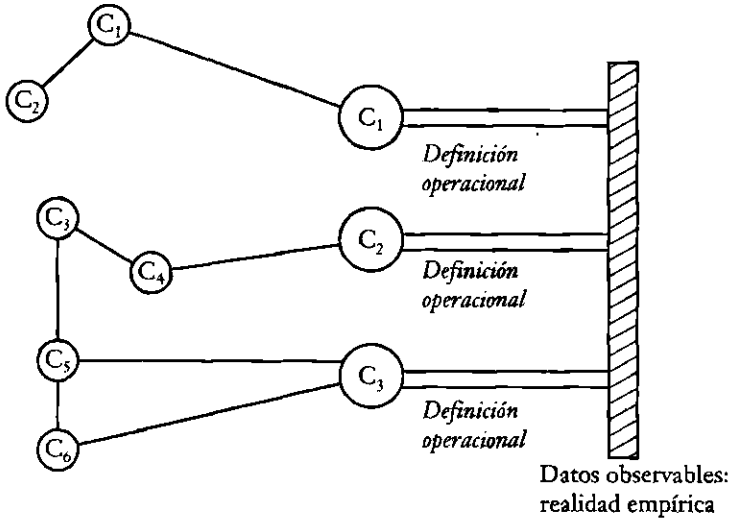
Aunque indispensables, las definiciones operacionales sólo aportan significados limitados de los constructos. Ninguna definición operacional puede expresar toda la riqueza y los diversos aspectos de algunas variables, como sucede con el prejuicio humano. Esto implica que las variables medidas por los científicos siempre tienen un significado limitado y específico. La "creatividad" que estudian los psicólogos no se refiere necesariamente a la "creatividad" de los artistas, aunque, por supuesto, tengan elementos comunes. Una persona que piensa en una solución creativa para un problema matemático puede mostrar una escasa creatividad como poeta (Barron y Harrington, 1981). Algunos psicólogos han definido operacionalmente a la creatividad como el desempeño en la prueba de Torrance de Pensamiento Creativo (Torrance, 1982). Los niños que obtienen una puntuación alta en esta prueba, tienen mayor probabilidad de exhibir logros creativos en la edad adulta.

Algunos científicos afirman que estos limitados significados operacionales son los únicos significados que "significan" algo, que todas las demás definiciones son disparates metafísicos. Señalan que las discusiones sobre ansiedad constituyen tonterías de tipo metafísico, a menos que se cuente con y se usen, definiciones operacionales adecuadas. Éste es un punto de vista extremo, aunque posee algunos aspectos saludables. Insistir en que cada término que usemos en el discurso científico sea definido operacionalmente sería demasiado reduccionista y restrictivo y, como se verá, científicamente erróneo. Northrop (1947/1983, p. 130) señala, por ejemplo: "La importancia de las definiciones operacionales estriba en que hacen posible la verificación y enriquecen el significado. Sin embargo, no agotan el significado científico". Margenau (1950/1977, p. 232) señala el mismo punto en su extensa discusión sobre los constructos científicos.

A pesar de los riesgos del operacionalismo extremo, parece seguro decir que ha sido y es una influencia saludable. Como señala Skinner (1945, p. 274):

> La actitud operacional, a pesar de sus inconvenientes, es positiva en cualquier ciencia, pero en especial en psicología, por la presencia de un vasto vocabulario de origen antiguo y no científico.

FIGURA 3.1



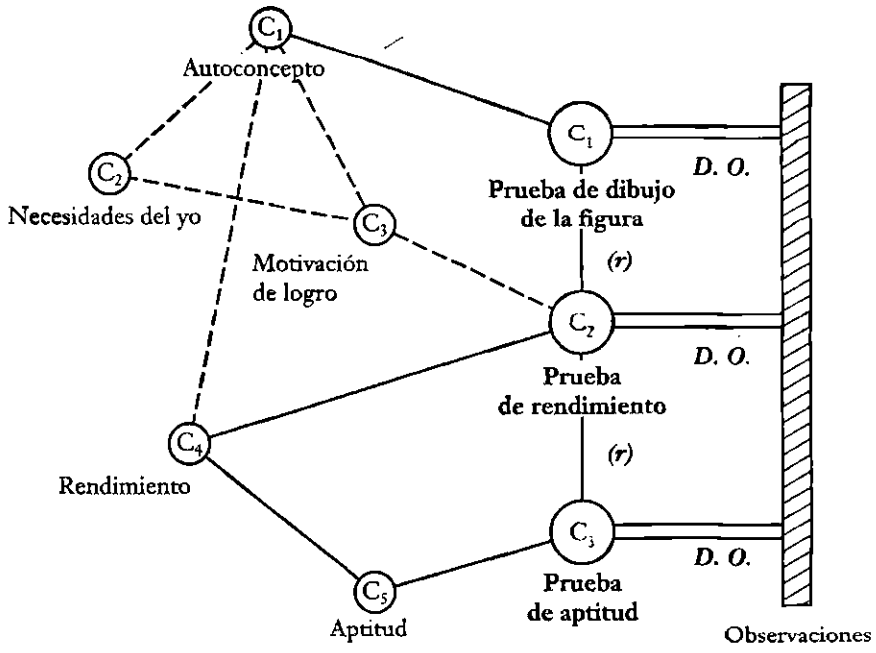
Cuando se consideran los términos usados en educación, es claro que esta disciplina también posee un vasto vocabulario de origen antiguo y no científico. Consideremos éstos: el niño integral, el enriquecimiento horizontal y vertical, la satisfacción de las necesidades del alumno, el tronco común, el ajuste emocional y el enriquecimiento curricular. Esto también se aplica al campo de la atención geriátrica. Aquí la enfermera especializada maneja términos como el proceso de envejecimiento, la autoimagen, el mantenimiento de la atención y la negligencia unilateral (Eliopoulos, 1993; Smeltzer y Bare, 1992).

Para aclarar las definiciones constitutivas y operacionales (así como la teoría) veamos la figura 3.1, que ha sido adaptada de Margenau (1950/1977) y de Torgerson (1958/1985). El diagrama intenta ilustrar una teoría bien desarrollada. Las líneas sencillas representan conexiones teóricas o relaciones entre constructos. Estos constructos, etiquetados con letras minúsculas, se definen de manera constitutiva; esto es, c_4 está definida de alguna forma por c_3 o viceversa. Las líneas dobles representan definiciones operacionales. Los constructos están ligados de forma directa a datos observables, y son vinculaciones indispensables con la realidad empírica. Sin embargo, no todos los constructos en una teoría científica se definen operacionalmente. De hecho, la teoría que tiene todos sus constructos así definidos es algo tenue.

Construyamos una "pequeña teoría" del bajo rendimiento para ilustrar estos conceptos. Supongamos que un investigador cree que el bajo rendimiento es en parte una función del autoconcepto de los alumnos. Piensa que los estudiantes que se perciben como inadecuados y que tienen percepciones negativas de sí mismos, también tienden a rendir menos que lo que su capacidad potencial y aptitudes indican. De aquí se sigue que las necesidades del yo (que no definiremos aquí) y la motivación de logro (que llamaremos necesidad de logro) están ligadas al bajo rendimiento. Como es natural, el investigador está consciente de la relación entre aptitud e inteligencia y logro en general. Un diagrama que ilustra esta "teoría" puede ser como el de la figura 3.2.

El investigador no cuenta con una medida *directa* de autoconcepto, pero supone que puede inferirlo a partir de una prueba de dibujo de la figura. Entonces define operacional-

FIGURA 3.2



mente el autoconcepto como ciertas respuesta a esa prueba. Éste es probablemente el método más común para medir constructos psicológicos (y educativos). La línea gruesa entre c_1 y C_1 indica la naturaleza relativamente directa de la relación que se asume entre el autoconcepto y la prueba. (La línea doble entre C_1 y el nivel de observación indica una definición operacional, como en la figura 3.1.)

De forma parecida, el constructo de rendimiento (c_4) se define operacionalmente como la discrepancia entre la medición realizada del rendimiento (C_2) y la de la aptitud (c_5). En este modelo el investigador no mide directamente la motivación de logro, ni tampoco posee una definición operacional de ella. En otro estudio se puede hipotetizar específicamente una relación entre rendimiento y motivación de logro, en cuyo caso se tratará de definir motivación para el logro en forma operacional.

Una línea continua sencilla entre conceptos, por ejemplo la que está entre el constructo rendimiento (c_4) y prueba de aprovechamiento (C_2), indica una relación relativamente bien establecida entre el rendimiento postulado y lo que miden las pruebas estandarizadas de rendimiento. La líneas continuas sencillas entre C_1 y C_2 y aquellas entre C_2 y C_3 indican relaciones obtenidas entre las puntuaciones de las pruebas de estas mediciones. (Las líneas entre C_1 y C_2, y entre C_2 y C_3 están marcadas como (r) para significar "relación" o "coeficiente de correlación".)

Las líneas discontinuas representan relaciones postuladas entre constructos que no están relativamente bien establecidas. Un buen ejemplo de esto es la relación postulada entre el autoconcepto y la motivación de logro. Uno de los propósitos de la ciencia es convertir estas líneas punteadas en continuas al cerrar la brecha entre definición-opera-

cional-medición. En este caso, es concebible que tanto el autoconcepto como la motivación de logro puedan ser definidas operacionalmente y medidas directamente.

En esencia, ésta es la forma en que el científico del comportamiento opera. Este especialista se desplaza de manera continua entre el nivel del constructo y el de la observación y lo logra al definir operacionalmente las variables de la teoría que pueden serlo. Luego, se estiman las relaciones entre la definición operacional y las variables medidas. De estas relaciones estimadas, el científico abstrae inferencias acerca de las relaciones entre los constructos. En el ejemplo anterior, el científico del comportamiento calcula la relación entre C_1 (la prueba de dibujo de la figura) y C_2 (prueba de rendimiento). Si la relación se establece en este nivel observacional, el científico infiere que existe una relación entre c_1 (autoconcepto) y c_4 (rendimiento).

Tipos de variables

Variables independientes y dependientes

Dejemos atrás los fundamentos de las definiciones para regresar a las variables. Es posible clasificar las variables de diversas formas. En este libro, tres tipos de variables son muy importantes y serán enfatizadas: 1) variables independientes y dependientes, 2) variables activas y atributo, y 3) variables continuas y categóricas.

La forma más útil de categorizar las variables es como independientes o dependientes. Esta taxonomía resulta muy útil debido a su aplicabilidad general, su simplicidad y su importancia especial tanto en la conceptualización como en el diseño de la investigación, así como en la comunicación de los resultados de ésta. Una *variable independiente* es la causa *supuesta* de la *variable dependiente*, el efecto *supuesto*. La variable independiente es el antecedente; la dependiente es el consecuente. Dado que uno de los objetivos de la ciencia es descubrir relaciones entre diferentes fenómenos, la búsqueda de las relaciones entre variables independientes y dependientes lo logra. Se asume que la variable independiente influye en la dependiente. En algunos estudios, la variable independiente "causa" cambios en la variable dependiente. Cuando decimos: si A, entonces B, tenemos una conjunción condicional de una variable independiente (A) y una variable dependiente (B).

Los términos "variable independiente" y "variable dependiente" proceden de las matemáticas, donde X es la variable independiente y Y es la dependiente. Ésta es probablemente la mejor forma de pensar en las variables independientes y dependientes porque no hay necesidad de utilizar la discutida palabra *causa* y las palabras afines a ella, y dado que el uso de tales símbolos se aplica a la mayoría de las situaciones de investigación. No hay restricción teórica alguna en cuanto a la cantidad de X y Y. Cuando más adelante consideremos el pensamiento y análisis multivariado, trataremos con diversas variables dependientes e independientes.

En los experimentos, el investigador manipula la variable independiente. Los cambios en los valores o niveles de la variable independiente generan cambios en la variable dependiente. Cuando investigadores del campo educativo estudiaron los efectos de diferentes métodos de enseñanza en el desempeño en una prueba de matemáticas, ellos variaron los métodos de enseñanza. En una condición pudieron tener "sólo exposición frontal", en la otra pudo haber "exposición frontal y video". El método de enseñanza es la variable independiente. La variable resultado, la puntuación en la prueba de matemáticas, es la variable dependiente.

La asignación de participantes a diferentes grupos con base en la existencia de alguna característica es un ejemplo de cuando el investigador no puede manipular la variable

independiente. En esta situación, los valores de la variable independiente son preexistentes. El participante tiene la característica o no la tiene. En este caso, no hay posibilidad de una manipulación experimental, pero se considera que "lógicamente" la variable tiene algún efecto en la variable dependiente. Las variables de características del sujeto constituyen la mayor parte de este tipo de variables independientes. Una de las variables independientes más comunes de este tipo es el género (femenino y masculino). Así, si un investigador desea determinar si hombres y mujeres difieren en las destrezas matemáticas, se aplicaría una prueba matemática a representantes de ambos grupos, y se compararían las puntuaciones de la prueba. La prueba de matemáticas sería la variable dependiente. Una regla general es que cuando el investigador manipula una variable o asigna participantes a los grupos según alguna característica, esa variable es la independiente. La tabla 3.1 muestra una comparación entre los dos tipos de variables independientes y su relación con la variable dependiente. La variable independiente debe tener al menos dos niveles o valores. Observe en la tabla 3.1 que ambas situaciones presentan dos niveles para la variable independiente.

La variable dependiente es, por supuesto, *hacia* la que se hace la predicción, mientras que la independiente es aquella *a partir de* la cual se predice. La variable dependiente, Y, es el efecto supuesto, que varía de manera concomitante a los cambios o variaciones en la variable independiente, X; es la variable que se observa para detectar variaciones como un resultado supuesto de la variación en la variable independiente. La variable dependiente es el resultado medido que el investigador usa para determinar si los cambios en la variable independiente tuvieron un efecto. Al predecir Y a partir de X, podemos tomar cualquier valor de X que deseemos, mientras que el valor de Y que predecimos *es* "dependiente" del valor de X que hemos elegido. En general, la variable dependiente es la condición que tratamos de explicar. Por ejemplo, la variable dependiente más común en educación, es "aprovechamiento" o "aprendizaje". Deseamos explicar o dar cuenta del aprovechamiento. Para ello tenemos un gran número de posibles X o variables independientes de dónde elegir.

Cuando se estudia la relación entre inteligencia y aprovechamiento escolar, la inteligencia es la variable independiente y el aprovechamiento es la dependiente. (¿Se podría concebir a la inversa?) Otras variables independientes que pueden estudiarse con relación a la variable dependiente aprovechamiento son: clase social, métodos de enseñanza, tipos de personalidad, tipos de motivación (recompensa y castigo), actitudes hacia la escuela y ambiente en el salón de clases, entre otros. Cuando se estudian los supuestos determinantes de la delincuencia, aquellos tales como condiciones de pobreza, hogares desintegrados, falta de amor de los padres y aspectos similares, constituyen las variables independientes y,

TABLA 3.1 *Relación de variables independientes manipuladas y no manipuladas con la variable dependiente*

Niveles de la variable independiente			
Método de enseñanza		**Género**	
Sólo exposición frontal Puntuaciones en la prueba de matemáticas	Exposición frontal y video Puntuaciones en la prueba de matemáticas	Femenino Puntuaciones en la prueba de matemáticas	Masculino Puntuaciones en la prueba de matemáticas
Variables dependientes		Variables dependientes	

como es natural, la delincuencia (o mejor aún, el comportamiento delictivo) es la variable dependiente. En la hipótesis frustración-agresión, frustración es la variable independiente, y agresión, la dependiente. En ocasiones, un fenómeno se estudia por sí mismo y ya sea la variable dependiente o la independiente están implícitas. Es el caso en que se estudian los comportamientos y características del maestro. La variable dependiente implícita común es el aprovechamiento o el comportamiento del niño, aunque el comportamiento del maestro puede, por supuesto, constituir una variable dependiente. Consideremos un ejemplo en el campo de la atención médica. Cuando se comparan medidas cognitivas y funcionales de pacientes con Alzheimer entre instituciones de internamiento tradicional y unidades de cuidado especial, la variable independiente es el lugar de atención. Las variables dependientes son las medidas cognitivas y funcionales (Swanson, Maas y Buckwalter, 1994).

La relación entre una variable independiente y una dependiente se puede entender mejor si trazamos dos ejes perpendiculares uno del otro. Uno representa a la variable independiente; el otro, a la dependiente. (Cuando dos ejes forman ángulos rectos entre sí se denominan ejes *ortogonales*.) De acuerdo a la tradición matemática, x, la variable independiente, es el eje horizontal y y, la dependiente, representa el eje vertical (x se denomina la abscisa y y la ordenada). Los valores para x se grafican en el eje de las x, y los valores de y en el eje de las y.

Una forma común y útil de "ver" e interpretar una relación es graficar un par de valores de xy, usando los ejes x y y como marco de referencia. En un estudio de desarrollo infantil, supongamos que tenemos dos grupos de medidas. Las medidas x de edad cronológica y las medidas y representan edad lectora. *La edad lectora* se denomina también edad de crecimiento. Las mediciones en serie de diferentes áreas del crecimiento de los individuos —por ejemplo estatura, peso o inteligencia— se expresan como la edad cronológica promedio en la que aparecen en la población estándar.

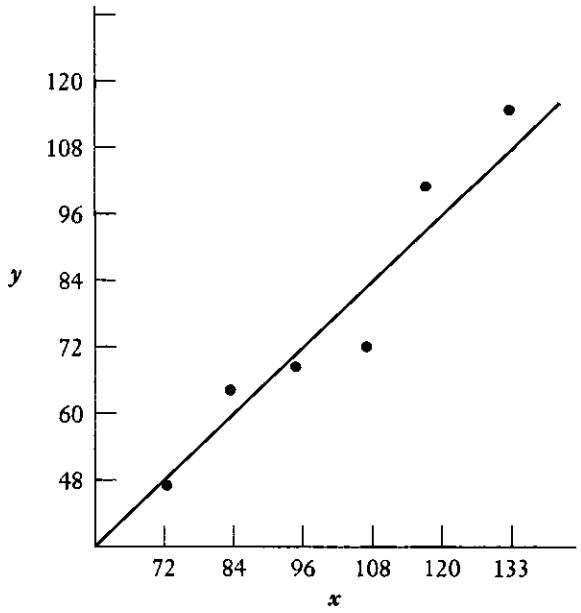
x: edad cronológica (en meses)	y: edad lectora (en meses)
72	48
84	62
96	69
108	71
120	100
132	112

Estas medidas se grafican en la figura 3.3.

La relación entre edad cronológica (EC) y edad lectora (EL), ahora puede "verse" y aproximarse de forma burda. Observe que hay una tendencia pronunciada (como se podría esperar), para que una mayor EC se asocie con una mayor EL, una EC media con una EL media, y una EC menor, con una EL menor. En otras palabras, la relación entre las variables independiente y dependiente, en este caso EC y EL, puede observarse en una gráfica como la que aparece en la figura 3.3. Se ha trazado una recta para "mostrar" esta relación: constituye un promedio aproximado de todos los puntos de la gráfica. Observe que si uno conoce las medidas de la variable independiente y una relación como la que se muestra en la figura 3.3, uno puede predecir, con considerable precisión, las medidas de la variable dependiente. Gráficas como ésta pueden usarse, desde luego, con cualesquier grupo de medidas para variables dependientes e independientes.

El estudiante debe estar atento a la posibilidad de que en un estudio una variable sea independiente, mientras en otro sea dependiente, e inclusive ambas en un mismo estudio.

FIGURA 3.3



Un ejemplo es la satisfacción laboral. La mayoría de los estudios sobre satisfacción laboral la utilizan como variable dependiente. Day y Schoenrade (1997) muestran el efecto de la orientación sexual en las actitudes laborales. Una de estas actitudes laborales es la satisfacción laboral. De la misma forma, Lekewise, Hodson (1989) estudia las diferencias de género en la satisfacción laboral. Scott, Moore y Miceli (1997) encuentran a la satisfacción laboral ligada a los patrones de comportamiento de los adictos al trabajo. Hay estudios en donde la satisfacción laboral es usada como una variable independiente: Meiksins y Watson (1989) muestran cuánto influye la satisfacción laboral en la autonomía profesional de los ingenieros. Estudios de Somers (1996); Francis-Felsen, Coward, Hogan y Duncan (1996); y Hutchinson y Turner (1988) evaluaron el efecto de la satisfacción laboral en la rotación del personal de enfermería.

Otro ejemplo es la ansiedad, que se ha estudiado como una variable independiente que afecta a la variable dependiente aprovechamiento. Oldani (1997) encontró que la ansiedad de la madre durante el embarazo influye en el aprovechamiento de los hijos (medido como éxito en el área musical). Capaldi, Crosby y Stoolmiller (1996) emplearon los niveles de ansiedad de varones adolescentes para predecir el momento de su primer encuentro sexual. Onwuegbuzie y Seaman (1995) estudiaron los efectos de la ansiedad ante los exámenes en la realización de la prueba, en un curso de estadística. La ansiedad también puede concebirse y usarse como una variable dependiente: por ejemplo, puede utilizarse para estudiar la diferencia entre tipos de cultura, nivel socioeconómico y género (véase Guida y Ludlow, 1989; Murphy, Olivier, Monson y Sobol, 1991). En otras palabras, la clasificación de la variable independiente y la dependiente es en realidad una taxonomía de los usos de la variable más que una distinción entre diferentes tipos de variables.

Variables activas y variables atributo

Una clasificación que nos será útil en nuestro estudio posterior del diseño de la investigación se basa en la distinción entre variables experimentales y medidas. Cuando se planea y ejecuta la investigación es importante distinguir entre estos dos tipos de variables. Las variables manipuladas se llamarán variables *activas*, mientras que las variables medidas se denominarán variables *atributo*. Por ejemplo, Colwell, Foreman y Trotter (1993) compararon dos métodos de tratamiento de las úlceras de presión de los pacientes encamados. Las variables dependientes fueron eficacia y efectividad costo. Los dos métodos de tratamiento fueron una compresa de gasa humedecida y una compresa con una cubierta de hidrocoloide. Los investigadores controlaron quién recibía qué tipo de tratamiento. Como tal, el tratamiento o variable independiente fue una variable activa o manipulada.

Así, cualquier variable manipulada, constituye una variable activa. "Manipulación" significa, en esencia, hacer cosas diferentes a distintos grupos de sujetos, como se verá con claridad en un capítulo posterior al discutir a profundidad las diferencias entre investigación experimental y no experimental. Se dice que existe manipulación cuando un investigador hace algo a un grupo (por ejemplo, reforzar positivamente cierta clase de comportamiento) y hace algo distinto con el otro grupo, o tiene a dos grupos siguiendo diferentes instrucciones. Cuando uno usa diferentes métodos de enseñanza o premia a los sujetos de un grupo y castiga a los de otro, o crea ansiedad a través de instrucciones que generan preocupación, uno está *activamente* manipulando las variables: métodos, reforzamiento y ansiedad.

Otra clasificación relacionada, usada principalmente por los psicólogos es la de las variables *estímulo* y *respuesta*. Una *variable estímulo* es cualquier condición o manipulación del ambiente realizada por el experimentador, que evoque una respuesta en un organismo. Una *variable de respuesta* es cualquier clase de comportamiento del organismo. El supuesto es que para cualquier clase de comportamiento siempre hay un estímulo. Por lo tanto, el comportamiento del organismo es una respuesta. Esta clasificación se refleja en la bien conocida ecuación: $R = f(O, E)$, que se lee: "las respuestas son una función del organismo y los estímulos", o "las variables de respuesta son una función de las variables organísmicas y de las variables estímulo".

Las variables que no pueden ser manipuladas son *las atributo o características del sujeto*. Es imposible, o al menos muy difícil, manipular muchas variables. Las variables consistentes en características humanas como inteligencia, aptitud, género, estatus socioeconómico, conservadurismo, dependencia de campo, necesidad de logro y actitudes son variables atributo. Los sujetos llegan a nuestro estudio con estas variables (atributos) ya presentes o preexistentes. El entorno temprano, la herencia, y otras circunstancias han hecho de los individuos lo que son. Se les llama también variables *organísmicas*. Cualquier propiedad, característica o atributo de un individuo constituye una variable organísmica, digamos que forma parte del organismo. En otras palabras, las variables organísmicas son aquellas características que los individuos poseen en diversos grados cuando ingresan a la situación de investigación. El término *diferencias individuales* implica variables organísmicas. Una de las más comunes variables atributo en las ciencias sociales y del comportamiento es el género: femenino-masculino. Los estudios diseñados para comparar diferencias de género involucran una variable atributo. Tomemos, por ejemplo, el estudio de Weerth y Kalma (1993). Estos investigadores compararon a hombres y mujeres en su respuesta a la infidelidad del cónyuge o pareja. La variable atributo aquí es género, que *no* es una variable manipulada. Hay estudios donde las puntuaciones de una o varias pruebas se usaron para dividir a un conjunto de personas en dos o más grupos. En este caso, las diferencias del grupo se reflejan como una variable atributo, como lo ilustra el estudio de Hart, Forth y

Hare (1990) quienes aplicaron una prueba de psicopatología a varones recluidos en una prisión. Con base en sus puntuaciones, los internos fueron asignados a uno de tres grupos: bajo, medio y alto. Después se comparó su puntuación en una batería de pruebas neuropsicológicas. El nivel de psicopatología preexiste y el investigador no lo manipula. Si un interno puntuaba alto, era asignado al grupo alto. Así, la psicopatología es una variable atributo en este estudio. Hay algunos estudios donde la variable independiente podría haber sido manipulada; sin embargo, por razones logísticas o legales no lo fue. Un ejemplo es el estudio de Swanson, Maas y Buckwalter (1994) quienes compararon diferentes formas de atención y su efecto en las medidas cognitivas y funcionales de los pacientes con Alzheimer. La variable atributo fue el tipo de atención. No se permitió a los investigadores asignar a sus pacientes a dos diferentes instituciones de atención (tradicional *vs.* unidad de cuidado especial). Los investigadores se vieron forzados a estudiar a los sujetos una vez que habían sido asignados al centro correspondiente. Por ello puede considerarse que la variable independiente es una variable no manipulada. Los investigadores heredaron grupos intactos.

La palabra ***atributo*** es lo suficientemente precisa cuando se usa con objetos o referentes inanimados. Sin embargo, las organizaciones, instituciones, grupos, poblaciones, casas y áreas geográficas también poseen atributos —*atributos activos*—. Las organizaciones son productivas de manera variable; las instituciones pasan de moda; los grupos difieren en su cohesión; las áreas geográficas varían ampliamente en sus recursos.

Esta distinción de atributo activo es general, flexible y útil. Veremos que algunas variables, por su propia naturaleza, son siempre atributos, mientras que otras variables que son atributos pueden también ser activas. Esta última característica hace posible investigar las "mismas" relaciones de diferentes formas. Y usando de nuevo el ejemplo de la variable ansiedad, podemos medirla: como es evidente en este caso, es una variable atributiva. Sin embargo, también puede ser manipulada al inducir diferentes grados de ansiedad: por ejemplo, si les decimos a los sujetos de un grupo experimental que la tarea que van a realizar va a ser muy difícil, que su inteligencia será evaluada y que su futuro depende de la puntuación que obtengan. A los sujetos del otro grupo experimental les decimos que lo hagan lo mejor posible, pero relajados, y que el resultado no es importante y que no va a influir en su futuro. En realidad no podemos asumir que la ansiedad medida (atributo) y la ansiedad manipulada (activa) sean la misma. Podemos suponer que ambas son "ansiedad" en un sentido amplio, pero ciertamente no son iguales.

Variables continuas y categóricas

Ya se ha hecho una distinción entre variables continuas y categóricas en especial útil para la planeación de la investigación y el análisis de datos. Sin embargo, su importancia justifica una consideración más amplia.

Una variable *continua* es capaz de asumir un conjunto ordenado de valores dentro de cierto rango. Esta definición significa, primero, que el valor de una variable continua refleja al menos un orden categórico, que un mayor valor de la variable implica más de la propiedad en cuestión que un menor valor. Los valores producidos por una escala para medir dependencia, por ejemplo, expresan diferentes cantidades de dependencia, desde la alta pasando por la media hasta la baja. Segundo, las medidas continuas en uso están contenidas en un rango, y cada individuo obtiene una "puntuación" dentro del mismo. Una escala para medir dependencia puede tener un rango de uno a siete. La mayoría de las escalas usadas en ciencias del comportamiento también tienen una tercera característica: hay un conjunto teóricamente infinito de valores en el rango. (Las escalas de rangos

ordenados son algo diferentes; se analizarán más adelante en el libro.) Esto quiere decir que una puntuación de un individuo en particular puede ser de 4.72 más que sólo de 4 o 5.

Las variables *categóricas*, como las llamaremos, pertenecen a una clase de mediciones llamadas nominales (se explicarán en el capítulo 25). En una medición nominal hay dos o más subconjuntos del grupo de objetos que se mide. Se categoriza a los individuos en razón de la posesión de las características que definen cualquier subgrupo. "Categorizar" significa asignar un objeto a una subclase (o subconjunto) de una clase (o conjunto) con base en que el objeto posea o no la característica que define al subconjunto. El individuo que está en proceso de ser categorizado posee o no la propiedad definitoria; esto es un asunto de todo o nada. Los ejemplos más simples son las variables categóricas dicotómicas: femenino-masculino, republicano-demócrata, correcto-incorrecto. Las variables politómicas, que son aquellas con más de dos subconjuntos, son bastante comunes, en especial en sociología y economía: preferencia religiosa, nivel educativo, nacionalidad y elección laboral, entre otras.

Las variables categóricas y la medición nominal tienen requisitos simples: se considera iguales a todos los miembros de un subconjunto y todos tienen asignado el mismo nombre (nominal) y el mismo número. Si la variable es preferencia religiosa, por ejemplo, todos los protestantes son iguales, todos los católicos son iguales y todos los "otros" son iguales. Si un individuo es católico (definido operacionalmente de una manera apropiada), la persona es asignada a la categoría "católica" y también se le asigna un "1" en esa categoría. Es decir, se contabiliza a ese individuo como "católico". Las variables categóricas son "democráticas". No hay un orden en cuanto a rango, ni mayor que o menor que, entre las categorías y se asigna un mismo valor a todos los miembros de una categoría.

La expresión "variables cualitativas" algunas veces se ha aplicado a las variables categóricas, en particular a las dicotómicas, probablemente en contraste con las "variables cuantitativas" (nuestras variables continuas). Este uso refleja un concepto distorsionado de lo que son las variables, ya que siempre son cuantificables; si no lo fueran, no serían variables. Si x tiene solamente dos subconjuntos y puede asumir sólo dos valores (1 y 0), éstos siguen siendo valores, y la variable varía. Si x es una variable politómica, como la afiliación política, cuantificamos nuevamente al asignar valores enteros a los individuos. Si un individuo dice ser demócrata, se ubica a esa persona en el subgrupo demócrata y se le asigna un 1. Todos los individuos en el subconjunto demócrata tendrán un valor de 1. Es en extremo importante comprender esto porque, por principio, son las bases para cuantificar muchas variables, incluso tratamientos experimentales, para llevar a cabo análisis complejos. En el análisis de regresión múltiple, como veremos más adelante, todas las variables —continuas y categóricas— son ingresadas como variables al análisis. En el ejemplo de género que se mencionó arriba, 1 era asignado a un género y 0 al otro. Diseñamos una columna de unos y ceros de la misma forma en que estableceríamos una columna de puntuaciones de dependencia. La columna de unos y ceros representa la cuantificación de la variable género. Aquí no hay misterio. Estas variables han sido llamadas variables *prototipo* (conocidas como dummy en inglés). Como son muy útiles y poderosas e incluso indispensables en el análisis de datos de la investigación moderna, requieren ser entendidas con claridad. Una explicación más a fondo puede encontrarse en Kerlinger y Pedhazur (1973) y en el capítulo 34 de este libro. El método se aplica con facilidad a las politomías. Una *politomía* es una división de los miembros de un grupo en tres o más subdivisiones.

Constructos observables y variables latentes

En gran parte de la discusión previa de este capítulo se ha implicado —pues no se ha declarado de forma explícita— que existe una diferencia fundamental entre constructos y

variables observadas. Más aún, podemos afirmar que los constructos no son observables; y que las variables, cuando se definen operacionalmente, son observables. Esta distinción es importante en tanto que si no estamos plenamente conscientes del nivel de discurso en que nos encontramos al hablar acerca de variables, es difícil ser claros sobre lo que hacemos.

Una expresión fructífera e importante que se encontrará y usará extensamente en este libro es "variable latente". Una variable latente es una "entidad" no observada, que se presume subyace a las variables observadas. El ejemplo mejor conocido de una variable latente importante es "inteligencia". Podemos decir que tres pruebas de habilidad —verbal, numérica y espacial— están relacionadas de manera positiva y sustancial. Esto significa, en general que las personas con puntuación alta en una, tienden a tener altas puntuaciones en las otras; de forma similar quienes obtienen bajas puntuaciones en una, tenderán a presentar bajas en las otras. Creemos que hay algo común a las tres pruebas o variables observadas, y lo llamamos "inteligencia", que es una variable latente.

Hemos encontrado muchos ejemplos de variables latentes en las páginas previas: aprovechamiento, creatividad, clase social, satisfacción laboral, preferencia religiosa, etcétera. De hecho, siempre que mencionamos los nombres de fenómenos en que varían las personas o los objetos, hablamos de variables latentes. En el campo de la ciencia, nuestro interés real está más en las relaciones entre variables latentes que entre variables observadas, ya que buscamos explicar fenómenos y sus relaciones. Cuando enunciamos una teoría, enunciamos en parte relaciones sistemáticas entre variables latentes. No estamos demasiado interesados en la relación entre el comportamiento observado de frustración y la conducta observada de tipo agresivo, por ejemplo, aunque debemos trabajar con ellos en el nivel empírico. En realidad, estamos interesados en la relación entre la variable latente frustración y la variable latente agresión.

Debemos ser precavidos, sin embargo, cuando tratamos con variables no observables. Los científicos que usan términos como "hostilidad", "ansiedad" y "aprendizaje", están conscientes de que hablan a cerca de constructos inventados. La "realidad" de estos constructos se infiere a partir del comportamiento. Si desean estudiar diferentes tipos de motivación, deben saber que "motivación" es una variable latente, un constructo inventado para dar cuenta de un comportamiento presumiblemente "motivado". Es necesario que sepan que esta "realidad" sólo está postulada. Sólo pueden juzgar si los jóvenes están motivados o no al observar sus comportamientos. Aun así para estudiar la motivación, deben medirla o manipularla. Pero no pueden medirla de forma directa por que, en pocas palabras, es una variable que está "en la cabeza", una entidad no observable, una variable latente. Se inventó el constructo "por algo" que *se presume que está* dentro de los individuos, "algo" que los impulsa a comportarse de tal y cual manera. Esto significa que los investigadores deben medir siempre supuestos indicadores de motivación y no a ella en sí misma. En otras palabras, deben medir algún tipo de comportamiento, sean marcas en un papel, palabras habladas, o gestos significativos, y después hacer inferencias sobre características supuestas o variables latentes.

Se han usado otros términos para expresar más o menos las mismas ideas. Por ejemplo, Tolman (1951 pp. 115-129) llamó a los constructos variables intervinientes. Las *variables intervinientes* representan un término inventado para dar cuenta de procesos psicológicos no observables, internos, que a su vez dan cuenta de la conducta. Una variable interviniente es una variable "en la cabeza": No se le puede ver, oír o tocar. Se infiere a partir del comportamiento. La "hostilidad" se infiere de actos presumiblemente hostiles o agresivos. La "ansiedad" se infiere de la puntuación en una prueba, respuesta en la piel, frecuencia cardiaca y ciertas manipulaciones experimentales. Otro término es "constructo hipotético", cuyo significado es bastante similar al de variable latente, aunque con un poco menos de generalidad; no es necesario detenernos en él. Debemos mencionar, sin embar-

go, que "variable latente" parece ser una expresión más general y aplicable que "variable interviniente" y que "constructo hipotético", en tanto que puede usarse virtualmente para cualquier fenómeno que se presume influye o es influido por otro. En otras palabras "variable latente" puede utilizarse con fenómenos psicológicos, sociológicos y de otro tipo. "Variable latente" parece ser un vocablo más útil por su generalidad y también por que ahora es posible, en el enfoque de análisis de estructuras de covarianza, evaluar el efecto de las variables latentes entre sí y las llamadas variables manifiestas u observadas. Está discusión que parece muy abstracta después se concretará para ser, esperamos, significativa. Veremos entonces que la idea de las variables latentes y las relaciones entre ellas son extremo importantes, fructíferas y útiles, y ayudan a cambiar los enfoques fundamentales para afrontar los problemas de investigación.

Ejemplos de variables y definiciones operacionales

Hemos aportado una cantidad de constructos y definiciones operacionales. Para ilustrar y aclarar la discusión previa, en especial en lo referente a la diferencia entre variables experimentales y variables medidas, y entre constructos y variables definidas operacionalmente, presentamos diversos ejemplos. Si la definición es experimental se etiqueta con (E); si es una definición medida se señala como (M).

Las definiciones operacionales difieren en su grado de especificidad. Algunas están ligadas de manera estrecha a las observaciones. Definiciones de "pruebas", como "la inteligencia se define como una puntuación x en una prueba de inteligencia" son muy específicas. Una definición como "la frustración consiste en no alcanzar una meta" son más generales y requieren una mayor especificación para que sean medibles.

Clase social. "...dos o más grupos de personas que se cree que están en una posición social superior e inferior y que son así categorizados por los miembros de una comunidad" (M) (Warner y Lunt, 1941, p. 82). Para ser operacional, está definición ha de especificarse a partir de preguntas dirigidas a las creencias de la gente sobre las posiciones de otras personas. Se trata de una definición subjetiva de clase social. La clase social, o el estatus social, también es define de forma más objetiva a través de índices como ocupación, ingreso y educación, o por combinaciones de tales índices. Por ejemplo, "...transformamos la información acerca de educación, ocupación e ingresos de los padres de jóvenes de una liga juvenil en un índice de nivel socioeconómico (NSE), en el que las puntuaciones altas indican una avanzada educación, una ocupación prestigiosa y un ingreso desahogado. Calificaciones menores reflejan pobreza, educación incompleta y los trabajos más modestos" (M) (Herrnstein y Murray, 1996, p. 131).

Aprovechamiento (escolar, aritmético y en ortografía). El aprovechamiento se acostumbra definir operacionalmente a partir de una prueba estandarizada de aprovechamiento (por ejemplo la prueba de Iowa de Destrezas Básicas, la prueba de aprovechamiento o la prueba elemental de la batería de Evaluación de Infantil de Kaufman [K-ABC]), por medio del promedio o por el juicio del maestro. "El aprovechamiento del estudiante se midió por una puntuación combinada de pruebas de lectura y matemáticas" (M) (Peng y Wright, 1994). En ocasiones, el aprovechamiento se presenta en forma de una prueba de ejecución. Silverman (1993) examinó en un grupo de estudiantes dos habilidades del juego de voleibol: la prueba de servicio y la prueba del pase con antebrazo. En la primera, los estudiantes recibían una puntuación entre 0 y 4, en función de donde fuera colocado el balón servido. La prueba del pase con antebrazo consistía en hacer rebotar la pelota en el antebrazo. El criterio usado fue contar el número de veces en que un estudiante pudo pasar el balón más allá de una línea de dos metros y medio contra la pared en un periodo de un minuto (M). En algunos estudios educativos también se usa una definición operacional del

concepto *percepción del aprovechamiento del estudiante.* Aquí se pide a los estudiantes que se evalúen a sí mismos. La pregunta usada por Shoffner (1990) fue: "¿Qué clase de estudiante crees que eres?". Las opciones eran "estudiante de 9 y 10" "estudiante de 8" y "estudiante de 7 y 6" (M).

Aprovechamiento (desempeño académico). "Como resultado se obtuvieron las calificaciones de todos los estudiantes en todas las secciones y se usaron para determinar la categoría de cada estudiante participante en el estudio. Se calculó el rango porcentilar por sección para cada uno y se usó como la medida dependiente del aprovechamiento en el análisis final de los datos" (M) (Strom, Hocevar y Zimmer, 1990).

Motivación intrínseca se define operacionalmente por Hom, Berger *et al.* (1994) como "la cantidad acumulada de tiempo que cada estudiante juega con un patrón de bloques sin un sistema de reforzamiento" (M).

Popularidad. La popularidad con frecuencia se define operacionalmente por el número de elecciones sociométricas que un individuo recibe de otros (en su clase, grupo de juego, etcétera). Se le pregunta a los sujetos: "¿con quién te gustaría trabajar?", "¿con quién te gustaría jugar?" y otras cuestiones similares. Cada sujeto debe elegir a uno, dos o más individuos de su grupo con base en esas preguntas de criterios (M).

Compromiso con el trabajo "...el comportamiento de cada niño durante una lección fue evaluada cada seis segundos como apropiadamente comprometida o no. La puntuación del compromiso con el trabajo, por lección, fue el porcentaje de las unidades de seis segundos en que los niños fueron calificados como apropiadamente comprometidos" (M) (Kounin y Doyle, 1975).

Reforzamiento. Las definiciones de reforzamiento tienen diversas formas. La mayoría incluye, de una forma u otra, el principio de la recompensa. Sin embargo, se puede utilizar tanto el reforzamiento positivo como negativo. A continuación se presentan definiciones experimentales específicas.

En los siguientes 10 minutos cada opinión que el sujeto (S) enunció fue registrada por el experimentador (E) y reforzada. Para dos grupos, el E estuvo de acuerdo con la opinión enunciada al decir "sí tienes razón", "así es", o algo similar, o al sentir o sonreír en caso de que no pudiera interrumpir (E).

> ...se administraron al modelo y al niño de manera alternativa doce diferentes grupos de reactivos de historias... Ante cada uno de ellos, el modelo expresaba de forma consistente respuestas en forma de juicios opuestas a la orientación moral del niño... Y el experimentador reforzaba el comportamiento del modelo con respuestas de aprobación verbal tales como "muy bien", "está bien" y "qué bueno". El niño fue reforzado de manera parecida cada vez que adoptaba el tipo de juicios morales del modelo en respuesta a su propio grupo de reactivos [esto se denomina "reforzamiento social"] (E) (Bandura y MacDonald, 1994).

El maestro otorga al niño un reconocimiento verbal cada vez que exhibe el comportamiento deseado. Éstos son: atender a la instrucción, cumplir con el trabajo escolar y contestar en voz alta. El registro se hace cada 15 segundos (E) (Martens, Hiralall y Bradley, 1997).

Actitudes hacia el SIDA se define en una escala de 18 reactivos, cada uno con un formato de tipo Likert para reflejar diversas actitudes hacia los pacientes con SIDA. Algunos ejemplos de los reactivos son: "a la gente con SIDA no se le debería permitir usar los sanitarios públicos", y "debería ser obligatorio para todos los estadounidenses hacerse una prueba para SIDA" (M) (Lester, 1989).

Personalidad limítrofe. Comrey (1993) la define como la presencia de una baja puntuación en tres escalas de la Escala de Personalidad de Comrey: confianza *vs.* defensividad, conformidad social *vs.* rebeldía, y estabilidad emocional *vs.* neuroticismo.

Delincuencia laboral se define operacionalmente como una combinación de tres variables: número de accidentes imputables al sujeto, número de cartas de advertencia y número de suspensiones (M) (Hogan y Hogan, 1989).

Religiosidad se define como una puntuación en la Escala de Francis de Actitudes hacia la Cristiandad que consta de 24 reactivos con una escala de respuesta tipo Likert. Algunos reativos son: "Orar me ayuda mucho" y "Dios me guía para conducir una vida mejor" (M) (Gillings y Joseph, 1996). La religiosidad no debe confundirse con la preferencia religiosa. La religiosidad se refiere a la fuerza de la devoción a la religión que uno ha elegido.

Autoestima es una variable independiente manipulada en el estudio de Steele, Spencer y Lynch (1993). En este caso, se aplica a los sujetos una prueba de autoestima, pero cuando se les retroalimenta la información en el reporte de retroalimentación, que parece ser el oficial es ambigua. Se divide a los sujetos del mismo nivel de autoestima medida en tres grupos diferentes de retroalimentación: positiva, negativa y ausente. En la condición de retroalimentación positiva (autoestima positiva), se describe a los sujetos con enunciados tales como "pensamiento claro". Aquéllos en el grupo negativo (autoestima negativa) reciben adjetivos como "pasivos al actuar". Al grupo "sin retroalimentación" se les indica que su perfil de personalidad (autoestima) no estuvo listo por demoras en la calificación e interpretación (E). La mayoría de los estudios sobre autoestima usan una definición operacional medida. En el ejemplo anterior, Steele, Spencer y Lynch también usaron la Escala de Sentimientos de Inadecuación de Autoestima de Janis-Field (M). En otro ejemplo Luhtanen y Crocker (1992) definieron la autoestima colectiva como una puntuación en una escala de 16 reactivos tipo Likert en los que se solicitaba pensar a cerca de una variedad de grupos sociales y características de sus miembros, tales como género, religión, raza y grupo étnico (M).

La *raza* es por lo general una variable medida. Sin embargo en un estudio de Annis y Corenblum (1986), un experimentador (E) ya sea de raza blanca o india preguntó a 83 niños indios canadienses de nivel preescolar y primer año sobre preferencias raciales e identidad personal. El interés se centraba en averiguar si la raza del experimentador influía o no en las respuestas.

Soledad. Una definición de ésta es la puntuación en la Escala de Soledad de UCLA que incluye reactivos tales como: "nadie me conoce realmente bien" o "carezco de compañía". También se cuenta con la Escala de Deprivación y Soledad que presenta reactivos como: "experimento una sensación de vacío" o "no hay quien muestre un interés particular en mí" (M) (Oshagan y Allen, 1992).

Halo. Se han postulado muchas definiciones operacionales del efecto de halo. Balzer y Sulsky (1992) encontraron y resumieron 108 definiciones que ajustaron en 6 categorías. Una establece que halo es "...la variaza promedio intratasa o la desviación estándar de las evaluaciones". Otra puede ser: "Comparar las evaluaciones obtenidas con las proporcionadas por jueces expertos" (M).

Memoria: recuerdo y reconocimiento "...recuerdo significa pedir al participante que repita lo que recuerda de los reactivos que le fueron mostrados, y asignar un punto por cada uno que corresponda a la lista de estímulos inicial"(M) (Norman, 1976, p. 97). "La prueba de reconocimiento consistió en 62 frases presentadas a todos lo sujetos... que fueron instruidos para evaluar cada frase de acuerdo a su grado de confianza de que la oración hubiera sido presentada en el conjunto inicial" (M) (Richter y Seay, 1987).

Habilidades sociales. Pueden ser definidas operacionalmente como una puntuación en la Escala de Evaluación de Destrezas Sociales (Gresham y Elliot, 1990). Existe la posibilidad de contar con información del estudiante de su padre y del maestro. Se evalúan los comportamientos sociales en términos de frecuencia de ocurrencia y también de acuerdo a su nivel de importancia. Algunos reactivos incluyen: "se lleva bien con gente que es

diferente (maestro)", "se ofrece a ayudar a miembros de la familia con sus tareas (papá)", y, "cuestiono cortésmente las reglas que pueden ser injustas (estudiante)" (M).

Congraciamiento. Una de las muchas técnicas de manejo de impresión (véase Orpen, 1996; Gordon, 1996). Se define operacionalmente como una puntuación en la Escala de Kumar y Beyerlein (1991). Que comprende 25 reactivos tipo Likert diseñados para medir la frecuencia en que el subordinado, en una relación superior-subordinado usa tácticas para congraciarse (M). Strutton, Pelton y Lumpkin (1995) modificaron la Escala de Kumar-Beyerlein, para medir el congraciamiento entre el vendedor y cliente (M).

Feminismo. Se define por la puntuación en el Cuestionario de Actitudes hacia las mujeres. Este instrumento consta de 18 enunciados en los que el entrevistado registra su acuerdo en una escala de cinco puntos. Los reactivos incluyen: "los hombres han mantenido el poder por demasiado tiempo"; "los concursos de belleza son degradantes para la mujer"; "los niños de madres trabajadoras tienden a sufrir" (Wilson y Reading, 1989).

Valores. "Ordene las 10 metas de acuerdo a la importancia que tienen para usted: 1) Tener éxito financiero; 2) Ser querido; 3) Tener éxito en el ámbito familiar; 4) Ser capaz en lo intelectual; 5) Vivir de acuerdo a principios religiosos; 6) Ayudar a otros; 7) Ser normal, bien ajustado; 8) Cooperar con los demás; 9) Trabajar con detalle; 10) Alcanzar el éxito laboral" (M) (Newcomb, 1978).

Democracia (democracia política). "El índice (de democracia política) consiste en tres indicadores de soberanía popular y tres de libertades políticas. Las medidas de soberanía popular son: 1) Elecciones limpias, 2) Selección ejecutiva efectiva, y 3) Selección legislativa. Los indicadores de libertades políticas son: 4) Libertad de prensa, 5) Libertad para que la oposición se asocie y 6) Sanciones gubernamentales" (M). Bollen (1979) proporciona detalles operacionales de seis indicadores sociales en un apéndice (pp. 585-586). Éste es un ejemplo particularmente bueno de la definición operacional de un concepto complejo. Más aún, constituye una excelente descripción de los ingredientes de la democracia.

Los beneficios del pensamiento operacional han sido grandiosos. De hecho el operacionalismo ha sido y es uno de los movimientos más significativos e importantes de nuestro tiempo. El operacionalismo extremo, por supuesto, puede ser peligroso porque oscurece el reconocimiento de la importancia de los constructos y definiciones constitutivas en la ciencia del comportamiento y porque puede restringir la investigación a problemas triviales. Sin embargo, hay poca duda de que sea una sana influencia. Resulta una clave indispensable para alcanzar la objetividad (sin la cual no hay ciencia), en tanto que demanda que las observaciones sean públicas y replicables, lo que ayuda a colocar las actividades de investigación más allá de los investigadores y sus predilecciones. Y, como dijo Underwood (1957, p. 53) en su texto clásico sobre investigación psicológica:

> Yo diría que el pensamiento operacional hace mejores científicos. El operacionalista se ve forzado a sacudir y aclarar sus conceptos empíricos... el operacionalismo facilita la comunicación entre científicos ya que el significado de los conceptos así definidos no es sujeto fácilmente a una mala interpretación.

RESUMEN DEL CAPÍTULO

1. Un *concepto* es una expresión de una abstracción formada a partir de la generalización de un particular, por ejemplo, peso. Esta expresión se deriva de observaciones de ciertos comportamientos o acciones.
2. Un *constructo* es un concepto que se ha formulado para ser usado en la ciencia. Se usa en esquemas teóricos y se define de tal manera que sea susceptible de ser observado y medido.

3. Una *variable* se define como una propiedad que puede tomar diferentes valores; es un símbolo al que se le asignan valores.
4. Los constructos y las palabras pueden ser definidas por
   a) otras palabras o conceptos.
   b) descripción de una acción o conducta implícita o explícita.
5. Una *definición constitutiva* se da cuando los constructos están definidos por otros constructos.
6. Una *definición operacional* se presenta cuando se aporta el significado al especificar las actividades u operaciones necesarias para medir y evaluar el constructo. Las definiciones operacionales sólo pueden dar un significado limitado al constructo; no pueden describir completo a un constructo o variable. Hay dos tipos de definiciones operacionales:
   a) de medida —nos dice cómo será medida la variable o constructo—.
   b) experimental —explica los detalles de cómo el experimentador manipula la variable (constructo)—.
7. Tipos de variables
   a) La *independiente* varía y es la causa supuesta de otra variable, la dependiente. En un experimento, constituye la variable manipulada, es la que está bajo el control del experimentador. En un estudio no experimental, es la variable que tiene un efecto lógico en la variable dependiente.
   b) El efecto de la variable *dependiente* se altera de forma concomitante con los cambios o variaciones en la variable independiente.
   c) Una variable *activa* se manipula. Manipulación significa que el experimentador tiene control sobre cómo cambian los valores.
   d) Una variable *atributiva* se mide y no puede ser manipulada, es decir, es aquella donde el experimentador no tiene control sobre los valores de la variable.
   e) Una variable *continua* es capaz de asumir un grupo ordenado de valores dentro de cierto rango. Entre dos valores hay un número infinito de otros valores. Esta variable refleja por lo menos una categoría ordinal.
   f) Las variables *categóricas* pertenecen a una clase de medición donde los objetos se asignan a subclases o a subgrupos, diferenciados y que no se traslapan. Se considera que todos los elementos de una misma categoría tienen la misma característica o características.
   g) Las variables *latentes* son entidades no observables y se asume que subyacen a las variables observadas.
   h) Las variables *intervinientes* son constructos que dan cuenta de procesos psicológicos internos no observables que explican el comportamiento. No pueden ser vistas pero se infieren a partir del comportamiento.

SUGERENCIAS DE ESTUDIO

1. Escriba las definiciones operacionales para 5 o 6 de los siguientes constructos. Cuando le sea posible, escriba dos definiciones: una experimental y una definición de medida.

reforzamiento	poder de castigo
aprovechamiento	habilidad lectora
bajo rendimiento	necesidades
liderazgo	interés

transferencia del entrenamiento	delincuencia
nivel de aspiración	necesidad de afiliación
conflicto organizacional	conformidad
preferencia política	satisfacción marital

Alguno de estos conceptos o variables —por ejemplo, necesidades y transferencia del entrenamiento— pueden resultar difíciles de definir operacionalmente. ¿Por qué?

2. ¿Podría alguna de las variables del inciso anterior ser tanto variable independiente como variable dependiente? ¿Cuáles?
3. Resulta instructivo y útil para los especialistas leer acerca de otros campos distintos del propio. Esto es en particular cierto para los estudiantes de investigación del comportamiento. Se sugiere que el estudiante de un campo en particular lea dos o tres estudios de investigación en una de las mejores revistas de otra disciplina. Si usted está en psicología lea una revista de sociología, por ejemplo, la *American Sociological Review*. Si usted está inmerso en el campo de la educación o la sociología, lea una revista de psicología como el *Journal of Personality y Social Psychology* o el *Journal of Experimental Psychology*. Los alumnos que no pertenecen al área de educación, pueden hojear el *Journal of Educational Psychology* o el *American Educational Research Journal*. Al leer, tome nota de las variables y compárelas con las de su propio campo. ¿Son primariamente variables activas o variables atributo? Observe, por ejemplo, qué variables psicológicas son más "activas" que las sociológicas. ¿Qué implican las variables de una disciplina para su investigación?
4. La lectura de los siguientes artículos será útil para entender y desarrollar definiciones operacionales.

Kinnier, R.T. (1995). A reconceptualization of values clarification: Values conflict resolution. *Journal of Counseling and Development, 74(1)*, 18-24.

Lego, S. (1988). Multiple disorder: An interpersonal approach to etiology, treatment and nursing care. *Archives of Psychiatric Nursing, 2(4)*, 231-235.

Lobel, M. (1994). Conceptualizations, measurement, and effects of prenatal maternal stress on birth outcomes. *Journal of Behavioral Medicine, 17(3)*, 225-272.

Navathe, P.D. & Singh, B. (1994). An operational definition for spatial disorientation. *Aviation, Space & Environmental Medicine, 65(12)*, 1153-1155.

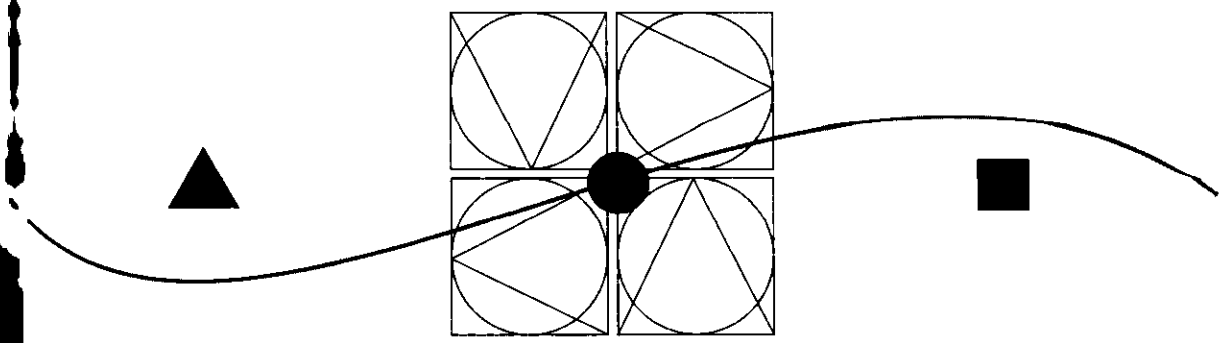
Sun, K. (1955). The definition of race. *American Psychologist, 50(1)*, 43-44.

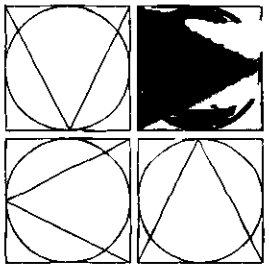
Talaga, J.A. & Beehr, T.A. (1995). Are the gender differences in predicting retirement decisions? *Journal of Applied Psychology, 80(1)*, 16-28.

Woods, D.W. Miltenberger, R.G. & Flach, A.D. (1996). Habits, tics and stuttering: Prevalence and relation to anxiety and somatic awareness. *Behavior Modification, 20(2)*, 216-225.

Parte Dos

Conjuntos, relaciones y varianza





CAPÍTULO 4

CONJUNTOS

- SUBCONJUNTOS
- OPERACIONES DE CONJUNTOS
- CONJUNTOS UNIVERSALES Y VACÍOS; LA NEGACIÓN DEL CONJUNTO
- DIAGRAMAS DE CONJUNTOS
- OPERACIONES CON MÁS DE DOS CONJUNTOS
- PARTICIONES Y PARTICIONES CRUZADAS
- NIVELES DEL DISCURSO

El concepto de "conjunto" es una de las ideas matemáticas más poderosas y útiles para comprender los aspectos metodológicos de la investigación. Los conjuntos y sus elementos son la materia prima que subyace a la forma en que opera la matemática. Aun si no estamos conscientes de ello, los conjuntos y la teoría de conjuntos son fundamentos de nuestro pensamiento descriptivo lógico y analítico así como de su operación. Virtualmente, constituyen la base de todo lo que se presenta en este libro y son los cimientos sobre los cuales erigimos complejos análisis numéricos, categóricos y estadísticos, incluso cuando no constituyen los fundamentos explícitos de nuestro pensamiento y trabajo. Por ejemplo, la teoría de conjuntos proporciona una definición no ambigua de relaciones, nos ayuda a aproximarnos y a entender la probabilidad y el muestreo, y es prima hermana de la lógica. También nos ayuda a entender el muy importante tema de las categorías y la categorización de los objetos del mundo. Más aún, el pensamiento de conjuntos puede apoyarnos para comprender ese difícil problema de la comunicación humana: la confusión causada por mezclar diferentes niveles del discurso.

La ciencia trabaja básicamente con conceptos de grupo, clase o conjunto. Cuando los científicos discuten sobre eventos u objetos individuales, los consideran como miembros de conjuntos de objetos. Pero esto es cierto también en el discurso humano en general. Decimos "ganso", pero la palabra *ganso* no tiene sentido sin el concepto de un grupo de tales individuos llamado "gansos". Cuando hablamos acerca de un niño y de sus problemas, resulta inevitable hablar de los grupos, clases o conjuntos de objetos a los que el niño pertenece. Esto incluiría a un niño de 7 años (primer conjunto), de segundo grado (segun-

do conjunto), brillante (tercer conjunto), saludable (cuarto conjunto) y varón (quinto conjunto).

De acuerdo con Farlow (1988) y Smith (1992), un *conjunto* es una colección bien definida de objetos. Un conjunto se encuentra bien definido cuando no hay duda sobre si un objeto dado pertenece o no al conjunto. Palabras como clase, escuela, familia, manada y grupo, indican conjuntos. Hay dos formas para definirlos: 1) a través de un listado de todos los miembros del conjunto, y 2) con una regla para determinar cuáles objetos pertenecen o no al conjunto. Llamaremos al 1) una definición de "lista" y a 2) una definición de "regla". En investigación se usa por lo general la definición de regla, aunque hay casos donde se enlista a todos los miembros de un conjunto por escrito o mentalmente. Por ejemplo, supongamos que estudiamos la relación entre la conducta del votante y la preferencia política. En *Estados Unidos la preferencia política* puede definirse como estar registrado como republicano o demócrata. Así, tenemos entonces un conjunto amplio de todas las personas con preferencias políticas en dos subconjuntos más pequeños: el subconjunto de los republicanos y el de los demócratas. Ésta es una definición de conjuntos a partir de una regla. Por supuesto, podríamos tener una lista de todos los demócratas y republicanos registrados para definir ambos subconjuntos, pero con frecuencia es difícil o imposible, además, resulta innecesario. La regla, en general, basta. Podría frasearse así: un republicano es cualquier persona que esté registrada en el partido republicano. Otra regla podría ser: un republicano es cualquier persona que diga que es republicano.

Subconjuntos

Un *subconjunto* de un conjunto es un conjunto que resulta de seleccionar conjuntos de un conjunto original. Cada subconjunto de un conjunto es parte del conjunto original. De forma más sucinta y precisa, el conjunto B es un subconjunto de un conjunto A siempre que todos los elementos de B son elementos de A (Kershner & Wilcox 1974). Designamos a los conjuntos con letras mayúsculas: A, B, K, L, X, Y, y así sucesivamente. Si B es un subconjunto de A, nosotros escribimos $B \subset A$, que significa "B es subconjunto de A", "B está contenido en A" o "todos los miembros de B son también miembros de A".

Cuando se muestrea una población, las muestras resultantes son subconjuntos de la población. Suponga que un investigador muestrea cuatro clases de 6o. grado, de todos los grupos de 6o. en una escuela grande. Las cuatro clases forman un subconjunto de la población de todas las clases de 6o. grado. Cada una de las cuatro clases de la muestra puede considerarse también subconjunto de las cuatro clases —y también de la población total de los grupos—. Todos los estudiantes de las cuatro clases pueden dividirse en dos subconjuntos: niños y niñas. Cuando un investigador fracciona o parte una población o una muestra en dos o más grupos, los subconjuntos se crean por medio de una "regla" o criterio para hacerlo. Los ejemplos son numerosos: dividir preferencia religiosa en protestante, católica, judía; inteligencia en alta y baja; etcétera. También podemos observarlo en condiciones experimentales: el modelo clásico de grupo experimental y grupo control es una idea conjunto-subconjunto. Hay individuos que se asignan al grupo experimental, que constituye un subconjunto de toda la muestra. Todos los demás individuos del experimento (los del grupo control) también forman un subconjunto.

Operaciones de conjuntos

Hay dos operaciones básicas de conjuntos: *intersección* y *unión*. Una operación consiste tan sólo en "hacer algo para". En aritmética sumamos, restamos, multiplicamos y dividimos.

En el ámbito de los conjuntos, "intersecamos" y "unimos". También los "negamos". Cuando tratamos con conjuntos, hay operadores lógicos involucrados. Para la intersección, el operador lógico es "y"; para la unión, el operador lógico apropiado es "o" y para la negación, el operador es "no". Si se desea saber más de operadores lógicos sugerimos leer a Udolf (1973).

La *intersección* es el traslape de dos o más conjuntos; consiste en los elementos compartidos en común por dos o más conjuntos. El símbolo para la intersección es ∩ (se lee "intersección") la intersección de los conjuntos A y B se escribe $A \cap B$, y $A \cap B$ es en sí misma un conjunto. Con mayor precisión, es el conjunto que contiene aquellos elementos de A y B que pertenecen a ambos A y B. La intersección también se escribe $A \bullet B$, o simplemente AB.

Sea $A = \{0, 1, 2, 3\}$; sea $B = \{2, 3, 4, 5\}$. (Note que las llaves "{ }" se usan para simbolizar conjuntos.) Entonces $A \cap B = \{2, 3\}$, como se muestra en la figura 4.1. $A \cap B$, o $\{2, 3\}$, es un nuevo conjunto compuesto por los elementos *comunes* a ambos conjuntos. Observe que $A \cap B$ también indica la *relación* entre los conjuntos —los elementos compartidos por A y B—.

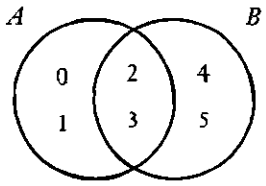
La *unión* de dos conjuntos se escribe $A \cup B$ que es un conjunto que contiene a todos los miembros de A y a todos los de B. Los matemáticos definen $A \cup B$ como un conjunto que contiene aquellos elementos que pertenecen a A o a B, o a ambos. En otras palabras, "sumamos" los elementos de A a aquellos de B para formar el nuevo conjunto $A \cup B$. Tomemos el ejemplo de la figura 4.1. A incluye a 0, 1, 2 y 3; B incluye a 2, 3, 4 y 5. $A \cup B = \{0, 1, 2, 3, 4, 5\}$. La unión de A y B en la figura 4.1 se indica por toda el área de ambos círculos. Observe que no contamos a los miembros de $A \cap B$ {2, 3} dos veces.

Algunos ejemplos de unión en investigación podrían ser agrupar juntos a hombres y mujeres, $M \cup F$ o a los republicanos y demócratas, $R \cup D$. Sea A todos los niños de escuelas primarias y B, todos los niños de escuelas secundarias de X distrito escolar. Entonces, $A \cup B$ es el conjunto de todos los niños de las escuelas en el distrito.

Conjuntos universales y vacíos; la negación del conjunto

El conjunto universal, que se escribe U, es el conjunto de todos los elementos bajo discusión. Puede llamarse *universo del discurso* o *nivel del discurso*. (Equivale al término *población* y *universo* en teoría de muestreo.) Esto implica que limitamos nuestra discusión a un conjunto definido de elementos —todos ellos— pertenecientes a la clase determinada, U. Si estudiamos los determinantes del aprovechamiento en la escuela primaria, por ejemplo, podemos definir U como todos los alumnos en grados 1o. al 6o. Podemos definir U, de forma alternativa, como las puntuaciones en una prueba de aprovechamiento de estos mismos alumnos. Los subconjuntos de U que pudieran estudiarse por separado, podrían ser las puntuaciones de los alumnos de 1er. grado, las de los estudiantes de 2o. grado, etcétera.

FIGURA 4.1



U puede ser grande o pequeño. Si regresamos al ejemplo de la figura 4.1, $A = \{0, 1, 2, 3\}$ y $B = \{2, 3, 4, 5\}$ y si $A \cup B = U$, entonces $U = \{0, 1, 2, 3, 4, 5\}$. Aquí U es muy pequeño. Supongamos que $A = \{$Juana, María, Felicia, Beatriz$\}$ y $B = \{$Tomás, Juan, Pablo$\}$. Si sólo hablamos de estos individuos, entonces $U = \{$Juana, María, Felicia, Beatriz, Tomás, Juan, Pablo$\}$ y por supuesto, $U = A \cup B$. Éste es otro ejemplo de un pequeño U. En investigación, los U son con frecuencia grandes. Si muestreamos las escuelas de un condado grande, entonces U sería todas las escuelas del condado, es decir, un U relativamente grande. U podría también estar constituido por todos los niños o todos los maestros en estas escuelas, un U aún mayor.

En investigación, es importante conocer el U que uno estudia. La ambigüedad en la definición de U puede llevarnos a conclusiones erróneas. Se sabe, por ejemplo, que las clases sociales difieren en cuanto a la incidencia de neurosis y psicosis (Murphy, Olivier, Monson y Sobol, 1991). Si estudiáramos determinantes supuestos de enfermedad mental y trabajamos sólo con sujetos de clase media, nuestras conclusiones, por supuesto, se limitarían tan sólo a ese sector social. Es fácil generalizar a todas las personas, pero esa práctica puede constituir un gran error. En ese caso, habríamos generalizado a todas las personas, U, cuando de hecho apenas hemos estudiado las relaciones en U_1, la clase media. Es muy posible, incluso probable, que las relaciones sean diferentes en U_2, la clase trabajadora.

El *conjunto vacío* es un conjunto sin miembros y se denota E. También se le llama conjunto *nulo*. Aunque pudiera parecer peculiar al estudiante que nos preocupemos por los conjuntos sin miembros, el concepto es muy útil y aun indispensable. Con ellos podemos transmitir ciertas ideas de una forma económica y sin ambigüedades. Para indicar que no hay relación entre dos conjuntos de datos, por ejemplo, podemos escribir la ecuación del conjunto $A \cap B = E$ que simplemente indica que la intersección de los conjuntos A y B está vacía, es decir, que no hay miembros de A que pertenezcan a B, y viceversa.

Sea $A = \{1, 2, 3\}$; sea $B = \{4, 5, 6\}$. Entonces, $A \cap B = E$. Es evidente que no hay miembros comunes a A y B. El conjunto de posibilidades de que tanto el candidato presidencial demócrata como el republicano ganen la elección nacional es un conjunto vacío (E). El conjunto de ocurrencias de lluvia sin nubes está vacío (E). El conjunto vacío es, pues, otra forma de expresar la falsedad de una proposición. En este caso podemos decir que el enunciado "lluvia sin nubes" es falso. En lenguaje de conjuntos esto puede expresarse $P \cap Q = E$ donde P es igual al conjunto de todas las ocurrencias de lluvia, Q es igual al conjunto de ocurrencias de nubes, y $\sim Q$ equivale al conjunto de todas las ocurrencias de no nubes.

La *negación o complemento* del conjunto A se escribe $\sim A$. Significa todos los miembros de U que no están en A. Si asumimos que A es igual a todos los hombres, cuando U equivale a todos los seres humanos, entonces $\sim A$ es igual a todas las mujeres (no-hombres). La dicotomización simple parece ser una base fundamental del pensamiento humano. La categorización es necesaria para pensar: uno debe, en el nivel más elemental, separar los objetos en aquellos que pertenecen a cierto conjunto y aquellos que no. Debemos distinguir en humanos y no-humanos, entre yo y no-yo, temprano y no-temprano, bueno y no-bueno.

FIGURA 4.2

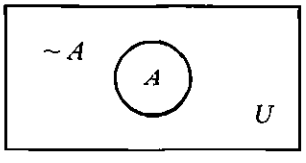
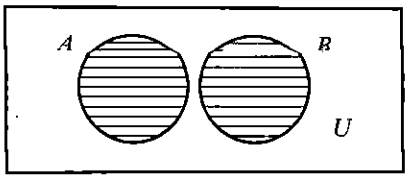


FIGURA 4.3



$A \cup B$

Si $U = \{0, 1, 2, 3, 4\}$ y $A = \{0, 1\}$ entonces $\sim A = \{2, 3, 4\}$. A y $\sim A$ son, por supuesto, subconjuntos de U. Una propiedad importante de los conjuntos y su negación se expresa en la ecuación de conjuntos: $A \cup \sim A = U$. Observe también que $A \cap \sim A = E$.

Diagramas de conjuntos

Ahora podemos recapitular e ilustrar las ideas básicas de los conjuntos que hemos presentado, a través de diagramas. Es posible representarlos con diversas figuras, pero los rectángulos y círculos son los más comunes. Se han adaptado de un sistema inventado por John Venn, un lógico del siglo XIX. En este libro usaremos rectángulos, círculos y óvalos. Observe la figura 4.2 donde U se representa con un rectángulo. Todos los elementos del universo bajo discusión están en U. Todos los miembros de U que no están en A forman otro subconjunto de U: $\sim A$. Observe una vez más que $A \cup \sim A = U$. Note también que $A \cap \sim A = E$; esto es, no hay elementos comunes a ambos, A y $\sim A$.

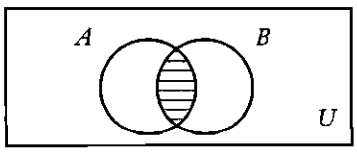
A continuación representamos (figura 4.3) dos conjuntos, A y B, ambos subconjuntos de U. A partir del diagrama puede observarse que $A \cap B = E$. Adoptamos una convención: cuando deseamos indicar un conjunto o un subconjunto, lo sombreamos ya sea de manera horizontal, vertical o diagonal. El conjunto $A \cup B$ ha sido sombreado en la figura 4.3.

La intersección, quizá el concepto más importante desde el punto de vista de este libro, se indica por la porción sombreada en la figura 4.4 y puede expresarse por la ecuación $A \cap B \neq E$; la intersección de los conjuntos A y B *no* está vacía.

Cuando dos conjuntos, A y B, son iguales, tienen los mismos elementos o miembros. El diagrama de Venn mostraría dos círculos congruentes en U. En efecto, sólo se vería un círculo. Cuando $A = B$, entonces $A \cap B = A \cup B = A = B$.

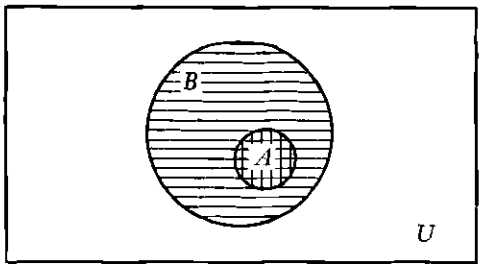
Trazamos el diagrama $A \subset B$; A es un subconjunto de B, en la figura 4.5. B se ha sombreado de forma horizontal, y A de manera vertical. Observe que $A \cup B = B$ (área sombreada completa) y $A \cap B = A$ (el área sombreada tanto horizontal como verticalmente). Todos los miembros de A están también en B, o todas las a son también b, si asumimos que a es igual a cualquier miembro de A y b es igual a cualquier miembro de B.

FIGURA 4.4



$A \cap B$

FIGURA 4.5



$A \subset B$

Operaciones con más de dos conjuntos

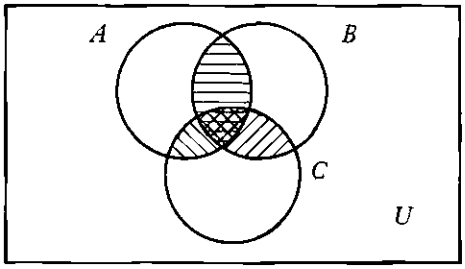
Las operaciones de conjunto no se limitan a dos subconjuntos de U. Sean A, B y C tres subconjuntos de U. Suponga que la intersección de estos tres subconjuntos de U no está vacía, como se muestra en la figura 4.6. El área con triple sombreado muestra $A \cap B \cap C$. Hay cuatro intersecciones, cada una sombreada de forma diferente: $A \cap B$, $A \cap C$, $B \cap C$ y $A \cap B \cap C$.

Aunque se pueden diagramar cuatro o más conjuntos, tales diagramas resultan difíciles de manejar, dibujar y revisar. No hay razón, sin embargo, para que las operaciones de intersección y unión no se apliquen simbólicamente a cuatro o más conjuntos.

Particiones y particiones cruzadas

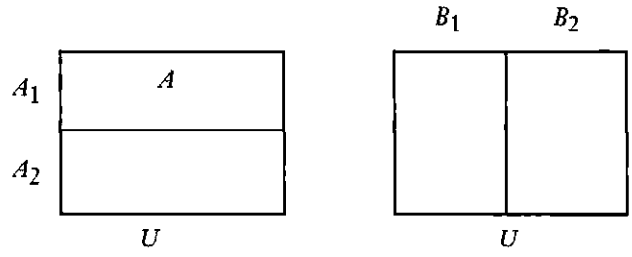
Nuestra discusión de conjuntos ha sido abstracta y quizás un poco monótona. Dejemos la discusión para examinar un aspecto de la teoría de conjuntos de gran importancia en la clarificación de los principios de categorización, análisis y diseño de investigación: la partición. U puede fraccionarse (partirse) en subconjuntos que no se intersequen y que agoten a U por completo. Cuando esto se hace, el proceso se denomina *partición*. Dicho formalmente, la partición fracciona un conjunto universal en subconjuntos *inconexos* y *exhaustivos* del conjunto universal.

FIGURA 4.6



$A \cap B \cap C$

FIGURA 4.7



Sea U un universo, y A y B los subconjuntos de U que son particiones. Llamamos a los subconjuntos de A: $A_1, A_2 \ldots, A_k$ y de B: $B_1 B_2 \ldots, B_m$. Las particiones por lo general se separan por corchetes mientras que los conjuntos y subconjuntos están marcados por paréntesis o llaves. Ahora, $[A_1 A_2]$ y $[B_1 B_2]$, por ejemplo son particiones si:

$$A_1 \cup A_2 = U \text{ y } A_1 \cap A_2 = E$$
$$B_1 \cup B_2 = U \text{ y } B_1 \cap B_2 = E$$

Los diagramas lo aclaran: la partición de U (representado por un rectángulo) por separado en los subconjuntos A_1 y A_2 y en B_1 y B_2 se muestra en la figura 4.7. Ambas particiones se han realizado en el mismo U. Hemos presentado algunos ejemplos de tales particiones: hombre-mujer, clase media-clase trabajadora, ingreso alto-ingreso bajo, demócrata-republicano, aprobado-reprobado, etcétera. Algunas son verdaderas dicotomías; otras no.

Es posible unir ambas particiones en una partición cruzada. Una *partición cruzada* es una nueva fracción que procede de sucesivas particiones del mismo conjunto U al formar todos los subconjuntos de la forma $A \cap B$. En otras palabras realizamos la partición A, y después la partición B en el mismo U, o en el mismo cuadrado, como se muestra en la figura 4.8. Cada casilla de la partición es una intersección de los subconjuntos A y B. Encontraremos en un capítulo posterior que tal partición cruzada es muy importante en el diseño de investigación y en el análisis de datos.

En forma anticipada a lo que se verá más adelante, he aquí un ejemplo de investigación de una partición cruzada. Dichos ejemplos se denominan tablas cruzadas o de contingencia. *Las tablas cruzadas* proveen la forma más elemental de mostrar una relación entre dos variables. El ejemplo es de un estudio de Miller y Swanson (1960), sobre prácticas de crianza infantil. Una de las tablas que usaron es una tabla cruzada cuyas variables son clase social (clase media y clase trabajadora) y destete (temprano y tardío). Los datos convertidos en porcentajes se muestran en la tabla 4.1. Las frecuencias reportadas por Miller y

FIGURA 4.8

		B	
		B_1	B_2
A	A_1	$A_1 \cap B_1$	$A_1 \cap B_2$
	A_2	$A_2 \cap B_1$	$A_2 \cap B_2$

TABLA 4.1 *Tabulación cruzada: relación entre clase social y destete (estudio de Miller y Swanson)*

		Destete	
		Temprano (B_1)	*Tardío* (B_2)
Clase social	*Clase media* (A_1)	60% (33)	40% (22)
	Clase trabajadora (A_2)	35% (17)	65% (31)

Swanson se pueden observar en el ángulo inferior derecho de cada casilla. Es evidente que existe una relación entre clase social y destete: las madres de clase social media muestran una tendencia a destetar a sus hijos más temprano que las de clase trabajadora. Se satisfacen las dos condiciones de inconexión y exhaustividad. La intersección de dos celdas cualesquiera está vacía. Por ejemplo, $(A_1 \cap B_1) \cap (A_1 \cap B_2) \cap (A_2 \cap B_1) \cap (A_2 \cap B_2) = E$. Las casillas agotan todos los casos: $(A_1 \cap B_1) \cup (A_1 \cap B_2) \cup (A_2 \cap B_1) \cup (A_2 \cap B_2) = U$.

La partición, por supuesto, va más allá de dos fracciones. En lugar de dicotomías, podemos tener politomías; en lugar de éxito-fracaso, por ejemplo, podemos tener éxito-éxito parcial-fracaso. Teóricamente una variable puede ser fraccionada en cualesquier número de subconjuntos, aunque por lo general hay limitaciones prácticas. Tampoco existe limitación teórica en el número de variables en una partición cruzada, pero las consideraciones prácticas por lo común limitan el número a tres o cuatro. El estudio de Foster, Dingman, Muscolino y Jankowski (1996) demuestra cómo se parte una variable en tres categorías. Su investigación es acerca de las decisiones relativas a contratación de empleados. Cada participante actuaba como gerente de recursos humanos y debía revisar tres currícula. El nombre en el documento determinaba el sexo del candidato. Los participantes debían recomendar un solicitante para la posición vacante. Los investigadores desarrollaron dos paquetes diferentes. En uno, un candidato masculino es el más calificado y la mujer es la de menores atributos. El segundo paquete es el reverso del anterior: la persona más calificada era mujer y el hombre era el menos apto. Había un tercer currículum de una persona cuyo sexo no podía ser determinado por el nombre. El cuadro que Foster y sus colaboradores presentan es una tabla cruzada en la que las variables son género del revisor/decisor (hombre y mujer) y tipo de currícula (altamente calificado masculino, menos apto masculino, altamente calificado femenino, menos apto femenino, altamente calificado sexo desconocido y menos apto sexo desconocido). Estos datos, convertidos en porcentajes, se muestran en la tabla 4.2. Las frecuencias obtenidas por Foster y colaboradores se presentan en la esquina inferior derecha de cada casilla, e ilustran una relación entre el género del revisor y el género del candidato. Las mujeres revisoras (quienes tomaban la decisión) tendían a seleccionar candidatos mujeres al hacer su recomendación. Los hombres revisores tendían a seleccionar candidatos masculinos aun cuando el candidato

TABLA 4.2 *Tabla de participación cruzada: relación entre género del revisor y calificación del solicitante (estudio de Foster et al.)*

Género del revisor	**Currículum A** Altamente calificado	**Currículum B** Menos calificado	**Currículum C** Menos calificado
(Paquete 1)	Jimena	Reme	Jorge
Femenino	50% (12)	33% (8)	17% (4)
Masculino	41% (7)	24% (4)	35% (6)
(Paquete 2)	Andrés	Michel	Carmen
Femenino	45% (9)	5% (1)	50% (10)
Masculino	53% (10)	26% (5)	21% (4)

femenino fuera superior en calificación. En un capítulo posterior se describirá con más detalle esta partición de variables.

Niveles del discurso

Cuando hablamos de cualquier asunto lo hacemos en un contexto o marco de referencia. Las expresiones, contexto y marco de referencia se relacionan de manera estrecha con *U*, el universo del discurso. Este último debe ser capaz de incluir cualesquier objetos de los que hablamos. Si vamos a otro *U* (otro nivel de discurso), el nuevo nivel no incluirá todos los objetos. De hecho, puede no incluir a objeto alguno. Si hablamos acerca de gente, por ejemplo, nosotros no hablamos —o mejor dicho, no deberíamos hablar— acerca de aves y sus costumbres a menos que de alguna forma relacionemos a las aves y sus costumbres con la gente, y que sea muy claro que esto es lo que hacemos. Hay dos niveles de discurso o universos (*U*) de discurso aquí: gente y aves. Cuando discutimos las implicaciones democráticas de la segregación, no debemos abordar de forma abrupta las preferencias religiosas, a menos que de alguna forma las relacionemos. Si lo hacemos, perderemos nuestro universo original de discurso, o no podremos asignar los objetos de un nivel (quizá religión) al otro nivel (la educación de los niños afroamericanos).

Para matizar esta presentación, cambiemos nuestro nivel de discurso a la música y el juicio y el entendimiento de diferentes géneros musicales. Una de las grandes dificultades al escuchar música moderna es que el sistema clásico de reglas que nuestros oídos han aprendido no se adecua al tipo de música de compositores como Bartók, Schöenberg o Ives. Uno tiene menos dificultades con Bartók y mucho más con Schöenberg e Ives porque el primero mantiene más bases clásicas que los otros dos. Examinemos la "Sonata Concordia" de Ives, en verdad un buen trabajo. La primera vez uno se sorprende por la aparente cacofonía y carencia de estructura. Después de escucharla varias veces, uno empieza a eliminar los juicios con base en el marco de referencia de la música clásica y a escuchar la belleza, significado y estructura del trabajo. El universo del discurso musical de Ives es simplemente muy diferente del universo del discurso clásico y resulta en extremo difícil desplazarse del *U* clásico al *U* de Ives. Algunas personas no pueden o incluso rechazan intentar este cambio. Encuentran la música de Ives extraña, inclusive repugnan-

te. Son incapaces de sacudirse el nivel de discurso de los conceptos estéticos y de juicio de la música clásica para hacer este desplazamiento.[1]

En la investigación debemos ser cuidadosos de no mezclar o desplazar nuestros niveles de discurso, y hacerlo sólo de manera consciente y con conocimiento. Pensar en términos de conjuntos nos ayuda a evitar estas mezclas y cambios. Como ejemplo extremo, supongamos que un investigador decide estudiar el entrenamiento para el control de esfínteres, el autoritarismo, las aptitudes musicales, la creatividad, la inteligencia, el aprovechamiento en lectura y el aprovechamiento escolar general de jóvenes de 3er. año de educación media. Aunque es concebible que se pueda extraer algún tipo de relaciones de este arreglo de variables, es más factible que se trate de una confusión intelectual. A cualquier costo, recuerde los conjuntos. Pregúntese: "Los objetos de los que discuto ¿pertenecen a un conjunto o conjuntos de mi presente discusión?" Si es así entonces usted está en un nivel de discurso. En caso contrario entra en la discusión otro nivel de discurso, otro conjunto, o conjunto de conjuntos. Si esto ocurre sin que se dé cuenta, el resultado es una confusión. En pocas palabras preguntemos: "¿Cuáles son U y los subconjuntos de U?"

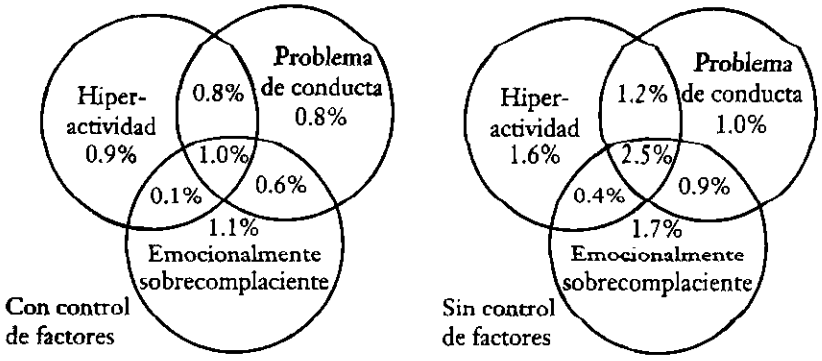
La investigación requiere definiciones precisas de los conjuntos universales. *Precisas* significa aportar una regla clara que nos diga cuándo un objeto es o no miembro de U. De manera similar, defina los subconjuntos de U y los subconjuntos de los subconjuntos de U. Si los objetos de U son personas, entonces no puede tener un subconjunto con objetos que no sean personas. (Aunque es posible tener un conjunto A de personas y el conjunto $\sim A$ de no–personas, esto nos lleva lógicamente a que U son personas. "No-personas" es, en este caso, un subconjunto de "personas", por definición o por convención.)

La idea de conjunto es fundamental en el pensamiento humano, en tanto que es probable que la mayor parte del pensamiento dependa de asignar cosas a categorías y de etiquetar esas categorías (véase Ross y Murphy, 1996; Smith, 1995). Lo que hacemos es agrupar clases de objetos —cosas, personas, eventos, fenómenos en general— y nombrar dichas clases. Tales nombres se convierten entonces en conceptos, etiquetas que no necesitamos aprender de nuevo y que pueden usarse para un pensamiento eficiente.

La teoría de conjuntos también constituye un instrumento general y ampliamente aplicable del pensamiento analítico y conceptual. Sus aplicaciones más importantes pertinentes a la metodología de la investigación son tal vez al estudio de las relaciones, la lógica, el muestreo, la probabilidad, la medición y el análisis de datos (véase Curtis, 1985; Hays, 1994). Sin embargo, la teoría de conjuntos puede aplicarse también a otras áreas y problemas no considerados técnicos en el sentido en que la probabilidad y la medición lo son. El uso de los conjuntos y los diagramas de Venn no es abundante en las ciencias de la conducta. Investigadores reconocidos a través de los años han empleado los conjuntos y diagramas de Venn en su trabajo de investigación. Piaget (1957), por ejemplo, usó el álgebra de conjuntos para tratar de explicar el pensamiento de los niños (véase también Piaget, García, Davidson y Easley, 1991). El trabajo clásico de Lewin (1935) en psicología Gestalt se apoya en conjuntos y diagramas de Venn para describir la interacción entre las personas y sus entornos, así como con ellos mismos. Más recientemente, Kubat (1993) aplicó conjuntos a estudios sobre aprendizaje de conceptos. Sheridan (1997) utilizó los diagramas de Venn para describir su marco conceptual para la intervención conjunta en cuestiones de conducta en psicología escolar. Sheridan establece que la identificación de problemas, el análisis de problemas, la implementación de planes y la evaluación del plan para la conducta de un niño puede explicarse como la intersección de los sistemas familiar, escolar y

[1] No queremos implicar que sea necesariamente deseable hacer el cambio, ni tampoco que toda la música moderna, aun toda la música de Ives, sea grandiosa o buena música. Tan sólo tratamos de mostrar la generalidad y aplicabilidad de las ideas de conjuntos y niveles de discurso.

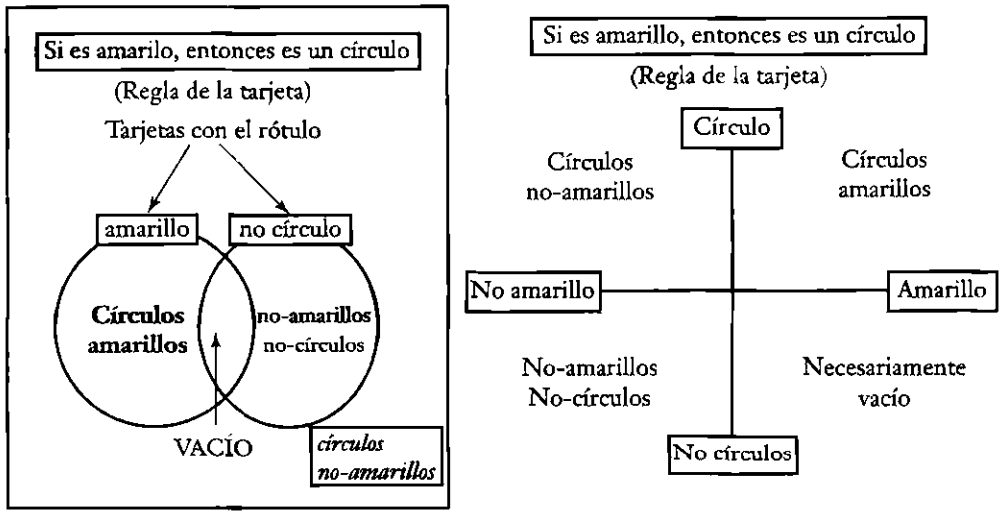
FIGURA 4.9



de apoyo del niño. Dayton (1976) emplea un diagrama de Venn para explicar cómo el individuo creativo confronta sus procesos mentales preconscientes y el mundo externo. Trites y Laprade (1983) los utilizan para ilustrar gráficamente un análisis de contingencias. Este análisis involucra un compuesto de puntuaciones factoriales en el estudio de la hiperactividad y el trastorno de conducta en niños (véase figura 4.9).

Bolman (1995) discute el papel y la necesidad del conocimiento de la ciencia de la conducta en la educación y práctica médicas. Para hacerlo, se apoya en diagramas de Venn para ilustrar lo que significa "fuerzas biopsicosociales", es decir, la intersección de las ciencias biológicas (anatomía, fisiología) con la psicología (sentimientos, identidad, metas) y la sociología y antropología (cultura, familia, ética). La combinación de estos tres factores constituye, en términos de Bolman, "la realidad clínica". Lane (1986) es un defensor de la enseñanza del razonamiento condicional (pensamiento lógico) a niños a través de diagramas de Venn y teoría de conjuntos. Lane condujo diversos estudios que comparaban diferentes materiales instruccionales para la enseñanza del pensamiento lógico. En cada uno de ellos

FIGURA 4.10



los diagramas de Venn (el enfoque de teoría de conjuntos) resultaron superiores a otros métodos en cuanto a desempeño inmediato, retención y transferencia del aprendizaje. La figura 4.10 presenta una muestra usada por Lane para comparar los diagramas de Venn con la tabla de lógica cartesiana. El concepto o la regla de la tarjeta bajo estudio es "si es amarillo, entonces es un círculo". Más adelante en este libro, se definirá medición a partir de una sola ecuación de teoría de conjuntos. Además, aclararemos los principios básicos del muestreo, del análisis y del diseño de investigación con base en conjuntos y teoría de conjuntos. Por desgracia, la mayor parte de los científicos sociales y educadores no están al tanto de la generalidad, poder y flexibilidad del pensamiento de conjuntos. Sin embargo, es posible predecir con certeza, que los investigadores en las ciencias sociales y educación encontrarán al pensamiento y teoría de conjuntos progresivamente más útiles para conceptualizar problemas teóricos de investigación.

RESUMEN DEL CAPÍTULO

1. Los conjuntos son útiles para entender los métodos de investigación. Son el fundamento del proceso descriptivo, lógico y del pensamiento y procesamiento analíticos. Constituyen la base del análisis numérico, categórico y estadístico.
2. Un conjunto es una colección bien definida de objetos o elementos. Dos formas de definir un conjunto son:
   a) Enlistar todos los miembros del conjunto.
   b) Aportar una regla para determinar si los objetos pertenecen o no al conjunto.
3. Los subconjuntos son partes del conjunto original. Si el conjunto entero constituye la población, entonces los subconjuntos son muestras de dicha población.
4. Las operaciones de conjuntos incluyen la intersección y la unión.
   a) La intersección está compuesta por los elementos comunes a dos o más conjuntos o subconjuntos; el símbolo es $\cap$.
   b) La unión es la combinación de los elementos no redundantes de dos o más conjuntos o subconjuntos; el símbolo es $\cup$.
5. El conjunto universal, U, se define como todos los elementos bajo consideración. Algunas veces se le llama población. El conjunto vacío, E, es el conjunto que no contiene miembros o elementos. También se llama conjunto *nulo*.
6. El conjunto de negación se simboliza por "~". Al colocar este símbolo antes de un conjunto se indica que contiene miembros del conjunto universal, U, que no están contenidos en el conjunto. Por ejemplo, $\sim A$ nos dice que este conjunto contiene todos los elementos en U que *no están en A*.
7. Una partición es la división de U en subconjuntos tales que cuando se combinan, U se restituye. Otra característica indispensable de una partición es que no haya elemento alguno de un subconjunto que se traslape con los elementos de otros subconjuntos.
8. Partición cruzada es la combinación de dos o más particiones diferentes. Las tablas de partición cruzada o tablas de contingencia son ejemplos de una partición cruzada que muestra la relación entre dos variables.

SUGERENCIAS DE ESTUDIO

1. Dibuje dos círculos traslapados, dentro de un rectángulo. Marque las siguientes partes: el conjunto universal U, los subconjuntos A y B, la intersección de A y B, y la unión de A y B.

a) Si estuviera trabajando en un problema de investigación con niños de 5o. grado, ¿qué parte del diagrama representaría a los niños de los que puede tomar su muestra?
b) ¿Qué representarían los conjuntos A y B?
c) ¿Qué podría significar la intersección de A y B?
d) ¿Cómo cambiaría el diagrama para representar un conjunto vacío? ¿Bajo qué condiciones un diagrama así tendría significado de investigación?

2. Considere la siguiente partición cruzada:

	Republicano (B_1)	Demócrata (B_2)
Masculino (A_1)		
Femenino (A_2)		

¿Cuál es el significado de los siguientes conjuntos?; es decir, ¿qué son, y cómo llamaríamos a cualquier objeto en estos conjuntos?
a) $(A_1 \cap B_1)$; $(A_2 \cap B_2)$
b) A_1; B_1
c) $(A_1 \cap B_1) \cup (A_1 \cap B_2) \cup (A_2 \cap B_1) \cup (A_2 \cap B_2)$
d) $(A_1 \cap B_1) \cup (A_2 \cap B_1)$

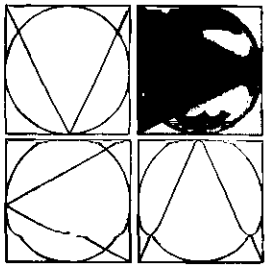
3. Genere una partición cruzada al utilizar las variables de estatus socioeconómico y preferencia de voto (demócrata y republicano). ¿Puede una muestra de individuos estadounidenses asignarse sin ambigüedad a las casillas de la partición cruzada? ¿Son exhaustivas las casillas? ¿No tienen traslapes? ¿Por qué son necesarias estas dos condiciones?
4. ¿Bajo qué condiciones puede ser verdadera la siguiente ecuación de conjuntos? [Observe que $n(A)$ representa el número de objetos en el conjunto A.]

$$n(A \cup B) = n(A) + n(B)$$

5. Suponga que un investigador en sociología quiere hacer un estudio sobre la influencia de la raza en el estatus ocupacional. ¿Cómo puede este investigador conceptualizar el problema en términos de conjuntos?
6. ¿Cómo están relacionados los conjuntos con las variables? ¿Podemos hablar sobre partición de variables? ¿Tiene sentido hablar de subconjuntos y variables? Explique.
7. Sea A = {Opus 101, Opus 106, Opus 109, Opus 110, Opus 111}, que es el conjunto de las cinco últimas sonatas para piano de Beethoven. Ésta es una definición de lista. He aquí una definición por regla:

$$A = \{a \mid a \text{ es una de las últimas cinco sonatas de Beethoven}\}$$

(El signo "|" se lee "dado"). ¿Bajo qué condiciones son mejores las definiciones de regla que las de lista?



CAPÍTULO 5

RELACIONES

Las relaciones son la esencia del conocimiento. Lo importante en ciencia no es el conocimiento de lo particular sino de las relaciones entre los fenómenos. Sabemos que las cosas son grandes sólo al compararlas con las más pequeñas. Así, establecemos las relaciones "mayor que" y "menor que". Los científicos educativos pueden "conocer" acerca del aprovechamiento sólo cuando estudian el aprovechamiento con relación al no aprovechamiento y a otras variables. Cuando saben que los niños de mayor inteligencia por lo general van bien en la escuela y que los niños de baja inteligencia con frecuencia lo hacen menos bien, ellos "conocen" una faceta importante del aprovechamiento. Cuando se percatan de que los niños de clase media tienden a desempeñarse mejor en la escuela que los de clase trabajadora, empiezan a comprender el "aprovechamiento". Conocen acerca de las relaciones que dan significado al concepto Aprovechamiento. Las relaciones entre Inteligencia y Aprovechamiento, entre Clase Social y Aprovechamiento y, de hecho, entre cualesquier variables constituyen el "meollo" básico de la ciencia.

La naturaleza relacional del conocimiento humano se ve con claridad aun cuando se analizan "hechos" en apariencia obvios. ¿Es dura una piedra? Para decir si este enunciado es cierto o falso primero debemos examinar conjuntos y subconjuntos de diferentes clases de piedras. Entonces, después de definir operacionalmente "duro" comparamos la "dureza" de piedras con otras "durezas". Los hechos más "simples" resultan, en el análisis, no

tan sencillos. Northrop (1947/1983, p. 317), al discutir sobre conceptos y hechos, afirma: "la única forma de obtener hechos puros, independientes de todo concepto y teoría, consiste tan sólo en observarlos para después permanecer perpetuamente mudo".

El diccionario nos dice que una *relación* es un vínculo, una conexión, un parentesco. Para la mayoría de las personas esta definición es suficientemente buena. Pero, ¿qué significan las palabras *vínculo*, *conexión* y *parentesco*? Otra vez, el diccionario dice que un *vínculo* es, una atadura, una fuerza que une; que una *conexión* es, entre otras cosas, una unión, una relación, una alianza. Pero una unión —un vínculo— ¿entre qué? Y ¿qué significa *unión*, *vínculo* y *fuerza que une*? Tales definiciones aunque intuitivamente útiles, son demasiado ambiguas para un uso científico.

Las relaciones como conjunto de pares ordenados

Las relaciones en ciencia siempre se dan entre clases o conjuntos de objetos. Uno no puede "conocer" la relación entre clase social y aprovechamiento escolar sólo al estudiar a un niño. "Conocer" la relación se logra sólo al abstraer la relación a partir de conjuntos de niños, o más precisamente, de conjuntos de características de niños. Permítanos tomar algunos ejemplos de relaciones y desarrollar de manera intuitiva un concepto de lo que es una relación.

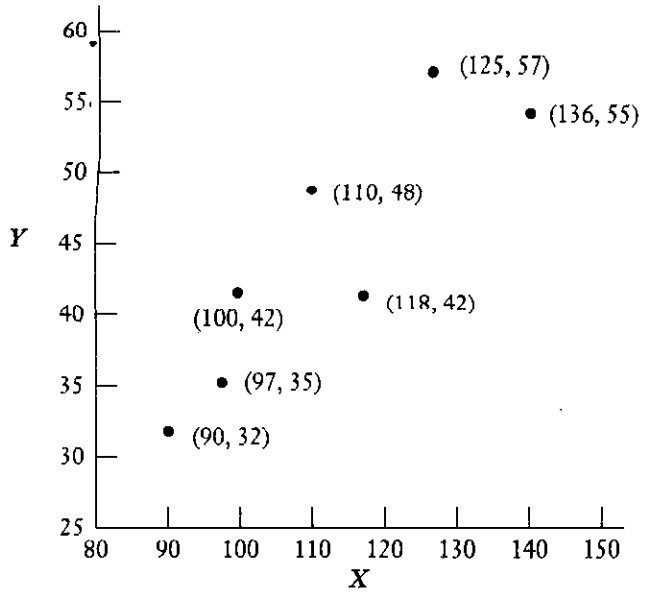
Sea A el conjunto de todos los papás, y B el de todos los hijos. Si pareamos a cada papá con su hijo (o hijos), tenemos la relación "papá-hijo". Podemos llamar a esta relación "crianza por parte del padre", aun cuando no se haya considerado a las hijas. De forma similar también podemos parear a los padres (elementos de A, en donde se considera a cada pareja de padres como un elemento individual) con sus hijos. Ésta constituirá la relación de "paternidad" o quizás, "familia". Sea A el conjunto de todos los esposos y B el conjunto de todas las esposas. El conjunto de parejas entonces define la relación "matrimonio". En otras palabras, un nuevo conjunto se ha formado, un conjunto de parejas en que se lista a los esposos siempre en primer término y a las esposas en segundo, y en donde cada esposo está aparejado sólo con su propia esposa.

Suponga que el conjunto A consiste de las puntuaciones de un grupo específico de niños en una prueba de inteligencia, y que B es la puntuación en una prueba de aprovechamiento. Si apareamos el coeficiente intelectual de cada niño con su puntuación de aprovechamiento, definimos una relación entre Inteligencia y Aprovechamiento. Observe que no podemos asignar con tanta facilidad un nombre como "paternidad" o "matrimonio" a está relación. Suponga que los conjuntos de puntuaciones son los siguientes:

Inteligencia	Aprovechamiento
136	55
125	57
118	42
110	48
100	42
97	35
90	32

Considere los dos conjuntos como un conjunto de pares, entonces este conjunto constituye una relación. Si graficamos ambos conjuntos de puntuaciones en los ejes X y Y, como se hizo en el capítulo 3 (figura 3.3), es más fácil "ver" la relación. Esto se hizo en la figura 5.1.

FIGURA 5.1

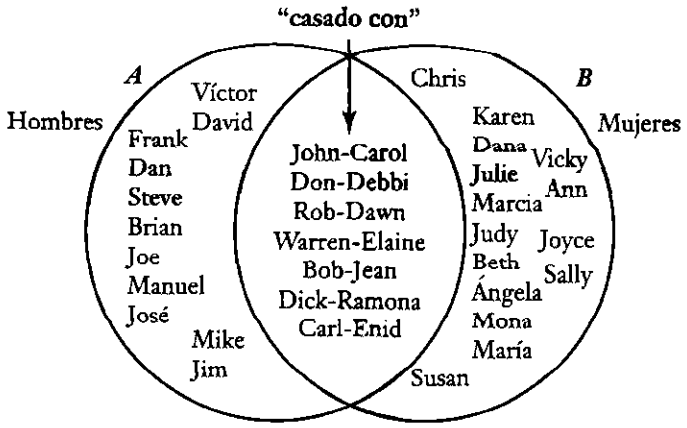


Cada punto se define por dos puntuaciones. Por ejemplo, el punto de la extrema derecha se define por (136, 55), y el de la extrema izquierda por (90, 32). Las gráficas como la figura 5.1 representan formas sucintas y muy útiles para expresar relaciones. Uno puede ver de inmediato, por ejemplo, que los valores más altos de X están acompañados por los valores más altos de Y, y los valores más bajos de X por los valores más bajos de Y. Como se verá en un capítulo posterior, también es posible trazar una línea a través de los puntos graficados de la figura 5.1, de la parte inferior izquierda a la parte superior derecha. (El lector debería tratar de hacerlo.) Esta línea, llamada *línea de regresión*, expresa la relación entre X y Y, entre inteligencia y aprovechamiento, pero también nos da, de una forma sucinta, considerablemente más información acerca de la relación: a saber: su dirección y magnitud.

Ahora estamos listos para definir "relación" de una manera formal: *una relación es un conjunto de pares ordenados*. Cualquier relación constituye un conjunto, un conjunto de cierta clase: un conjunto de pares ordenados. Un *par ordenado* está formado por dos objetos, o por un conjunto de dos elementos, en el que hay un orden fijo de aparición de los objetos. En realidad, hablamos de pares ordenados que significa (como antes se indicó) que los miembros de *cada par* siempre aparecen en un orden determinado. Si los miembros de los conjuntos A y B están pareados, entonces debemos especificar si los miembros de A o los de B aparecen primero en cada par. Si definimos la relación de matrimonio, por ejemplo, especificamos el conjunto de pares ordenados con, digamos, los esposos siempre en primer lugar en cada par. En otras palabras, el par (a, b) no equivale al par (b, a). Los pares ordenados están encerrados en paréntesis y marcamos: ( ), y un conjunto de pares ordenados se indica: $\{(a, k)\ (b, l), (c, m)\}$.

Por fortuna hemos dejado atrás la definición ambigua del diccionario. La definición de relaciones como conjuntos de pares ordenados, aunque puede parecer un poco extraña

FIGURA 5.2



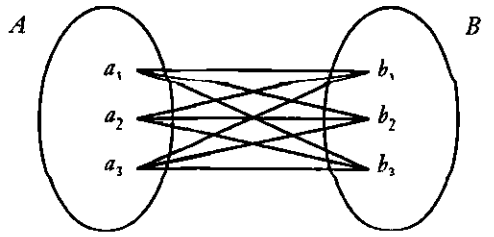
y aun curiosa al lector, no es ambigua y resulta general. Más aún, el científico, como el matemático, puede trabajar con ella.

Al discutir sobre relaciones, hay dos tipos especiales de conjuntos que juegan un papel importante. Uno se llama el *dominio* y el otro se llama la *imagen*. En lugar de definirlos de manera formal de inmediato, puede resultar más claro si consideramos un ejemplo: sea A el conjunto de todos los hombres y B el conjunto de todas las mujeres. Digamos que queremos definir la relación "casado con". Podemos hacerlo al formar la intersección apropiada de A y B (es decir, $A \cap B$) de manera que cada pareja ordenada en la intersección consistiera en las parejas casadas (mostrado en la figura 5.2). La intersección consiste en las parejas casadas. En este caso, el dominio en esta relación son los hombres que están casados. La imagen serían todas las mujeres que están casadas. Así, el dominio sería el conjunto de hombres que están en la intersección $A \cap B$ y la imagen sería el conjunto de mujeres que están en la intersección. El dominio es {John, Don, Rob, Warren, Bob, Dick y Carl}. La imagen es {Carol, Debbi, Dawn, Elaine, Jean, Ramona y Enid}. Observe que el dominio de la relación es siempre un subconjunto de A, y la imagen de la relación es un subconjunto de B.

De manera formal, si permitimos que RL represente la relación, que a sean los elementos del conjunto A, y b los elementos del conjunto B, entonces el dominio de RL es el conjunto de todas las cosas a tal que, para algunos b, el par ordenado (a, b) está en RL. La imagen (que también se llama contradominio) de la relación RL es el conjunto de todas las cosas b tal que, para algunas x, el par ordenado (a, b) está en RL.

Definir el dominio y la imagen en una relación es importante porque juegan un papel clave al definir una *función*. Hays (1994) considera a la función como uno de los conceptos más importantes en matemáticas y ciencia. Las funciones y las relaciones son muy similares. Se puede considerar que una función es una clase especial de relación. Una relación es una función cuando cada elemento del dominio está pareado con un miembro y sólo con uno de la imagen. La mayoría de las personas concibe a la función en términos numéricos, pero esto no es necesariamente así. En la sociedad estadounidense, por ejemplo, la relación de ser esposo es una función, en tanto que un hombre tiene a lo más una esposa en cualquier momento dado. Sin embargo, la relación de ser una madre no es una función ya que una persona puede ser madre de varios niños. Si lo vemos con cuidado, la relación de ser la hija de una madre sí es una función, en tanto que esa niña puede tener sólo una

FIGURA 5.3



madre biológica. Así, una función es un conjunto de pares ordenados, en donde no hay dos pares distintos que posean los mismos elementos.

Determinación de relaciones en la investigación

Aunque hemos evitado la ambigüedad con nuestra definición de relaciones, no hemos aclarado en especial el problema práctico de "determinar" relaciones. Existe otra forma de definir una relación que nos puede ayudar. Sean A y B conjuntos. Si pareamos de manera individual cada miembro de A con cada miembro de B, obtendremos *todos los pares posibles* entre ambos conjuntos, lo que se denomina el *producto cartesiano* de los dos conjuntos y se enuncia $A \times B$. Una relación se define como un subconjunto de $A \times B$; es decir, cualquier subconjunto de pares ordenados tomados de $A \times B$ constituye una relación. (véase Kershner y Wilcox, 1974, para un excelente análisis de relaciones).

Para ilustrar esta idea de manera sencilla, sea el conjunto $A = \{a_1, a_2, a_3\}$ y el conjunto $B = \{b_1, b_2, b_3\}$.[1] Entonces, el producto cartesiano, $A \times B$, puede diagramarse como se muestra en la figura 5.3. Esto es, generamos nueve pares ordenados: (a_1, b_1), (a_1, b_2), ..., (a_3, b_3). Con conjuntos grandes, por supuesto, habrá muchos pares, de hecho tendremos mn pares, donde m y n son la cantidad de elementos en A y B, respectivamente.

Esto no es muy interesante —al menos en este contexto—. ¿Qué hacemos para determinar o "descubrir" una relación? De manera empírica determinamos qué elementos de A "van con" qué elementos de B, de acuerdo con algún criterio. Es obvio que hay muchos subconjuntos de pares de $A \times B$, la mayoría de los cuales no "tienen sentido" o no nos interesan. Kershner y Wilcox (1974) afirman que una relación es "un método para distinguir ciertos pares ordenados de otros; es un esquema para señalar determinados pares de todos los demás". De acuerdo con esta forma de concebir las relaciones, la relación de "matrimonio" es un método o procedimiento para distinguir las parejas casadas de todos los posibles pares de hombres y mujeres. De esta forma podemos incluso considerar a la religión como una relación. Sea $A = \{a_1, a_2, \ldots, a_n\}$ el conjunto de todas las personas en los Estados Unidos y sea B = {Católica, Protestante, Judía, etcétera} el conjunto de religiones. Si ordenamos los pares, en este caso cada persona con una religión, entonces tenemos la "relación" de religión o, para ser más precisos, la "afiliación religiosa". Para evitar que el estudiante esté confundido por la extraña sensación de definir una relación como un subconjunto de $A \times B$, diremos otra vez que es natural que muchos de los posibles subconjuntos de pares ordenados $A \times B$ no tengan sentido. Quizá lo más importante

[1] Los subíndices sólo etiquetan y distinguen a los miembros individuales de los conjuntos. No implican orden alguno. Observe también que no es necesario que haya igual número de miembros en ambos conjuntos.

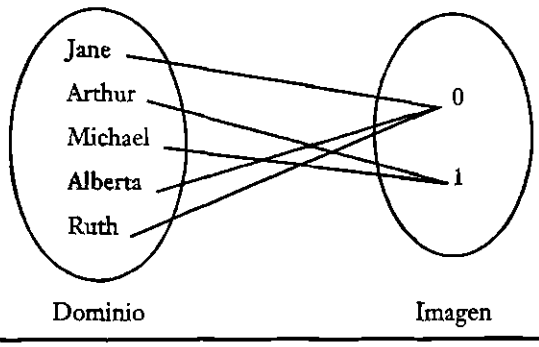
sea que nuestra definición de relación no presenta ambigüedad y es general por completo. No importa qué conjuntos de pares ordenados elijamos, constituyen una relación. Nos corresponde decidir cuáles conjuntos tienen sentido científico según el dictado de los problemas para los que buscamos respuestas y cuáles no.

El lector puede preguntarse por qué nos tomamos tantas molestias en definir las relaciones. La respuesta es simple: casi toda la ciencia busca y estudia relaciones. Literalmente no existe forma empírica para "conocer" nada, excepto a través de sus relaciones con otras cosas, como lo indicamos antes. Si, como en el caso de Behling y Williams (1991), nos interesamos en la percepción de la inteligencia y las expectativas del aprovechamiento escolar, es necesario que relacionemos la percepción y la expectativa con otras variables. Para explicar un fenómeno como la percepción de la inteligencia debemos "descubrir" sus determinantes —las relaciones que tiene con otras variables pertinentes—. Behling y Williams "explicaron" las percepciones de los maestros sobre la inteligencia y las expectativas del aprovechamiento escolar hacia los estudiantes al relacionarlos con el tipo y estilo de vestimenta de los estudiantes de educación media superior. Encontraron que el estilo de vestimenta influye en las percepciones tanto de los maestros como de los compañeros. Es evidente que, si las relaciones son fundamentales en la ciencia, debemos saber con claridad qué son, al igual que cómo se estudian. Se ha descuidado la definición de "relación" en la investigación del comportamiento. Parece que se asume que es un concepto cuyo significado todos conocen. También se le confunde con "relación" que es una conexión de alguna clase entre la gente, o entre la gente y los grupos como una relación madre-hijo. *No es lo mismo que una relación.*

Reglas de correspondencia y mapeo

Cualquier objeto —gente, números, resultados en juegos de azar, puntos en el espacio, símbolos, etcétera— pueden ser miembros de conjuntos y pueden relacionarse como pares ordenados. Se dice que los miembros de un conjunto son *mapeados* en los miembros de otro conjunto al usar una regla de correspondencia. Una *regla de correspondencia* es una receta o una fórmula que nos dice cómo mapear los objetos de un conjunto en los objetos de otro conjunto. Nos indica en pocas palabras, cómo debe realizarse la correspondencia entre los miembros de los conjuntos. Estudie la figura 5.4, que muestra la relación entre los nombres de cinco individuos y los símbolos 1 y 0, que representan masculino (1) y femenino (0). Tenemos aquí un mapeo de sexo (1 y 0) en los nombres. Esto es por supuesto, una relación donde cada nombre tiene asignado un 1 o un 0, masculino o femenino.

FIGURA 5.4



En una relación, los dos conjuntos cuyos "objetos" se relacionan se llaman *el dominio* y la *imagen* o *contradominio*, o D y C. D es el conjunto de los primeros elementos y C el conjunto de los segundos elementos. En la figura 5.4, le asignamos 1 a masculino y 0 a femenino. A cada miembro del dominio se asigna el miembro apropiado de la imagen. D = {Jane, Arthur, Michael, Alberta, Ruth}, y C = {0, 1}. La regla de correspondencia, dice: si el objeto de D es femenino asigne un "0", si es masculino un "1".

En otras palabras, los objetos, especialmente los números, se asignan a otros objetos —personas, lugares, números, etcétera— de acuerdo a la regla. El proceso es muy variado en sus aplicaciones pero simple en su concepción. En lugar de pensar en todas las diferentes formas de expresar las relaciones por separado, caemos en la cuenta de que todos son conjuntos de pares ordenados y que los objetos de un conjunto simplemente se mapean en los objetos del otro conjunto. Todas las variadas formas de expresar relaciones —como mapeos, correspondencias, ecuaciones, conjuntos de puntos, tablas, o índices estadísticos— pueden reducirse a conjuntos de pares ordenados.

Algunas formas de estudiar relaciones

Podemos expresar las relaciones de varias formas. En el análisis previo, se ilustraron algunas. Una de ellas consiste simplemente en listar y parear los miembros de los conjuntos, como en las figuras 5.3 y 5.4. De hecho, este método no se usa con frecuencia en la literatura de investigación. Ahora se examinarán algunos procedimientos más útiles.

Gráficos

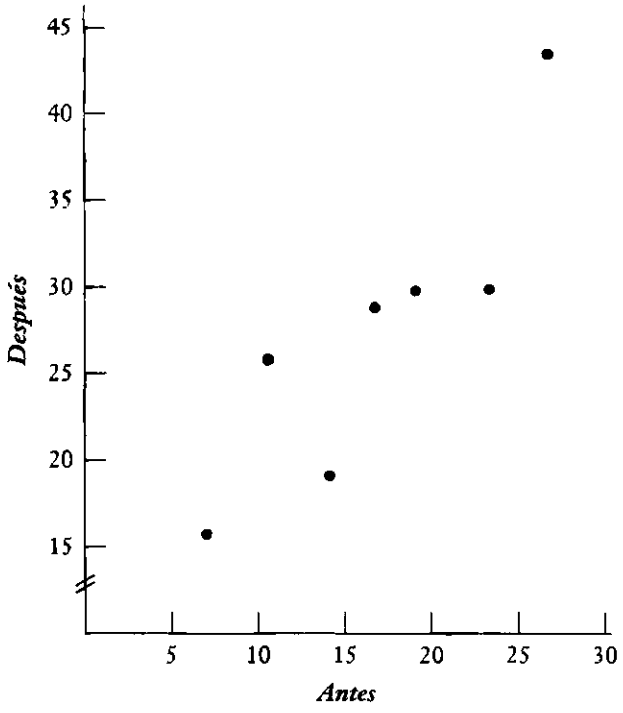
Un gráfico es un dibujo en el que los dos miembros de cada par ordenado de una relación se trazan en dos ejes, X y Y (o cualquier otra designación apropiada). La figura 5.1 es un gráfico de pares ordenados de las puntuaciones ficticias de inteligencia y aprovechamiento que se dieron antes. Podemos ver en la gráfica que los pares ordenados tienden a "ir juntos": los valores altos de Y van con los valores altos de X, y los valores bajos de Y van con los valores bajos de X.

Un conjunto más interesante de pares ordenados está graficado en la figura 5.5. Los números usados para este gráfico fueron tomados de un estudio fascinante de Miller y DiCara (1968), en el que se entrenó a siete ratas para secretar orina. (Dado que la secreción de orina es una función autónoma está normalmente más allá de control y por lo tanto, del entrenamiento y aprendizaje.) El "antes" o eje de las X de la gráfica indica valores de secreción de orina antes del entrenamiento; el "después" o eje de las Y indica valores después del entrenamiento. Utilizaremos los mismos datos en otro contexto más adelante en el libro (y describiremos el estudio más detenidamente), por ahora no habrá más detalles. La relación entre los dos conjuntos de valores de secreción de orina es pronunciada. Otra vez, valores elevados antes del entrenamiento se acompañan por valores altos después del entrenamiento, y sucede algo similar con los valores bajos. El gráfico y la relación que expresa refleja las diferencias individuales en la secreción de orina. El significado completo de este enunciado será claro cuando más tarde describamos el análisis estadístico de estos datos.

Tablas

Quizás la forma más común de presentar datos para mostrar relaciones es a través de tablas. Las variables de las relaciones presentadas se señalan por lo general en la parte alta

FIGURA 5.5



	Antes	Después
1	23	30
2	14	19
3	16	29
4	18	30
5	7	16
6	26	44
7	12	26

y a los lados de la tabla, mientras que los datos están contenidos en su interior. Los datos estadísticos son casi siempre medias, frecuencias y porcentajes. Considere la tabla 5.1, que es una presentación resumida de los datos de frecuencias presentados por Freedman, Wallington y Bless (1967). Estos investigadores sometieron a prueba la noción de que el acatamiento se relaciona con la culpa: a mayor sentimiento de culpa, mayor será el acatamiento a las peticiones. Los experimentadores indujeron a la mitad de los sujetos a mentir, y supusieron que hacerlo produciría culpa (al parecer así fue). La variable Culpa o Mentira está localizada en la parte superior de la tabla. Ésta es la variable independiente, por supuesto. La variable dependiente fue el acatamiento a las peticiones hechas. Esta variable está etiquetada a un lado de la tabla. Los datos en las casillas de la tabla son frecuencias; esto es, el número de sujetos que cayeron dentro de los subconjuntos o subcategorías. De los 31 sujetos inducidos a mentir, 20 acataron las demandas del

TABLA 5.1 *Resultados en frecuencias del experimento para estudiar la relación entre culpa y acatamiento (estudio de Freedman, Wallington y Bless).*

	Mentir (Culpa)	Sin mentir (No culpa)
Acatamiento	20	11
No acatamiento	11	20
	31	31

experimentador y 11 no. De los 31 sujetos que no fueron inducidos a mentir, 11 acataron y 20 no lo hicieron. Los datos son consistentes con la hipótesis. En un capítulo posterior se estudiará en detalle cómo analizar e interpretar datos de frecuencia y tablas de esta naturaleza.

El punto esencial de la tabla 5.1 es que la relación y la evidencia de la naturaleza de la relación se expresan en ella. En este caso los datos se presentan en forma de frecuencias. (Una frecuencia es el número de miembros de conjuntos y subconjuntos. Un porcentaje es una razón o proporción por ciento. Se calcula al multiplicar 100 veces la razón de un subconjunto a un conjunto, o de un subconjunto a otro subconjunto.) La tabla en sí constituye una partición cruzada, con frecuencia llamada tabulación cruzada o tabla de contingencia en la que una variable de la relación se entrecruza contra otra variable de la relación. Los nombres de ambas variables aparecen en la parte superior y al lado de la tabla, como se indicó antes. La dirección y magnitud de la relación en sí misma se expresa por los tamaños relativos de las frecuencias en las casillas de la tabla. En la tabla 5.1 muchos más sujetos (20 de 31) inducidos a mentir acataron la petición que los sujetos no inducidos a mentir (11 de 31).

Se presenta un ejemplo más complicado en la tabla 5.2. Se trata de una presentación resumida de los datos de frecuencia de un estudio de Mays y Arrington (1984). Estos investigadores sometieron a prueba la noción de que la violación de la territorialidad o límites espaciales está relacionada con características demográficas. Las personas con características de un bajo estatus (raza o género) tienden a que se les viole su espacio con mayor frecuencia. Los experimentadores crearon 10 condiciones para representar 10 configuraciones diádicas. Estas diadas se generaron al cruzar dos niveles de raza (caucásico y afroamericano) con dos niveles de sexo (femenino y masculino); por ejemplo, "afroamericano femenino con caucásico masculino". Los confederados con dichas especificaciones demográficas para cada una de 10 condiciones planteadas se mantuvieron a una distancia confortable para conversar y sostuvieron una discusión casual. Dos observadores ocultos registraron la ruta que siguieron unas 210 personas que ingresaban en cada condición. Anotaron las tendencias para atravesar (y así penetrar los límites de las diadas) o pasar alrededor de ellas. También registraron el grupo étnico y el género de los que ingresaban. La variable independiente es la configuración diádica. La variable dependiente es la penetración en los límites de la diada o la no penetración (pasar alrededor de ella). Aunque Mays y Arrington consideraron muchas variables en el estudio, por ejemplo, aquí sólo usaremos la combinación diádica de sexo (masculino-masculino, femenino-femenino, femenino-masculino). Las combinaciones se encuentran etiquetadas en la parte superior de la tabla 5.2 y la variable dependiente está al lado. La tabla muestra tanto las frecuencias como los porcentajes.

Los porcentajes totales para las combinaciones masculino-masculino y femenino-femenino fueron muy similares (31% *vs.* 29%). También, los porcentajes de la conducta de rodear para las diadas masculino y femenino fueron muy parecidos (30% *vs.* 28%). Sin

TABLA 5.2 *Resultados en frecuencias y porcentajes del experimento para estudiar la relación entre diadas sexuales y violación del espacio (estudio de Mays y Arrington)*

	Masculino-Masculino	Femenino-Femenino	Femenino-Masculino	
Alrededor	650 (30)	600 (28)	898 (42)	2148 (86)
A través de	106 (33)	119 (37)	98 (30)	323 (14)
	756 (31)	719 (29)	996 (40)	2471

embargo, la tabla muestra que hay una tendencia de las diadas femenino-masculino para rodear (42%). Uno puede especular, a partir de ello, que hay un mayor respeto por el espacio compartido por parte de combinaciones de sexo diferente que del mismo. Mays y Arrington dan crédito a esta noción al señalar los porcentajes que aparecen en la violación o en los datos de "a través de" (segundo reglón de la tabla 5.2). En términos de porcentajes, la porción de espacio de las diadas femeninas se viola con más frecuencia que las masculinas (37% *vs.* 33%). Para las diadas con diferente combinación de sexo (femenino-masculino), el espacio se invade menos que para las otras (30%). En un capítulo posterior se estudiará en detalle cómo analizar e interpretar los datos de frecuencia (porcentaje) y las tablas de esta clase.

La tabla 5.2 muestra la naturaleza del formato de la tabla de relación. En este caso, los datos están en forma de frecuencias y porcentajes. Usar los porcentajes de la tabla 5.2 para un análisis visual simple, resulta más fácil que utilizar las frecuencias puras. Con los porcentajes, en la tabla 5.2, el valor máximo es 100 y el mínimo es 0. Hay más invasores de la diada femenino-femenino (119 de 323 o 37%) que en la diada masculino-masculino (106 de 323 o 33%).

Una clase diferente de tabla presenta medias, promedios aritméticos, en el cuerpo del cuadro. Las medias expresan la variable dependiente. Si hay sólo una variable independiente, sus categorías estarán etiquetadas en la parte superior de la tabla. Si hay dos o más variables independientes, sus categorías pueden presentarse de varias formas en la parte superior y a los lados, como se verá en capítulos posteriores. Se da un ejemplo en la tabla 5.3, que constituye la forma más simple que una tabla puede asumir. Hyatt y Tingstrom (1993) estudiaron el efecto del uso de jerga conductual en la percepción de los maestros hacia dos intervenciones conductuales: reforzamiento y castigo. Los hallazgos en investigaciones previas en el efecto de la jerga de los maestros no fueron concluyentes. En este estudio, se presentó a los maestros con un estudiante hipotético con un problema conductual. Luego se les dio la descripción de dos tipos de intervenciones. La descripción contenía o bien jerga conductual tal como "condicionado de manera operante y comportamiento apropiado incompatible", o bien, una descripción sin jerga con palabras tales como "se le premió por sentarse correctamente". Los maestros que recibieron la jerga conductual pueden considerarse como del grupo experimental, mientras que los que recibieron instrucciones sin jerga constituyeron el grupo control. Todas las percepciones de los maestros fueron medidas por medio del Inventario de Evaluación del Tratamiento (también referido como TEI), que permite a los maestros evaluar las intervenciones en términos de cómo perciben la aceptabilidad, adecuación, justicia y efectividad. Las puntuaciones variaron de entre 15 a 105. Las puntuaciones altas indicaban una mayor aceptabilidad. Como puede verse en la tabla 5.3, la media del grupo experimental es mayor que la del grupo control. ¿Es "grande" o "pequeña" la diferencia entre las medias? Más adelante se verá cómo evaluar el tamaño y el significado de tales diferencias. Por ahora, sólo nos interesa por qué la tabla expresa una relación.

En tablas de esta clase siempre se expresa o implica una relación. Las que son tan simples como ésta, rara vez se usan en la literatura. Basta sólo mencionar ambas medias en

TABLA 5.3 *Medias de los grupos con y sin jerga (estudio de Hyatt y Tingstrom)*[a]

Experimental (jerga)	Control (sin jerga)
79.38	73.68

[a] Las medias se calcularon a partir del Inventario de Evaluación del Tratamiento.

el texto de un informe. Más aún, pueden compararse más de dos medias. Sin embargo, el principio es el mismo; las medias "expresan" la variable dependiente y las diferencias entre ellas, el efecto supuesto de la variable independiente. En este caso hay dos variables relacionadas: la jerga y la percepción. La rúbrica "experimental y control" expresa la jerga recibida por el grupo experimental pero no por el control. Ésta es la variable independiente. Las dos medias en el cuadro expresan la percepción del maestro con relación al método de intervención, medida por el TEI: ésta es la variable dependiente. Si las medias difieren lo suficiente, puede asumirse que la jerga conductual tuvo un efecto en la percepción de los maestros o su aceptabilidad.

Las tablas de medias son en extremo importantes en la investigación del comportamiento, en especial en investigación experimental. Puede haber dos, tres o más variables independientes, y pueden expresar los efectos individuales y combinados de estas variables en una variable dependiente, o incluso en dos o más variables dependientes. El punto central es que siempre se estudian las relaciones, aun cuando no siempre es fácil conceptualizarlas y establecerlas.

Gráficos y correlación

Aunque antes examinamos de forma breve las relaciones y gráficas, puede ser provechoso ahondar en este tópico. Suponga que tenemos dos conjuntos de puntuaciones de los mismos individuos en dos pruebas, X y Y:

X	Y
1	1
2	1
2	2
3	3

Los dos conjuntos forman un conjunto ordenado de pares. Este conjunto constituye, por supuesto, una relación. También puede escribirse, usando R para denotar relación, $R = \{(1,1), (2,1), (2,2), (3,3)\}$. Se diagrama en la gráfica de la figura 5.6.

Con frecuencia podemos darnos una idea aproximada de la dirección y grado de la relación al inspeccionar la lista de pares ordenados, pero es un método impreciso. Las

FIGURA 5.6

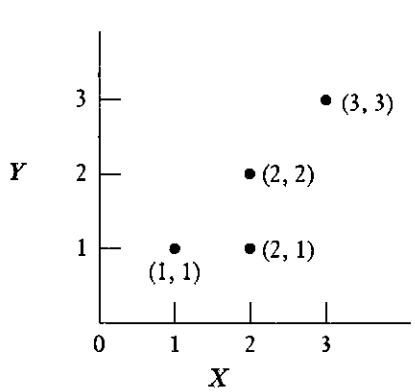
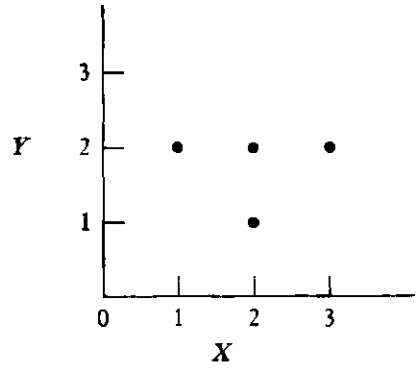


FIGURA 5.7



gráficas, como la de las figuras 5.1 y 5.6, nos dicen más. Es más fácil "ver" que los valores de Y "se mueven" con los valores de X. Los valores más altos de X acompañan a los valores más altos de Y, y viceversa. En este caso, la relación —o correlación, como también se le llama— es positiva. (Si tuviéramos la ecuación $R = \{(1,3), (2,1), (2,2), (3,1)\}$, la relación sería negativa. El lector debe graficar estos valores y observar su dirección y significado). Si la ecuación fuera $R = \{(1,2), (2,1), (2,2), (3,2)\}$, la relación sería nula o cero; esto se grafica en la figura 5.7. Puede verse que los valores de Y no "se mueven" con los valores de X de ninguna forma sistemática. Esto no significa que "no haya" relación. Siempre existe una relación —por definición— en tanto que existe un conjunto de pares ordenados. Sin embargo, es común que se diga que "no hay" relación. Resulta más preciso decir que la relación es nula o cero.

Los científicos sociales por lo general calculan índices de relación, con frecuencia llamados coeficientes de correlación, entre conjuntos de pares ordenados para obtener estimaciones más precisas de la dirección y grado de las relaciones. Si calculamos uno de estos índices, el coeficiente de correlación producto-momento, o r, para los pares ordenados de la figura 5.6, obtenemos $r = .85$. Para los pares de $R = \{(1,3), (2,1), (2,2), (3,1)\}$ dijimos que la relación era negativa, $r = -.85$. Para los pares de la figura 5.7, el conjunto de pares mostró una relación nula o cero, $r = 0$.[2]

El coeficiente producto-momento y otros coeficientes de correlación, están basados en la variación concomitante de los miembros de conjuntos de pares ordenados. Si *covarían* —varían juntos altos valores con altos valores, valores medios con valores medios, y valores bajos con valores bajos, o valores altos con valores bajos, etcétera— se dice que hay una relación positiva o negativa, en su caso. Si no covarían, se dice que "no hay" relación. Los índices más útiles varían de +1.00 a través de 0 hasta –1.00. Un +1.00 indica una relación positiva perfecta, –1.00 indica una relación negativa perfecta, y un 0 indica relación no discernible (o relación cero). Algunos índices van sólo de 0 a +1.00. Otros índices pueden tomar valores distintos.

La mayoría de los coeficientes de relación nos dicen qué tan similares son los órdenes de posición de los dos conjuntos de medidas. La tabla 5.4 presenta tres ejemplos para ilustrar estos órdenes de posición o formas de "moverse" juntos. Se presentan los coefi-

[2] Los métodos para calcular estas r y otros coeficientes de correlación se discuten en los textos de estadística, que incluyen un análisis de mayor profundidad sobre la interpretación de coeficientes de correlación de la que es posible en este libro.

TABLA 5.4 *Tres conjuntos de pares ordenados con diferentes direcciones y grados de correlación.*

(*I*) $r = 1.00$		(*II*) $r = -1.00$		(*III*) $r = 0$	
X	*Y*	*X*	*Y*	*X*	*Y*
1	1	1	5	1	2
2	2	2	4	2	5
3	3	3	3	3	3
4	4	4	2	4	1
5	5	5	1	5	4

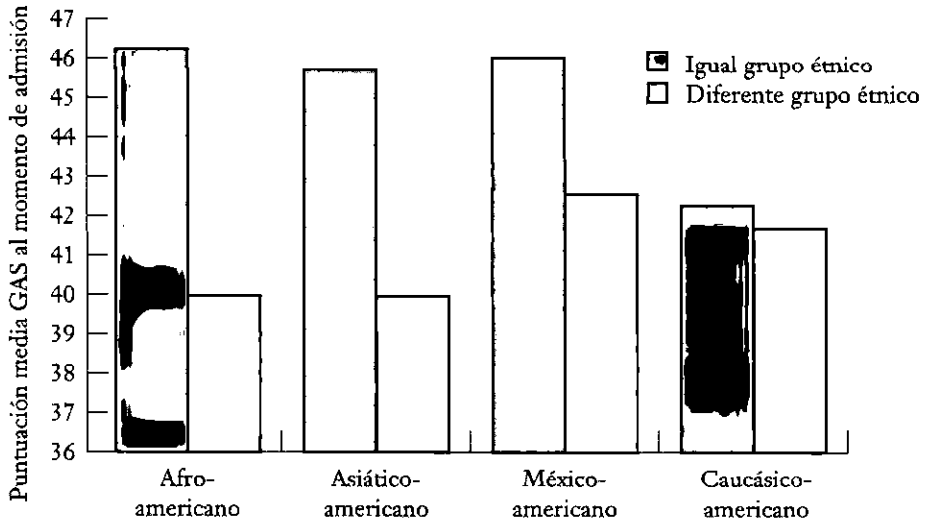
cientes de correlación con cada uno de los conjuntos de pares ordenados. *I* es obvio: el orden de las puntuaciones *X* y *Y* de *I* se mueven perfectamente en conjunto. Así lo hacen las puntuaciones *X* y *Y* de *II*, pero en la dirección opuesta. En *III*, no hay relación discernible entre la posición del orden. En *I* y *II*, uno puede predecir a la perfección el valor de *Y* dado el valor de *X*, pero en *III* uno no puede predecir los valores de *Y* a partir de *X*. Rara vez los coeficientes de correlación son 1.00 o 0; por lo común toman valores intermedios.

Ejemplos de investigación

Para ilustrar nuestra abstracta discusión sobre las relaciones, veamos dos ejemplos interesantes de relaciones y correlación. Russell, Fujino, Sue, Cheung y Snowden (1996) examinaron los efectos de la etnicidad en las diadas terapeuta-cliente en la evaluación del funcionamiento mental. Estos investigadores usaron un gran conjunto de datos de clientes adultos atendidos en los servicios de consulta externa de una importante clínica metropolitana de salud mental. De estos datos, los investigadores extrajeron cuatro grupos étnicos para el estudio: asiático-americanos, afro-americanos, méxico-americanos, y caucásico-americanos. La Escala de Evaluación Global (GAS por sus siglas en inglés) obtenida al momento de la admisión, se usó como medida del funcionamiento mental. El terapeuta del caso asignó la puntuación del GAS al cliente. Las puntuaciones altas indicaban un buen funcionamiento general, mientras que las bajas señalaban una severa disfunción.

Russell y sus colaboradores examinaron las puntuaciones del GAS para aquellos clientes cuyo origen étnico era el mismo del terapeuta (por ejemplo, terapeuta asiático-americano con cliente asiático-americano; terapeuta afro-americano con cliente afro-americano, etcétera) contra las puntuaciones del GAS de aquellos clientes que difirieron en cuanto a origen étnico con relación a su terapeuta. La figura 5.8 muestra la relación entre unos y otros, y sus puntajes GAS. Observe que aquí la relación se presenta en forma de una gráfica diferente a las que hemos visto antes. Este formato se denomina *histograma* o *gráfico de barras*. El gráfico muestra consistentemente que las puntuaciones GAS fueron más altas (mejor salud mental) cuando los terapeutas fueron del mismo origen étnico que los clientes, que cuando no lo fueron. Esto significa que el terapeuta percibió al cliente con un mayor nivel del funcionamiento en cuanto a salud mental cuando el cliente tenía el mismo origen étnico, que cuando el cliente era de un grupo étnico diferente. También para las minorías étnicas, cuando había coincidencia de terapeuta-cliente en este sentido, se presentaba una mayor puntuación GAS que en el caso de caucásicos. Sin embargo, para los clientes en cuyos casos no había correspondencia étnica, la puntuación GAS fue la más alta para los méxico-americanos y la más baja para los afro-americanos y asiático-americanos. La mayor discrepancia entre las relaciones con y sin coincidencia étnica en la puntuación de GAS fue en los grupos minoritarios.

FIGURA 5.8



Nuestro segundo ejemplo no es cuantitativo, aunque implica cantidad y no sería difícil cuantificar las variables. Hardy (1974) estudió, entre otras cosas, la relación entre afiliación religiosa y productividad doctoral. ¿Qué grupo religioso produce doctorados más sobresalientes y cuál los menos? (Hardy estudiaba en realidad valores y su influencia en la escolaridad.) Los resultados se presentan en la tabla 5.5. Requieren comentario: es aparente que la relación es fuerte; mientras más liberal es un grupo religioso, mayor producción de grados doctorales. Los pares ordenados de grupos religiosos y su evaluación de productividad se ven fácilmente.

Nuestro último ejemplo es un estudio de Little (1997). Constituye una variante del estudio de Hardy presentado antes. A diferencia de esa investigación, el estudio de Little incluye un alto nivel de cuantificación: estudió la relación entre las universidades que ofrecen grados y la productividad académica en el campo de la psicología escolar. La pregunta fue: ¿qué posgrado universitario produce a los graduados más sobresalientes? La respuesta es importante porque aportará información más allá de los resultados de la investigación previa, que se basó sobre todo en la afiliación institucional de los autores sin

TABLA 5.5 *La relación entre afiliación religiosa y resultados de doctores sobresalientes en los Estados Unidos (estudio de Hardy).*

Tipo de religión	Evaluación de productividad
Liberal, protestante secularizado y judíos	Alta productividad
Liberal moderado, disidente, protestante antitradicional	Productividad arriba del promedio
Protestante tradicional	Productividad moderada
Fundamentalistas y protestantes conservadores	Baja productividad
Católicos	Productividad muy baja

TABLA 5.6 *La relación entre la educación a nivel posgrado y el número de graduados en psicología escolar, de 1987 a 1995 (estudio de little).*

Orden	Universidad	Número de graduados	Total ponderado
1	Georgia	19	65.98
2	Indiana	10	52.48
3	Minnesota	13	40.88
4	Texas	12	38.61
5	Wisconsin	11	27.92
6	Columbia	10	27.60
7	California, Berkeley	7	27.24
8	South Carolina	4	22.39
9	Oregon	5	21.48
10	Ball State	7	20.32
11	Ohio State	7	19.56
12	Kent State	5	19.36
13	Nebraska	8	18.31
14	Arizona State	2	16.28
15	Utah	5	15.90
16	Temple	5	15.88
17	Indiana State	4	15.61
18	Illinois	2	13.43
19	Southern Mississipi	7	13.36
20	Connecticut	4	12.75
21	Michigan State	5	12.38
22	Pittssburgh	1	11.97
23	Cincinnati	5	11.91
24	Pennsylvania	3	11.03
25	Penn State	4	10.09

proporcionar información acerca de dónde fueron educados. El estudio Little provee esa información al presentar los datos sobre el sitio en que los autores recibieron su grado terminal. Little establece que esta medida constituye un mejor indicador de la calidad de los programas de posgrado en psicología escolar. Una reproducción parcial de los resultados totales se muestra en la tabla 5.6. Little muestra que la mayoría de los programas de grado en Estados Unidos se concentra en particular en las regiones del medio oeste, sureste y en la costa este. Entre los datos de Little está un conjunto de pares ordenados que relacionan universidad y productividad. Little, en este estudio, encontró un buen número de discrepancias entre sus hallazgos y los publicados por *US News & World Report* (1995) sobre los mejores programas de Estados Unidos en esta disciplina. Los resultados de Little se basaron en datos empíricos reunidos en las seis revistas más importantes de investigación en psicología escolar. *US News & World Report* basó sus ordenamientos en la reputación de la universidad.

Relaciones multivariadas y regresión

En nuestro análisis sobre relaciones pudimos haber dado la impresión de que los científicos e investigadores están siempre preocupados por las relaciones entre dos variables. Cuando, por ejemplo, hablamos acerca de las relaciones entre coincidencia de origen étnico y salud mental, jerga y percepción, institución de posgrado y producción de trabajo académico, es

posible que hayamos generado la idea errónea de que los científicos se preocupan por estudiar sólo relaciones de dos variables. Esto no es así. De hecho, ha habido un gran número de investigación de dos variables, pero en las ciencias del comportamiento esto ha cambiado de forma dramática. La preocupación de los investigadores de la conducta en la actualidad tiene más que ver con las relaciones múltiples. Mientras que los investigadores modernos saben que la relación entre inteligencia y aprovechamiento es sustancial y positiva, también asumen que existen muchos determinantes para una y otra variable. Saben, por ejemplo, que la clase social tiene una influencia decisiva en ambas. También creen, aunque hay evidencia contradictoria, que la autoestima afecta a los dos. Más aún, los metodólogos han desarrollado poderosos enfoques y métodos analíticos para manejar lo que llamaremos problemas multivariados. Revisemos de manera breve la lógica y sustancia de estos problemas.

Algo de lógica de la investigación multivariada

La estructura oculta de nuestro argumento hasta ahora ha sido resumida por la expresión "si p, entonces q": "si inteligencia, entonces aprovechamiento"; "si estatus bajo, entonces violación del espacio"; "si este estilo de vestimenta, entonces esta percepción de inteligencia". Representan, desde luego, relaciones implicadas. Pero van más allá: también implican una dirección: de las variables independientes a las variables dependientes. Pueden concebirse como "si p, entonces q", que en lógica se denomina enunciado condicional. Es posible conceptualizar a la mayoría de los problemas de investigación y estudiar la estructura de los argumentos científicos al usar enunciados condicionales y otros relacionados (Kerlinger, 1969). Sin embargo, las relaciones de la investigación del comportamiento son más complejas que un simple enunciado "si p, entonces q". Es más probable que los investigadores contemporáneos digan algo como "si p, entonces q bajo las condiciones r y t". Este enunciado condicional puede escribirse como: $p \rightarrow q \mid r, t$, que se lee como en la oración precedente (" $\mid$ " significa "bajo las condiciones" o "dado"). De forma más simple podemos escribir: $(p_1, p_2, p_3) \rightarrow q$ que significa "si p_1 y p_2 y p_3, entonces q". De manera más concreta, esto significa que las variables p_1 y p_2 y p_3 influyen en la variable q en ciertas formas. Podemos decir, por ejemplo, que la inteligencia, clase social y autoestima afectan el rendimiento escolar de tal y cual formas.

La forma más simple de demostrar las relaciones en forma gráfica es a través de los diagramas de ruta. Un diagrama de ruta para el enunciado anterior se puede observar en la figura 5.9. En él, donde usamos x_1, x_2 y x_3 para las variables independientes y y para la

FIGURA 5.9

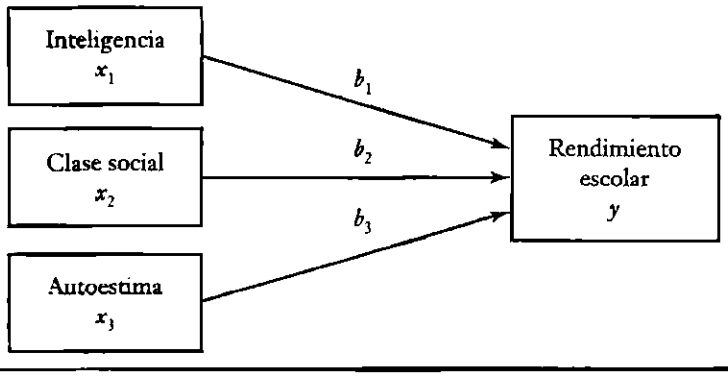
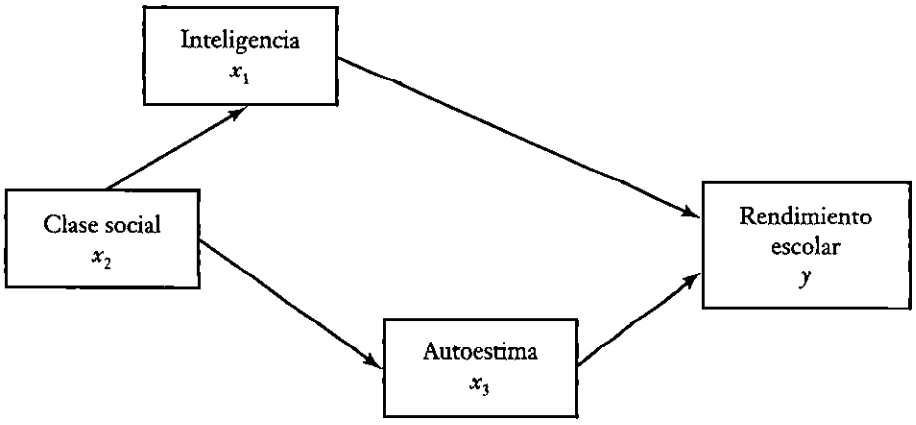


FIGURA 5.10



variable dependiente, se especifica en efecto, que las tres variables independientes afectan en forma directa a la variable dependiente. Esto se llama un problema de regresión múltiple directa (ver abajo) en el que k(=3) variables independientes influyen mutuamente en una variable dependiente. Este enfoque también ha cambiado en forma dramática en la década pasada. Los investigadores están ahora preparados para hablar de y probar tanto influencias directas como indirectas. Un modelo alternativo y un diagrama de ruta analítico se presentan en la figura 5.10. Aquí la Inteligencia y la Autoestima influyen en el Rendimiento Escolar en forma directa, pero la clase Social, no. En su lugar, influye de manera indirecta en el Rendimiento Escolar *a través* de la Inteligencia y la Autoestima, lo cual es un concepto bastante diferente.

Relaciones múltiples y regresión

La situación de investigación descrita en la figura 5.9 es un problema de regresión múltiple: k (=3) variables independientes influyen de forma mutua y simultánea en una variable dependiente. Más adelante se mostrará cómo se resuelve un problema de esta naturaleza. (El método es técnicamente complejo aunque conceptualmente simple, pero nos costará algo de trabajo.) Por ahora, el problema es encontrar inicialmente la relación entre las tres variables independientes, tomadas de forma simultánea, y la variable dependiente. El segundo punto consiste en determinar en qué medida cada variable independiente, x_1, x_2 y x_3, influye en la variable dependiente, y. Aunque es ahora mucho más complejo, el problema aún es una relación, un conjunto de pares ordenados.

Lo que el método hace en esencia —y bellamente— es encontrar la mejor combinación posible de x_1, x_2 y x_3 dada y, así como las relaciones entre las cuatro variables, de tal forma que la correlación entre la combinación de las tres variables y y sea máxima. En el problema mostrado en la figura 5.9, la regresión múltiple encuentra los valores de b_1, b_2 y b_3 tales que hagan la correlación entre x_1, x_2 y x_3 tomadas juntas, y y tan alta como sea posible. (El estudiante de matemáticas reconocerá que se trata de un problema de cálculo.) Los pesos de b, llamados pesos de regresión o coeficientes, se usan entonces con las tres variables para predecir la variable dependiente, y. El método, en efecto, crea una nueva variable que es una combinación de x_1, x_2 y x_3. Llamemos a esta variable y'. Así, la

correlación múltiple es entre y, la variable dependiente observada, y y', la variable dependiente predicha a partir del conocimiento de x_1, x_2 y x_3.

El lector que está pendiente y atento observará que las relaciones y correlaciones son simétricas: con frecuencia no importa cuál es la variable dependiente y cuál la independiente. En el análisis de regresión, sin embargo, sí implica una diferencia en tanto que la regresión es asimétrica. Nosotros decimos: si x entonces y o: si x_1, x_2 y x_3, entonces y. Muchos autores hablan de "análisis causal", en especial cuando se refieren a problemas como el de las figuras 5.9 y 5.10. Nosotros preferimos evitar las palabras *causa y causal* porque son ideas excesivamente difíciles —por ejemplo, ¿qué es una causa?— y porque no resulta necesario usarlas. Comrey y Lee (1992, p. 338) establecen que "las inferencias causales no pueden hacerse con certeza. Lo más que se puede decir es que los datos son consistentes con la inferencia causal propuesta..." Así, podemos operar de manera adecuada con enunciados condicionales, aunque no siempre con facilidad.[3]

La regresión en otras palabras, trata con relaciones, aunque en general es un camino de una sola vía, de variables independientes a dependientes. Para anticipar una discusión posterior, veamos una ecuación de regresión:

$$Y = a + b_1X_1 + b_2X_2$$

Si ignoramos a —que no es importante para el argumento— podemos ver que Y es la suma de X_1 y X_2, cada una ponderada por su propia b. Cuando resolvemos la ecuación para las bs (y por supuesto, la a), las usamos para producir un resultado de Y' para cada persona en la muestra. Y y Y' (conserve en la mente que Y y Y' representan valores para cada persona en la muestra) son entonces un conjunto de pares ordenados y, por lo tanto, una relación. La correlación entre ellas es meramente un coeficiente de correlación ordinario, r. Pero se le denomina R y se le llama coeficiente múltiple de correlación o coeficiente de correlación múltiple. Más adelante se examinará el uso e interpretación de la regresión múltiple, el coeficiente de correlación múltiple, así como los pesos de la regresión con mayor detalle y con ejemplos de investigación reales. En este momento el asombro natural del lector a partir de los misterios supuestos del pensamiento multivariado deben haberse disipado y reemplazado por admiración y quizá un poco de sobrecogimiento y excitación por lo atractivo y poderoso de estas ideas y métodos.

RESUMEN DEL CAPÍTULO

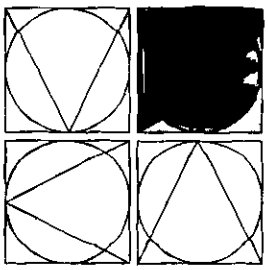
1. Las relaciones son la esencia del conocimiento. Casi toda ciencia busca y estudia relaciones.
2. Las relaciones en la ciencia se dan entre clases o conjuntos de objetos.
3. Las relaciones pueden expresarse como conjuntos de pares ordenados.
4. Los pares ordenados son conjuntos de elementos con un orden fijo de aparición.
5. Hay dos conjuntos especiales: dominio e imagen de una relación.
6. La función es un tipo especial de relación que conecta elementos del dominio y de la imagen.
7. Los miembros de un conjunto se mapean en los miembros del otro conjunto por medio de una regla de correspondencia.

[3] El lenguaje está saturado con palabras que implican causa, por ejemplo, "influencia" y "dependencia". Evitamos al máximo el uso de expresiones causales por la sola razón de que nunca es posible decir sin ambigüedades que una cosa causa otra. De forma más pragmática, no necesitamos la palabra o concepto "causa"; los enunciados condicionales de si p, entonces q son suficientes para los propósitos científicos.

8. La regla de correspondencia es una receta o fórmula que muestra cómo se mapea los objetos.
9. Formas de estudiar las relaciones:
   a) Gráficos (de dos dimensiones)
   b) Tablas
   c) Gráficos y correlación (donde la correlación es un valor estadístico/numérico).
10. La regresión múltiple es un método estadístico que relaciona una variable dependiente a una combinación lineal de una o más variables independientes. Este procedimiento incluso puede decirle al investigador en qué medida cada variable independiente explica o se relaciona a la variable dependiente.

SUGERENCIAS DE ESTUDIO

1. Los análisis sobre relaciones parecen estar confinados a los textos de matemáticas. El mejor análisis que hemos encontrado, aunque abstracto y algo difícil, está en Kershner y Wilcox (1974).
2. A continuación se presentan seis ejemplos de relaciones. Suponga que el primer conjunto es el dominio y el segundo, la imagen. ¿Por qué son todas estas relaciones?
   a) Páginas del libro y número de las páginas
   b) Números de los capítulos y páginas de un libro
   c) Encabezados o categorías en una tabla de población y las cantidades de población en un informe de censos
   d) Niños de una clase de 3er. grado y sus puntuaciones en una prueba estandarizada de rendimiento.
   e) $Y = 2x$
   f) $Y = a + b_1X_1 + b_2X_2$
3. Un investigador educativo ha estudiado la relación entre ansiedad y rendimiento escolar. Exprese la relación en el lenguaje de conjuntos.
4. Suponga que desea estudiar las relaciones entre las siguientes variables: inteligencia, estatus socioeconómico, necesidad de logro y rendimiento escolar. Elabore dos modelos alternativos que "expliquen" el rendimiento escolar. Dibuje diagramas de ruta para cada uno de los modelos.
5. Determine cuáles de las siguientes relaciones son funciones:
   a) $(\neq Q, R \neq, \neq S, T \neq)$
   b) $(\neq w, j \neq, \neq k, l \neq, \neq p, q \neq)$
   c) $(\neq a, b \neq, \neq 102, 103 \neq, \neq a, c \neq)$
   d) Dado $A = \{a, b, c\}$ y $B = \{4, 5, T\}$, ¿el producto cartesiano cruzado $A \times B$ es una función? Explique por qué sí o por qué no.



CAPÍTULO 6

VARIANZA Y COVARIANZA

Para estudiar problemas científicos y responder preguntas científicas debemos estudiar las diferencias entre fenómenos. En el capítulo 5 examinamos relaciones entre variables; en un sentido, estudiábamos similitudes. Ahora nos concentraremos en las diferencias, ya que sin diferencias ni variaciones no hay manera técnica de determinar las relaciones entre variables. Si queremos estudiar la relación entre raza y aprovechamiento, por ejemplo, estamos indefensos si tan sólo contamos con medidas de aprovechamiento de los niños caucásicos estadounidenses. Debemos tener medidas de aprovechamiento de más de una raza. En síntesis, es necesario que la raza varíe; debe tener varianza. Se requiere explorar el concepto de varianza de manera analítica y con cierta profundidad. Para hacerlo adecuadamente es necesario también retirar algo de la crema de la leche de las estadísticas.

Estudiar conjuntos de números como tales resulta pesado. En general es necesario reducir los conjuntos en dos formas: 1) por medio del cálculo de promedios o medidas de tendencia central, y 2) por medio del cálculo de medidas de variabilidad. La medida de tendencia central usada en este libro es la media. La medida de variabilidad más usada es la *varianza*. Ambas clases de medidas sintetizan conjuntos de puntuaciones, pero de diferentes maneras. Ambas constituyen "resúmenes" de conjuntos completos de puntuaciones y expresan dos importantes facetas de dichos conjuntos: 1) su tendencia o promedio central,

y 2) su variabilidad. Resolver problemas de investigación sin estas medidas es en extremo difícil. Empezaremos el estudio de la varianza con algunos cálculos simples.

Cálculo de medias y varianzas

Tomemos el conjunto de números $X = \{1, 2, 3, 4, 5\}$. La media se define:

$$M = \frac{\Sigma X}{n} \tag{6.1}$$

n equivale al número de casos en el conjunto de puntuaciones; Σ significa "la suma de" o "súmelos" y X representa a cualquiera de las puntuaciones (cada puntuación es una X). Entonces, la fórmula se lee: "sume las puntuaciones y divida entre el número de casos en el conjunto".

$$M = \frac{1 + 2 + 3 + 4 + 5}{5} = \frac{15}{5} = 3$$

La media del conjunto X es 3. En este libro "M" representará la media. Otros símbolos que se usan con frecuencia son $\bar{X}$ y μ.

El cálculo de la varianza, aunque no tan sencilla como el de la media, aún es simple. La fórmula es:

$$V = \frac{\Sigma x^2}{n} \tag{6.2}$$

donde V es la varianza; n y Σ son lo mismo que en la ecuación 6.1. Σx^2 se denomina suma de cuadrados (esto necesita algo de explicación). Las puntuaciones se enlistan en una columna:

	X	x	x^2
	1	−2	4
	2	−1	1
	3	0	0
	4	1	1
	5	2	4
ΣX:	15		
M:	3		
Σx^2:			10

En este cálculo, x es una desviación de la media. Se define como:

$$x = X - M \tag{6.3}$$

Así, para obtener x tan sólo reste de X la media de todas las puntuaciones. Por ejemplo si $X = 1$, $x = 1 - 3 = -2$; en el caso de $X = 4$, $x = 4 - 3 = 1$; etcétera. Esto se hizo en la tabla anterior. La ecuación 6.2, sin embargo, indica que se elevó al cuadrado cada x. Esto también se hizo antes (recuerde que el cuadrado de un número negativo siempre es positivo). En otras palabras, $\sum x^2$ señala que hay que restar la media a cada puntuación para obtener x, elevar al cuadrado cada x para obtener x^2, y entonces sumar todas las x^2. Por último el promedio de las x^2 se genera al dividir $\sum x^2$ entre n, el número de casos. $\sum x^2$, la *suma de cuadrados*, es un estadístico muy importante que usaremos con frecuencia.

La varianza en el presente caso es

$$V = \frac{(-2)^2 + (-1)^2 + (0)^2 + (1)^2 + (2)^2}{5} = \frac{4 + 1 + 0 + 1 + 4}{5} = \frac{10}{5} = 2$$

"V" representará la varianza en este libro. Otros símbolos comúnmente usados son σ^2 y s^2. El primero también se llama valor poblacional; el segundo es un valor muestral. N se usa para representar el número total de casos en una muestra total o en una población. (Se definirá "muestra" y "población" en un capítulo posterior.) n se usa para una submuestra o subconjunto de U de una muestra total. Los subíndices apropiados se anexarán y explicarán conforme sea necesario. Por ejemplo, si queremos indicar el número de elementos en el conjunto A, subconjunto de U, podemos escribir n_A o n_a. De manera similar agregaremos subíndices para x, V, etcétera. Cuando se use doble subíndice, como r_{xy}, el significado por lo general resultará obvio.

La varianza se llama también *cuadrado medio* (cuando se calcula de una forma ligeramente diferente). Se llama así porque es evidente que equivale a la media de las x^2. Es claro que no es difícil calcular la media y la varianza.[1]

La pregunta es: ¿Por qué calcular la media y la varianza? El fundamento para el cálculo de la media se explica fácilmente. La media expresa el nivel general, el centro de gravedad, de un conjunto de medidas. Es un buen representante del nivel de las características o rendimiento de un grupo. También posee ciertas propiedades estadísticas deseables y es el estadístico más frecuente en las ciencias del comportamiento. En gran parte de este tipo de investigación, por ejemplo, se compara la media de diferentes grupos experimentales para estudiar relaciones como se señaló en el capítulo 5. Podemos probar la relación entre climas organizacionales y productividad, por ejemplo. Pudimos usar tres tipos de climas y estar interesados en conocer qué clima tiene el mayor efecto en la productividad. En tales casos, las medias por lo general se comparan. Por ejemplo, de tres grupos, cada uno operando bajo alguno de los tres climas A_1, A_2 y A_3, ¿cuál tiene la mayor media en, digamos, una medida de productividad?

La base lógica para los cálculos y uso de la varianza en investigación es más difícil de explicar. En el caso usual de puntuaciones ordinarias, la varianza es una medida de dispersión del conjunto de puntuaciones: nos dice qué tanto se dispersan los valores. Si un grupo de alumnos es muy heterogéneo en cuanto a rendimiento en lectura, la varianza de sus puntuaciones en este rubro será muy grande comparada con la varianza de un grupo que es homogéneo en este aspecto. La varianza, entonces, es una medida de la dispersión de las puntuaciones; describe la medida en que las puntuaciones difieren entre sí. Para propósi-

[1] El método para calcular la varianza usado en este capítulo difiere de otros que se usan de ordinario. De hecho, el presentado aquí es impracticable en la mayoría de las situaciones. Nuestro propósito no es aprender estadística como tal. Lo importante es ir tras las ideas básicas. Se han construido métodos para realizar cálculos, ejemplos y demostraciones para ayudar en esta búsqueda de ideas básicas.

tos descriptivos, por lo general se usa la raíz cuadrada de la varianza, y se denomina *desviación estándar*. Algunas propiedades matemáticas, sin embargo, hacen a la varianza más útil en investigación. Se sugiere que el estudiante complemente este tópico con las secciones apropiadas de un texto elemental de estadística (véase Comrey y Lee, 1995). No es posible discutir en este libro todas las facetas del significado e interpretación de medias, varianzas y desviaciones estándar. El resto de este capítulo y las partes posteriores de este libro explorarán otros aspectos del uso del estadístico varianza.

Tipos de varianza

La varianza viene en varias formas. Cuando lea literatura especializada y de investigación, con frecuencia se topará con el término, algunas veces con algún adjetivo calificativo, otras no. Para entender la literatura es necesario tener una idea clara de las características y propósitos de las diferentes varianzas. Para diseñar y hacer investigación, uno debe tener un profundo entendimiento del concepto de varianza así como un dominio amplio de los conceptos y manipulaciones estadísticas de la varianza.

Varianza poblacional y muestral

La *varianza poblacional* es la varianza de U, un universo o población de medidas. En general, se usan símbolos griegos para representar los parámetros o medidas poblacionales. Para la varianza poblacional, se usa el símbolo σ^2 (sigma cuadrada). El símbolo σ se utiliza para la desviación estándar poblacional. La media poblacional es μ (mu). Si se conocen todas las medidas de un conjunto universal definido, U, entonces se conoce la varianza. Con frecuencia, sin embargo, la totalidad de medidas de U no están disponibles. En tales casos, la varianza se estima al calcular la varianza de una o más de las muestras de U. Una buena parte de la energía estadística se dirige a esta importante actividad. Quizá surja una pregunta: ¿Qué tanto varía la inteligencia de los ciudadanos de Estados Unidos? Se trata de una pregunta poblacional o de U. Si hubiera una lista completa de todos los millones de gente en Estados Unidos —y también un listado completo de las puntuaciones de pruebas de inteligencia de estas personas— la varianza podría calcularse en forma simple, aunque pesada. Tal lista no existe. Entonces se prueban las muestras, muestras representativas, de estadounidenses y se calculan las medias y varianza. Se utilizan muestras para estimar la media y varianza de toda la población. Estos valores estimados se llaman estadísticos (en la población se denominan parámetros). La media muestral se denota por el símbolo M y la varianza muestral por SD^2 o s^2. Muchos libros de estadística usan $\bar{X}$ (X-barra) para representar la media muestral.

La varianza de muestreo es la varianza de los estadísticos calculados a partir de muestras. Las medias de cuatro muestras aleatorias tomadas de una población diferirán. Si el muestreo es aleatorio y las muestras lo suficientemente grandes, las medias no deben variar mucho; esto es, la *varianza de las medias* deberá ser relativamente pequeña.[2]

[2] Por desgracia, en la mayor parte de la investigación actual sólo está disponible una muestra —y con frecuencia es pequeña—. Podemos, sin embargo, estimar la varianza muestral de las medias por medio de *la varianza estándar de la media*. (El término "error estándar de la media" se usa con frecuencia y es la raíz cuadrada de la varianza estándar de la media.) La fórmula es $V_M = V_S/n_S$ donde V_M es la varianza estándar de la media, V_S es la varianza de la muestra, y n_S es el tamaño de la muestra. Observe una conclusión importante que puede rescatarse

Varianza sistemática

Quizás la forma más general de clasificar la varianza es en varianza sistemática y *varianza del error*. La *varianza sistemática* es la variación en las medidas debidas a influencias conocidas o desconocidas que "causan" que las puntuaciones se inclinen a una dirección más que a otra. Cualquier influencia natural o generada por el hombre que cause que los eventos sucedan de una forma predecible son influencias sistemáticas. Las puntuaciones de rendimiento en pruebas de niños de una adinerada escuela suburbana tienden a ser sistemáticamente más altas que las de los alumnos de una escuela en un barrio pobre. Los maestros expertos pueden de manera sistemática influir en el aprovechamiento de los niños, en contraste con aquellos que reciben una enseñanza deficiente.

Hay muchas causas de la varianza sistemática. Los científicos buscan separar aquellas que les interesan de aquellas que no. También intentan separar la varianza aleatoria de la varianza sistemática. De hecho, la investigación puede definirse de forma precisa y técnica, como el estudio controlado de varianzas.

Varianza entre grupos (experimental)

Un tipo importante de varianza sistemática en la investigación es la varianza entre grupos o varianza experimental. La *varianza entre grupos* o *experimental*, como su nombre lo indica, es aquella que refleja diferencias sistemáticas entre *grupos* de medidas. La varianza discutida previamente como varianza de puntuaciones refleja las diferencias entre individuos en un grupo. Podemos decir por ejemplo, que, con base en la evidencia y pruebas actuales, la varianza en la inteligencia de una muestra aleatorizada de niños de 11 años de edad es de cerca de 225 puntos. (Se obtiene al elevar al cuadrado la desviación estándar derivada del manual de una prueba. La desviación estándar de la Prueba California de Madurez Mental para niños de 11 años de edad, por ejemplo, es de alrededor de 15, y $15^2 = 225$.) Esta cantidad es un estadístico que nos indica la medida en que los individuos difieren uno de otro.

La varianza entre grupos, por otro lado, es aquella que se debe a las diferencias entre *grupos* de individuos. Si se mide el aprovechamiento de niños de la región norte y de la región sur en escuelas comparables, habría diferencias entre los grupos del norte y del sur. Los grupos al igual que los individuos difieren o varían, y es posible y apropiado calcular la varianza entre estos grupos.

La varianza entre grupos y la experimental son en esencia iguales. Ambas provienen de las diferencias entre grupos. La varianza entre grupos es un término que abarca todos los casos de diferencias sistemáticas entre grupos, tanto experimentales como no experimentales. La varianza experimental con frecuencia se asocia con la varianza originada por la manipulación activa de las variables independientes por parte de los investigadores.

He aquí un ejemplo de varianza entre grupos —en este caso, varianza experimental—. Suponga que un investigador prueba la eficacia relativa de tres diferentes clases de reforzamiento en el aprendizaje. Después de reforzar a los tres grupos de sujetos de forma diferencial, el experimentador calcula la media de los grupos. Suponga que son 30, 23 y 19. La media de las tres medias es 24, y calculamos la varianza *entre las medias* o *entre los grupos*:

---

de esta ecuación: si se incrementa el tamaño de la muestra, V_M disminuye. En otras palabras, para hacer que la muestra sea más confiable y cercana a la media poblacional, haga la n más grande. Por el contrario, mientras más pequeña sea la muestra, más riesgosa se hace la estimación (ver las sugerencias de estudio 5 y 6).

	X	x	x^2
	30	6	36
	23	−1	1
	19	−5	25
ΣX:	72		
M:	24		
Σx^2:			62

$$V_e = \frac{62}{3} = 20.67$$

En este experimento, se presume que los diferentes métodos de reforzamiento tienden a "sesgar" las puntuaciones en un sentido u otro, lo cual es, por supuesto, el propósito del investigador. El objetivo del método A es aumentar todas las puntuaciones, relativas al aprendizaje de un grupo experimental. El experimentador puede creer que el método B no tendrá efecto en el aprendizaje y que el método C tendrá un efecto depresor. Si el experimentador está en lo correcto, las puntuaciones del método A tenderán a aumentar, mientras que las del método C tenderán a disminuir. Así, las puntuaciones de los grupos, como un todo —y por supuesto sus medias— difieren de forma sistemática. El reforzamiento es una variable *activa*, manipulada a propósito por el experimentador con la intención consciente de "sesgar" las puntuaciones en una forma diferencial. Prokasy (1987), por ejemplo, ayudó a consolidar este punto al resumir el número de variaciones de reforzamiento en el paradigma pavloviano en el estudio de respuestas de tipo esquelético. Así, cualesquier variables manipuladas por el experimentador están íntimamente asociadas con la varianza sistemática. Camel, Withers y Greenough (1986) dieron a su grupo experimental de ratas diferentes grados de experiencia temprana —ambiental (experiencias enriquecidas tal como una jaula grande con otras ratas y oportunidades para explorar), y al grupo control una condición de experiencia reducida (aislamiento y jaulas individuales)—. Ellos intentaron de forma deliberada construir una varianza sistemática en sus resultados medidos (patrón y número de ramificaciones dendríticas [las dendritas son las estructuras ramificadas de una neurona]). La idea básica atrás del famoso "diseño clásico" de investigación científica en el que se usan grupos experimental y control es que a través del control y manipulación cuidadosos, se hacen variar de forma sistemática las medidas del resultado del grupo experimental (también llamadas "medidas de criterio") para aumentar o disminuir, mientras que las medidas del grupo control se mantienen por lo general al mismo nivel. La varianza por supuesto, es entre los dos grupos, es decir, se hace que ambos grupos difieran. Por ejemplo, Braud y Braud (1972) manipularon grupos experimentales en una forma por demás inusual. Entrenaron ratas de un grupo experimental para elegir el mayor de dos círculos en una tarea de elección; el grupo control de ratas no recibió entrenamiento. Después se inyectó en los cerebros de dos nuevos grupos de ratas extractos de cerebro de los animales de ambos grupos. Desde el punto de vista estadístico intentaron aumentar la varianza entre grupos y tuvieron éxito: ¡Los sujetos de los nuevos "grupos experimentales", superaron a los del nuevo "grupo control" en la tarea de elegir el mayor círculo en la misma tarea de elección!

Esto es claro y fácil de ver en experimentos. En la investigación no experimental, cuando existen diferencias entre los grupos estudiados, no siempre es tan sencillo y directo el visualizar la varianza entre grupos que uno estudia. Pero la idea es la misma. El principio puede establecerse de forma algo diferente: a mayor diferencia entre grupos, más puede asumirse que la o las variables independientes han operado. Si hay una pequeña diferencia entre grupos, debe presumirse que la o las variables independientes *no* han operado. En otras palabras, o bien sus efectos son demasiado débiles para ser aparentes, o

bien diversas influencias se han cancelado mutuamente. Así, juzgamos los efectos de las variables independientes manipuladas o que han trabajado en el pasado a través de la varianza entre grupos. Ya sea que se hayan manipulado o no las variables independientes, el principio es el mismo.

Para ilustrar el principio, usaremos el bien estudiado problema del efecto de la ansiedad en el aprovechamiento escolar. Es posible manipular la ansiedad al contar con dos grupos experimentales e inducir ansiedad en uno y en el otro no. Esto se puede lograr al aplicar a cada grupo la misma prueba con diferentes instrucciones. Decimos a los miembros de un grupo que su calificación dependerá por completo de la prueba, mientras que al otro grupo le decimos que el examen no tiene una importancia en particular y que el resultado no afectará sus notas. Por otro lado, la relación entre ansiedad y aprovechamiento puede también estudiarse al comparar grupos de individuos en los cuales se presupone que diferentes circunstancias ambientales y psicológicas han actuado para producir ansiedad. (Por supuesto, se asume que la ansiedad inducida de manera experimental y la ansiedad preexistente —la variable de estímulo y la variable organísmica— no son iguales.) Guida y Ludlow (1989) condujeron un estudio para probar la hipótesis de que diferentes circunstancias ambientales y psicológicas actúan para producir diferentes niveles de ansiedad al resolver exámenes. Estos investigadores hipotetizaron que los estudiantes en la cultura de Estados Unidos mostrarían un nivel más bajo de ansiedad ante las pruebas que los alumnos de la cultura chilena. Para usar el lenguaje de este capítulo, los investigadores hipotetizaron una mayor varianza entre grupos que la esperada debido al azar a causa de la diferencia entre las condiciones ambientales, educativas y psicológicas chilenas y estadounidenses. (Se encontró apoyo para la hipótesis. Los estudiantes chilenos exhibieron un mayor nivel de ansiedad al resolver exámenes que los alumnos de Estados Unidos. Sin embargo, al considerar sólo los grupos socioeconómicos más bajos de ambas culturas, los aprendices de Estados Unidos tuvieron un mayor nivel de ansiedad que los chilenos.)

Varianza del error

La *varianza del error* es la fluctuación o variación de medidas que no se pueden explicar. Las fluctuaciones de las mediciones en la variable dependiente en un estudio de investigación donde todos los participantes fueron tratados de igual forma se considera varianza del error. Algunas de estas variaciones se deben al azar: en este caso, la varianza del error es varianza aleatoria. Consiste en la variación en las medidas debida a fluctuaciones usualmente pequeñas y autocompensadoras —ahora aquí, ahora allá; ahora arriba, ahora abajo—. La varianza muestral antes discutida, por ejemplo, es varianza del error o aleatoria.

Hagamos una breve digresión en tanto que en este capítulo y el siguiente se usa el concepto de "azar" o "aleatoriedad". Las ideas de aleatoriedad y aleatorización se discutirán con mucho más detalle en el capítulo 8. Por ahora, *aleatoriedad* significa que no hay una forma conocida de describir correctamente o explicar los eventos y sus resultados en términos de lenguaje. En otras palabras, los eventos azarosos no pueden predecirse. Una muestra aleatoria es un subconjunto de un universo; se selecciona a sus miembros de tal forma que cada miembro del universo tiene igual posibilidad de ser elegido. Ésta es otra forma de decir que si los miembros son seleccionados aleatoriamente, no hay forma de predecir qué miembro será elegido en cada oportunidad, al mantener constantes las demás condiciones.

Sin embargo, no debe pensarse que la varianza aleatoria es la única fuente posible de varianza del error. La varianza del error puede constar también de otros componentes como lo señaló Barber (1976). Todo lo que pudiera estar incluido en el término "varianza del error" puede incluir errores de medición en el instrumento usado, errores de procedi-

miento llevados a cabo por el investigador, registro erróneo de las respuestas y la expectativa que el investigador tiene de los resultados. Es posible que "sujetos iguales" difieran en la variable dependiente porque uno de ellos puede estar experimentando un funcionamiento fisiológico o psicológico distinto al momento en que las mediciones fueron tomadas.

Para regresar a nuestra discusión principal, puede decirse que la varianza del error es la varianza en las mediciones debida a ignorancia. Imagine un gran diccionario en el que todo en el mundo —cada ocurrencia, evento, pequeña cosa, asunto importante— se presenta con todo detalle. Para entender cualquier evento que ha ocurrido, sucede ahora o pasará, todo lo que se necesita hacer es mirar el diccionario. Con él es evidente que no hay posibilidad de ocurrencias azarosas. Todas las cosas están explicadas. Es decir, no hay varianza del error; todo es varianza sistemática. Por desgracia (o más bien, por fortuna) no contamos con ese diccionario. Así, muchos eventos y ocurrencias no pueden explicarse. Una gran parte de la varianza elude la identificación y el control. Ello constituye la varianza del error mientras se nos escape su identificación y control.

Aunque parezca un poco extraño y aun bizarro, este modo de razonamiento es útil, mientras sepamos que una parte de la varianza del error del hoy puede no serlo mañana. Suponga que conducimos un experimento sobre la enseñanza de solución de problemas en el cual asignamos alumnos a tres grupos de forma aleatoria. Al terminar el experimento, estudiamos las diferencias entre los tres grupos para ver si la enseñanza tuvo algún efecto. Sabemos que las puntuaciones y medias de los grupos siempre mostrarán fluctuaciones menores, a veces un punto o dos o tres más, y en otras uno o dos o tres menos, que probablemente nunca se controlarán. Algo siempre hace variar así las puntuaciones. De acuerdo a la postura que se analiza, no fluctúan porque sí: es probable que no exista la "aleatoriedad absoluta". Desde una postura determinista, debe haber alguna causa (o causas) para las fluctuaciones. Así es, podemos identificar algunas y quizá controlarlas. Cuando hacemos esto, sin embargo, tenemos varianza sistemática.

Descubrimos, por ejemplo, que el género "causa" la fluctuación de la puntuaciones, en tanto que hay hombres y mujeres mezclados en los grupos experimentales. (Por supuesto, hablamos de forma figurativa. Es evidente que el género no hace que las puntuaciones fluctúen.) Así que conducimos el experimento y controlamos el género al incluir, por ejemplo, sólo a hombres. Las puntuaciones aún varían, pero menos. Entonces retiramos otra supuesta causa de las alteraciones: la inteligencia. La puntuación todavía fluctúa, aunque algo menos. Continuamos retirando otras fuentes de varianza. Así, controlamos la varianza sistemática al tiempo de identificar y controlar cada vez más la varianza desconocida de forma gradual.

Ahora observe que antes de controlar o retirar estas varianzas sistemáticas, antes de "saber" acerca de ellas, tendríamos que haberlas denominado "varianza del error" —en parte por ignorancia y en parte por nuestra incapacidad para controlar o para hacer algo acerca de ella—. Podemos hacer esto una y otra vez, y aún así habrá varianza residual: finalmente nos rendimos, ya no "sabemos" nada más; hemos hecho todo lo posible y todavía queda varianza. Así, una definición práctica de varianza del error sería: La *varianza del error* es aquella que persiste en un conjunto de medidas después de que todas las fuentes conocidas de varianza sistemática se han retirado de dichas medidas. Esto es tan importante que merece un ejemplo numérico.

Un ejemplo de varianza sistemática y varianza del error

Supongamos que estamos interesados en conocer si la cortesía al frasear instrucciones para una tarea afecta la memoria de las palabras amables. Llamamos "cortesía" y "descor-

tesía" a la variable A dividida en A_1 y A_2 (esta idea es de Holtgraves, 1997). Se asigna a los estudiantes al azar a dos grupos. Se define al azar qué grupo recibe el tratamiento A_1 y cuál A_2. En este experimento, los alumnos en A_1 recibieron instrucciones fraseadas sin cortesía, tales como, "usted debe escribir el nombre completo de cada estado que recuerde". Los estudiantes en A_2, por su parte, leyeron instrucciones con igual significado pero presentadas en forma cortés: "sería útil que usted escribiera el nombre completo de cada estado que recuerde". Después de leer las instrucciones, los sujetos tuvieron una tarea distractora consistente en recordar los 50 estados de la Unión Americana. Después se les aplicó una prueba de memoria de reconocimiento. Se usa este examen para determinar el recuerdo general de todas las palabras corteses. Las puntuaciones fueron:

	A_1	A_2
	3	6
	5	5
	1	7
	4	8
	2	4
M	3	6

Las medias son diferentes; éstas varían. Hay varianza entre grupos. Al tomar la diferencia entre las medias con el valor aparente —más adelante lo veremos con más precisión— podemos concluir que la vaguedad en la forma de expresarse tuvo un efecto. Al calcular la varianza entre grupos como lo hicimos antes, tenemos:

		x	x^2
	3	1.5	2.25
	6	1.5	2.25
M:	4.5		
Σx^2:			4.50

$$V_e = \frac{4.5}{2} = 2.25$$

En otras palabras, calculamos la varianza entre grupos como antes lo hicimos para las cinco puntuaciones 1, 2, 3, 4 y 5. Tan sólo tratamos las dos medias como si fueran puntuaciones individuales, y seguimos adelante con un cálculo ordinario de varianza. La varianza entre grupos, V_e, es entonces, 2.25. Una prueba estadística apropiada mostraría que la diferencia entre las medias de ambos grupos es lo que se llama "estadísticamente significativa". (Su significado se analizará en otro capítulo.)[3] Resulta evidente que el uso de palabras corteses en las instrucciones ayudó a incrementar las puntuaciones de memoria de los estudiantes.

[3] El método de cálculo usado aquí no es el que se utiliza para una prueba de significancia estadística; aquí tiene fines pedagógicos. Observe también que la escasa cantidad de casos en los ejemplos y las cifras pequeñas tienen el objetivo de simplificar la demostración. Los datos reales de una investigación, por supuesto, son en general más complejos y requieren de muchos más casos. En el análisis de varianza real, la expresión correcta para la suma de cuadrados-entre es: $SS_b = n\Sigma x_b^2$. Por simplicidad pedagógica, sin embargo, hemos conservado solamente Σx_b^2, para reemplazarla después por SS_b.

Si acomodamos las diez puntuaciones en una columna y calculamos la varianza tendremos:

X	x	x^2
3	−1.5	2.25
5	.5	.25
1	−3.5	12.25
4	−.5	.25
2	−2.5	6.25
6	1.5	2.25
5	.5	.25
7	2.5	6.25
8	3.5	12.25
4	−.5	.25
M: 4.5		
Σx^2:		42.50

$$V_t = \frac{42.5}{10} = 4.25$$

Ésta es la varianza total, V_t. $V_t = 4.25$ contiene todas las fuentes de variación en las puntuaciones. Ya sabemos que una de ellas es la varianza entre grupos, $V_e = 2.25$. Ahora calculemos otra varianza. Lo logramos al calcular la varianza de A_1 sola y la varianza de A_2 sola, para después promediar ambas:

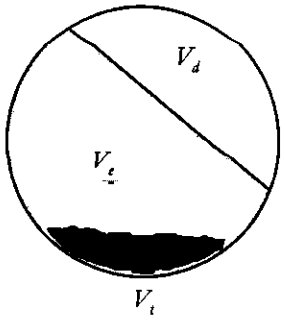
A_1	x	x^2	A_2	x	x^2
3	0	0	6	0	0
5	2	4	5	−1	1
1	−2	4	7	1	1
4	1	1	8	2	4
2	−1	1	4	−2	4
ΣX: 15			30		
M: 3			6		
Σx^2:		10			10

$$V_{A_1} = \frac{10}{5} = 2 \qquad V_{A_2} = \frac{10}{5} = 2$$

La varianza de A_1 es 2, y la varianza de A_2 es 2. El promedio es 2. Dado que cada una de las varianzas fue calculada por *separado* y después promediada, podemos llamar al promedio de la varianza calculada la "varianza dentro de grupos" (o intragrupos) y la denominamos V_d que significa varianza dentro de grupo, o varianza intragrupos. De ese modo, $V_d = 2$. *Esta varianza no se afecta por la diferencia entre las dos medias.* Se demuestra con facilidad al restar una constante de 3 a las puntuaciones de A_2; con ello, la media de A_2 es 3. Entonces, si se calcula la varianza de A_2, será la misma que antes: 2. Como es obvio la varianza intragrupos será la misma: 2.

Ahora escriba una ecuación: $V_t = V_e + V_d$. Esta ecuación indica que la varianza total está integrada por la varianza entre grupos y la varianza intragrupos. Pero, ¿así es? Sustituya

FIGURA 6.1



los valores numéricos: 4.25 = 2.25 + 2.00. Nuestro método funciona —y muestra, también, que las varianzas son aditivas (como se calculó)—.

Las ideas sobre varianza bajo discusión pueden quizás aclararse con un diagrama. En la figura 6.1 se divide un círculo en dos partes. Sea el área total del círculo la varianza total de las 10 puntuaciones o V_t. La porción más grande y sombreada representa la varianza entre grupos o V_e. El área más pequeña sin sombrear representa la varianza del error o varianza dentro de grupos o V_d. En el diagrama podemos ver que $V_t = V_e + V_d$. (Observe la similitud con el razonamiento de conjuntos y la operación de unión.)

V_t representa una medida de todas las fuentes de varianza V_e y a la medida de la varianza entre grupos (o una medida del efecto del tratamiento experimental). Pero ¿qué es V_d, la varianza intragrupos? Dado que, de la varianza total, hemos dado cuenta de una fuente conocida de varianza a través de la varianza entre grupos, podemos asumir que la varianza remanente se debe a factores derivados del azar que llamamos *varianza del error*. Pero, usted se preguntará seguramente si hay otras fuentes de varianza. ¿Qué hay de las diferencias individuales en inteligencia, género, etcétera? Ya que asignamos a los estudiantes a los grupos experimentales de manera azarosa, suponga que esas fuentes de varianza se distribuyen de igual forma, o casi igual, entre A_1 y A_2. Y, debido a la asignación al azar, no podemos aislar ni identificar otras fuentes de varianza. Esta varianza remanente se denomina *varianza del error*, con lo que sabemos muy bien que es probable que haya otras fuentes de varianza pero bajo el supuesto (y esperamos estar en lo correcto) de que están distribuidas de forma equitativa entre ambos grupos.

Una demostración sustractiva: remoción de la varianza entre grupos de la varianza total

Demostremos otra forma de retirar del conjunto original de puntuaciones la varianza entre grupos, a través de un sencillo procedimiento de sustracción. Primero, sea cada una de las medias de A_1 y A_2 igual a la media total; retiramos la varianza entre grupos. La media total es 4.5. (Véase arriba donde la media de las 10 puntuaciones fue calculada.) Segundo, ajuste cada puntuación individual de A_1 y A_2 al restar o sumar, según el caso, una constante apropiada. Dado que la media de A_1 es 3, sumamos 4.5 − 3 = 1.5 a cada una de las puntuaciones de A_1. La media de A_2 es 6 y 6 − 4.5 = 1.5 es la constante que será *restada* a cada una de las puntuaciones de A_2.

Estudie las puntuaciones "corregidas" y compárelas con las originales. Observe que variaron menos de lo que lo hicieron antes. Retiramos la varianza entre grupos, una por-

ción considerable de la varianza total. La varianza que permanece es la parte de la varianza total debida, presumiblemente, al azar. Calculamos la varianza de las puntuaciones "corregidas" de A_1, A_2 y la total, y observamos estos resultados sorprendentes:

Corrección: +1.5 −1.5

A_1	A_2
3 + 1.5 = 4.5	5 − 1.5 = 3.5
5 + 1.5 = 6.5	5 − 1.5 = 3.5
1 + 1.5 = 2.5	7 − 1.5 = 5.5
4 + 1.5 = 5.5	8 − 1.5 = 6.5
2 + 1.5 = 3.5	4 − 1.5 = 2.5
ΣX: 22.5	22.5
M: 4.5	4.5

A_1	x	x^2	A_2	x	x^2
4.5	0	0	4.5	0	0
6.5	2	4	3.5	−1	1
2.5	−2	4	5.5	1	1
5.5	1	1	6.5	2	4
3.5	−1	1	2.5	−2	4
ΣX: 22.5			22.5		
M: 4.5			4.5		
Σx^2:		10			10

$$V_{A_1} = \frac{10}{5} = 2 \qquad V_{A_2} = \frac{10}{5} = 2$$

La varianza intragrupos es la misma de antes. No se afecta por la operación de corrección. Como es evidente, la varianza entre grupos ahora es 0. Y, ¿qué sucede con la varianza total, V_t? Al calcularla, obtenemos $\Sigma x_t^2 = 20$, y $V_t = 20 \div 10 = 2$. Así la varianza intragrupos es ahora igual a la varianza total. El lector debe estudiar este ejemplo con cuidado hasta que entienda con claridad lo que ha sucedido y *por qué*.

Aunque el ejemplo anterior quizá es suficiente para destacar los puntos esenciales, puede consolidarse la comprensión del estudiante sobre estas ideas básicas de la varianza si ampliamos el ejemplo al señalar otra fuente de varianza. El lector recordará que la varianza intragrupos contiene la variación debida a diferencias individuales. Ahora suponga que, en lugar de asignar a los alumnos a los dos grupos de forma aleatoria, los hemos apareado con base en su inteligencia, que está relacionada con la variable dependiente. Es decir, hemos asignado pares de miembros con aproximadamente igual puntuación en pruebas de inteligencia en ambos grupos. El resultado del experimento podría ser:

A_1	A_2
3	6
1	5
4	7
2	4
5	8
M: 3	6

Observe con detalle que la única diferencia entre este arreglo y el anterior es que el apareamiento ha causado que las puntuaciones covaríen. Las medidas de A_1 y A_2 ahora tienen casi igual orden. De hecho, el coeficiente de correlación entre ambos conjuntos de puntuaciones es de 0.90. Tenemos aquí otra fuente de varianza: la debida a diferencias individuales en inteligencia, que se refleja en el orden de los pares de medidas de criterio. (La relación precisa entre las ideas de orden y de apareamiento y sus efectos en la varianza se revisarán en otro capítulo. Por el momento el estudiante deberá tomar como acto de fe que aparear genera varianza sistemática.)

Esta varianza puede calcularse y extraerse como se hizo antes, excepto que hay una operación adicional. Primero iguale las medias de A_1 y A_2 y "corrija" las puntuaciones como antes. El resultado es:

Corrección:	+1.5	−1.5
	4.5	4.5
	2.5	3.5
	5.5	5.5
	3.5	2.5
	6.5	6.5
M:	4.5	4.5

Segundo, al igualar los renglones (haciendo la media de cada renglón igual a 4.5 y "corrigiendo" las puntuaciones de los renglones consecuentemente) encontramos los siguientes datos:

Corrección	A_1	A_2	Medias originales	Medias corregidas
0	4.5 + 0 = 4.5	4.5 + 0 = 4.5	4.5	4.5
+1.5	2.5 + 1.5 = 4.0	3.5 + 1.5 = 5.0	3.0	4.5
−1.0	5.5 − 1.0 = 4.5	5.5 − 1.0 = 4.5	5.5	4.5
+1.5	3.5 + 1.5 = 5.0	2.5 + 1.5 = 4.0	3.0	4.5
−2.0	6.5 − 2.0 = 4.5	6.5 − 2.0 = 4.5	6.5	4.5
M	4.5	4.5	4.5	4.5

Las medidas doblemente corregidas ahora muestran muy poca varianza. La varianza de las 10 puntuaciones doblemente corregidas es 0.10, en verdad muy pequeña. Por supuesto, no queda varianza entre grupos (columnas) o entre individuos (renglones) en las medidas. Después de una doble corrección, la varianza total completa es varianza del error. (Como se verá más adelante, cuando las varianzas tanto de columnas como de renglones se extraen de esta forma ——aunque con un método más rápido y eficiente— se elimina la varianza intragrupos.)

Una recapitulación de la remoción de la varianza entre grupos de la varianza total

Ésta ha sido una larga operación. Una breve recapitulación de los principales puntos puede resultar útil. Cualquier conjunto de medidas tiene una varianza total. Si las medidas a

partir de las cuales se calcula esta varianza se han derivado de respuestas de seres humanos, siempre habrá al menos dos fuentes de varianza. Una se deberá a fuentes sistemáticas de variación como las diferencias individuales de los sujetos cuyas características o logros se han medido, y las diferencias entre los grupos o subgrupos involucrados en la investigación. La otra se derivará del error aleatorio, fluctuaciones de las medidas de las que no se puede dar cuenta en la actualidad. Las fuentes de varianza sistemática tienden a hacer que las puntuaciones se inclinen hacia una dirección u otra, lo que implica diferencias en las medias. Si el género es una fuente sistemática de varianza en un estudio sobre rendimiento escolar, por ejemplo, entonces la variable género tenderá a actuar de manera tal que las puntuaciones de aprovechamiento de las niñas tiendan a ser mayores que las de los niños. Las fuentes del error aleatorio, por otro lado, tienden a hacer que las medidas fluctúen en un sentido en cierto momento, y en otra forma al siguiente. Los errores aleatorios, en otras palabras, se autocompensan; tienden a balancearse (o cancelarse) uno al otro.

En cualquier experimento o estudio, la o las variables independientes constituyen una fuente de varianza sistemática —al menos así debiera ser—. El investigador "quiere" que los grupos experimentales difieran sistemáticamente y con frecuencia busca maximizar tal varianza al controlar o minimizar otras fuentes de varianza, tanto sistemáticas como del error. El ejemplo experimental de antes ilustra la idea adicional de que estas varianzas son aditivas, y debido a esta propiedad aditiva es posible analizar un conjunto de puntuaciones en sus varianzas sistemática y del error.

Componentes de la varianza

La discusión hasta este punto pudo convencer al estudiante de que cualquier varianza total tiene "componentes de varianza". El caso que acabamos de considerar, sin embargo, incluye un componente experimental debido a la diferencia entre A_1 y A_2, un componente resultado de diferencias individuales, y un tercer componente referido al error aleatorio. Ahora estudiaremos el caso de dos componentes de varianza experimental sistemática. Para hacerlo, sintetizamos las medidas experimentales, y las creamos a partir de componentes *conocidos* de varianza. En otras palabras, damos marcha atrás: empezamos con fuentes "conocidas" de varianza ya que no habrá varianza del error en las puntuaciones sintetizadas.

Tenemos una variable X con tres valores. Sea $X = \{0, 1, 2\}$. También tenemos otra variable, Y, con tres valores. Sea $Y = \{0, 2, 4\}$. X y Y son, entonces, fuentes *conocidas* de varianza. Asumimos una condición experimental idónea con dos variables independientes que actúan *concertadamente* para producir efectos en la variable dependiente, Z. Esto es, cada puntuación de X opera con cada puntuación de Y para producir una puntuación Z de la variable dependiente. Por ejemplo, la puntuación X, 0, no tiene influencia. La puntuación X, 1, opera con Y como sigue: $\{(1 + 0), (1 + 2), (1 + 4)\}$. De forma similar, la puntuación X, 2, opera con Y: $\{(2 + 0), (2 + 2)$ y $(2 + 4)\}$. Todo esto es fácil de visualizar si generamos Z de una forma clara.

El conjunto de las puntuaciones en una matriz de 3×3 (una matriz es cualquier conjunto o tabla rectangular de números) es el conjunto de las puntuaciones Z. El propósito de este ejemplo se perderá a menos que el lector recuerde que en la práctica no conocemos las puntuaciones de X y Y; sólo conocemos las puntuaciones de Z. En una situación experimental real, manipulamos o establecemos X y Y y esperamos que sean efectivas. Esto puede no resultar así. En otras palabras, los conjuntos $X = \{0, 1, 2\}$ y $Y = \{0, 2, 4\}$ nunca podrán conocerse así. Lo más que podemos hacer es estimar su influencia a partir de estimar la cantidad de varianza en Z debida a X y a Y.

		Y				Z		
		0	2	4		0	2	4
	0	0 + 0	0 + 2	0 + 4	0	0	2	4
X	1	1 + 0	1 + 2	1 + 4	= 1	1	3	5
	2	2 + 0	2 + 2	2 + 4	2	2	4	6

Los conjuntos X y Y tienen las siguientes varianzas:

X	x	x^2
0	−1	1
1	0	0
2	1	1

Y	y	y^2
0	−2	4
2	0	0
4	2	4

ΣX:	3		ΣY:	6	
M:	1		M:	2	
Σx^2:		2	Σy^2:		8

$$V_x = \frac{2}{3} = .67 \qquad V_y = \frac{8}{3} = 2.67$$

El conjunto Z tiene la varianza como sigue:

Z	z	z^2
0	−3	9
2	−1	1
4	1	1
1	−2	4
3	0	0
5	2	4
2	−1	1
4	1	1
6	3	9

ΣZ:	27	
M:	3	
Σz^2:		30

$$V_z = \frac{30}{9} = 3.33$$

Ahora .67 + 2.67 = 3.34, o $V_z = V_x + V_y$, con los errores propios del redondeo.

Este ejemplo ilustra que, bajo ciertas condiciones, las varianzas operan de forma aditiva para producir las medidas experimentales que analizamos. Aunque el ejemplo es "puro" y por lo tanto irreal, es razonable. Es posible concebir a X y Y como variables independientes; pudieran ser el nivel de aspiración y las actitudes de los alumnos. Z puede ser el aprovechamiento verbal, una variable dependiente. El hecho de que las puntuaciones reales no se comportan exactamente de esta forma no modifica la idea. Se comportan así de manera aproximada. Planeamos la investigación para hacer este principio tan verídico como sea posible, y analizamos los datos como si fuera verdadero. ¡Y funciona!

Covarianza

La covarianza en realidad no es nada nuevo. Recordemos que en una discusión previa de conjuntos y correlación hablamos acerca de la relación entre dos o más variables análogas a la intersección de conjuntos. Sea X {0, 1, 2, 3}, un conjunto de medidas de actitud de cuatro niños. Sea $Y = \{1, 2, 3, 4\}$, un conjunto de medidas de aprovechamiento de los mismos niños, pero no en el mismo orden. Sea R un conjunto de pares ordenados de los elementos de X y Y, donde la regla de apareamiento sería: se parea cada medida de actitud con cada medida de aprovechamiento del sujeto, con la medida de actitud colocada primero. Suponga que resulta que $R = \{(0, 2), (1, 1), (2, 3), (3, 4)\}$. De acuerdo a nuestra definición previa de relación, este conjunto de pares ordenados constituye una relación, en este caso, entre X y Y. El resultado del cálculo de la varianza de X y de la varianza de Y es:

	X	x	x^2	Y	y	y^2
	0	−1.5	2.25	2	−.5	.25
	1	−.5	.25	1	−1.5	2.25
	2	.5	.25	3	.5	.25
	3	1.5	2.25	4	1.5	2.25
ΣX:	6			10		
M:	1.5			2.5		
Σx^2:			5			5

$$V_x = \frac{5}{4} = 1.25 \qquad V_y = \frac{5}{4} = 1.25$$

Ahora nos planteamos un problema. (Observe con atención en lo que sigue y que trabajaremos con desviaciones de la media, x y y, y no con las puntuaciones brutas originales.) Hemos calculado las varianzas de X y Y usando las x y y; esto es, las desviaciones de las medias respectivas de X y Y. Si podemos calcular la varianza de cualquier conjunto de puntuaciones, ¿no es posible calcular la relación *entre* dos conjuntos cualesquiera de puntuaciones en una forma similar? ¿Es concebible que podamos calcular la varianza de dos conjuntos de forma simultánea? Y si lo hacemos, ¿será ésta una medida de la varianza de ambos conjuntos unidos? ¿Será esta varianza también una medida de la relación entre los dos conjuntos?

Lo que deseamos es usar alguna operación estadística análoga a la operación intersección de conjuntos, $X \cap Y$. **Para calcular la varianza de X o de Y, elevamos al cuadrado las** desviaciones de la media, las x o y, y después las sumamos y promediamos. Una respuesta natural a nuestro problema es realizar una operación análoga en las x y y *juntas*. Para calcular la varianza de X, lo primero que hicimos fue: $(x_1 \cdot x_1), \ldots, (x_4 \cdot x_4) = x_1^2, \ldots, x_4^2$. ¿Por qué no hacemos esto tanto con las x como con las y, multiplicando los pares ordenados así: $(x_1 \cdot y_1), \ldots, (x_4 \cdot y_4)$? Así, en lugar de escribir Σx^2 o Σy^2 escribiríamos Σxy, como sigue:

x	y	=	xy
−1.5	−.5	=	.75
−.5	−1.5	=	.75
.5	.5	=	.25
1.5	1.5	=	2.25

$$\Sigma xy = 4.00$$

$$V_{xy} = CoV_{xy} = \frac{4}{4} = 1.00$$

Vamos a darle nombre a Σxy y a V_{xy}. Σxy se llama *producto cruzado* o la suma de los productos cruzados. V_{xy} es llamada *covarianza*, que denotaremos con CoV con los subíndices apropiados. Si calculamos la varianza de estos productos, simbolizados como V_{xy} o CoV_{xy}, obtendríamos 1.00, como se indicó antes. Este 1.00, entonces, puede considerarse como un índice de la relación entre ambos conjuntos. Pero se trata de un índice no satisfactorio porque su tamaño fluctúa con los márgenes y escalas de diferentes X y Y; esto es, puede ser 1.00 en este caso y 8.75 en otro caso, lo que hace que las comparaciones de caso-con-caso sean difíciles de manejar. Necesitamos una medida que sea comparable de un problema a otro. Una medida así —y excelente— se obtiene tan sólo al escribir una fracción o tasa. Es la covarianza, CoV_{xy}, dividida entre un promedio de las varianzas de X y Y. En general, el promedio se presenta en la forma de una raíz cuadrada del producto de Vx y Vy. La fórmula completa para nuestro índice de relación entonces sería:

$$R = \frac{CoV_{xy}}{\sqrt{V_x \cdot V_y}}$$

Ésta es una forma del bien conocido coeficiente de correlación producto-momento. Al usarlo con nuestro pequeño problema obtenemos:

$$R = \frac{CoV_{xy}}{\sqrt{V_x \cdot Vy}} = \frac{1.00}{1.25} = .80$$

Este índice, que se denota por lo general como r, puede ir de +1.00 pasando por el 0 hasta −1.00, como aprendimos en el capítulo 5. Así, tenemos otra importante fuente de variación en los conjuntos de puntuaciones, siempre y cuando los elementos de los conjuntos, las X y Y, hayan sido ordenados en pares después de convertirlos en puntuaciones de desviación. Esta variación se llama de forma acertada *covarianza* y es una medida de la relación entre los conjuntos de puntuaciones.

Puede verse que la definición de relación como conjunto de pares ordenados nos lleva a varias formas de definir la relación del ejemplo anterior:

R = {(x, y); x y y son números, x siempre viene primero}
xRy = igual que arriba, o "x está relacionada con y"
R = {(0, 2),(1, 1),(2, 3),(3, 4)}
R = {(−1.5, −.5), (−.5, −1.5), (.5, .5), (1.5, 1.5)}

$$R = \frac{CoV_{xy}}{\sqrt{V_x \cdot V_y}} = \frac{1.00}{1.25} = .80$$

La varianza y la covarianza son conceptos de máxima importancia en investigación y en el análisis de los datos de investigación, por dos razones: primero, digamos que sintetizan la variabilidad de variables y la relación entre ellas. Esto se visualiza con facilidad al darnos cuenta de que las correlaciones son covarianzas que se han estandarizado para tener valores entre −1 y +1. Pero el término también implica la variación conjunta de las variables en general. En la mayor parte de nuestra investigación, literalmente perseguimos y estudiamos la covariación de los fenómenos. En segundo lugar, la varianza y covarianza forman la columna vertebral estadística del análisis multivariado, como se verá

hacia el final de este libro. La mayoría de las discusiones del análisis de datos está basada en varianzas y covarianzas. El análisis de varianza, por ejemplo, estudia diversas fuentes de variación de observaciones, en general en experimentos, como se indicó antes. El análisis factorial es en efecto el estudio de la covarianza, uno de cuyos propósitos es aislar e identificar fuentes comunes de variación. El análisis contemporáneo más reciente, el enfoque multivariado más poderoso y avanzado concebido hasta ahora, se llama *análisis de estructuras de covarianza* porque el sistema estudia conjuntos complejos de relaciones a partir del análisis de las covarianzas entre variables. Es evidente que la varianza y covarianza serán el centro de gran parte de nuestra discusión y preocupación a partir de este momento.

Anexo computacional

Uno de los mayores problemas que tienen los escritores de libros de texto en la actualidad al introducir el uso de programas de cómputo es la rapidez con que el material se vuelve obsoleto. En un periodo de un año o menos diversos fabricantes de programas estadísticos de cómputo pueden tener actualizaciones y cambios en los programas. Estas actualizaciones y modificaciones causarán una diferencia entre el programa y lo que está escrito sobre cómo usarlo. Por ejemplo, cuando la revisión de este libro de texto se inició, un programa muy popular, el paquete estadístico para ciencias sociales (SPSS) for Windows estaba en la versión 6.0. Al momento de escribir esto la versión en circulación es 8.0, y la 9.0 está por salir. Por lo anterior, presentar cualquier conjunto específico de enunciados de programación para tales programas muy pronto puede resultar inútil para los estudiantes e investigadores. Así que el objetivo es exhibir algunas características generales subyacentes a todos los programas estadísticos, que puedan tener una gran aplicación con los programas más nuevos y con los anteriores. También es importante elegir un programa estadístico que perdure por el lapso de tiempo entre las revisiones del libro (¡demasiado optimismo!). Por ejemplo, en la tercera edición de este libro publicada en 1986, las computadoras personales estaban todavía en su infancia: fuera de algunos programas de hoja de cálculo,

FIGURA 6.2

Untitled - SPSS Data Editor

File Edit View Data Transform Statistics Graphs Utilities Windows Help

	var	var	var	var	var	var	
1							
2							
3							
4							
5							
6							
7							

había un escaso desarrollo en términos de lenguajes de programación y paquetería estadística. Entre los primeros para computadoras personales o de escritorio están: SPSS-PC, STATA, NCSS y Anderson-Bell STAT. La mayor parte no eran competitivos en términos de flexibilidad y poder computacional al compararlos con el software estadístico disponible para computadoras de gran escala (macrocomputadoras). Algunos programas estadísticos especializados fueron desarrollados por un gran número de investigadores que no usaban computadoras grandes. Algunos de estos programas se escribieron en BASIC, FORTRAN, C y COBOL.

Sin embargo, el investigador actual goza la "bendición" de contar con programas estadísticos muy poderosos y flexibles que pueden usarse con facilidad para fines de análisis de datos. De aquí al final del libro, usaremos en general el SPSS for Windows para demostrar los cálculos estadísticos. ¿Por qué? En años recientes, muchas compañías competidoras de programas de cómputo (incluyendo la favorita del segundo autor) han sido adquiridas por SPSS Inc. Aún puede conseguirse paquetería de la "competencia" pero el desarrollo que conlleva es dudoso. Otra razón es que el SPSS está disponible en la mayoría de las grandes universidades y hay versiones del programa para estudiantes que pueden adquirir e instalar en sus computadoras personales. A pesar de las críticas dirigidas a SPSS, (desde su concepción) se han introducido en el mercado programas fáciles de utilizar para investigadores y estudiantes. Una vez que se clarifican algunas ideas generales en relación a cómo cargar los datos e ingresarlos en el programa, la solicitud de ciertas rutinas estadísticas se facilita mucho.

La discusión del SPSS en este libro repara sobre la versión Windows disponible para computadoras personales, y no será válida necesariamente para las versiones de macrocomputadoras o plataformas que no son Windows (tales como las versiones DOS). Esta discusión asume que el lector utiliza Windows 95 o uno más actual y que está familiarizado con los comandos y funciones de Windows. También implica el conocimiento del uso del ratón (el dispositivo apuntador) que es imperativo cuando se realizan operaciones con Windows ya que todas las operaciones se realizan al señalar con el ratón (*mouse*), haciendo clic (o resaltando) un objeto para seleccionarlo.

FIGURA 6.3

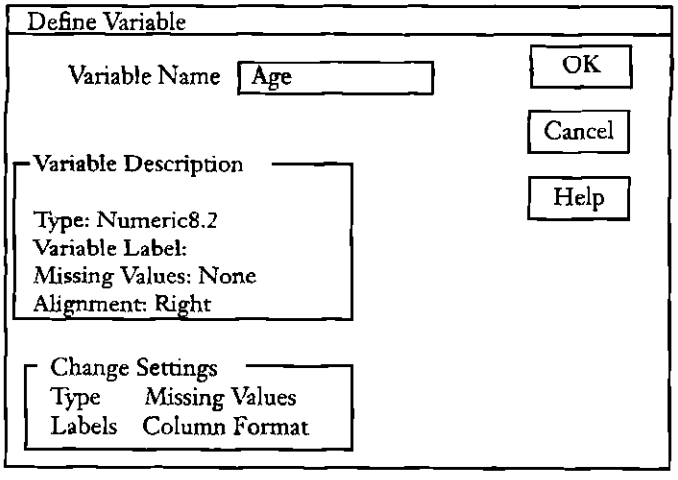


FIGURA 6.4

Untitled - SPSS Data Editor

File Edit View Data Transform Statistics Graphs Utilities Windows Help

	Age	Gender	Score	var	var	var	
1	12	2	60				
2	13	1	75				
3	15	1	45				
4	14	2	80				
5	14	1	85				
6	12	2	39				
7	13	1	62				

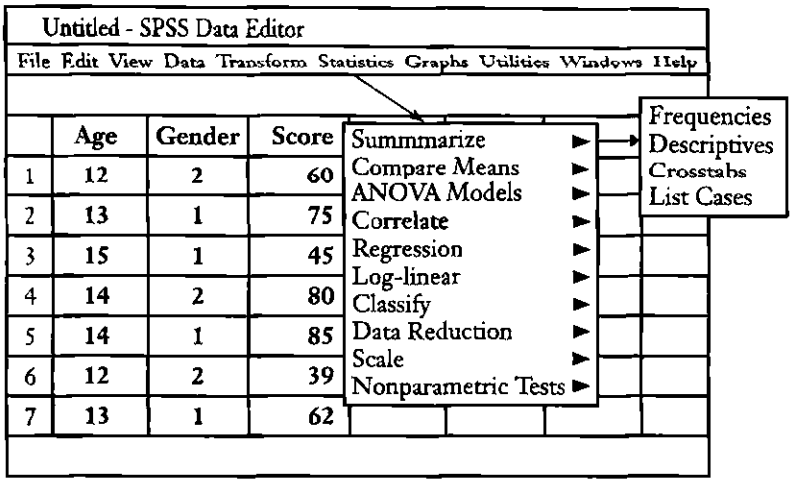
Al ejecutar el SPSS for Windows (en lo sucesivo SPSSWIN), la primera pantalla que aparece es una tabla para ingresar datos, similar a la de la figura 6.2, y tiene forma de hoja de cálculo en la que el usuario deberá incorporar los datos. Si el investigador previamente ha creado un banco de datos compatible con SPSSWIN, puede usarlo en forma directa, sin que tenga que reingresar los datos.

El formato general de datos para casi todos los programas estadísticos de computadora es una tabla donde las variables son las columnas y las observaciones (personas, individuos) son los renglones. Si tenemos el siguiente conjunto de datos, su ingreso al SPSSWIN será fácil.

	Variable		
	Año	**Género**	**Resultado de la prueba**
Personas			
1	12	M	60
2	13	F	75
3	15	F	45
4	14	M	80
5	14	F	85
6	12	M	39
7	13	F	62

Su primer paso es definir las variables para SPSSWIN. Donde usted vea la etiqueta "var" de la hoja de cálculo, usted puede ingresar el nombre para que haga referencia a sus variables. Para hacerlo, use el ratón y haga doble clic sobre la celda en la hoja de cálculo marcada con "var". Al hacerlo aparece otra pantalla que le permite especificar el nombre de la variable y sus atributos (v. gr. datos o caracteres numéricos). En la primera columna

FIGURA 6.5



escriba "Age", después haga clic en el botón "OK" (véase figura 6.3). Repita esta operación para cada una de sus variables. Para la variable "Gender" use un "1" para "F" y un "2" para "M". Después ingrese los datos de su tabla: cada valor ocupa una celda de la hoja de cálculo. Al ingresar todos los datos, su hoja de cálculo deberá ser similar a la que se muestra en la figura 6.4.

Usted puede guardarlo como un conjunto de datos al hacer clic en "FILE" y seleccionar "SAVE". Cuando lo haya hecho, se le pedirá un nombre para su conjunto de datos. Su siguiente paso es realizar un análisis estadístico. En este capítulo sólo llevaremos a cabo estadística descriptiva, que incluye medias y desviaciones estándar. La figura 6.5 muestra las pantallas usadas del SPSS. Primero haga clic en "Statistics", lo que hará aparecer un nuevo menú, del cual elija "Sumarize". Al hacerlo aparecerá un tercer menú en el cual podrá usted elegir (haciendo clic) "Descriptives", que genera una pantalla (ventana) de

FIGURA 6.6

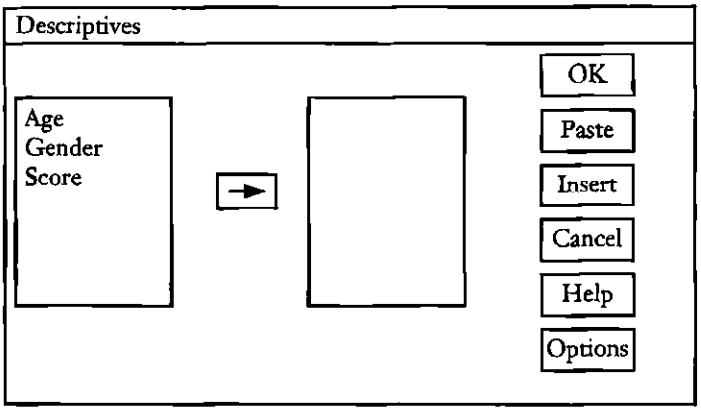
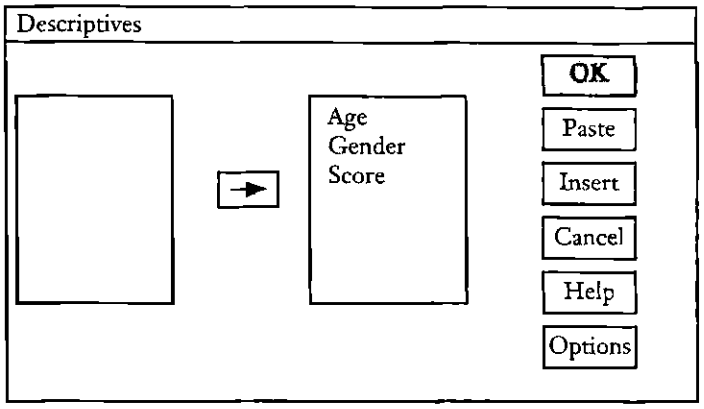


FIGURA 6.7



"estadísticos descriptivos" como se muestra en la figura 6.6. Su siguiente acción será mover las variables del cuadro de la izquierda al de la derecha. Para ello, seleccione las variables con el ratón y haga clic en el botón de la derecha. Este botón se localiza entre los dos cuadros. Se calcularán las estadísticas descriptivas sólo para aquellas que estén en el cuadro de la derecha.

Las variables que sean de poco o ningún interés para el investigador pueden quedarse en el cuadro de la izquierda. La figura 6.7 muestra la pantalla después de que las tres variables han sido desplazadas al cuadro de la derecha. El propósito de esta ventana es permitir al investigador elegir qué variables deben usarse en el análisis. Una vez realizado, seleccione el botón "OK"; entonces el programa realizará todos los cálculos necesarios y los desplegará en la pantalla de salida del SPSS.

El resultado del análisis se muestra abajo. Puede salvar los resultados en un archivo, si lo desea. Con el SPSS for Windows, puede realizar muchos análisis diferentes con el mismo conjunto de datos.

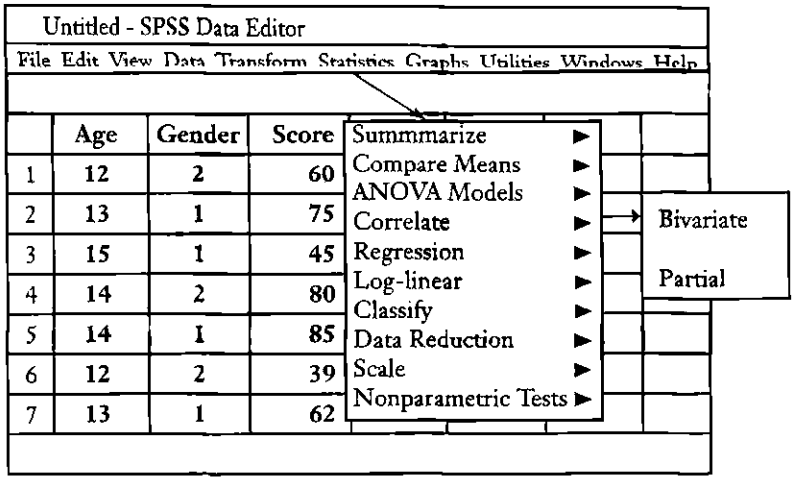
Variable	Mean	Std Dev	Variance	Minimum	Maximum	N
Gender	1.43	.53	.29	1	2	7
Age	13.29	1.11	1.24	12.00	15.00	7
Score	63.71	13.43	303.90	39.00	85.00	7

Si desea calcular las covarianzas entre variables, seleccione "Correlate" del menú, lo que generará una nueva alternativa donde puede elegir "Bivariate" (figura 6.8). Para obtener las covarianzas, seleccione el botón "options" y haga una marca en el cuadro para solicitar que se desplieguen los resultados de covarianzas.

Los resultados que da la computadora son como sigue:

Variables		Cases	Croos-Prod Dev	Variance-Covar
Age	Gender	7	-1.8571	-.3095
Age	Score	7	28.5714	4.7619
Gender	Score	7	-12.1429	-2.0238

FIGURA 6.8



En los capítulos subsiguientes cuando nos refiramos a cálculos, nos basaremos en esta demostración. La información resulta fundamental e importante para trabajar con eficiencia con el SPSS for Windows. Sin embargo, esta breve introducción no pretende ser sustituto del manual del SPSS que está disponible. El usuario de paquetería estadística debe estar consciente de que la computadora sólo calcula lo que se le pide y no puede interpretar el resultado si se han cometido algunos errores lógicos.

RESUMEN DEL CAPÍTULO

1. Las diferencias entre las mediciones son necesarias para estudiar las relaciones entre variables.
2. La varianza es una medida estadística usada para estudiar diferencias.
3. La varianza, junto con la media, se usan para resolver problemas de investigación.
4. Clases de varianza:
   - *a*) La variabilidad de una variable o característica en el universo o población es la varianza poblacional.
   - *b*) La muestra es un subconjunto del universo y también tiene variabilidad que se denomina varianza muestral.
   - *c*) La estadística calculada de una muestra a la otra varía y se llama varianza muestral.
   - *d*) La varianza sistemática es la variación de la que se puede dar cuenta. Puede ser explicada. Cualquier influencia de origen natural o humano que cause que se den eventos en alguna manera predecible es varianza sistemática.
   - *e*) Un tipo de varianza sistemática es la llamada varianza entre grupos. Cuando hay diferencias entre grupos de sujetos, y se conoce la causa de esa diferencia, se denomina varianza entre grupos.
   - *f*) Otro tipo de varianza sistemática es la varianza experimental, que es un poco más específica que la varianza entre grupos en tanto que está asociada con la varianza originada por la manipulación activa de la variable independiente.

g) La varianza del error es la fluctuación o variación de las medidas en la variable dependiente que no puede ser explicada en forma directa por las variables bajo estudio. Una parte de la varianza del error se debe al azar, lo que se conoce como varianza aleatoria. La fuente de esta fluctuación generalmente es desconocida. Otras posibles fuentes de varianza del error incluyen al procedimiento del estudio, el instrumento de medición y las expectativas del investigador.

5. Las varianzas pueden fraccionarse en sus componentes. En este caso, la palabra *varianza* se refiere como varianza total. La partición de la varianza total en sus componentes de varianza sistemática y varianza del error juega un papel importante en los análisis estadísticos de los datos de investigación.
6. La covarianza es la relación entre dos o más variables:
   a) Es un coeficiente de correlación no estandarizado;
   b) La covarianza y la varianza son los fundamentos estadísticos de las estadísticas multivariadas (que se presentarán en capítulos posteriores).

SUGERENCIAS DE ESTUDIO

1. Un psicólogo social ha hecho un experimento en el que un grupo, A_1, tuvo una tarea frente a un auditorio, y otro grupo, A_2, debió realizarla sin público. Las puntuaciones de ambos grupos en la tarea que evaluaba habilidad con los dedos fueron:

A_1	A_2
5	3
5	4
9	7
8	4
3	2

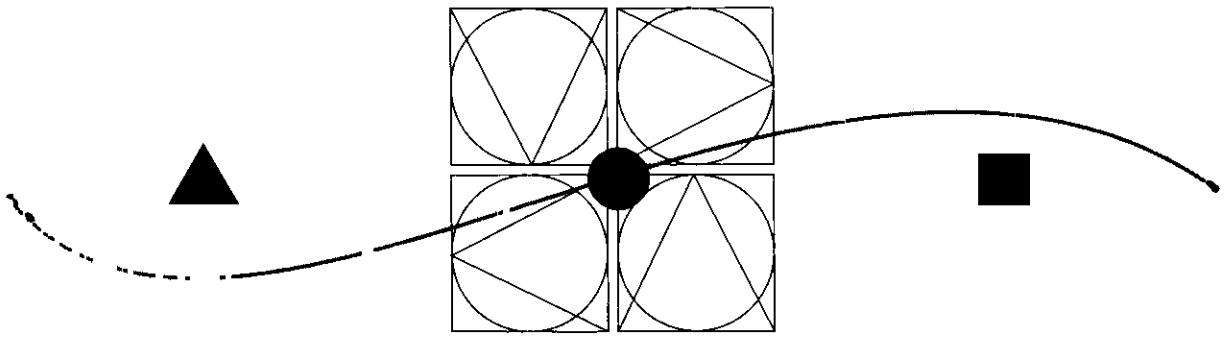
   a) Calcule las medias y varianzas de A_1 y A_2, usando el método descrito en el texto.
   b) Calcule la varianza entre grupos, V_e y también la varianza intragrupos, V_d.
   c) Acomode todas las 10 puntuaciones en una columna y calcule la varianza total, V_t.
   d) Sustituya los valores calculados obtenidos en ***b***) y en ***c***) en la ecuación: $V_t = V_e + V_d$. Interprete los resultados.
   [Respuestas: ***a***) $V_{A1} = 4.8$; $V_{A2} = 2.8$; ***b***) $V_e = 1.0$; $V_d = 3.8$; ***c***) $V_t = 4.8$.]
2. Para el ejercicio 1 anterior, sume 2 a cada una de las puntuaciones de A_1, y calcule V_t, V_e y V_d. ¿Cuál (o cuáles) de las varianzas cambió? ¿Cuál (o cuáles) permaneció igual? ¿Por qué?
   [Respuestas: $V_t = 7.8$; $V_e = 4.0$; $V_d = 3.8$.]
3. Para el ejercicio 1 anterior, iguale las medias de A_1 y A_2, al sumar una constante de 2 a cada una de las puntuaciones de A_2. Calcule V_t, V_e y V_d. ¿Cuál es la principal diferencia entre este resultado y el de la pregunta 1? Explique por qué.
4. Suponga que un investigador social obtuvo medias de conservadurismo (A), actitud hacia la religión (B) y antisemitismo (C) de 100 individuos. Las correlaciones entre variables fueron: $r_{ab} = .70$; $r_{ac} = .40$; $r_{bc} = .30$. ¿Qué significan estas correlaciones? [Consejo: Eleve al cuadrado las *rs* antes de tratar de interpretar las relaciones. También piense en pares ordenados.]

5. El propósito de esta sugerencia de estudio y de la número 6 es dotar al estudiante de intuición acerca de la variabilidad de los estadísticos muestrales, la relación entre las varianzas poblacionales y muestrales y las varianzas entre grupos y del error. El apéndice C contiene 40 conjuntos de 100 números aleatorios entre 0 y 100, con sus medias, varianzas y desviaciones estándar. Extraiga 10 conjuntos de 10 números cada uno a partir de 10 lugares diferentes en la tabla.
   a) Calcule la media, varianza y desviación estándar de cada uno de los 10 conjuntos. Encuentre la media más alta y la más baja, así como la varianza más alta y la más baja. ¿Difieren mucho una de la otra? ¿Qué valor "deberían" tener las medias (50)? Una vez hecho esto, aparte los 10 totales y calcule la media de los 100 números. ¿Difieren mucho las 10 medias de la media total? ¿Difieren mucho de las medias reportadas en la tabla de medias, varianzas y desviaciones estándar que se presenta después de los números aleatorios?
   b) Cuente los números pares y nones en cada uno de los 10 conjuntos. ¿Son los que "deberían ser"? Cuente los números pares e impares de los 100 números. ¿El resultado es "mejor" que los resultados de los 10 conteos? ¿Por qué debería serlo?
   c) Calcule la varianza de las 10 medias. Ésta es, por supuesto, la varianza entre grupos V_b. Calcule la varianza del error usando la fórmula: $V_e = V_t - V_b$.
   d) Analice el significado de sus resultados después de repasar la explicación en el texto.
6. Tan pronto como sea posible, los estudiantes de investigación deben empezar a entender y utilizar la computadora. La sugerencia de estudio 5 se puede realizar mejor y más fácilmente con la computadora. Sería mejor, por ejemplo, extraer 20 muestras de 100 números cada una. ¿Por qué? En cualquier caso, los estudiantes deben aprender cómo realizar operaciones estadísticas simples utilizando el equipo de cómputo y los programas que existen en sus instituciones. Todas las instituciones poseen programas de cómputo para calcular medias y desviaciones estándar (las varianzas se obtienen elevando al cuadrado las desviaciones estándar)[4] y para generar números aleatorios. Si usted tiene acceso al equipo de su institución, utilícelo para llevar a cabo la sugerencia de estudio 5, pero incremente el número de muestras y sus n.

[4] Puede hacer pequeñas discrepancias entre las desviaciones estándar y las varianzas calculadas a mano y aquellas obtenidas a partir de la computadora, porque los programas existentes y las rutinas de las calculadoras manuales generalmente usan una fórmula con N menos 1 en lugar de colocar N en el denominador de la fórmula. Sin embargo, las discrepancias serán pequeñas, sobre todo si la N es grande. (La razón para que existan diferentes fórmulas se explicará más adelante cuando nos ocupemos de muestreo y otros temas.)

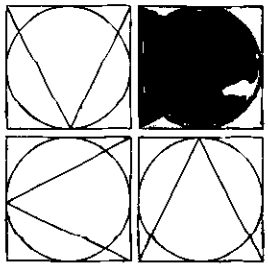
Parte Tres

Probabilidad y muestreo



Capítulo 7
Probabilidad

Capítulo 8
Muestreo y aleatoriedad



CAPÍTULO 7

PROBABILIDAD

- DEFINICIÓN DE PROBABILIDAD
- ESPACIO MUESTRAL, PUNTOS MUESTRALES Y EVENTOS
- DETERMINACIÓN DE PROBABILIDADES CON MONEDAS
- UN EXPERIMENTO CON DADOS
- ALGO DE TEORÍA FORMAL
- EVENTOS COMPUESTOS Y SU PROBABILIDAD
- INDEPENDENCIA, EXCLUSIÓN MUTUA Y EXHAUSTIVIDAD
- PROBABILIDAD CONDICIONAL
  - **Definición de probabilidad condicional**
  - **Un ejemplo académico**
  - **Teorema de Bayes: revisión de las probabilidades**

El tema de la probabilidad es simple y obvio, confuso y complejo. Es un tema sobre el que sabemos mucho y a la vez es un tema del cual no sabemos nada. Los alumnos de jardín de niños y los filósofos pueden estudiar probabilidad. Es tedioso y es interesante. Tales contradicciones son características de la probabilidad.

Si se utiliza la expresión "leyes del azar", dicha expresión, por sí misma, es particularmente contradictoria. Casualidad o azar, por definición, es la ausencia de ley. Si los eventos pueden explicarse por medio de una ley, entonces no se deben al azar, por lo tanto, ¿por qué decimos "leyes del azar"? La respuesta es también peculiarmente contradictoria: es posible obtener conocimiento de la ignorancia si se ve el azar como ignorancia. Esto es debido a que los eventos aleatorios, en conjunto, ocurren obedeciendo leyes con regularidad monótona. A partir del desorden del azar, el científico unifica el orden de la predicción y el control científicos.

No es fácil explicar estos conceptos desconcertantes, de hecho, los filósofos discrepan en sus respuestas. Afortunadamente, no hay desacuerdo sobre los eventos probabilísticos empíricos —o al menos hay muy poco—. Casi todos los científicos y filósofos concordarían en que si dos dados se lanzan un número de veces, probablemente habrá más 7 que 2 o 12.

También estarían de acuerdo en que ciertos eventos, como encontrarse un billete de 100 dólares o ganar una combinación de apuestas, son extremadamente improbables.

Definición de probabilidad

¿Qué es la probabilidad? Al hacer esta pregunta surge inmediatamente un problema complejo. Wang (1993), Brady y Lee (1989) y Cowles (1989) han establecido que históricamente parece no haber acuerdo sobre la respuesta. Esto puede deberse a que existen dos escuelas principales de pensamiento: la de frecuencia y la de no frecuencia. Además, en la escuela de frecuencia hay por lo menos dos definiciones, entre otras, que parecen ser irreconciliables: la *a priori* y la *a posteriori*. La definición *a priori* se debe al controvertido Pierre Laplace y al distinguido matemático Augustus DeMorgan (Cowles, 1989). Aquí, la probabilidad de un evento es igual al número de casos favorables dividido entre el número total de casos (igualmente posibles), o $p = f \div (f + u)$, donde p es la probabilidad, f es el número de casos favorables y u el número de casos no favorables. El método para calcular la probabilidad, implicado en la definición, es *a priori* en el sentido de que la probabilidad está dada de tal manera que se pueden determinar las probabilidades de eventos antes de la investigación empírica. Las personas frecuentemente hacen afirmaciones respecto de las probabilidades sin datos empíricos que las avalen. Aunque estas afirmaciones reflejan, más bien, un punto de vista propio. La interpretación de la probabilidad de Laplace y DeMorgan se considera una definición clásica, la cual es la base de la probabilidad teórica matemática.

La definición *a posteriori*, o de frecuencia relativa a largo plazo, es empírica por naturaleza. Ésta explica que, en una serie real de pruebas, la probabilidad es la razón del número de veces en que un evento ocurre en el número total de ensayos. Con esta definición, uno se aproxima a la probabilidad de forma empírica aplicando una serie de pruebas, contando el número de veces en que un cierto evento ocurre y después calculando la relación. El resultado del cálculo es la probabilidad de dicha clase de evento. Tienen que usarse definiciones de frecuencia cuando no es posible la enumeración teórica sobre clases de eventos. Por ejemplo, para calcular la probabilidad de la longevidad y la que tiene un caballo en una carrera, se tienen que usar tablas actuariales y calcular la probabilidad a partir de conteos y cálculos pasados. La afirmación de que un cortador de diamantes es 95% preciso, indica que de cada cien diamantes que esta persona ha cortado en el pasado, 95 de ellos fueron cortados correctamente.

Hablando prácticamente (y para los propósitos del texto), la distinción entre la definición *a priori* y la *a posteriori* no es demasiado importante. Siguiendo a Margenau (1950/1977, p. 264), unimos las dos definiciones diciendo que el enfoque *a priori* provee una definición constitutiva de probabilidad, mientras que el enfoque *a posteriori* ofrece una definición operacional de probabilidad. Se requiere usar ambos enfoques, ya que se necesita complementar una con la otra.

El planteamiento de la no frecuencia es atribuida a John Maynard Keynes (1921/1979). Keynes, un economista de fama mundial, escribió un número importante de publicaciones citadas frecuentemente. Existe una teoría económica completa que está basada en las contribuciones de Keynes. Aquellos que trabajan en tal clase de investigación son llamados *keynesianos*. La contribución de Keynes a la probabilidad y la estadística generalmente no se menciona en la mayoría de los libros de texto de probabilidad, pero aún así es importante para aquellos que hacen investigación en las ciencias del comportamiento (Brady y Lee, 1989a). En este enfoque hay dos valores: 1) el valor de la probabilidad en sí misma y 2) el peso de una evidencia asociada a ella. El peso de la evidencia es subjetiva, involucra

la percepción de quien toma la decisión sobre la cualidad y cantidad de información alrededor del valor de la probabilidad, obtenido empíricamente. En esencia Keynes establece que quienes toman las decisiones se confrontan con la probabilidad de los eventos, y también con la cantidad y/o cualidad de la información asociada a ellos. Quienes toman las decisiones utilizan la información aunada a la probabilidad para tomar una decisión. Keynes define un coeficiente de peso y riesgo; dicho coeficiente asigna, esencialmente, un peso a un valor empírico de probabilidad. Si el peso de la evidencia es fuerte, a la probabilidad se le da un mayor peso. Se asigna un peso cercano a cero si el peso de la evidencia respecto de esa probabilidad es débil. Según Brady y Lee (1989b, 1991) el enfoque de Keynes explica algunas de las llamadas paradojas de la toma de decisiones, que la teoría de frecuencia no puede explicar adecuadamente. Bakan (1974) afirma que la teoría de la probabilidad de Keynes capta la esencia del proceso que enfrentan los psicólogos clínicos que tratan con problemas de relevancia. Al llevar a cabo una terapia, el psicólogo escucha, lee y ve muchos indicios e información, pero selectivamente ubica a algunos de ellos como más relevantes que otros. La teoría de Keynes posee un mayor alcance en las explicaciones probabilísticas y puede utilizarse para explicar el resultado de un estudio de Rosenthal y Gaito, reportado en Bakan (1974), donde se le pidió a un grupo de profesores psicólogos doctorados juzgar dos estudios diferentes: *A* y *B*. En cada uno de los estudios *A* y *B* se había realizado la misma prueba estadística y se había obtenido el mismo valor de *p*. Sin embargo, el tamaño de la muestra del estudio *A* era de 10 y el del estudio *B* era de 100. A cada profesor de la facultad se le preguntó cuál de los estudios le inspiraba mayor confianza o credibilidad. La mayoría de ellos le otorgó mayor confianza a los resultados del estudio *B*. Keynes explicaría esto a la luz del hecho de que en su juicio, estos individuos daban mayor peso a un tamaño muestral de 100 que a un tamaño muestral de 10.

En resumen, la frecuencia relativa a largo plazo es la teoría prevalente en la investigación de las ciencias del comportamiento (Cowles, 1989). La mayoría de los científicos del comportamiento confían la manipulación estadística de sus datos, siguen la escuela de la frecuencia relativa del pensamiento. En casi todos los textos elementales de estadística que cubren el tema de la probabilidad, solamente se discute la teoría de la frecuencia relativa y sus efectos en los métodos estadísticos.

Espacio muestral, puntos muestrales y eventos

Para calcular la probabilidad de cualquier resultado primero es necesario determinar el número total de resultados posibles. En un dado, los resultados posibles son 1, 2, 3, 4, 5, 6. Llamemos a este conjunto *U*, ya que es el espacio muestral o universo de posibles resultados. El espacio muestral incluye todos los posibles resultados en un "experimento" que son de interés para el experimentador. Los elementos primarios de *U* son llamados elementos o puntos muestrales. Se escribe, entonces, $U = \{1, 2, 3, 4, 5, 6\}$, para unificar este capítulo con el razonamiento y método de conjuntos empleados en los capítulos 4, 5 y 6. Si x_j es igual a cualquier punto muestral o elemento en *U*, se escribe ahora $U = \{x_1, x_2, \ldots, x_n\}$. Todos los posibles resultados de lanzar dos dados son ejemplos de diferentes *U* (véase tabla 7.1); todos los niños de un jardín de niños en tal o cual sistema escolar; todos los votantes elegibles en *X* país.

Algunas veces la determinación de un espacio muestral es fácil, pero algunas veces es difícil; el problema es análogo a la definición de conjuntos del capítulo 4. Los conjuntos pueden definirse listando a todos lo miembros del conjunto y estableciendo una regla para la inclusión de los elementos en él. En la teoría de la probabilidad, ambas definiciones se utilizan. ¿Qué valor tendría *U* al lanzar al aire 2 monedas? He aquí una lista de todas las

TABLA 7.1 *Matriz de posibles resultados con dos dados.*

		Segundo dado					
		1	2	3	4	5	6
	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
Primer	3	4	5	6	7	8	9
dado	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

posibilidades: $U = \{(C, C), (C, X), (X, C), (X, X)\}$.* Ésta es una definición listada de U; una definición por regla —aunque no la usaremos— podría ser $U = \{x\}$; donde x representa todas las combinaciones de C y X. En este caso U es un producto cartesiano. Que sea $A_1 = \{C_1, X_1\}$, el resultado de la primera moneda y $A_2 = \{C_2, X_2\}$, el de la segunda moneda. Recordando que un producto cartesiano de dos conjuntos es el conjunto de todos los pares ordenados, cuya primera entrada es un elemento de un conjunto y la segunda entrada un elemento de otro conjunto, podemos esquematizar la generación del producto cartesiano de este caso, $A_1 \times A_2$, como en la figura 7.1. Note que hay cuatro líneas conectando a A_1 y a A_2, por lo que hay cuatro posibilidades: $\{(C_1, C_2), (C_1, X_2), (X_1, C_2), (X_1, X_2)\}$. Este esquema de pensamiento y procedimiento puede utilizarse para definir muchos espacios muestrales de *Us*, aunque el procedimiento real puede ser tedioso.

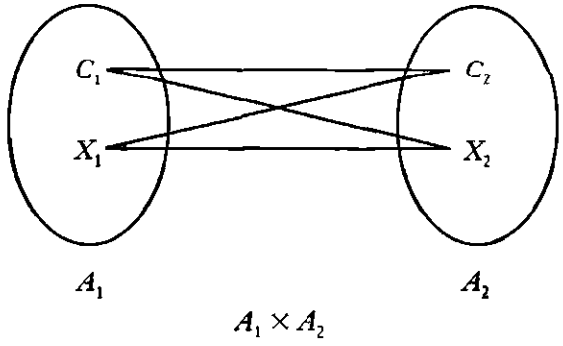
En caso de usar dos dados, ¿qué sería U? Si se piensa en el producto cartesiano de dos conjuntos, probablemente surja un poco de problema. Suponga que A_1 son los resultados o los puntos del primer dado: {1, 2, 3, 4, 5, 6} y que A_2 son los resultados o los puntos del segundo dado, entonces $U = A_1 \times A_2 = \{(1, 1), (1, 2), \ldots, (5, 6), (6, 6)\}$. Esto se puede ilustrar como se hizo con el ejemplo de las monedas, pero contar las líneas es más difícil debido a que habrá demasiadas. Se puede conocer el número de posibles resultados simplemente realizando la operación $6 \times 6 = 36$, o en una fórmula: mn, donde m es el número de posibles resultados del primer conjunto y n es el número de posibles resultados del segundo conjunto.

A menudo es posible resolver problemas difíciles de probabilidad usando diagramas de árbol. Éstos definen espacios muestrales y posibilidades lógicas con claridad y precisión. Un árbol es un diagrama que da todas las posibles alternativas o resultados en las combinaciones de conjuntos, al proporcionar rutas y puntos del conjunto. Esta definición es un poco difícil de manejar por lo que se requiere ilustrarla: tomemos el ejemplo de las monedas (se coloca el árbol sobre un costado). El diagrama de árbol se muestra en la figura 7.2.

Para determinar el número de posibles alternativas se cuenta el número de alternativas o puntos ubicados en la "cima" del árbol. En este caso hay cuatro alternativas y para nombrarlas se leen, para cada punto final, los puntos que llevan a él. Por ejemplo, la primera alternativa es (C_1, C_2). Obviamente tres, cuatro o más monedas pueden ser usadas, el único problema es que el procedimiento es tedioso por el gran número de alternativas. El diagrama para tres monedas se ilustra en la figura 7.3; aquí existen ocho posibles alternativas, resultados o puntos muestrales: $U = \{(C_1, C_2, C_3), (C_1, C_2, X_3), \ldots, (X_1, X_2, X_3)\}$ (los elementos de este conjunto se llaman *tríos ordenados*).

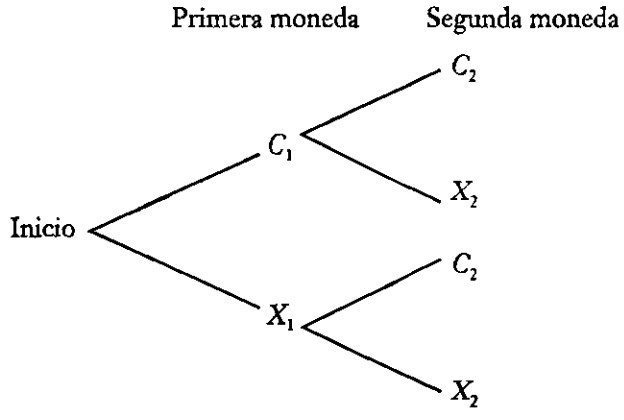
* C y X corresponden a los nombres dados a los lados de una moneda, "cara" y "cruz". En lo sucesivo se utilizará C para referirse a "cara" y X para referirse a "cruz". (*N. del T.*)

FIGURA 7.1



Los puntos muestrales de un espacio muestral pueden parecer un poco confusos al lector porque se han discutido dos clases de puntos sin haber hecho alguna diferenciación. Esta confusión puede aclararse por medio del uso de un término: Un evento es un subconjunto de U; cualquier elemento de un conjunto es también un subconjunto del conjunto. Recuerde que con el conjunto $A = \{a_1, a_2\}$ por ejemplo, ambos $\{a_1\}$ y $\{a_2\}$ son subconjuntos de A, así como lo son $\{a_1, a_2\}$ y $\{\ \}$, es el conjunto vacío. Igualmente, todos los resultados de las figuras 7.2 y 7.3, por ejemplo, (C_1, X_2), (X_1, C_2) y (X_1, C_2, X_3) son subconjuntos de sus respectivos U, por lo tanto, por definición también son eventos. Pero en el uso cotidiano, un evento abarca un poco más que los puntos. Todos los puntos son eventos (subconjuntos), pero no todos los eventos son puntos, es decir, un punto o resultado es una clase especial de evento: la clase más simple. Siempre que se establece una proposición, se describe un evento. Por ejemplo: "si se lanzan dos monedas al aire, ¿cuál es la probabilidad de tener dos caras?" Las "dos caras" es un evento, que si sucede, en este caso, es también un punto muestral. Pero si la pregunta fuera: "¿Cuál es la probabilidad de obtener al menos una cara?". "Al menos una cara" es un evento, pero no un punto muestral porque incluye, en este caso, tres puntos muestrales: (C_1, C_2), (C_1, X_2) y (X_1, C_2) (véase figura 7.2).

FIGURA 7.2



Determinación de probabilidades con monedas

Si se lanza una moneda recién acuñada 3 veces y se anota $p(C) = 1/2$ y $p(X) = 1/2$, que significa que la probabilidad de caras es 1/2 y similar para las cruces, se supone, entonces, equiprobabilidad. El espacio muestral para los tres lanzamientos de monedas (o un lanzamiento de tres monedas) es: $U = \{(C, C, C), (C, C, X), (C, X, C), (C, X, X), (X, C, C), (X, C, X), (X, X, C), (X, X, X)\}$. Note que si no se presta atención al orden de caras y cruces, se obtiene un caso de tres caras, un caso de tres cruces, tres casos de dos caras y una cruz, y tres casos de dos cruces y una cara. La probabilidad de cada uno de los ocho resultados es obviamente 1/8, la probabilidad de tres caras es 1/8 y la de tres cruces es 1/8. La probabilidad de dos caras y una cruz es, por otro lado, 3/8 y de forma similar para la probabilidad de dos cruces y una cara.

Las probabilidades de todos los puntos en el espacio muestral deben sumar 1.00, y las probabilidades siempre son positivas. Si se realiza un diagrama de árbol de probabilidad para el experimento de los tres lanzamientos de moneda, se parecerá al de la figura 7.3. Cada rama completa del árbol (desde el inicio hasta el tercer lanzamiento) es un punto muestral y todas las ramas comprenden el espacio muestral. Las secciones de cada rama (o ruta) están etiquetadas con la probabilidad; en este caso todas son etiquetadas con 1/8. Esto conduce naturalmente al enunciado de un principio básico: si los resultados en diferentes puntos del árbol (en el primero, segundo y tercer lanzamiento) son independientes uno de otro (esto es, si un resultado no influye a otro en forma alguna), entonces la probabilidad de cualquier punto muestral (*CCC*, quizá) es el producto de las probabilidades de los resultados aislados. Por ejemplo la probabilidad de tres caras es $1/2 \times 1/2 \times 1/2 = 1/8$.

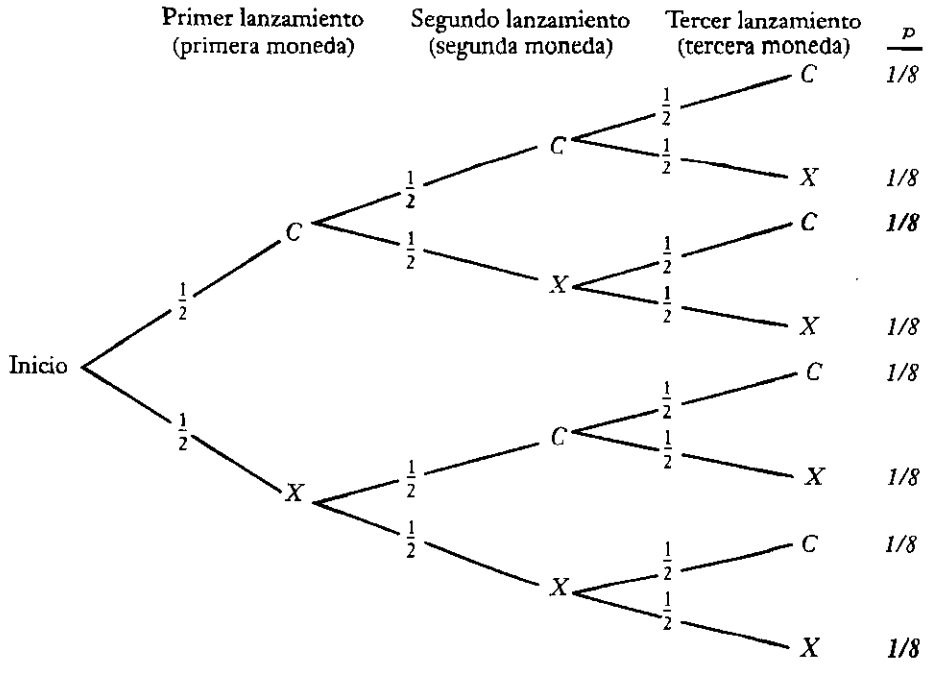
Otro principio implica que para obtener la probabilidad de cualquier evento se sumen las probabilidades de los puntos muestrales que componen dicho evento. Por ejemplo, ¿cuál es la probabilidad de obtener dos caras y una cruz? Si se buscan en el árbol las ramas que tienen dos caras y una cruz existen tres de ellas, que se observan en la figura 7.3. Así, $1/8 + 1/8 + 1/8 = 3/8$. En el lenguaje de conjuntos, se encuentran los subconjuntos (eventos) de U y se observan sus probabilidades. Los subconjuntos de U del tipo "2 caras y 1 cruz" son, partiendo del árbol o de la definición previa de U: $\{(C, C, X), (C, X, C), (X, C, C)\}$. Si se llama A a este conjunto o evento, entonces $p(A_1) = 3/8$.

Este procedimiento puede seguirse con un experimento laborioso de cien lanzamientos de moneda al aire, pero en su lugar, para obtener las expectativas teóricas, solamente se multiplica el número de lanzamientos por la probabilidad de cualquiera de ellas para tener el número esperado de caras (o cruces). Esto puede hacerse porque *todas* las probabilidades son iguales. Una pregunta importante que debe hacerse ahora es: en experimentos reales en los que cien monedas fueran lanzadas al aire, ¿obtendríamos exactamente 50 caras, suponiendo que las monedas están niveladas? No, no muy a menudo, serían aproximadamente 8 veces en 100 repeticiones de esta naturaleza. Esto puede escribirse $p = 8/100$, o 0.08. (Las probabilidades pueden escribirse en forma decimal o de fracción, aunque es más frecuente usar la forma decimal.)

Un experimento con dados

Si lanzamos dos dados recién fabricados 72 veces bajo condiciones cuidadosamente controladas; si sumamos el número de puntos en los dos dados en los 72 lanzamientos, obtendremos un conjunto de sumas que van de 2 a 12. Algunos de estos resultados (sumas) serán más frecuentes que otros, simplemente porque hay más formas de obtener ciertos resultados. Por ejemplo, solamente hay una forma de obtener 2 o 12: 1 + 1 y 6 + 6, pero

FIGURA 7.3



existen 3 maneras de obtener 4: 1 + 3, 3 + 1 y 2 + 2. Si esto es cierto, entonces las probabilidades para obtener diferentes sumas deben ser diferentes. El juego de los dados está basado en estas diferencias en las frecuencias esperadas.

*Para resolver un problema de probabilidad a priori, primero se debe definir el espacio muestral: U = {(1, 1), (1, 2), (1, 3), ..., (6, 4), (6, 5), (6, 6)}; es decir, se aparea cada número del primer dado con cada número del segundo dado en turno (otra vez el producto cartesiano). Esto puede verse fácilmente si se coloca este procedimiento en una matriz (véase tabla 7.1). Suponiendo que queremos conocer la probabilidad del evento "obtener 7", simplemente contamos el número de 7 en la tabla. Se encuentran seis de ellos a lo largo de la diagonal central. Existen 36 puntos muestrales en U, obtenidos por medio de enumerarlos, como se hizo anteriormente, o usando la fórmula **mn**. Esta fórmula implica multiplicar el número de posibilidades del primer evento por el número de posibilidades del segundo evento. Este método puede definirse de la siguiente forma: suponga que hay **m** formas de hacer algo, A, y que hay **n** formas de hacer otra cosa, B; si las **n** formas de hacer B son independientes de las **m** formas de hacer A, entonces hay $m \times n$ formas de hacer ambas, A y B. Este principio puede extenderse para más de dos cosas. Si, por ejemplo, hay tres cosas A, B y C, donde hay **r** formas de hacer C, entonces la fórmula es **mnr**.*

Si se aplica esta fórmula al problema de los dados, entonces $mn = 6 \times 6 = 36$, y suponiendo equiposibilidad otra vez, la probabilidad de cualquier resultado simple es 1/36. La probabilidad de obtener un 12, por ejemplo, es de 1/36. Sin embargo, la probabilidad de obtener un 4 es diferente, dado que el 4 ocurre tres veces en el cuadro anterior. Se deben sumar las probabilidades para cada uno de estos elementos del espacio muestral: 1/36 + 1/36 + 1/36 = 3/36; así $p(4) = 3/36 = 1/12$. Como se vio anteriormente, la probabilidad de

un 7 es $p(7) = 6/36 = 1/6$; la probabilidad de un 8 es $p(8) = 5/36$. Note también que podemos calcular las probabilidades de combinaciones de eventos. Los jugadores frecuentemente apuestan a tales combinaciones. Por ejemplo, ¿cuál es la probabilidad de obtener un 4, o un 10? En lenguaje de conjuntos ésta es una pregunta de unión: $p(4 \cup 10)$. Se cuenta el número de 4 y de 10 en el cuadro; hay tres 4 y tres 10. Así $p(4 \cup 10) = 6/36$.

En la tabla 7.1, por medio del conteo de las probabilidades de cada tipo de resultado se puede extraer una tabla de frecuencias esperadas (f_e) para 36 lanzamientos; después se duplican estas frecuencias para obtener las frecuencias esperadas (*a priori*) para 72 lanzamientos; se confrontan las frecuencias esperadas (f_e) contra las frecuencias obtenidas (f_o) cuando dos dados son tirados realmente 72 veces, y finalmente aparecen las diferencias absolutas entre las frecuencias esperadas y las frecuencias obtenidas (los resultados se encuentran en la tabla 7.2). Las discrepancias no son grandes, de hecho, con una prueba estadística no difieren significativamente de lo esperado por efectos del azar. El método *a priori* parece tener sus virtudes.

Algo de teoría formal

Se tiene un espacio muestral U, con los subconjuntos A, B,... Los elementos de U (y de A, B,...) son a_i, b_i,... esto es $a_1, a_2, \ldots, a_n$ y $b_1, b_2, \ldots b_m$, etcétera. A, B y los demás son eventos. En realidad, aunque se ha hablado de la probabilidad de una ocurrencia aislada, de hecho se refiere a la probabilidad de un tipo de ocurrencia. Se puede hablar acerca de la probabilidad de cualquier evento aislado de U, por ejemplo, porque cualquier miembro particular de U es concebido como representativo de todo U, y también para las probabilidades de los subconjuntos A, B, ..., K de U. La probabilidad de U es 1; la probabilidad de E (el conjunto vacío) es 0 (cero); o $p(U) = 1.00$; $p(E) = 0$. Para determinar la probabilidad de cualquier subconjunto de U, debe asignarse una medida del conjunto. Para asignar tal medida, se da un peso a cada elemento de U y también a cada elemento de los subconjuntos de U. Un *peso* es definido por Kemeny, Snell y Thompson (1974) como sigue:

> Un peso es un número positivo asignado a cada elemento, x, en U, y escrito $w(x)$, de tal manera que la suma de todos estos pesos, $\Sigma w(x)$, es igual a 1.

Ésta es una noción de función; w es llamada una función del peso. Es una regla que asigna pesos a los elementos de un conjunto U, en forma tal que la suma de los pesos es igual a 1; esto es, $w_1 + w_2 + w_3 + \ldots + w_n = 1.00$, y $w_i = 1/n$. Los pesos son iguales, asumiendo equiprobabilidad; cada peso es una fracción con 1 en el numerador y el número de casos n en el denominador. En el experimento previo de lanzar una moneda (figura 7.3), los pesos asignados a cada elemento de U, siendo U *todos* los resultados, son igual a 1/8. La suma de todas las funciones del peso $w(x)$, es $1/8 + 1/8 + \ldots + 1/8 = 1$. En la teoría de la probabilidad, la suma de los elementos del espacio muestral debe ser siempre igual a 1.

TABLA 7.2 *Frecuencias esperadas y obtenidas de las sumas de dos dados lanzados 72 veces*

Suma de los dados	2	3	4	5	6	7	8	9	10	11	12
$f_e(36)$	1	2	3	4	5	6	5	4	3	2	1
$f_e(72)$	2	4	6	8	10	12	10	8	6	4	2
$f_o(72)$	4	2	6	6	10	15	7	11	6	4	1
Diferencia	2	2	0	2	0	3	3	3	0	0	1

Obtener la medida de un conjunto a partir de los pesos es fácil: la medida de un conjunto es la suma de los pesos de los elementos de dicho conjunto.

$$\sum_{x \text{ en } U} w(x) \quad \text{o} \quad \sum_{x \text{ en } A} w(x)$$

(Observe que la suma de los pesos en un subconjunto A, de U, no tiene que ser igual a 1. De hecho, usualmente es menor que 1.)

Escribir $m(A)$, significa "La medida del conjunto A". Esto simplemente habla de la suma de los pesos de los elementos del conjunto A.

Si se selecciona una muestra al azar, a partir de 400 alumnos de cuarto grado de un sistema escolar, entonces U son todos los 400 alumnos. Cada alumno es un punto muestral de U. Cada alumno es una x en U. La probabilidad de seleccionar cualquier niño al azar es 1/400. Suponga que A es igual a los niños en U, y B es igual a las niñas en U, y que hay 100 niños y 300 niñas. A cada niño le es asignado un peso de 1/400 y a cada niña un peso de 1/400. Si se desea una muestra de 100 alumnos en total, la expectativa es, entonces, 25 niños y 75 niñas en la muestra. La medida del conjunto A, $m(A)$, es la suma de los pesos de todos los elementos en A. Dado que hay 100 niños en U, sumamos los 100 pesos: 1/400 + 1/400 + ... + 1/400 = 100/400 = 1/4 o:

$$m(A) = \sum_{x \text{ en } A} w(x) = \frac{1}{4}$$

De manera similar:

$$m(B) = \sum_{x \text{ en } B} w(x) = \frac{3}{4}$$

Para el conjunto B (las niñas), se suman 300 pesos, siendo cada uno de ellos de 1/400. En pocas palabras, la suma de los pesos representa las probabilidades, es decir que la medida de un conjunto es la probabilidad de que un miembro del conjunto sea elegido. Así, se puede decir que la probabilidad de que un miembro de la muestra de 400 alumnos sea niño es de 1/4, y la probabilidad de que el miembro seleccionado sea niña es de 3/4. Para determinar las frecuencias esperadas, se multiplica el tamaño de la muestra por estas probabilidades: $1/4 \times 100 = 25$ y $3/4 \times 100 = 75$.

La probabilidad tiene tres propiedades fundamentales:

1. La medida de cualquier conjunto, como se definió anteriormente, es mayor o igual a 0 y menor o igual a 1, es decir, las probabilidades (medidas de conjuntos) pueden ser 0, 1 o alguna cantidad entre ellos.
2. La medida de un conjunto $m(A)$, es igual a 0, si y sólo si no hay miembros en A; esto es, si A está vacío.
3. Suponga que A y B son los conjuntos. Si A y B están separados, esto es, $A \cap B = E$, entonces $m(A \cup B) = m(A) + m(B)$.

Esta ecuación dice que cuando A y B no tienen miembros en común, entonces tanto la probabilidad de A como de B, o de ambos es igual a las probabilidades combinadas de A y B.

No hay necesidad de un ejemplo para ilustrar el punto 1), ya se han dado varios anteriormente. Para ilustrar 2), suponga que en el ejemplo de niños y niñas queremos saber cuál es la probabilidad de tener en la muestra a un maestro, pero U no incluye maestros. Si

C es el conjunto de los maestros de cuarto grado, en este caso, dicho conjunto está vacío y $m(C) = 0$. Con el mismo ejemplo de alumnos niños y niñas se ilustra 3): si A es el conjunto de niños y B el conjunto de niñas, entonces $m(A \cup B) = m(A) + m(B)$ pero $m(A \cup B) = 1.00$ porque ellos son los únicos subconjuntos de U. Debido a que $m(A) = 1/4$ y que $m(B) = 3/4$, la ecuación se sostiene.

Eventos compuestos y su probabilidad

Anteriormente se dijo que un evento es un subconjunto de U, pero es necesario detallar esto. Un evento es un conjunto de posibilidades: es un conjunto de eventos posibles; es el resultado de un "experimento" de probabilidad. Un evento compuesto es la co-ocurrencia de dos o más eventos aislados (o compuestos). Las dos operaciones de conjuntos de intersección y unión —las operaciones que más interesan aquí— implican eventos compuestos. Si se lanza una moneda y un dado, el resultado es un evento compuesto y se puede calcular la probabilidad de tal evento. Aún más interesante, se puede preguntar cómo están relacionadas ciertas variables demográficas. Una forma de hacer esto es buscar respuestas a preguntas tales como: "¿Cuál es la probabilidad de detectar a un usuario de drogas que elige días específicos para usar la droga, sin considerar ninguna estrategia de prueba de drogas?" (véase Borack, 1997) o, "¿Cuál es la probabilidad de que dos estudiantes en el mismo salón de clases tengan el mismo mes y día de nacimiento?"(véase Nunnikhoven, 1992), o "¿Cuál es la probabilidad de que un estudiante de posgrado abandone sus estudios?"(véase Cooke, Sims & Peyrefitte, 1995).

Los eventos compuestos son más interesantes que los eventos aislados —y más útiles en investigación—. Con ellos pueden estudiarse las relaciones. Para entender esto, primero proceda a definir e ilustrar qué son los eventos compuestos y después a examinar ciertos problemas de conteo y las formas en que el conteo está relacionado con la teoría de conjuntos y la teoría de la probabilidad. Se encontrará que si la teoría básica es entendida, la aplicación de la teoría de probabilidad a los problemas de investigación se facilita considerablemente. Además, la interpretación de los datos se ve menos sujeta a errores.

Suponga que se ha estudiado un grupo de niños de una escuela primaria y que hay 100 niños en total en dicho grupo: 60 de cuarto grado y 40 de sexto grado. La función numérica es útil ya que asigna a cualquier conjunto el número de miembros en el conjunto. El número de miembros en A es $n(A)$. En este caso $n(U) = 100$, $n(A) = 60$, y $n(B) = 40$, donde A es el conjunto de los alumnos de cuarto grado y B es el conjunto de los alumnos de sexto grado, ambos subconjuntos de U, que representa los 100 alumnos de la escuela primaria. Si no hay traslape entre ambos conjuntos, $A \cap B = E$, entonces la siguiente ecuación se sostiene:

$$n(A \cup B) = n(A) + n(B) \tag{7.1}$$

Recuerde que anteriormente la definición de frecuencia de probabilidad fue dada como sigue:

$$p = \frac{f}{f + u} \tag{7.2}$$

donde f es el número de casos favorables y u el número de casos desfavorables. El numerador es $n(f)$ y el denominador es $n(U)$, el número total de casos posibles. De manera similar se pueden dividir los términos de la ecuación 7.1 entre $n(U)$:

$$\frac{n(A \cup B)}{n(U)} = \frac{n(A)}{n(U)} + \frac{n(B)}{n(U)} \tag{7.3}$$

Esto se reduce a probabilidades análogamente a la ecuación 7.2:

$$p(A \cup B) = p(A) + p(B) \tag{7.4}$$

Usando el ejemplo de los 100 alumnos de escuela y sustituyendo los valores en la ecuación 7.3, se obtiene:

$$\frac{100}{100} = \frac{60}{100} + \frac{40}{100},$$

lo que se sustituye por la ecuación 7.4:

$$1.00 = .60 + .40$$

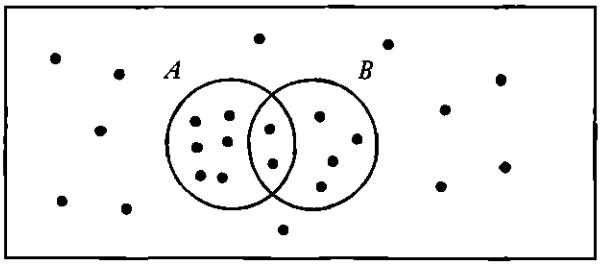
En muchos casos, dos (o más) conjuntos en los que nos interesamos, no están separados sino que de hecho se traslapan. Cuando esto sucede, entonces $A \cap B \neq E$ y no es cierto que $n(A \cup B) = n(A) + n(B)$. Observe la figura 7.4, en ésta A y B son subconjuntos de U; los puntos muestrales están indicados por puntos. El número de puntos muestrales en A es 8; el número de puntos muestrales en B es 6. Hay dos puntos muestrales en $A \cap B$. por lo que la ecuación anterior no se sostiene. Si se calculan todos los puntos en $A \cup B$ con la ecuación 7.1, se obtiene 8 + 6 = 14 puntos; pero existen solamente 12 puntos, por lo que esta ecuación debe modificarse en una ecuación más general que abarque todos los casos:

$$n(A \cup B) = n(A) + n(B) - n(A \cap B) \tag{7.5}$$

Debe quedar claro que cuando se usa la ecuación 7.1, el error resulta de contar los dos puntos de $A \cap B$ dos veces. Por lo tanto, se substrae una vez $n(A \cap B)$, para corregir la ecuación. Ahora se ajusta a cualquier posibilidad. Si, por ejemplo, $n(A \cap B) = E$, el conjunto vacío, la ecuación 7.5 se reduce a la ecuación 7.1. La ecuación 7.1 es un caso especial de la ecuación 7.5. Si se calcula el número de puntos muestrales en $n(A \cap B)$ de la ecuación 7.4, entonces se obtiene $n(A \cup B) = 8 + 6 - 2 = 12$. Si se divide la ecuación 7.5 entre $n(U)$, como en la ecuación 7.3, resulta:

$$p(A \cup B) = p(A) + p(B) - p(A \cap B) \tag{7.6}$$

FIGURA 7.4



Si se sustituyen las marcas o puntos muestrales, se encuentra que:

$$\frac{12}{24} = \frac{8}{24} + \frac{6}{24} - \frac{2}{24}$$

$$.50 = .33 + .25 - .08$$

Entonces en una muestra aleatoria de U, las probabilidades de que un elemento sea miembro de A, B, $A \cap B$ y $A \cup B$, son respectivamente: .33, .25, .08 y .50.

Independencia, exclusión mutua y exhaustividad

Considere las siguientes preguntas, variantes de las que deben hacerse los investigadores. ¿La ocurrencia de este evento A impide la posibilidad de ocurrencia de este otro evento B? ¿La ocurrencia del evento A tiene alguna influencia en la ocurrencia del evento B? ¿Están relacionados los eventos A, B y C? ¿Cuando A ocurre, esto tiene influencia en los resultados de B y quizás C? ¿Los eventos A, B, C y D agotan las posibilidades? ¿O hay quizás, otras posibilidades E, F, etcétera? Por ejemplo, un investigador está estudiando las decisiones de un comité de educación y su relación con la preferencia política, preferencia religiosa, educación y otras variables. Para relacionar estas variables con las decisiones del comité, el investigador ha de tener algún método para clasificar las decisiones. Una de las primeras preguntas que deben hacerse es "¿Mi sistema de clasificación agota todas las posibilidades?" Una pregunta adicional es "Si el comité toma una clase de decisión, ¿excluirá esto la posibilidad de que se tome otra decisión?" Quizás la pregunta más importante que el investigador puede hacer es "¿Si el comité toma una decisión particular, influirá ésta en cualquier otra decisión?"

Se ha hablado de exhaustividad, exclusión mutua e independencia. Ahora se definen estas ideas de una manera más detallada y se usan en ejemplos de probabilidad. Su aplicabilidad general e importancia se harán evidentes en los capítulos donde se estudia el análisis de datos.

Suponga que A y B son subconjuntos de U. ¿Hay otros subconjuntos de U (además del conjunto vacío)? ¿Agotan A y B el espacio muestral? ¿Están todos los puntos muestrales del espacio muestral U incluidos en A y B? Un ejemplo simple es: Establezca que $A = \{C, X\}$; y que $B = \{1, 2, 3, 4, 5, 6\}$. Si se lanzan simultáneamente una moneda y un dado, ¿cuáles son las posibilidades resultantes? A menos que todas las posibilidades estén agotadas, no se puede resolver el problema de probabilidad. Hay 12 posibilidades (2×6). Los conjuntos A y B agotan el espacio muestral (esto, por supuesto, es obvio dado que A y B generan el espacio muestral). Ahora tomemos un ejemplo más real: un investigador está estudiando las preferencias religiosas y utiliza el siguiente sistema para categorizar a los individuos: protestantes, católicos y judíos. Implícitamente, U es establecido de tal forma que incluya a toda la gente (con o sin preferencia religiosa) y a todos los subconjuntos de U: A es igual a protestantes, B es igual a católicos y C es igual a judíos. La pregunta de conjuntos es: ¿$A \cup B \cup C = U$? ¿Se han agotado todas las preferencias religiosas? ¿Que hay de los budistas? ¿Y de los musulmanes? ¿Y de los ateos?

La *exhaustividad*, entonces significa que los subconjuntos de U agotan todo el espacio muestral, o que $A \cup B \cup \ldots \cup K = U$, donde A, B, ..., K son subconjuntos de U, el espacio muestral. En lenguaje de probabilidad, esto significa: $p(A \cup B \cup \ldots \cup K) = 1.00$. A menos que el espacio muestral U, haya sido agotado, por así decirlo, las probabilidades no pueden calcularse adecuadamente. Por ejemplo, en el caso de la preferencia religiosa,

suponga que $A \cup B \cup C = U$ pero que de hecho hubiera un gran número de individuos sin una preferencia religiosa en particular. Así que en realidad, $A \cup B \cup C \cup D = U$, donde D es el subconjunto de individuos sin una preferencia religiosa. Las probabilidades calculadas en el supuesto de esta ecuación serían muy diferentes de aquellas basadas en el supuesto de la ecuación previa.

Dos eventos, A y B, son *mutuamente excluyentes* cuando no están unidos, o cuando $A \cap B = E$, es decir, cuando la intersección de dos (o más) conjuntos es el conjunto vacío (o cuando dos conjuntos no tienen elementos en común), se dice que ambos son mutuamente excluyentes. Esto es igual a decir, otra vez en lenguaje de probabilidad, $p(A \cap B) = 0$. Es más conveniente para los investigadores que los eventos sean mutuamente excluyentes, porque entonces pueden sumar las probabilidades de los eventos. Un principio, en términos de conjuntos y probabilidades es el siguiente: si los eventos (conjuntos) A, B y C son mutuamente excluyentes, entonces $p(A \cup B \cup C) = p(A) + p(B) + p(C)$. Éste es el caso especial de un principio más general que se discutió en una sección previa (véase ecuaciones 7.1, 7.4, 7.5 y 7.6 y la argumentación que las explica).

Uno de los principales propósitos del diseño de investigación es establecer las condiciones de independencia de los eventos, de tal manera que las condiciones de dependencia de éstos puedan ser estudiadas adecuadamente. Dos eventos A y B, son estadísticamente independientes si se ajustan a la ecuación:

$$p(A \cap B) = p(A) \cdot p(B) \tag{7.7}$$

que dice que la probabilidad de que A y B ocurran, es igual a la probabilidad de A por la probabilidad de B. Ejemplos fáciles y claros de eventos independientes son el lanzamiento de monedas y dados. Si A es el evento de lanzar un dado y B es el evento de lanzar una moneda al aire, y $p(A) = 1/6$ y $p(B) = 1/2$, entonces, si $p(A) \cdot p(B) = 1/6 \cdot 1/2 = 1/12$, entonces A y B son independientes. Si lanzamos una moneda 10 veces, un lanzamiento no tiene influencia en ningún otro lanzamiento; los lanzamientos son independientes y sucede lo mismo cuando se lanza un dado. De manera similar, cuando lanzamos una moneda al aire y al mismo tiempo se lanza un dado, los eventos de lanzar el dado, A, y lanzar la moneda, B, son independientes. El resultado de lanzar el dado no tiene influencia en el lanzamiento de la moneda, y viceversa. Desafortunadamente, este claro modelo no siempre se puede aplicar a situaciones de investigación.

La creencia, basada en el sentido común, de la llamada ley de promedios es tremendamente errónea, pero ilustra el poco entendimiento que existe respecto del concepto de independencia. Esta dice que si hay un gran número de ocurrencias de un evento, la posibilidad de que ese evento ocurra en el siguiente ensayo es mínima. Supongamos que una moneda es lanzada al aire, y que resultan caras en cinco veces seguidas. La "ley de los promedios" haría creer que existe una mayor posibilidad de obtener cruz en la siguiente lanzada, pero no es así. La probabilidad sigue siendo 1/2. Cada lanzamiento es un evento independiente.

Suponga que los estudiantes de una clase universitaria están tomando un examen, y que lo están realizando bajo las condiciones usuales de no comunicación, no ver el documento del compañero, etcétera. Las respuestas de cada uno pueden ser consideradas independientes de las respuestas de cualquier otro estudiante. ¿Pueden considerarse independientes las respuestas a los reactivos dentro de cada examen? Suponga que la respuesta a un reactivo posterior en el examen está insertada en reactivo previo en el mismo examen. Digamos que la probabilidad de tener correcto el reactivo posterior debido al azar es 1/4, pero el hecho de que la respuesta fuese dada antes puede cambiar esta probabilidad. Para algunos estudiantes puede llegar a ser de 1.00. Lo que es importante para el

investigador es conocer que esa independencia es frecuentemente difícil de lograr y que la carencia de independencia, cuando la investigación supone que la hay, puede afectar seriamente la interpretación de los datos.

Suponga que se ordenan por rangos los exámenes y luego se les asignan calificaciones con base en dicho orden. Éste es un procedimiento perfectamente legítimo y útil, pero debe reconocerse que las calificaciones dadas con el método de ordenar por rangos no son independientes (si acaso pudieran serlo). Si se toman cinco de esos exámenes y después de leerlos, uno de ellos, es ordenado en primer lugar (como el mejor), el siguiente se ordena como segundo, y así con los cinco exámenes. Se le asigna el número "1" al primero, el número "2" al segundo, el número "3" al tercero, el número "4" al cuarto y el número "5" al quinto. Después de usar el número 1, sólo quedan, 2, 3, 4 y 5. Después del número 2, quedan solamente el 3, 4 y 5. Cuando se asigna el 4, obviamente debe asignarse el 5 al examen restante; en pocas palabras, la asignación del 5 se ve influida por la asignación del 4 (y también por las del 1, 2 y 3). La asignación de los eventos no fue independiente. Aquí podría surgir la pregunta, "¿Y esto, importa?". Suponga que tomamos los rangos, y se los usamos como puntuaciones, para luego hacer inferencias acerca de las diferencias de las medias entre los grupos, digamos, entre dos clases. La prueba estadística usada para hacer esto está probablemente basada en el paradigma de la moneda y los dados, con su independencia original, pero aquí no se ha seguido este modelo (uno de sus más importantes supuestos, la independencia, ha sido ignorada).

Cuando investigamos eventos que carecen de independencia, las pruebas estadísticas carecen de cierta validez. Una prueba de χ^2, por ejemplo, supone que los eventos (respuestas de los individuos a una pregunta de entrevista) registrados en las casillas de una tabla de contingencia, son independientes una de otra. Si los eventos registrados no son independientes uno de otro, entonces las bases de la prueba estadística y las inferencias hechas a partir de ésta, han sido alteradas.

Considere la investigación de las relaciones entre la autoeficacia percibida y el comportamiento. En este ejemplo, los investigadores intentan mostrar la congruencia entre el juicio de la autoeficacia y el comportamiento real. A los participantes generalmente se les aplican escalas de autoeficacia que describen un conjunto de tareas bien definidas. A cada participante se le pide juzgar si puede completar cada tarea. La percepción y el comportamiento son congruentes cuando la percepción del sujeto iguala a la ejecución real, esto es, "dijo que puede hacerlo y lo hizo" y "dijo que no puede hacerlo y en realidad no lo hizo". Cervone (1987) afirma que un gran número de investigadores que trabajan en esta área han reportado congruencias excepcionalmente altas. Numerosos estudios han encontrado congruencias del 80-90%. Cervone señala, sin embargo, que los datos obtenidos de las escalas de autoeficacia no son independientes ya que cada individuo participante contribuye con más de una observación al análisis. Cervone (1987, p.710) afirma que "como en cualquier área de investigación, uno no debe suponer que las observaciones múltiples de un mismo sujeto sean independientes".

Kramer y Schmidhammer (1992) encontraron problemas de independencia similares en investigaciones de conducta animal. Algunos estudios sobre conducta humana y conducta animal dependen de datos etológicos de frecuencia. Este tipo de datos generalmente cuentan el número de encuentros entre organismos (animales o humanos) o de la ejecución de un comportamiento. Kramer y Schmidhammer usan el ejemplo de la medición del comportamiento de la rana. Para obtener observaciones independientes del comportamiento de vocalización o no vocalización de la rana macho a lo largo de una sección de la orilla de un lago, los investigadores necesitan comprobar si la ausencia o presencia de la vocalización de una rana no tiene efecto en otras ranas. Kramer y Schmidhammer observaron que muchos patrones de comportamiento de interés para el etólogo tienden a

ocurrir en conglomerados y no en forma independiente. Kramer y Schmidhammer citan muchos estudios que pueden tener un problema potencial de independencia.

Otro estudio es el realizado por Keane (1990), en el cual examinó la preferencia de ratas macho de patas blancas por hembras en etapa de estro. Keane registró el número de encuentros que cada rata hembra en etapa de estro tenía con cada ratón macho; dichos encuentros fueron clasificados como agresivos (pelea o persecución) o amistosos (acicalar u olfatear). El origen de la rata macho fue documentado a partir del momento del nacimiento, para que el experimentador supiera cómo cada rata hembra estaba emparentada con cada macho. Keane deseaba saber si la hembra en etapa de estro prefería a un macho emparentado o a uno que no lo fuera. Keane encontró que la rata hembra mostraba una conducta más amistosa y menos conductas agresivas hacia sus primos hermanos que hacia los machos con quienes no estaba emparentada. Si se toman en consideración los puntos del artículo de Kramer y Schmidhammer, el estudio de Keane sería imperfecto en el hecho de que sus observaciones pueden no ser independientes. Una rata hembra pudo haber tenido un especial "disgusto" hacia un no pariente o un "gusto" especial por un primo hermano, y los conteos de frecuencia podrían estar inclinados a favor de ese par de machos.

Ahora consideremos un estudio antiguo, pero importante, sobre el comportamiento agresivo de los simios, realizado por Hebb y Thompson en 1968. Los datos de su investigación se presentan en la tabla 7.3. El problema trataba acerca de la relación entre sexo y agresión. Se tomaron muestras del comportamiento de 30 chimpancés adultos con el fin de estudiar diferencias individuales en el temperamento de los simios. Sin entrar en detalles, puede decirse que un análisis de las observaciones mostró que tanto los machos como las hembras presentaron un comportamiento amistoso casi con igual frecuencia, pero los machos eran más agresivos. Los resultados de Hebb y Thompson parecerían decir: "¡cuidado con los machos!", pero los autores señalaron que esto difiere de la experiencia que tienen los cuidadores de simios. ¡19 de 20 arañazos y cortaduras fueron infligidos por hembras! Entonces Hebb y Thompson buscaron la interesante, aunque desconcertante idea de tabular la incidencia de actos agresivos de dos maneras: cuando éstos eran precedidos por una cuasi–agresión, es decir, por una alerta de ataque, y cuando los actos agresivos eran precedidos por una conducta amistosa. Las incidencias resultantes del comportamiento buscado parecen indicar: ¡cuidado con las hembras cuando son amistosas!". Los machos fueron los únicos en mostrar actos cuasi–agresivos antes de actos verdaderamente agresivos (37 actos de machos y 0 actos de hembras), sin embargo, sólo las hembras actuaron en forma agresiva después de mostrar una conducta amistosa (15 actos de hembra y 0 actos de machos).

Estos datos no pueden ser analizados estadísticamente en forma válida dado que los números indican la frecuencia de tipos de actos, así, todos los 37 actos de los machos podrían haber sido realizados sólo por uno o dos de ellos. Si un simio hubiera cometido todos los 37 actos, entonces debería quedar claro que los actos no eran independientes uno de otro; dicho simio podría tener mal carácter y las conductas por mal carácter notoriamente carecen de independencia en los actos humanos y animales.

El siguiente ejemplo es hipotético: un investigador realiza un muestreo de 100 decisiones de un comité de educación. Hay una gran variedad de formas de hacer esto. Muchas decisiones pueden ser muestreadas de pocos comités, o muchas decisiones pueden ser muestreadas de muchos comités, o ambos. Si el investigador desea asegurarse de la independencia de las decisiones, entonces la mayoría de las decisiones deberán ser muestreadas de muchos comités de educación. Teóricamente, sólo una decisión debería ser tomada de cada comité. Esto nos da alguna seguridad de independencia —al menos tanta como sea posible—. Tan pronto como más de una decisión es tomada de un mismo comité, el inves-

tigador debe considerar el hecho de que las decisiones de clase A pueden influir en las decisiones de clase B. La decisión A puede influir en la decisión B, por ejemplo, porque los miembros del comité deseen parecer consistentes. Ambas decisiones pueden involucrar gastos de equipo instruccional, y si el comité adopta una política liberal respecto de A, entonces deberá adoptar una política similar con B.

Suponga que un investigador ha calculado la probabilidad de la diferencia entre dos medias, en donde dicha diferencia fue debida al azar. La probabilidad fue de 5/100, o 0.05, lo que indica que hubo aproximadamente 5 oportunidades en 100 de que el resultado obtenido fuera debido al azar, es decir, que si la condición experimental es repetida 100 veces sin manipulación experimental, aproximadamente 5 de estas 100 veces podría dar una diferencia de medios tan grande como la obtenida con la manipulación experimental. Sintiéndose dudoso acerca del resultado —después de todo existen 5 oportunidades en 100 de que el resultado pueda ser debido al azar— el investigador repitió cuidadosamente el experimento completo, y se obtuvo el mismo resultado (¡suerte!). Habiendo controlado todo cuidadosamente para asegurarse de que los dos experimentos fueran independientes, la probabilidad calculada para los dos resultados fue debida al azar. Esta probabilidad fue aproximadamente de 0.02. Así se pueden observar tanto los valores de independencia en el experimento como la importancia de la replicación de los resultados.[1]

Hay que hacer notar, finalmente, que la fórmula para la independencia trabaja en dos sentidos. 1) indica la probabilidad de que ambos eventos ocurran al azar, si los eventos son independientes y se conocen las probabilidades de los eventos por separado. Si se encontrara que los dados repetidamente muestran, digamos 12, entonces probablemente algo ande mal con los dados. Si un jugador observa que otro jugador parece ganar siempre, el jugador que va perdiendo, por supuesto, tendrá sospechas. Las posibilidades de ganar continuamente un juego limpio son pocas; puede suceder, por supuesto, pero es muy poco probable que así sea. En investigación es poco probable que se obtengan dos o tres resultados significativos por azar; en ese caso es probable que algo más allá del azar esté operando (la variable independiente, se espera). 2) La fórmula para independencia puede invertirse, por así decirlo; puede decírsele a los investigadores qué hacer para sacar ventaja de las probabilidades multiplicativas. El investigador debe, si esto es posible, planear la investigación para que los eventos sean independientes, aunque esto es más fácil decirlo que hacerlo, lo cual se hará evidente antes de finalizar este libro.

Probabilidad condicional

En toda investigación —y quizá especialmente en la investigación científica social y educacional— los eventos a menudo no son independientes. Se puede ver la independencia de otra manera. Cuando dos variables están relacionadas, éstas no son independientes. La discusión previa sobre conjuntos aclara: si $A \cap B = E$, entonces no hay relación (específicamente una relación cero), o A y B son independientes. Si $A \cap B \neq E$, entonces hay una relación, o A y B no son independientes. Cuando los eventos no son independientes,

[1] El método para calcular estas probabilidades combinadas fue propuesto por Fisher y se describe en Mosteller y Bush (1954). El estudiante perspicaz puede preguntarse por qué el principio de conjuntos aplicado a la probabilidad, $p(A \cap B) = p(A) \cdot p(B)$, no es aplicable, es decir, ¿por qué no calcular $.05 \times .05 = 0.0025$? Mosteller y Bush explican este punto, pero dado que éste es punto difícil y discutible no se incluye en el presente texto. Todo lo que el lector necesita hacer es recordar que la probabilidad de obtener una diferencia sustancial entre las medias en la misma dirección en experimentos repetidos es considerablemente más pequeña que tener tal diferencia una sola vez. Así, uno puede estar más seguro de los datos y conclusiones propias que los de otros datos que resulten iguales.

los científicos pueden afinar sus inferencias probabilísticas. El significado de esta afirmación puede explicarse, hasta cierto punto, al estudiar la probabilidad condicional.

Cuando los eventos no son independientes, el enfoque de la probabilidad debe ser modificado. Un ejemplo simple es el siguiente: ¿cuál es la probabilidad de que de cualquier pareja casada, tomada al azar, ambos miembros sean republicanos? Primero, suponiendo que existe equiprobabilidad y que todo lo demás es igual, el espacio muestral U (todas las posibilidades) es $\{RR, RD, DR, DD\}$, donde la esposa aparece primero en cada posibilidad o punto muestral; de este modo la probabilidad de que ambos, esposo y esposa sean republicanos es $p\{RR\} = 1/4$. Pero si ya sabe que uno de ellos es republicano, ¿cuál es la probabilidad de que ambos sean republicanos ahora? U se reduce a $\{RR, RD, DR\}$. El conocimiento de que uno de ellos es republicano elimina la posibilidad DD, y de esta forma se reduce el espacio muestral, por lo tanto, $p(RR) = 1/3$. En caso de tener, además, la información de que la esposa es republicana, ¿cuál es la probabilidad de que ambos esposos sean republicanos? Ahora $U = \{RR, RD\}$, y por lo tanto $p(RR) = 1/2$. Las nuevas probabilidades son, en este caso "condicionadas" por un conocimiento previo de los hechos.

Definición de probabilidad condicional

Suponga que A y B son los eventos en el espacio muestral, U, como lo hemos estado haciendo. La probabilidad condicional se expresa: $p(A \mid B)$, que se lee: "la probabilidad de A, dado B". Por ejemplo, se podría decir: "La probabilidad de que un esposo y su esposa sean, ambos, republicanos, siendo que el esposo es republicano," o, mucho más difícil de contestar, aunque más interesante: "la probabilidad de una labor docente universitaria altamente efectiva, dado el grado académico de doctor". Por supuesto, puede escribirse también $p(B \mid A)$. La fórmula para la probabilidad condicional cuando se involucran dos eventos es:

$$p(A \mid B) = \frac{p(A \cap B)}{p(B)} \tag{7.8}$$

La fórmula toma una noción anterior de probabilidad y la altera por las situaciones de probabilidad condicional. (Note por favor que la teoría de la probabilidad condicional se extiende a más de dos eventos, pero esto no se analizará en este libro.) Es importante recordar que en problemas de probabilidad el denominador debe ser el espacio muestral. La fórmula anterior cambia el denominador de la razón y por lo tanto cambia el espacio muestral, el cual ha sido reducido de U a B. Para demostrar este punto tomemos dos ejemplos: uno de independencia o probabilidad simple y otro de dependencia o probabilidad condicional.

Si lanzamos al aire una moneda dos veces, los eventos son independientes. ¿Cuál es la probabilidad de obtener cara en el segundo lanzamiento, si la cara apareció en el primero?

TABLA 7.3 *Matriz de probabilidad que muestra las probabilidades conjuntas de dos eventos independientes.*

		Segundo lanzamiento		
		Cara_2	Cruz_2	
Primer lanzamiento	Cara_1	1/4	1/4	1/2
	Cruz_1	1/4	1/4	1/2
		1/2	1/2	

TABLA 7.4 *Matriz de probabilidades conjuntas de eventos.*

		Segundo lanzamiento		
		$Cara_2$	$Cruz_2$	
Primer lanzamiento	$Cara_1$	.30	.20	.50
	$Cruz_1$	.30	.20	.50
		.60	.40	1.00

Esto ya se sabe: 1/2. Si se calcula la probabilidad usando la ecuación 7.8, primero se hace una matriz de probabilidad (véase la tabla 7.3). Para las probabilidades de cara (C) y cruz (X) en el primer lanzamiento, se leen las entradas marginales en el lado derecho de la matriz; igualmente para las probabilidades del segundo lanzamiento, que están en la parte inferior de la matriz. De este modo $p(C_1) = 1/2$, $p(C_2) = 1/2$, y $p(C_1 \cap C_2) = 1/4$. Por lo tanto:

$$p(C_2 \mid C_1) = \frac{p(C_2 \cap C_1)}{p(C_1)} = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2}$$

Los resultados concuerdan con el razonamiento simple previo. Sin embargo, si se hace el problema un poco más complejo, quizás la fórmula sea más útil. Suponga que la probabilidad de obtener cara en el segundo lanzamiento fuera .60 en lugar de .50, y que los eventos continúan siendo independientes. ¿Cambia esto la situación? Esta nueva situación se ilustra en la tabla 7.4 (El .30 en la celda $C_1 \cap C_2$ se calcula con las probabilidades en los márgenes: .50 × .60 = .30. Esto es permitido dado que sabemos que los eventos son independientes. Si no fueran independientes, los problemas de probabilidad condicional no podrían resolverse sin el conocimiento de por lo menos uno de los valores.) La fórmula nos da:

$$p(C_2 \mid C_1) = \frac{p(C_2 \cap C_1)}{p(C_1)} = \frac{.30}{.50} = .60$$

Pero este .60 es el mismo que la probabilidad simple de $Cara_2$. Cuando los eventos son independientes, se obtienen los mismos resultados. Esto es, en este caso: $p(C_2 \mid C_1) = p(C_2)$ y en el caso general:

$$p(A \mid B) = p(A) \tag{7.9}$$

Se obtiene otra definición o condición de independencia. Si se mantiene la ecuación 7.9, los eventos son independientes.

Un ejemplo académico

Hay más ejemplos interesantes de probabilidad condicional que los de monedas y otros mecanismos del azar. Tomemos el problema desconcertante y frustrante de predecir el éxito de los estudiantes de doctorado en una universidad. ¿Podría ser usado el modelo de moneda y dados en una situación tan compleja? Sí, bajo ciertas condiciones. Desafortunadamente estas condiciones son difíciles de ajustar, aunque ha habido éxito limitado. En caso de tener cierta información empírica, el modelo puede ser muy útil. Suponga que los administradores de una universidad están interesados en predecir el éxito de sus estudiantes de doctorado, ya que se encuentran preocupados por el pobre rendimiento de

TABLA 7.5 *Probabilidades conjuntas, problema de los graduados universitarios.*

		Éxito (E)	Fracaso (F)	
MAT	≥65	.20	.10	.30
MAT	<65	.20	.50	.70
		.40	.60	1.00

muchos de ellos y desean establecer un sistema de selección. La universidad continúa admitiendo a todos los aspirantes al doctorado, como en el pasado, pero por tres años todos los estudiantes aceptados presentaron el Miller Analogies Test (MAT), un examen que ha probado ser útil al predecir el éxito en la universidad en muchas áreas (por ejemplo, psicología, educación, economía). Esta prueba también ha sido usada para evaluar personal para puestos de alto nivel en la industria. Se selecciona un punto de corte arbitrario en el puntaje bruto de 65.

La administración universitaria encuentra que el 30% de todos los candidatos del periodo de tres años tuvo una puntuación de 65 o más. Cada uno se categoriza como éxito (E) o fracaso (F). El criterio es simple: ¿Obtuvo el grado el estudiante? Si lo obtuvo, entonces es definido como éxito. Se encuentra que 40% del número total fueron exitosos. Para determinar la relación entre la puntuación MAT y el éxito o fracaso, la administración, empleando nuevamente el punto de corte de 65, determina las proporciones mostradas en la tabla 7.5. El MAT divide al grupo exitoso en dos (.20 y .20) pero diferencia marcadamente en el grupo de fracaso (.10 y .50). Ahora las preguntas que se hacen son: ¿cuál es la probabilidad de obtener el grado de doctorado si el candidato recibe una puntuación de 65 o mayor en el MAT? ¿Cuál es la probabilidad de que un candidato obtenga el grado si la puntuación MAT es menor de 65? Los cálculos son

$$p(E \mid \geq 65) = \frac{p(E \cap \geq 65)}{p(\geq 65)} = \frac{.20}{.30} = .67$$

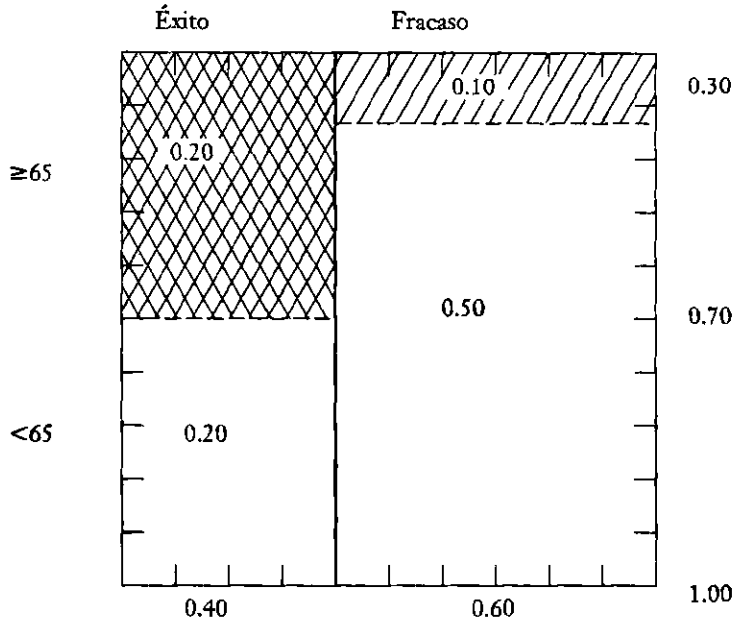
$$p(E \mid < 65) = \frac{p(E \cap < 65)}{p(< 65)} = \frac{.20}{.70} = .29$$

Claramente puede verse que el MAT es un buen predictor de éxito en el programa.

Note cuidadosamente lo que sucede en todos estos casos: cuando escribimos $p(A \mid B)$ en lugar de la sola $p(A)$ en efecto reducimos el espacio muestral de U a B. Si se toma el ejemplo anterior, la probabilidad de éxito sin ningún otro conocimiento es un problema de probabilidad en todo el espacio muestral U. Esta probabilidad es de .40, pero conociendo la puntuación MAT, el espacio muestral se reduce de U a un subconjunto $U \geq 65$. El número real de ocurrencias de un evento exitoso, por supuesto, no cambia; el mismo número de personas tiene éxito, pero la fracción de probabilidad tiene un nuevo denominador; en otras palabras, el estimado de la probabilidad es refinado por el conocimiento de subconjuntos "pertinentes" de U. En este caso, ≥ 65 y < 65 son subconjuntos "pertinentes" de U. Al decir subconjuntos "pertinentes" se quiere decir que la variable implicada está relacionada con la variable criterio: éxito y fracaso.

La siguiente forma de ver el problema puede ayudar. Una interpretación por áreas, del problema de los estudiantes graduados, ha sido representada en la figura 7.5. Aquí se emplea la idea de una medida de un conjunto. Recuerde que una medida de un conjunto o subconjunto es la suma de los pesos de dicho conjunto o subconjunto, y que tales pesos

FIGURA 7.5



son asignados a los elementos del conjunto o subconjunto. La figura 7.5 es un cuadrado con 10 partes iguales en cada lado y cada parte es igual a 1/10 o .10. El área de todo el cuadrado es el espacio muestral *U*, y la medida de *U*, *m* (*U*), es igual a 1.00. Esto significa que todos los pesos asignados a todos los elementos del cuadrado suman en total 1.00. Las medidas de los subconjuntos han sido insertadas: $m(F) = .60$, $m(<65) = .70$, $m(S \cap \geq 65) = 0.20$. Las medidas de estos subconjuntos pueden ser calculadas multiplicando las longitudes de sus lados. Por ejemplo, el área de la caja superior izquierda (con doble trazado diagonal) es de $.5 \times .4 = .20$. Recuerde que la probabilidad de cualquier conjunto o subconjunto es la medida de dicho conjunto (o subconjunto). Así que la probabilidad de cualquiera de las cajas en la figura 7.5 es la que se indica en cada una da ellas. Se puede calcular la probabilidad de cualesquiera de las dos cajas sumando las medidas de los conjuntos; por ejemplo, la probabilidad de éxito es .20 + .20 = .40.

Estas medidas (o probabilidades) están definidas en el área total, o $U = 1.00$. La probabilidad de éxito es igual a .40/1.00. Como se conoce el rendimiento de los estudiantes en el MAT, las áreas que indican las probabilidades asociadas con ≥65 y <65 están delimitadas por los guiones horizontales. La probabilidad simple de ≥65 es igual a .20 + .10 = .30, o .30/1.00. Toda el área sombreada de la parte superior indica esta probabilidad, mientras que las áreas de las medidas de éxito y fracaso están indicadas por las líneas gruesas que las separan en el cuadrado.

El problema de probabilidad condicional es el siguiente: ¿Cuál es la probabilidad de éxito, dado el conocimiento de la puntuación del MAT, o dado ≥65 (también podría ser <65 por supuesto)? Se tiene un pequeño nuevo espacio muestral, indicado por toda el área sombreada en la parte superior del cuadrado; *U* ha sido reducido a este espacio más pequeño porque se conoce la "verdad" del pequeño espacio. En lugar de dejar que este

pequeño espacio sea igual a .30, ahora supone que es igual a 1.00. (Podría parecer que tenemos un nuevo *U*). En consecuencia, las medidas de las cajas que constituyen el nuevo espacio muestral deben ser recalculadas. Por ejemplo, en lugar de calcular la probabilidad de $p(\geq 65 \cap S) = .20$ porque es de 2/10 del área total del cuadrado, se debe calcular la probabilidad con base en el área de ≥65 (el área sombreada en la parte superior del cuadrado), debido a que se sabe que los elementos en el conjunto ≥65 tienen puntajes mayores o iguales a 65. Una vez hecho esto, se obtiene .20/.30 = .67 [el área sombreada para éxito ÷ el área sombreada para éxito + el área sombreada para fracaso], que es exactamente lo que obtuvimos cuando se usó la ecuación 7.8.

Lo que sucede es que el conocimiento adicional hace que *U* ya no sea tan relevante como espacio muestral. *Todos los enunciados de probabilidad son relativos a los espacios muestrales.* La cuestión básica, entonces, es definir adecuadamente los espacios muestrales. En el problema anterior de los esposos y esposas se hizo la pregunta: ¿Cuál es la probabilidad de que ambos, esposo y esposa sean republicanos? El espacio muestral era *U* = {*RR, RD, DR, DD*], pero cuando se agregó el conocimiento de que uno de ellos era republicano y se elaboró la misma pregunta, se hizo al *U* original, irrelevante para el problema y un nuevo espacio muestral llamado *U'*, es requerido. Consecuentemente, la probabilidad de que ambos sean republicanos es diferente cuando tenemos más conocimiento.

Se pueden calcular otras probabilidades de forma similar. Si se deseara conocer la probabilidad de fracaso, dada una puntuación MAT menor a 65, viendo la figura 7.5, la probabilidad que se busca es la caja más grande a la derecha, etiquetada con .50. Dado que se sabe que la puntuación es <65, se utiliza este conocimiento para establecer un nuevo espacio muestral. Las dos cajas inferiores cuya área es igual a .20 + .50 = .70 representan este espacio muestral. De este modo se calcula una nueva probabilidad: .50/.70 = .71; la probabilidad de fracaso para obtener el grado si uno tiene una puntuación MAT menor a 65 es de .71.

Teorema de Bayes: revisión de las probabilidades

Ninguna discusión sobre las probabilidades condicionales podría estar completa sin mencionar brevemente el teorema de Bayes y su utilidad en la investigación de las ciencias conductuales aplicadas. Mediante la manipulación de la fórmula de la probabilidad condicional, el reverendo Thomas Bayes fue capaz de desarrollar una fórmula para calcular probabilidades condicionales especiales. Con el teorema de Bayes, se pueden actualizar o revisar probabilidades en curso, basadas en nueva información o datos. Esta información puede usarse para reestructurar la incertidumbre. Numerosos científicos e investigadores recomiendan el uso del teorema de Bayes (véase Wang, 1993).

Rara vez los datos empíricos son concluyentes. Por ejemplo, algunos estudiantes muy buenos que buscan ser admitidos en la universidad tendrán una pobre ejecución en el examen de admisión. Sin embargo, habrá algunos malos estudiantes que tendrán una buena puntuación en dicho examen. No obstante, un resultado desfavorable en un examen puede incrementar las oportunidades de rechazar a un mal estudiante, y un resultado favorable puede incrementar la probabilidad de seleccionar a un buen estudiante. De la misma forma, en la vida diaria, nosotros constantemente ajustamos viejas creencias, basados en nueva información. El teorema de Bayes, expresado cómo una fórmula numérica, muestra cómo puede hacerse esto.

$$p(C_i \mid A) = \frac{p(C_i)p(A \mid C_i)}{\sum_{j=1}^{k} p(C_j)p(A \mid C_j)}$$

Ejemplo

Basado en los datos que actualmente existen, un investigador ha determinado que el 10% de la población padece un trastorno alimenticio. Tales datos generalmente son publicados y provienen de fuentes establecidas. Lo que esto nos dice es que, sin ninguna información adicional, si 100 personas fueran elegidas al azar, 10 de ellas tendrían un trastorno alimenticio. Si T es usada para designar que una persona tiene un trastorno alimenticio, entonces $\sim T$ se usa para indicar la ausencia de dicho trastorno; por lo tanto, $p(T) = .10$ y $p(\sim T) = .90$. Éstas son llamadas "viejas" probabilidades. Algunos psicólogos se refieren a ellas como tasas base, y usar estas probabilidades por sí mismas puede llevar a resultados infructuosos.

Con la intención de mejorar la habilidad para detectar trastornos alimenticios, se desarrolló una prueba psicológica que, aunque imperfecta, da nueva información y puede ayudar a un terapeuta practicante a hacer un diagnóstico correcto. El investigador escogió individuos que se suponía presentaban este trastorno y también seleccionó un grupo de sujetos que no parecían presentarlo, y les aplicó la prueba. El número de personas con resultado positivo en la prueba, es decir, quienes en realidad tenían el trastorno, pueden representarse como una probabilidad condicional $p(+|T)$. El número de sujetos con resultado negativo en la prueba, quienes en realidad no tenían el trastorno, puede escribirse como $p(-|\sim T)$. Estas dos probabilidades condicionales son llamadas clasificaciones correctas. Empíricamente, $p(+|T) = .91$ y $p(-|\sim T) = .95$. De aquí que, con base en datos conocidos, la prueba es capaz de detectar adecuadamente al 91% de los que padecen el trastorno y al 95% de aquellos que no lo presentan. El número de personas que obtiene un resultado positivo, pero que en realidad no tiene el trastorno se designa como $p(+|T)$ y es llamado un falso positivo; el número de sujetos que recibe un resultado negativo pero que en realidad tiene el trastorno se anota como $p(-|T)$ y representa un falso negativo. Estas dos últimas probabilidades condicionales dan el nivel de imperfección de la prueba, y por lo tanto $p(+|\sim T)$ es .05 y $p(-|T) = .09$. Ahora, usando el teorema de Bayes, se pueden responder las siguientes preguntas: ¿Cuál es la probabilidad de que la persona realmente padezca el trastorno si su resultado fue negativo: $p(T|-)$? ¿Cuál es la probabilidad de que una persona padezca el trastorno y que haya resultado positivo en la prueba: $p(T|+)$? Así que usando las probabilidades condicionales, uno puede actualizar las "viejas" probabilidades. Si se utilizan estos números, la ecuación para el teorema de Bayes resulta:

$$p(T|+) = \frac{p(+|T)p(T)}{p(+|T)p(T) + p(+|\sim T)p(\sim T)} = \frac{.91(.10)}{.91(.10) + (.05)(.90)}$$
$$= \frac{.091}{.091 + .045} = \frac{.091}{.136} = 0.669 = 0.67$$

De manera similar:

$$p(T|-) = \frac{p(-|T)p(T)}{p(-|T)p(T) + p(-|\sim T)p(\sim T)}$$
$$= \frac{.09(.10)}{.09(.10) + (.95)(.90)} = \frac{.009}{.009 + .855} = \frac{.009}{.864} = 0.01$$

Con el uso de la prueba, si una persona obtiene una puntuación positiva, entonces existe un 67% de probabilidad de que padezca un trastorno alimenticio. Basándose en el teorema de Bayes, se ha ajustado la probabilidad de que la persona presente este trastorno,

de .10 a .67. Existe una probabilidad del 1% de que la persona que obtenga una puntuación negativa, tenga un trastorno alimenticio.

Doscher y Bruno (1981) utilizaron el teorema de Bayes en lugar de la fórmula habitual para hacer correcciones respecto de exámenes contestados mediante adivinación. Desarrollaron distribuciones de probabilidad de niveles de verdadero conocimiento, basados en el teorema de Bayes, de tal modo que con una calificación real de un examen, las tablas de calibración desarrolladas a partir del teorema, darían una estimación probabilística de la puntuación verdadera. Aquí, dicho teorema fue usado para ajustar las calificaciones de la prueba dadas por adivinación. Doscher y Bruno encontraron que el método de Bayes era más efectivo que la fórmula usada habitualmente para realizar la corrección. Estos mismos investigadores concluyeron que para niños urbanos el uso de una calificación sin ajustar, generalmente sobrestima el conocimiento del niño, lo cual puede resultar en que se le ubique en una situación de aprendizaje con tareas demasiado difíciles. Encontraron que con el uso de la fórmula habitual de corrección para la adivinación, el ajuste era demasiado grande y generalmente se ubicaba al niño en situaciones de aprendizaje de poco reto para él. Doscher y Bruno (1981, p. 488) dicen:

> Un procedimiento analítico basado en el teorema de Bayes permite la estimación probabilística de las calificaciones verdaderas de una prueba, usando información previa acerca de la distribución probable de calificaciones verdaderas y el patrón de adivinación específico de la población estudiada.

De manera similar al estudio de Doscher y Bruno, Jones (1991) introdujo el uso del teorema de Bayes en conjunción con las decisiones del consejero. Jones recomienda el análisis bayesiano, ya que una afirmación probabilística se hace sobre la persona que está siendo evaluada en lugar de sólo dar una puntuación y una interpretación. La investigación de Jones, basada en el teorema de Bayes, se enfocó en la selección de operadores, para emplearlos en un programa de rehabilitación para débiles visuales. Jones estableció los pasos que un consejero debía seguir en el programa, para usar el teorema de Bayes. El consejero iniciaría con un examen de los registros de la agencia y localizaría los datos psicométricos de los candidatos contratados como operadores. También determinaría cuáles serían eventualmente clasificados como exitosos y no exitosos. Jones etiquetó éstas como "creencias previas". Los registros también daban al consejero las puntuaciones de la Prueba de Destrezas Cognitivas (Cognitive Skills Test), y a partir de lo anterior, el consejero determinaría cuántos de los aspirantes exitosos tenían una puntuación de experto o superior y cuántos de los aspirantes exitosos tenían una puntuación inferior. Datos similares se obtendrían respecto de aquellos aspirantes que no fueran exitosos. Armado con esta información, el consejero podía ahora dar un estimado de probabilidad (usando el teorema de Bayes) del éxito de una persona a partir de que tuviera o no una puntuación de experto en la prueba. Jones continúa mostrando cómo integrar perfiles de personalidad en el proceso de clasificación.

El marco bayesiano abarca muchas áreas de investigación. Muchos de los métodos estadísticos más avanzados, tales como el análisis factorial confirmatorio, los modelos estructurales y el análisis discriminante, se basan en el enfoque bayesiano. El lector debería leer al menos uno de los siguientes artículos para adquirir más información. Estes (1991) proporciona algunos detalles adicionales, no mencionados aquí, sobre el uso del teorema de Bayes en casos criminales. Smith, Penrod, Otto y Park (1996) condujeron un experimento para evaluar la conducta de los jurados que aprendieron el uso del teorema de Bayes y la evidencia probabilística en casos legales de crímenes. Wang (1993) muestra la superioridad del teorema de Bayes sobre otros métodos en el pronóstico de negocios. Bierman, Bonini y Hausman (1991) dan detalles sobre el uso del teorema de Bayes en mercadotecnia e investigación de negocios.

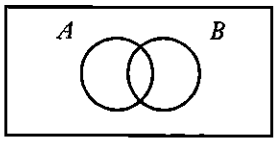
RESUMEN DEL CAPÍTULO

1. La probabilidad no es fácil de definir.
2. Hay tres definiciones amplias:
   a) La ***a priori***, implica la capacidad de la gente para dar un estimado de probabilidad en ausencia de datos empíricos.
   b) La *a posteriori* es la definición más común usada en estadística, basada en frecuencias relativas a largo plazo.
   c) El peso de la evidencia de Keynes implica dos números de probabilidad: uno es subjetivo y está basado en la cantidad de información disponible.
3. El espacio muestral es el número total de posibles resultados de un experimento. Cualquier resultado aislado es llamado punto muestral o elemento. Un evento es uno o más elementos dispuestos de tal forma que toman un significado.
4. La probabilidad de un espacio muestral es de 1.00. Los puntos muestrales y eventos son menores de 1.00. La suma de todas las probabilidades de los elementos es 1.00. A cada punto muestral se le puede asignar un peso. La suma total de los pesos usados debe ser 1.00.
5. La probabilidad tiene tres propiedades fundamentales:
   a) La medida de cualquier conjunto, como se definió antes, es mayor o igual a 0 (cero) y menor o igual a 1. En pocas palabras, las probabilidades (medidas de los conjuntos) son 0, 1, o una cifra entre ellos.
   b) La medida de un conjunto, $m(A)$ es igual a 0 si y solamente si no hay miembros en A; esto es, que A esté vacío.
   c) Suponga que A y B son conjuntos. Si A y B no están unidos —esto es $A \cap B = E$— entonces $m(A \cup B) = m(A) + m(B)$.
6. Un evento compuesto es la co-ocurrencia de dos o más eventos aislados (o compuestos).
7. La exhaustividad se refiere a la partición de un espacio muestral en subconjuntos. Cuando estos subconjuntos se combinan, cubren todo el espacio muestral.
8. Si la intersección de dos o más conjuntos resulta en un conjunto vacío, estos conjuntos son llamados mutuamente excluyentes.
9. Dos eventos se consideran independientes si la probabilidad de que ambos eventos ocurran es igual a la probabilidad de un evento multiplicado por la probabilidad del otro.
10. A veces se desea calcular la probabilidad de un evento después de recibir información adicional que puede alterar el espacio muestral. Esta probabilidad se llama probabilidad condicional.
11. El teorema de Bayes implica probabilidad condicional. Las probabilidades de Bayes son efectivas para revisar probabilidades usando información adicional o nueva.

SUGERENCIAS DE ESTUDIO

1. Suponga que usted está seleccionando una muestra de jóvenes de noveno grado para realizar una investigación. Hay 250 estudiantes de noveno grado en el sistema escolar, 130 niños y 120 niñas.
   a) ¿Cuál es la probabilidad de seleccionar a cualquier joven?
   b) ¿Cuál es la probabilidad de seleccionar a una niña? ¿y a un niño?

FIGURA 7.6



c) ¿Cuál es la probabilidad de seleccionar a un niño o una niña? ¿Cómo escribiría este problema en lenguaje de conjuntos? [Pista: ¿Es equivalente a una intersección o unión de conjuntos?]

d) Suponga que usted selecciona una muestra de 100 niños y niñas. Usted tiene 90 niños y 10 niñas. ¿Qué conclusiones podría sacar?

[Respuestas: *a)* 1/250; *b)* 120/250, 130/250; *c)* 1.]

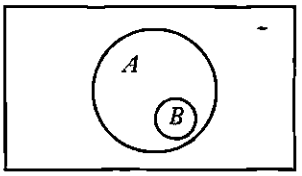
2. Lance una moneda y un dado una vez. ¿Cuál es la probabilidad de obtener cara en la moneda y 6 en el dado? Dibuje un árbol que muestre todas las probabilidades. Etiquete las ramas del árbol con los pesos o probabilidades apropiados. Ahora conteste algunas preguntas. ¿Cuál es la probabilidad de tener:
   a) cruz y un 1, un 3 o un 6?
   b) cara y un 2 o un 4?
   c) cara o cruz y un 5?
   d) cara y cruz y un 5 o un 6?

   [Respuestas: *a)* 1/4; *b)* 1/6; *c)* 1/6; *d)* 1/3.]
3. Lance una moneda y ruede un dado 72 veces. Escriba los resultados, uno al lado del otro, en una hoja cuadriculada, tal como ocurran. Verifique los resultados obtenidos contra las frecuencias esperadas teóricamente. Ahora compare sus respuestas con cada una de las respuestas de la pregunta 2. ¿Se acercan los resultados obtenidos a los resultados esperados? (Por ejemplo, suponga que usted calculó cierta probabilidad para la pregunta 2 *a)*. Ahora cuente el número de veces que la cruz se apareó con un 1, un 3, o un 6 ¿Es igual la fracción obtenida a la fracción esperada?)
4. Suponga que en la figura 7.6 hay 20 elementos en U, cuatro de los cuales están en A, seis en B y dos en $A \cap B$. Si usted selecciona aleatoriamente un elemento, ¿cuál es la probabilidad de que éste:
   a) sea de A?
   b) sea de B?
   c) sea de $A \cap B$?
   d) sea de A o de B? [Pista: recuerde la ecuación: $p(A \cup B) = p(A) + p(B) - p(A \cap B)$.]
   e) no sea ni de A ni de B?
   f) sea de B, pero no de A?

   [Respuestas: *a)* 1/5; *b)* 3/10; *c)* 1/10; *d)* 2/5; *e)* 3/5; *f)* 1/5.]
5. Observe la figura 7.6 y conteste las siguientes preguntas:
   a) Dado B, ¿cuál es la probabilidad de A?
   b) Dado A, ¿cuál es la probabilidad de B?

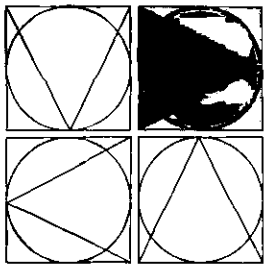
   [Respuestas: *a)* 1/3; *b)* 1/2.]
6. Considere la figura 7.7. Hay 20 elementos en U, 4 en B, y 8 en A. Si un elemento de U es seleccionado al azar, ¿cuál es la probabilidad de que el elemento sea de:
   a) A?
   b) B?
   c) $A \cap B$?
   d) $A \cup B$?
   e) U?

FIGURA 7.7



[Respuestas: *a*) 2/5; *b*) 1/5; *c*) 1/5; *d*) 2/5; *e*) 1.]

7. Con base en la figura 7.7, conteste las siguientes preguntas:
   a) Dado A (sabiendo que un elemento de la muestra vino de A), ¿cuál es la probabilidad de B?
   b) Dado B, ¿Cuál es la probabilidad de A?
   [Respuestas: *a*) 1/2; *b*) 1.]
8. Suponga que usted tuvo un examen de opción múltiple con dos reactivos de cuatro opciones, con las cuatro opciones de cada reactivo enumeradas *a*, *b*, *c* y *d*. Las respuestas correctas a los dos reactivos son *c* y *a*.
   a) Describa el espacio muestral. (Dibuje un diagrama de árbol; véase la figura 7.3.)
   b) ¿Cuál es la probabilidad de que cualquier examinado tenga ambos reactivos correctos por adivinación?
   c) ¿Cuál es la probabilidad de que responda correctamente al menos uno de los reactivos, por adivinación?
   [Pista: Esto puede ser un poco problemático. Dibuje un diagrama de árbol y piense en las probabilidades. Cuéntelas.]
   d) ¿Cuál es la probabilidad de contestar de forma incorrecta ambos reactivos, por adivinación?
   e) Dado que un examinado responde correctamente el primer reactivo, ¿cuál es la probabilidad de que esa persona responda correctamente el segundo reactivo, por adivinación?
   [Respuestas: *b*) 1/16; *c*) 7/16; *d*) 9/16; *e*) 1/4.]
9. La mayoría de la discusión en el texto ha sido basada en el supuesto de equiprobabilidad. Sin embargo, a menudo este supuesto no está justificado. Por ejemplo, ¿cuál es el error en el siguiente argumento? La probabilidad de que uno muera mañana es de un medio. ¿Por qué? Porque uno puede morir mañana o no morir mañana. Dado que hay dos posibilidades, cada una tiene una probabilidad de ocurrencia de un medio. ¿Cómo manejan las compañías de seguros este razonamiento? Suponga ahora, que un investigador de ciencias políticas estudia la relación entre las preferencias religiosas y políticas y asume que las probabilidades de que un católico sea republicano o demócrata fueran iguales. ¿Qué pensaría usted de sus resultados? ¿Tienen estos ejemplos implicaciones para los investigadores que conocen algo sobre los fenómenos que están estudiando? Explique.



CAPÍTULO 8

MUESTREO Y ALEATORIEDAD

Existen muchas situaciones en las que se desea saber algo acerca de las personas, de los eventos, y de las cosas. Para aprender algo acerca de la gente, por ejemplo, se selecciona gente que conocemos —o que no conocemos— y se estudia. Después del "estudio", se obtienen ciertas conclusiones, frecuentemente acerca de la gente en general. Parte de este método se basa en la sabiduría popular. Las observaciones basadas en el sentido común acerca de la gente, sus motivaciones y su comportamiento derivan, la mayoría de las veces, de observaciones y experiencias con pocas personas; se afirman hechos tales como: "la gente hoy en día no tiene valores ni sentido moral"; "los políticos son corruptos" y "los alumnos de escuelas públicas no aprenden lo suficiente". El fundamento para tales afirmaciones es: la gente saca conclusiones acerca de otras personas y acerca de su entorno, basadas principalmente en experiencias limitadas. Para llegar a tales conclusiones, las personas deben *muestrear* sus "experiencias" acerca de la gente; en realidad toman muestras relativamente pequeñas de todas las experiencias posibles. La palabra *experiencias* tiene que ser entendida en un sentido amplio, puede significar experiencia directa con otras personas, por ejemplo, una interacción de primera mano con, digamos, musulmanes o asiáticos o puede significar una experiencia indirecta, como haber escuchado hablar de musulmanes o asiáticos a los amigos o conocidos. Sin embargo, que la experiencia sea directa o indirecta, no importa mucho por ahora, ya que se asume que toda esa experiencia es directa. Un individuo afirma conocer algo acerca de los asiáticos y dice "yo sé que ellos son gregarios porque he tenido experiencia directa con muchos asiáticos" o "algunos de

mis mejores amigos son asiáticos, y yo sé que ...". El punto es que las conclusiones de esta persona están basadas en una muestra de asiáticos, o en una muestra de las conductas de los asiáticos, en o ambas. Este individuo nunca podrá "conocer" a todos los asiáticos y en su análisis debe depender de las muestras. De hecho, la mayoría del conocimiento sobre el mundo está basado en muestras, que la mayoría de las veces son inadecuadas.

Muestreo, muestreo aleatorio y representatividad

Muestrear significa tomar una porción de una población o de un universo como representativa de esa población o universo. Esta definición no dice que la muestra tomada —o extraída, como algunos investigadores dicen— sea representativa, más bien que se toma una porción de la población y ésta se *considera* representativa. Cuando un administrador escolar visita ciertos salones de clases del sistema escolar para evaluarlo, ese administrador está muestreando clases de todo el sistema escolar. Esta persona puede suponer que al visitar, digamos, de 8 a 10 salones de clases "al azar" de un total de 40, él obtendrá una clara idea de la calidad de la enseñanza que se está dando en el sistema. Otra forma podría ser visitar 2 o 3 veces la clase de un maestro para muestrear su desempeño al enseñar, con lo cual el administrador está muestreando conductas, en este caso, conductas de enseñanza, a partir del universo de todas las posibles conductas del maestro. Esta forma de muestreo es necesaria y legítima; sin embargo, pueden surgir situaciones donde el universo entero puede ser medido, entonces ¿para qué molestarse en hacer muestreos? ¿Por qué no medir cada uno de los elementos del universo? ¿Por qué no hacer un censo? Una de las principales razones es la económica. El segundo autor (HBL) trabajó en el departamento de investigación de mercado de una gran cadena de abarrotes en el sur de California, la cual constaba de 100 tiendas. La investigación de clientes y productos ocasionalmente se llevaba a cabo mediante una prueba controlada de tienda. Estos estudios fueron conducidos en una operación real diaria de una tienda de abarrotes. Quizás el interés era probar un nuevo alimento para perros, así ciertas tiendas serían elegidas para recibir el nuevo producto mientras que otro conjunto de tiendas no lo incluirían. La confidencialidad es muy importante en este tipo de estudios, ya que si un fabricante de alimento para perros de la competencia obtuviera información de que se está practicando una prueba de mercado en tal tienda, podrían contaminarse los resultados. Para llevar a cabo en una tienda pruebas controladas sobre cupones de descuento, nuevos productos o colocación en los anaqueles, podría efectuarse una investigación con dos grupos de 50 tiendas cada uno; sin embargo, el trabajo y los costos administrativos serían prohibitivos. Tendría más sentido usar muestras representativas de dicha población. Elegir 10 tiendas para cada grupo reduciría los costos de la realización del estudio. Estudios más pequeños son más manejables y controlables. Un estudio que utilice muestras puede ser realizado en un tiempo predecible. En algunas disciplinas, tales como el control de calidad y la educación (evaluación instruccional), el muestreo es esencial. En el control de calidad existe un procedimiento llamado prueba destructiva. Una manera de determinar si un producto cumple con sus especificaciones es someterlo a una prueba de rendimiento real. Cuando el producto se destruye (falla), éste puede ser evaluado. Por ejemplo, al probar neumáticos, no tendría sentido destruir cada uno de ellos para determinar si el fabricante ha cumplido con un adecuado control de calidad. De la misma manera, un maestro que quiera determinar si un niño ha aprendido el material, le aplicará un examen; sería difícil elaborar un examen que cubriera cada aspecto de lo enseñado y la retención del conocimiento del niño.

El muestreo aleatorio es el método de obtener una porción (o muestra) de una población o universo, de tal manera que cada miembro de esa población o universo tenga la

misma posibilidad de ser seleccionado. Esta definición tiene la virtud de comprenderse con facilidad; desafortunadamente, no es del todo satisfactoria debido a que es limitada. Una mejor definición es la de Kirk (1990, p. 8):

> El método de extracción de muestras a partir de una población, de manera que toda muestra posible de un tamaño particular tiene la misma posibilidad de ser seleccionada, se llama *muestreo aleatorio* y las muestras resultantes son *muestras aleatorias*.

Esta definición es general y, por lo tanto, más satisfactoria que la definición previa.

Defina un universo de estudio de todos los niños de 4° grado en cualquier sistema escolar. Suponga que son 200 niños. Ellos comprenden la población (o universo). Se selecciona un niño al azar; su posibilidad de ser seleccionado es 1/200, si el procedimiento de muestreo es aleatorio. De la misma manera se seleccionan otros niños. Suponiendo que después de seleccionar un niño, ese niño (o un símbolo asignado al niño) es regresado a la población, entonces la posibilidad de seleccionar cualquier segundo niño es también de 1/200; si no se regresa este niño a la población, entonces la posibilidad para cada uno de los niños restantes es, por supuesto, de 1/199. Esto es llamado *muestreo sin reemplazamiento*. Cuando los elementos de la muestra son regresados a la población después de haber sido elegidos, el procedimiento se llama ***muestreo con reemplazamiento.***

Suponga que de la población de los 200 niños de 4° grado en cualquier sistema escolar, se obtiene una muestra aleatoria de 50 niños. Si la muestra es aleatoria, todas las muestras posibles de 50 niños tienen la misma probabilidad de ser seleccionadas —un número muy grande de muestras posibles—. Para hacer esta idea comprensible, suponga una población consistente en 4 niños *a*, *b*, *c* y *d* de donde se extrae una muestra de 2 niños al azar; entonces la lista de todas las posibilidades o ***el espacio muestral*** es: (*a*, *b*), (*a*, *c*), (*a*, *d*), (*b*, *c*) (*b*, *d*), (*c*, *d*). Existen 6 posibilidades. Si la muestra de 2 es seleccionada al azar, entonces su probabilidad es de 1/6, es decir, que cada uno de los pares tienen la misma probabilidad de ser elegido. Este tipo de razonamiento es necesario para resolver muchos problemas de investigación, pero generalmente tiende a limitarse a la idea más simple de muestreo asociada con la primera definición. La primera definición, entonces, es un caso particular de la segunda definición general —el caso especial en el que $n = 1$—.

Desafortunadamente, nunca se puede estar seguro de que una muestra aleatoria sea representativa de la población de la cual fue seleccionada. Recuerde que cualquier muestra particular de tamaño n tiene la misma probabilidad de ser seleccionada que cualquier otra muestra del mismo tamaño, por lo que una muestra particular puede no ser representativa en absoluto. Pero, ¿qué significa "representativa"? Por lo general, *representativo* significa que es típico de una población, es decir, que ejemplifica las características de la población. Desde el punto de vista de la investigación, *representativo* debe ser definido con mayor precisión, aunque con frecuencia es difícil ser preciso. Es necesario preguntar ¿de qué características se está hablando? Por lo tanto, en investigación, una *muestra representativa* es aquella muestra que tiene aproximadamente las mismas características de la población, relevantes a la investigación en cuestión. Si sexo y clase socioeconómica son variables (características) relevantes a la investigación, una muestra representativa tendrá aproximadamente la misma proporción de hombres y mujeres, y de individuos de clase media y clase trabajadora que la población. Cuando se selecciona una muestra al azar, se espera que sea representativa; se espera que las características relevantes de la población estén presentes en la muestra en, aproximadamente, la misma forma en que están presentes en la población, pero nunca se puede estar seguro. No hay garantía.

Como señala Stilson (1966), uno se basa en el hecho de que las características típicas de una población son aquellas más frecuentes y, por lo tanto, las que tienen mayor probabilidad de estar presentes en cualquier muestra aleatoria. Cuando el muestreo es al azar, la

variabilidad muestral es predecible. En el capítulo 7 se aprendió que, por ejemplo, si se lanzan dos dados un cierto número de veces, la probabilidad de que salga un 7 es mayor que la de que salga un 12 (véase tabla 7.1).

Una muestra extraída al azar no está sesgada, en el sentido de que ningún miembro de la población tiene más posibilidad de ser seleccionado que cualquier otro miembro; es democrática ya que todos los miembros son iguales al momento de la selección. En un ejemplo de investigación, diferente al de monedas y dados, suponga que tenemos una población de 100 niños. Los niños son diferentes en cuanto a su inteligencia, una variable relevante para el estudio. Se busca conocer la media de la puntuación de inteligencia de la población, pero por alguna razón, solamente se puede muestrear a 30 de los 100 niños. Si se muestrea aleatoriamente, hay un gran número de posibles muestras de 30 alumnos y las muestras tienen las mismas posibilidades de ser seleccionadas. Las medias de la mayoría de las muestras estarán relativamente cerca de la media de la población y algunas pocas no lo estarán. Si el muestreo fue aleatorio, la probabilidad de seleccionar una muestra con una media cercana a la media poblacional es mayor que la probabilidad de seleccionar una muestra con una media alejada de la media poblacional.

Si la muestra no es seleccionada al azar, algún factor o factores desconocidos pueden predisponer a la selección de una muestra sesgada. En este caso, quizás una muestra con una media que no esté cercana a la media poblacional. La inteligencia promedio de esta muestra será, entonces, una estimación sesgada de la media poblacional. Si se conociera a los 100 niños, inconscientemente podría tenderse a seleccionar a los niños más inteligentes: no es tanto que *nosotros lo hiciéramos* así, sino que nuestro método nos *permite* hacerlo así. Los métodos aleatorios de selección impiden que operen los propios sesgos o cualquier otro factor sistemático de selección. El procedimiento es objetivo, ya que es ajeno a las propias predilecciones y sesgos.

El lector puede estar experimentando una sensación vaga e inquietante de intranquilidad. Si no se puede estar seguro de que las muestras aleatorias sean representativas, ¿cómo confiar en los resultados de la investigación y en su aplicabilidad a las poblaciones de donde se obtuvieron las muestras? ¿Por qué no seleccionar las muestras de manera sistemática, para que sean representativas? La respuesta es compleja: primero —y de nuevo— nunca se puede estar seguro: y segundo, es más probable que las muestras aleatorias incluyan las características típicas de la población si las características son frecuentes en dicha población. En investigaciones reales se seleccionan muestras aleatorias siempre que sea posible y se espera y supone que las muestras sean representativas. Uno aprende a vivir con incertidumbre, pero procura reducirla siempre que sea posible, tal como uno lo hace en la vida diaria, pero de manera más sistemática y con suficiente conocimiento y experiencia respecto al muestreo aleatorio y a los resultados del azar. Por fortuna la falta de certeza no impide que la investigación funcione.

Aleatoriedad

El concepto de aleatoriedad está en el centro de los métodos probabilísticos modernos en las ciencias naturales y del comportamiento, pero es difícil definir *aleatorio*. El concepto del diccionario de fortuito, accidental, sin dirección o propósito, no ayuda mucho. De hecho los científicos son muy sistemáticos acerca de la aleatoriedad, ya que seleccionan con cuidado muestras aleatorias y planean procedimientos aleatorios.

Se puede asumir la postura de que nada sucede al azar, que para cualquier evento hay una causa. La única razón, de acuerdo a esta postura, para que uno use la palabra *aleatorio*, es que el ser humano no sabe lo suficiente. Para el sabio nada es aleatorio. Suponga que un

sabio tiene un periódico sabio; dicho periódico es gigantesco, y en él cada evento está cuidadosamente descrito hasta el último detalle —para mañana, para el siguiente día, y así hasta un tiempo indefinido (véase Kemeny, 1959, p. 39)—. Aquí nada es desconocido y, por supuesto, no hay aleatoriedad. Desde este punto de vista, la aleatoriedad es ignorancia.

Basándose en este argumento, la aleatoriedad se define de forma inversa: se dice que los eventos son aleatorios si no se pueden predecir sus resultados. Por ejemplo, no hay forma conocida para ganar un volado con una moneda. Debido a que no hay un sistema para jugar un juego que asegure ganarlo (o perderlo), entonces los eventos (resultados del juego) son aleatorios. Dicho de manera formal, *aleatoriedad* significa que no hay ley conocida capaz de ser expresada en lenguaje, que describa o explique correctamente los eventos y sus resultados. En otras palabras, cuando los eventos son aleatorios no pueden predecirse individualmente. Sin embargo, y por extraño que suene, se puede predecir con éxito en su conjunto; esto es, que se pueden predecir los resultados de un número grande de eventos. No se puede saber si el resultado de lanzar una moneda al aire será cara o cruz, pero si se lanza una moneda nivelada mil veces, se puede predecir con bastante certeza el número total de caras y cruces.

Un ejemplo de muestreo aleatorio

Para dar al lector una idea de aleatoriedad y muestras aleatorias, se utilizará una tabla de números aleatorios. Una tabla de números aleatorios contiene números generados mecánicamente, de tal manera que no tienen ningún orden perceptible o sistemático en ellos. Se dijo antes que si los eventos son aleatorios no pueden ser predichos. Pero ahora se hará una predicción de la *naturaleza general* de los resultados de un experimento. Mediante un cuadro de números aleatorios se seleccionan 10 muestras de 10 números cada una. Dado que los números son aleatorios, cada muestra "deberá" ser representativa del universo de números. El universo puede definirse de varias formas. Aquí se define como un conjunto completo de números en la tabla de la Rand Corporation.[1] Ahora, se extraen muestras de la tabla. Las medias de las 10 muestras serán, por supuesto, diferentes. Sin embargo, deberían fluctuar dentro de un rango relativamente estrecho, con la mayoría de ellas cerca de la media de los 100 números y de la media teórica de toda la población de números aleatorios. La cantidad de números pares en cada muestra de 10, debe ser aproximadamente igual al número de pares. Con toda seguridad habrá fluctuaciones, algunas quizás extremas, pero comparativamente la mayoría serán modestas. Las muestras se presentan en la tabla 8.1.

La medias de las muestras se encuentran debajo de cada muestra. La media de U, que es la media teórica de toda la población de los números aleatorios de Rand, {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}, es 4.5. La media de los 100 números, que puede ser considerada como una muestra de U, es 4.56. Ésta es, por supuesto, muy cercana a la media de U. Puede notarse que las medias de las 10 muestras varían alrededor de 4.5, siendo 2.4 la más baja y 5.7 la más alta. Solamente dos de estas medias difieren por más de 1 del 4.5. Una prueba estadística (posteriormente se aprenderán los fundamentos de tales pruebas) demuestra que las 10 medias no difieren significativamente una de otra. (La expresión "no difieren significa-

[1] La fuente de los números aleatorios fue Rand Corporation (1955). Ésta es una tabla de números aleatorios grande y cuidadosamente construida. Estos números *no* fueron generados por computadora. Existen, sin embargo, muchas otras tablas, que son suficientemente buenas para propósitos prácticos. Muchos textos actuales de estadística tienen tablas como éstas. El apéndice C, al final de este libro, contiene 4 000 números aleatorios generados por computadora.

TABLA 8.1 *Diez muestras de números aleatorios*

	1	2	3	4	5	6	7	8	9	10	
	9	0	8	0	4	6	0	7	7	8	
	7	2	7	4	9	4	7	8	7	7	
	6	2	8	1	9	3	6	0	3	9	
	7	9	9	1	6	4	9	4	7	7	
	3	3	1	1	4	1	0	3	9	4	
	8	9	2	1	3	9	6	7	7	3	
	4	8	3	0	9	2	7	2	3	2	
	1	4	3	0	0	2	6	9	7	5	
	2	1	8	8	4	5	2	1	0	3	
	3	1	4	8	9	2	9	3	0	1	
Media	5.0	3.9	5.3	2.4	5.7	3.8	5.2	4.4	5.0	4.9	Media total = 4.56

tivamente una de otra" quiere decir que las diferencias no son mayores que las que podrían ocurrir debido al azar.) Por medio de otra prueba estadística, nueve son "buenos" estimados de la media poblacional (4.5) y uno (2.4) no lo es.

Cambiando el problema de muestreo, se puede definir un universo consistente en números pares y números impares. Suponga que en el universo entero hay un número igual de ambos. En la muestra de 100 números debería haber aproximadamente 50 números nones y 50 números pares. En realidad hay 54 números nones y 46 números pares. Una prueba estadística muestra que la desviación de 4 para los números nones y 4 para los números pares no se aparta significativamente de lo esperado por el azar.[2]

De manera similar, si se muestrean seres humanos, el número de hombres y mujeres en las muestras debe aproximarse en proporción al número de hombres y mujeres en la población (si el muestreo es aleatorio y las muestras son lo suficientemente grandes). Si medimos la inteligencia de los miembros de una muestra y la puntuación media de inteligencia de la población es 100, entonces la media de la muestra debe acercarse a 100, aunque se debe tener siempre en mente la posibilidad de seleccionar una muestra desviada, como sería una con una media de 80 o menos, o de 120 o más. Las muestras desviadas ocurren, pero es menos probable que ocurran. El razonamiento es similar al utilizado para demostraciones de lanzamientos de monedas al aire. Si lanzamos una moneda al aire tres veces, es menos probable que resulten tres caras (C) o tres cruces (X), que dos caras y una cruz o dos cruces y una cara. Esto es porque U = {*CCC*, *CCX*, *CXC*, *CXX*, *XCC*, *XCX*, *XXC*, *XXX*}. Solamente hay un punto de *CCC* y uno de *XXX*, mientras que hay tres puntos con dos C y hay tres con dos X.

Aleatorización

Suponga que un investigador desea probar la hipótesis de que el consejo psicológico puede ayudar a los estudiantes con bajo rendimiento. La prueba incluye utilizar dos grupos de estudiantes con bajo rendimiento: uno que recibirá consejo psicológico y otro que no lo

[2] La naturaleza de tales pruebas estadísticas, así como el razonamiento que las sostiene, se explicará con detalle en la parte 4. El estudiante no deberá preocuparse demasiado si no capta completamente las ideas estadísticas expresadas aquí. De hecho, uno de los propósitos de este capítulo es introducir algunos elementos básicos de tales ideas.

recibirá. Naturalmente, lo desable es tener dos grupos iguales respecto a otras variables independientes que tengan un posible efecto en el rendimiento. Una forma de hacer esto es asignar aleatoriamente a los niños a ambos grupos, por ejemplo, lanzando una moneda para cada niño. Si resulta cara, el niño es asignado a un grupo, y si resulta cruz, se le asigna al otro. Note que si hubiera tres grupos experimentales, probablemente no sería útil la moneda, sino que podría utilizarse un dado con 6 caras: si resultara un 1 o un 2, se colocaría al niño en el grupo 1, los números 3 y 4 lo pondrían en el grupo 2 y los números 5 y 6 ubicarían al niño en el grupo 3. También se podría usar una tabla de números aleatorios para signar a los niños a los grupos, de manera que si se obtiene un número non, el niño se asigna al grupo 1, y si el número es par el niño es colocado en el otro grupo. El investigador puede suponer ahora que los grupos son aproximadamente iguales respecto de todas las variables independientes posibles. Mientras más grande es el grupo, más segura es tal suposición. Así como no hay garantía de no seleccionar una muestra desviada (como se explicó antes), tampoco la hay de que los grupos sean iguales o casi iguales en todas las posibles variables independientes. No obstante, puede decirse que el investigador ha usado la aleatorización para igualar los grupos, o, dicho de otro modo, para controlar las influencias sobre la variable dependiente, que no sean las de la variable independiente manipulada. Aunque se usará el término "aleatorización", muchos investigadores prefieren usar las palabras *asignación aleatoria*; este proceso implica asignar a los participantes a las condiciones experimentales en forma aleatoria. Aunque algunos creen que la asignación al azar elimina la variación, en realidad sólo la distribuye.

Un experimento "ideal" sería aquel en que todos los factores o variables que pudieran afectar el resultado experimental fueran controlados. Si todos estos factores se conocieran, en primer lugar, y se *pudieran* hacer esfuerzos para controlarlos, en segundo lugar, entonces se tendría un experimento ideal. Sin embargo, la realidad es que no podemos conocer todas las variables pertinentes, ni podemos controlarlas (aún si las conociéramos). Sin embargo, la aleatorización puede ayudar.

La *aleatorización* es la asignación de miembros de un universo a los tratamientos experimentales, de manera que para cualquier asignación a un tratamiento, cada miembro del universo tiene la misma probabilidad de ser elegido para dicha asignación. El propósito básico de la *asignación aleatoria*, como se indicó anteriormente, es repartir sujetos (objetos, grupos) a tratamientos. Individuos con diferentes características son distribuidos de manera aproximadamente igual entre los tratamientos, de modo que las variables que puedan afectar a la variable dependiente (que no sean las variables experimentales), tengan "igual" efecto en los diferentes tratamientos. No hay garantía de que esta situación deseable se alcance, pero es más probable lograrla con la aleatorización que de otra forma. La aleatorización también tiene una razón y propósito estadísticos. Si se usa la asignación alcatoria, entonces es posible distinguir entre la varianza sistemática o experimental y la varianza del error. Las variables que pueden producir un sesgo son distribuidas a los grupos experimentales de acuerdo al azar. Las pruebas de significancia estadística (que se estudiarán más adelante) lógicamente dependen de la asignación al azar. Estas pruebas son usadas para determinar si el fenómeno observado es estadísticamente diferente al efecto del azar; sin asignación aleatoria las pruebas de significancia carecen de fundamento lógico. La idea de la aleatorización parece haber sido descubierta o inventada por Sir Ronald Fisher (véase Cowles, 1989). Fue Fisher quien virtualmente revolucionó el diseño y los métodos estadísticos y experimentales usando conceptos aleatorios como parte de su influencia. Él ha sido llamado "el padre del análisis de varianza". En cualquier caso, la aleatorización y lo que se refiere como el principio de aleatorización es uno de los mayores logros intelectuales de nuestro tiempo, y no es posible sobrestimar la importancia tanto de la idea como de las medidas prácticas que vinieron a mejorar la experimentación y la inferencia.

La aleatorización quizás puede ser aclarada en tres sentidos: estableciendo los principios de la aleatorización, describiendo cómo se usan en la práctica y demostrando cómo trabaja con objetos y números. Los tres merecen la misma importancia.

El *principio de aleatorización* puede ser enunciado como sigue: dado que en los procedimientos aleatorios cada miembro de una población tiene la misma posibilidad de ser seleccionado, si se eligen sujetos que presentan ciertas características distintivas —masculino o femenino, con alta o baja inteligencia, conservador o liberal, etcétera— probablemente a la larga serán compensados por la elección de otros miembros de la población con características, en cantidad o calidad, diferentes a las de ellos. Se puede decir que es un principio práctico que sucede en general; no se puede decir que sea una ley de la naturaleza, sino que es simplemente una afirmación de lo que ocurre con mayor frecuencia cuando se usan procedimientos aleatorios.

Se dice que los sujetos son asignados al azar a grupos experimentales y que los tratamientos experimentales son asignados al azar a grupos. Por ejemplo, en el experimento citado antes, donde se prueba la efectividad del consejo psicológico en el rendimiento escolar, los sujetos pueden ser asignados aleatoriamente a dos grupos, usando números aleatorios o lanzando una moneda al aire. Cuando los sujetos han sido asignados, los grupos pueden, a su vez, ser aleatoriamente designados como grupo experimental y grupo control usando un procedimiento similar. Se encontrarán varios ejemplos de aleatorización más adelante.

Una demostración de aleatorización senatorial

Para demostrar cómo funciona el principio de aleatorización, a continuación se describe un experimento de muestreo y diseño. Se tiene una población de 100 miembros del senado de los Estados Unidos, de los que se puede obtener una muestra. En esta población (en 1993) hay 56 demócratas y 44 republicanos. Se seleccionan dos votos importantes: el asunto 266, una enmienda para prohibir el incremento a las cuotas por pastoreo; y el asunto 290, una enmienda sobre el financiamiento para abortos. Los datos usados en este ejemplo fueron tomados del *Congressional Quarterly* de 1993. Estos votos fueron importantes porque cada uno reflejaba propuestas presidenciales. Un voto en contra en el asunto 266 y un voto a favor en el asunto 290, indicaban apoyo del presidente. Aquí se ignora la sustancia y se considera a los votos o mejor dicho, a los senadores que emitieron los votos, como poblaciones de las que tomamos la muestra.

Suponga que hacemos un experimento usando tres grupos de senadores, con 20 en cada grupo. La naturaleza del experimento no es relevante aquí. Se desea que los tres grupos de senadores sean aproximadamente iguales en todas las características posibles. Se generan números de aproximación aleatoria que van del 1 al 100, usando un programa de computadora escrito en BASIC (la referencia de los programas GWBASIC o QUICKBASIC de Microsoft se incluye en el final del capítulo). Los primeros 60 números tomados, no repetidos (muestreo sin reemplazamiento) son registrados en grupos de 20 cada uno. La afiliación política para los demócratas (d) y para los republicanos (r) se anota junto al nombre del senador. También se incluyen los votos de los senadores a las dos enmiendas: " f " para voto a favor y "c" para voto en contra. Estos datos se presentan en la tabla 8.2.

¿Qué tan "iguales" son los grupos? En la población total de 100 senadores, 56 son demócratas y 44 son republicanos, o 56% y 44%. En la muestra total de 60 hay 34 demócratas y 26 republicanos, es decir, 57% son demócratas y 43% son republicanos. Hay una diferencia de 1% para lo esperado de 56% y 44%. Las frecuencias obtenidas y las espera-

TABLA 8.2 *Voto senatorial por grupos de n = 20 en el asunto 266 del senado y el asunto 290*

#	Nombre-partido	266	290	#	Nombre-partido	266	290	#	Nombre-partido	266	290
73	hatfield-r	f	c	78	chafee-r	f	f	58	smith-r	f	c
27	coats-r	f	c	20	coverdell-r	f	c	95	byrd-d	c	c
54	kerrey-d	f	f	25	mosley-brown-d	c	f	83	matthews-d	f	c
93	murray-d	c	f	42	kerry-d	c	f	52	burns-r	f	c
6	mccain-r	f	c	68	dorgan-d	f	c	80	thurmond-r	f	c
26	simon-d	c	f	57	gregg-r	f	c	13	dodd-d	f	f
7	bumpers-d	c	f	11	campbell-d	f	f	88	hatch-r	f	c
81	daschle-d	c	f	31	dole-r	f	c	63	moynihan-d	f	f
76	specter-r	c	f	37	mitchell-d	c	f	89	leahy-d	c	f
38	cohen-r	c	f	30	grassley-r	f	c	75	wofford-d	c	f
32	kasselbaum-r	f	c	22	inouye-d	f	f	92	warner-r	f	c
44	riegel-d	c	f	99	simpson-r	f	c	91	robb-d	c	f
98	khol-d	c	f	8	pryor-d	c	?	34	mcconnell-r	f	c
77	pell-d	c	f	4	stevens-r	f	c	96	rockefeller-d	c	f
61	bingaman-d	f	c	23	craig-r	f	c	28	lugar-r	f	c
16	roth-r	c	c	12	brown-r	f	c	43	levin-d	c	f
24	kempthoner-r	f	c	10	feinstein-d	f	f	59	bradley-d	c	f
100	wallop-r	f	c	87	bennett-r	f	c	69	glenn-d	c	f
15	biden-d	c	c	19	nunn-d	c	c	9	boxer-d	c	f
14	lieberman-d	c	f	45	wellstone-d	c	f	67	conrad-d	f	c

das de los demócratas en los tres grupos (I, II y III) y la muestra total se presentan en la tabla 8.3. Las desviaciones respecto de lo esperado son pequeñas. Los tres grupos no son exactamente "iguales" en el sentido de que tengan igual número de senadores republicanos y demócratas. El primer grupo tiene 11 demócratas y 9 republicanos; el segundo grupo tiene 10 demócratas y 10 republicanos, y el tercer grupo tiene 13 demócratas y 7 republicanos. Éste no es un resultado "inusual" cuando se emplea el muestreo aleatorio. Posteriormente se verá que las discrepancias no difieren estadísticamente.

Recuerde que se está demostrando tanto el muestreo aleatorio como la aleatorización, pero especialmente la aleatorización, por lo tanto la pregunta es si la asignación aleatoria de los senadores a los tres grupos da como resultado la "igualación" de los grupos en todas sus características. Por supuesto que no se pueden probar todas las características, sino solamente las que están disponibles. En el presente caso solamente se tiene la afiliación al partido político, que se estudió anteriormente, y los votos en los dos asuntos: prohibición al incremento a las cuotas de pastoreo (asunto 266) y prohibición de fondos para ciertos tipos de aborto (asunto 290). ¿Cómo funcionó la asignación aleatoria en los dos asuntos? Los resultados se presentan en la tabla 8.4. De la votación original sobre el asunto 266, 99 de los senadores votaron a favor y 40 en contra. Estos votos totales produjeron frecuencias esperadas de votos a favor en el grupo total, de $59 \div 99 = .596$, o 60%; por lo tanto, la expectativa es de $20 \times .60 = 12$ en cada grupo experimental. El voto original de los 99 senadores que votaron sobre el asunto 290 fue de 40 a favor, o 40% ($40 \div 99 = .404$). Las frecuencias esperadas para el grupo que votó a favor, entonces, son: $20 \times .40 = 8$. Las frecuencias obtenidas y las esperadas y las desviaciones respecto de lo esperado, para los tres grupos de 20 senadores y para la muestra total de 60 en el asunto 266 y el asunto 290 se pueden ver en la tabla 8.4.

Es notorio que todas las desviaciones de lo esperado al azar son pequeñas. Los tres grupos son aproximadamente "iguales" en el sentido de que la incidencia de los votos

TABLA 8.3 *Frecuencias obtenidas y esperadas de partido político (Demócratas) en muestras aleatorias de 20 senadores de Estados Unidos*[a]

	Grupos			*Totales*
	I	**II**	**III**	
Obtenidas	11	10	13	34
Esperadas[b]	11.2	11.2	11.2	33.6
Desviación	.2	1.2	1.8	.4

[a] Solamente se reporta mayor de las dos expectativas de la división republicano-demócrata: los demócratas (.56).
[b] Las frecuencias esperadas fueron calculadas como sigue: $20 \times .56 = 11.2$. De la misma forma se calculó el total: $60 \times .56 = 33.6$.

sobre los dos asuntos es aproximadamente la misma en cada uno de los grupos. Las desviaciones de lo esperado al azar respecto de los votos a favor (y por supuesto de los votos en contra) es pequeña. Hasta donde se puede ver, la aleatorización ha sido "exitosa". Esta demostración también puede ser interpretada como un problema de muestreo aleatorio. Se podría preguntar, por ejemplo, si las tres muestras con 20 sujetos, y la muestra total de 60 son representativas. ¿Reflejan con precisión las características de la población de 100 senadores?. ¿Reflejan las muestras la proporción de demócratas y republicanos en el senado? Las proporciones en las muestras fueron .55 y .45 (I), .50 y .50 (II), .65 y .35 (III). Las proporciones reales son .56 y .44. Aunque hay una desviación de 1%, 6% y 9% respectivamente en las muestras, estas desviaciones están dentro de lo esperado al azar. Se puede decir, por lo tanto, que las muestras son representativas en lo que respecta a la pertenencia al partido político. Un razonamiento similar se puede aplicar a las muestras y a los votos en los dos asuntos.

Ahora se puede realizar el experimento suponiendo que los tres grupos son "iguales"; por supuesto que podrían no serlo, pero las probabilidades están a favor. Y como se ha visto, el procedimiento por lo general trabaja bien. La verificación de las características de los senadores en los tres grupos mostró que los grupos son bastante "iguales" respecto a la preferencia política y a los votos a favor (y en contra) en los dos asuntos. Así se puede tener mayor confianza en que si los grupos son desiguales, las diferencias probablemente se deban a la manipulación experimental y no a las diferencias entre los grupos antes de haber iniciado la investigación.

Sin embargo, un experto como Feller (1967, p. 29), escribió:

TABLA 8.4 *Frecuencias esperadas y obtenidas de votos a favor sobre el asunto 226 y el asunto 290 en grupos aleatorios de senadores*

	Grupos							
	I		*II*		*III*		*Total*	
	266	**290**	**266**	**290**	**266**	**290**	**266**	**290**
Obtenidas	9	10	13	9	11	9	33	28
Esperadas[a]	12	8	12	8	12	8	36	24
Desviación	3	2	1	1	1	1	3	4

[a] Las frecuencias esperadas fueron calculadas para el grupo I (asunto 266), de la siguiente manera: hubo 59 votos a favor, de un total de 99 votos o 59/99 = .60; $20 \times .60 = 12$. Para el grupo total el cálculo es: $60 \times .60 = 36$.

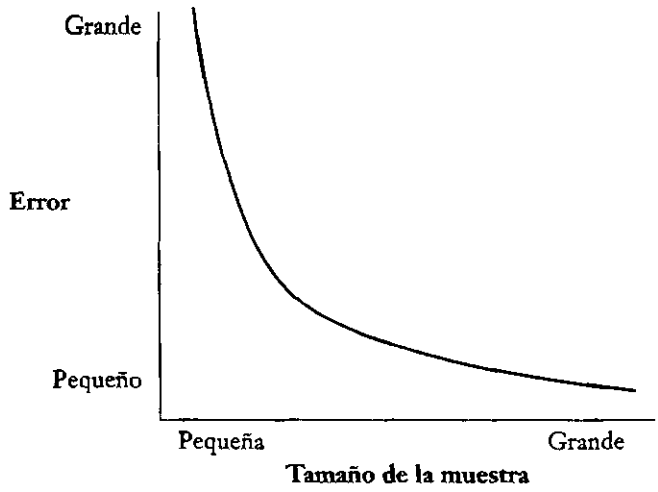
> En el muestreo de poblaciones humanas, el estadístico encuentra dificultades considerables y frecuentemente impredecibles, y la amarga experiencia ha mostrado que es difícil obtener, tan siquiera, una ordinaria imagen del azar.

Williams (1978) expone muchos ejemplos donde la "aleatorización" no funcionaba en la práctica. Uno de tales ejemplos, que influyó la vida de un gran número de hombres fue el sorteo del servicio militar en 1970. Aunque nunca fue probado del todo, parecía que los números en el sorteo no estaban aleatorizados; se metieron en una cápsula el mes y día de nacimiento, incluyendo todos los 366 días del año. Las cápsulas se colocaron en una tómbola, la cual se giraba muchas veces de modo que las cápsulas quedaran bien mezcladas. La primera cápsula seleccionada indicaba a los primeros que serían reclutados, la segunda cápsula seleccionada indicaba a los siguientes y así sucesivamente. Los resultados mostraron que las fechas para los últimos meses del año tenían una mediana menor que los primeros meses, por lo que los hombres con fecha de nacimiento hacia el final del año fueron llamados al servicio más pronto. Si la selección hubiese sido completamente al azar, las medianas para cada mes debían haber sido mucho más parecidas. El punto importante aquí es que muchos análisis estadísticos dependen de una aleatorización exitosa. Pero lograr esto en la práctica no es tarea fácil.

Tamaño de la muestra

Un regla dura pero eficaz, que se enseña a estudiantes principiantes de investigación es: utilizar una muestra tan grande como sea posible. Siempre que se calcula una media, un porcentaje o cualquier otro estadístico a partir de una muestra, se está estimando un valor poblacional. Una pregunta que debe hacerse es: ¿Qué tanto error es posible que resulte en estadísticos calculados de muestras de diferentes tamaños? La curva de la figura 8.1 expresa aproximadamente las relaciones entre el tamaño de la muestra y el error (el error es la desviación respecto a los valores poblacionales). La curva dice que a menor tamaño de la muestra, mayor será el error, y que a mayor tamaño de la muestra, menor será el error resultante.

FIGURA 8.1



Considere el siguiente ejemplo extremo. El Dr. Stanley Sue facilitó al segundo autor de este texto el puntaje de admisión de la Escala de Evaluación Global (GAS, por las siglas en inglés del Global Assessment Scale), y los días totales en terapia, de 3166 niños del condado de Los Ángeles (en las instalaciones del Centro Mental de la localidad), en los años 1983 a 1988. El doctor Sue es profesor de psicología y Director del Centro Nacional de Investigación en Salud Mental para asiático-americanos en la Universidad de California, Davis. El Dr. Sue otorgó permiso para utilizar sus datos. La información contenida en las tablas 8.5 y 8.6 se originó a partir de estos datos. El autor agradece al Dr. Stanley Sue. El GAS es un puntaje asignado por un terapeuta a cada paciente basado en su funcionamiento psicológico, social y ocupacional. La puntuación del GAS usada en este ejemplo es la que el paciente recibió al momento de su admisión o en la primera visita a la unidad.

De esta "población" se seleccionaron aleatoriamente 10 muestras de 2 niños cada una. La selección aleatoria de estas muestras y de otras se hizo usando la función "muestra" del SPSS ([paquete estadístico para ciencias sociales] [Norusis, 1992]). Las medias muestrales fueron calculadas por medio de la rutina "descriptivos" del SPSS y se pueden observar en la tabla 8.5. Las desviaciones de las medias, respecto de la media poblacional también se incluyen en ese mismo cuadro.

Las medias del GAS van de 42.5 a 60.5 y las medias de los días totales en terapia van de 14 a 303. Las dos medias totales (calculadas para las 20 puntuaciones del GAS y para las 20 puntuaciones de días totales) son 49.9 y 106.8. Estas medias calculadas a partir de muestras pequeñas varían considerablemente. Las medias de las puntuaciones de GAS y de días totales de la población (N = 3 166) fueron de 49.29 y 83.54. Las desviaciones (Des.) de las medias del GAS presentan un amplio rango: de –6.79 a 11.21. Las desviaciones de los días totales van de –69.54 a 219.46. Con muestras muy pequeñas como éstas no se puede depender de una sola media para estimar el valor de la población. Sin embargo se puede confiar más en las medias calculadas de las 20 puntuaciones, aunque ambas tienen un sesgo hacia arriba.

TABLA 8.5 *Muestras (n = 2) de las puntuaciones GAS y de días totales en terapia de 3 166 niños; media de las nuestras y desviaciones de las medias muestrales de la población (datos del Dr. Sue)*

GAS

Muestra	1	2	3	4	5	6	7	8	9	10
	61	46	65	50	51	35	45	44	43	60
	60	50	35	55	55	50	41	47	50	55
Media	60.5	48	50	52.5	53	42.5	43	45.4	46.5	57.5
Des.	11.21	–1.29	.71	3.21	3.71	–6.79	–6.29	–3.79	–2.79	8.21

Media total (20) = 49.9
Media poblacional (3 166) = 49.29

Días totales en terapia

Muestra	1	2	3	4	5	6	7	8	9	10
	92	9	172	0	3	141	28	189	28	17
	57	58	38	70	603	110	0	51	72	398
Media	74.5	33.5	105	35	303	125.5	14	120	50	207.5
Des.	–9.04	–50.04	21.46	–48.54	219.46	41.96	–69.54	36.46	–33.54	123.96

Media total (20) = 106.80
Media poblacional (3 166) = 85.54

TABLA 8.6 *Medias y desviaciones de la media poblacional de cuatro muestras de GAS y cuatro muestras de días totales (n = 20), muestra total (n = 89) y población (N = 3 166) (datos del Dr. Sue)*

	Muestras (n = 20)				Total (n = 80)		Población (N = 3 166)
GAS	49.35	48.90	49.85	50.6	49.68		49.29
Des.	.06	.39	.56	1.31		.385	
Días totales	69.40	109.95	89.55	103.45	93.08		83.54
Des.	14.14	26.41	6.01	19.91		9.540	

Cuatro muestras aleatorizadas más, de 20 puntuaciones GAS y 20 de días totales fueron tomadas de la población: las medias de estas muestras se presentan en la tabla 8.6. Las desviaciones (Des.) de cada una de las medias de las muestras de 20 puntuaciones, en relación con la media poblacional, también aparecen en la tabla, así como las medias de la muestra de 80 puntuaciones y de la población total. Las desviaciones del GAS van de .06 a 1.31, y las desviaciones de los días totales van de 6.01 a 26.41. La media de las 80 puntuaciones de GAS es de 49.68, y la media de las 3166 puntuaciones del GAS es de 49.29. Las medias comparables de los días totales son 93.08 (n = 80) y 83.54 (N = 3 166). Estas medias son mucho mejores estimados de las medias poblacionales.

Ahora se pueden sacar algunas conclusiones: primero, los estadísticos calculados a partir de muestras grandes son más precisos (si las demás características son iguales) que los calculados a partir de las muestras pequeñas. Un vistazo a las desviaciones de las tablas 8.5 y 8.6 mostrará que las medias de las muestras de 20 puntuaciones se desviaron mucho menos de la media poblacional que las medias de las muestras de dos puntuaciones. Por otro lado, las medias de la muestra de 80 sujetos se desvió muy poco de las medias poblacionales (0.39 y 9.54).

Debe ser muy claro ahora por qué el principio de la investigación y del muestreo es: use muestras grandes.[3] Las muestras grandes no se recomiendan sólo porque los números grandes sean mejores en sí mismos, sino para permitir que el principio de la aleatorización, o simplemente del azar "funcione". Con muestras pequeñas, la probabilidad de seleccionar muestras sesgadas es mayor que con muestras grandes; por ejemplo, en una muestra aleatoria de 20 senadores seleccionada hace algunos años, los primeros 10 senadores obtenidos (de 20) fueron todos ¡demócratas! Una serie de 10 demócratas sería inusual, *¡pero puede y de hecho sucede!* Digamos que se decide realizar un experimento con sólo dos grupos de 10 sujetos cada uno. Uno de los grupos tiene 10 demócratas y el otro tiene tanto demócratas como republicanos. Los resultados podrían estar seriamente sesgados, especialmente si el experimento tuviera relación con su preferencia política o sus actitudes sociales. Con grupos grandes, digamos 30 o más sujetos, hay menos riesgo. Muchos departamentos de psicología de grandes universidades tienen un requisito de investigación para los estudiantes inscritos en una clase de introducción a la psicología. Para tales situaciones, puede ser relativamente fácil obtener muestras grandes. Sin embargo, para ciertos estudios de investigación (como los de ingeniería humana o de investigación de mercados), el costo de reclutar participantes es alto. Recuerde el estudio de Williams y Adelson comentado por Simon (1987) en el capítulo 1. Así, la regla de tener muestras grandes puede no ser apropiada para todas las situaciones de investigación. En algunos estudios, 30 o más elementos, participantes o sujetos pueden ser muy pocos, especialmente en estudios que son de

[3] La situación es más compleja de lo que este simple enunciado indica. Las muestras que son demasiado grandes pueden traer otros problemas; las razones se explicarán en un capítulo posterior.

naturaleza multivariada. Comrey y Lee (1992), por ejemplo, afirman que muestras de 50 o menos sujetos poseen una inadecuada confiabilidad para los coeficientes de correlación. De aquí que puede ser más apropiado obtener una aproximación al tamaño de muestra que se necesita. La determinación estadística del tamaño muestral para los diferentes tipos de muestras se explicará en el capítulo 12.

Tipos de muestras

La discusión del muestreo ha sido hasta ahora confinada al muestreo aleatorio simple. El propósito es ayudar al estudiante a entender los principios fundamentales, por lo que se ha hecho énfasis en la idea del muestreo aleatorio simple, ya que es el fundamento de gran parte del pensamiento y de los procedimientos de la investigación moderna. El estudiante deberá comprender, sin embargo, que el muestreo aleatorizado simple no es la única clase de muestreo usado en la investigación del comportamiento; de hecho, es poco utilizado al menos para describir las características de poblaciones y las relaciones entre tales características. Sin embargo, es el modelo en el que todo muestreo científico se basa.

Otras clases de muestras pueden clasificarse de forma general en probabilísticas y no probabilísticas (y ciertas formas mixtas). Las *muestras probabilísticas* usan alguna forma de muestreo aleatorizado en una o más de sus etapas. Las *muestras no probabilísticas* no usan el muestreo aleatorizado, por lo que carecen de las virtudes que se han discutido, pero con frecuencia son necesarias e imprescindibles. Su debilidad puede, en cierta medida, ser mitigada con el uso del conocimiento, la experiencia y el cuidado al seleccionar las muestras, y replicando los estudios con diferentes muestras. Es importante que el estudiante sepa que el muestreo probabilístico no es necesariamente superior al muestreo no probabilístico en todas las situaciones y que el muestreo probabilístico tampoco garantiza muestras más representativas del universo en estudio. En el muestreo probabilístico el énfasis radica en el método y en la teoría que lo sustenta, mientras que en muestreo no probabilístico el énfasis reside en la persona que hace el muestreo y que puede acarrear consigo complicaciones enteramente nuevas e importantes. La persona que hace el muestreo debe ser conocedora de la población que se estudia, así como del fenómeno en estudio.

Una forma de muestreo no probabilístico es el *muestreo por cuotas*, donde el conocimiento de los estratos de la población —sexo, raza, región, etcétera— se utiliza para seleccionar a los miembros de la muestra que sean representativos, "típicos" y apropiados para ciertos propósitos de investigación. Un *estrato* es la partición del universo o población en dos o más grupos que no se traslapan (mutuamente excluyentes). Se toma una muestra de cada fracción. El muestreo por cuotas deriva su nombre de la práctica de asignar cuotas o proporciones de clases de personas a los entrevistadores. Este tipo de muestreo ha sido muy utilizado en encuestas de opinión pública. Para realizar este muestreo correctamente, el investigador necesita tener una lista muy completa de las características de la población, y luego debe conocer las proporciones de cada cuota. Después de esto, el siguiente paso es recolectar los datos. Dado que las proporciones pueden diferir de cuota a cuota, se les asigna un peso a los elementos de la muestra. El muestreo por cuotas es difícil de realizar porque requiere información precisa de las proporciones de cada cuota, y esta información pocas veces está disponible.

Otra forma de muestreo no probabilístico es el *muestreo propositivo*, que se caracteriza por el uso de juicios e intenciones deliberadas para obtener muestras representativas al incluir áreas o grupos que se presume son típicos en la muestra. El muestreo propositivo es usado con mucha frecuencia en la investigación de mercados. Para probar la reacción de los consumidores ante un nuevo producto, el investigador puede distribuir el nuevo producto entre personas que se ajustan al concepto que el investigador tiene del universo.

Otro ejemplo del uso del muestreo propositivo son las encuestas políticas. Con base en los resultados de votaciones pasadas y registros de partidos políticos existentes en cierta región, el investigador, propositivamente selecciona un grupo de distritos electorales. El investigador cree que esa selección comparte las características de todo el electorado. Una presentación muy interesante de cómo esta información fue usada para ayudar a elegir a un senador de Estados Unidos por el estado de California, se puede revisar en Barkan y Bruno (1972).

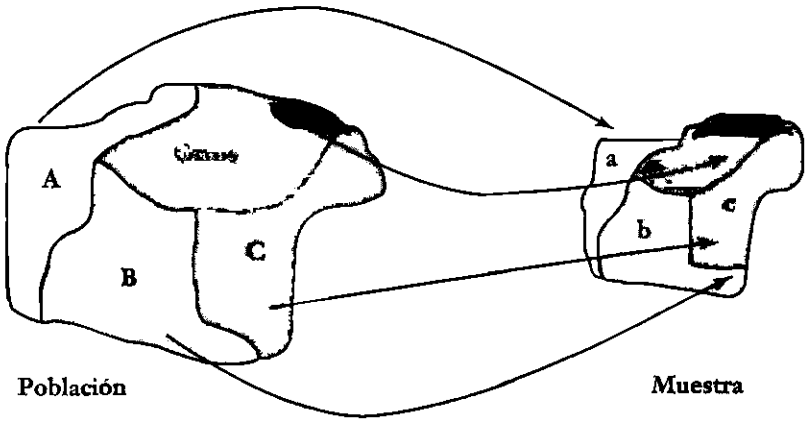
El llamado *muestreo accidental*, la forma más débil de muestreo, es quizá el más utilizado. Aquí se toman muestras disponibles a mano —estudiantes del último año de preparatoria, estudiantes universitarios de segundo año, una asociación de padres y maestros, etcétera—. Esta práctica es difícil de defender, aunque, usada con conocimiento razonable y cuidado, probablemente no merezca la mala reputación que tiene. El mejor consejo sería: evite las muestras accidentales a menos que no tenga otra opción (el muestreo aleatorio generalmente es caro y difícil de realizar); si se utilizan muestras accidentales es necesario ser extremadamente precavido en el análisis e interpretación de los datos.

El *muestreo probabilístico* incluye una variedad de formas. Cuando se explicó el muestreo aleatorio simple, se habló sobre una versión de muestreo probabilístico. Otras formas comunes de muestreo probabilístico son el muestreo estratificado, el muestreo por racimos, el muestreo por racimos de dos etapas y el muestreo sistemático. Otros métodos menos convencionales, incluyen el enfoque bayesiano o secuencial. La superioridad de un método de muestreo sobre otro se evalúa por lo regular en términos de la cantidad de variabilidad reducida en los parámetros estimados y en términos de costos. El costo es algunas veces interpretado como la cantidad de trabajo en la recolección de datos y en el análisis de datos.

En el *muestreo estratificado*, primero se divide a la población en estratos tales como hombres y mujeres, afro-americanos y mexico-americanos, etcétera. Después se seleccionan muestras aleatorias de cada estrato. Si la población consta de 52% de mujeres y 48% de hombres, una muestra estratificada de 100 participantes consistirá en 52 mujeres y 48 hombres. Las 52 mujeres se seleccionarían al azar del grupo disponible de mujeres y los 48 hombres se obtendrían aleatoriamente del grupo de hombres. Esto es llamado también distribución proporcional. Cuando este procedimiento se realiza correctamente, es superior al muestreo aleatorio simple. Comparado con el muestreo aleatorio simple, el muestreo estratificado generalmente reduce tanto la cantidad de variabilidad como el costo de recolección y análisis de datos. El muestreo estratificado saca provecho de las diferencias entre estratos. La figura 8.2 ilustra la idea básica del muestreo estratificado, el cual añade control al proceso de muestreo disminuyendo la cantidad de error de muestreo. Este diseño se recomienda cuando la población está compuesta de conjuntos de grupos desiguales. El muestreo estratificado aleatorizado ayuda a estudiar las diferencias en los estratos; permite dar especial atención a ciertos grupos que, de otra forma, podrían ser ignorados a causa de su tamaño. El muestreo aleatorizado estratificado se lleva a cabo con procedimientos de distribución proporcional (PDP). Al utilizar estos procedimientos, la división proporcional de la muestra asemeja la de la población. La mayor ventaja de usar PDP es que proveen de una muestra "auto-ponderada".

El *muestreo por racimos* —el método usado con mayor frecuencia en encuestas— es el muestreo aleatorio consecutivo de unidades o conjuntos y subconjuntos. Un racimo puede ser definido como un grupo de cosas de la misma clase; es un conjunto de elementos muestrales unidos por alguna(s) característica(s) en común. En el muestreo por racimos, el universo es fraccionado en racimos, y después los racimos son muestreados aleatoriamente. Entonces cada elemento en el racimo elegido es medido. En investigación sociológica, el investigador puede usar las manzanas de la ciudad como racimos: las manzanas de la ciu-

FIGURA 8.2



dad son elegidas aleatoriamente y los entrevistadores entonces hablan o entrevistan con todas las familias de las manzanas seleccionadas. A este tipo de muestreo algunas veces se le llama *muestreo de área*. Si un investigador utilizara el muestreo aleatorio simple o el muestreo aleatorio estratificado, necesitaría una lista completa de las familias o casas habitación para tomar su muestra. Dicha lista sería muy difícil de obtener en una gran ciudad, y aun cuando se tuviera, el costo del muestreo sería muy alto ya que implicaría medir casas habitación en una gran área de la ciudad. El muestreo por racimos es más efectivo si se usa un gran número de pequeños racimos. En investigación educacional por ejemplo, los distritos escolares de un estado o de un condado pueden usarse como racimos, tomando una muestra aleatoria de dichos distritos; cada escuela dentro del distrito escolar sería medida. Sin embargo, el distrito escolar puede conformar un racimo tan grande, que sería mejor usar las escuelas como racimos.

El *muestreo por racimos de dos etapas*, se inicia con un muestreo por racimos, como se describió antes; después, en lugar de medir cada elemento de los racimos elegidos al azar, se selecciona una muestra aleatoria de los elementos y se miden estos elementos. En el ejemplo educativo expuesto antes, se identificaría cada distrito escolar como un racimo y después se elegirían k distritos escolares al azar. De estos k distritos escolares, en lugar de medir cada escuela en los distritos elegidos (como se haría en el muestreo por racimos regular), se tomaría otra muestra aleatoria de las escuelas de cada distrito y se medirían solamente las escuelas elegidas.

Otra clase de muestreo probabilístico —si en realidad puede llamarse muestreo probabilístico— es el *muestreo sistemático*. Este método es una ligera variación del muestreo aleatorio simple. Este método supone que el universo o población consiste en elementos que están ordenados de alguna forma. Si la población consta de N elementos y se desea elegir una muestra de tamaño n, primero es necesario formar la razón N/n. Esta razón se redondea a un número entero k, el cual se usa como el intervalo del muestreo. El primer elemento muestral es elegido aleatoriamente de entre 1 y k, y los elementos subsecuentes se eligen cada k intervalos. Por ejemplo, si el elemento seleccionado aleatoriamente de los que van del 1 al 10, es 6, entonces los elementos subsecuentes son 16, 26, 36, etcétera. La representatividad de la muestra elegida de esta manera depende del ordenamiento de los N elementos de la población.

El estudiante que más adelante realizará investigación, deberá conocer mucho más sobre estos métodos, por lo que se le invita a que consulte una o más de las excelentes referencias sobre el tema, presentadas al final de este capítulo. Williams (1978) da una interesante presentación y demostración de cada método de muestreo usando datos artificiales.

Otro tema relacionado con la aleatorización y el muestreo, son las pruebas de aleatorización o permutación. Este tema se abordará de nuevo cuando se hable del análisis de datos para los diseños cuasi experimentales. Edgington (1980, 1996) fue quien propuso este método en psicología y en las ciencias del comportamiento. Él propone el uso de las pruebas de aleatorización aproximada para el análisis estadístico de los datos de muestras no aleatorias y para diseños de investigación de caso único. Ahora se explicará brevemente cómo funciona este procedimiento. Tomemos el ejemplo de Edgington de correlacionar las puntuaciones de CI de padres adoptivos y sus hijos adoptados. Si la muestra no se selecciona aleatoriamente, podría estar sesgada a favor de los padres que deseaban que se midiera su CI y el de sus hijos adoptados. Es probable que algunos padres adoptivos bajaran sus puntuaciones para hacerlas coincidir intencionalmente con el CI de sus hijos adoptados. Una forma de manejar datos no aleatorios como éste, es primero calcular la correlación entre los padres y los hijos; entonces se podrían aparear aleatoriamente las puntuaciones de los padres con las de los hijos: esto es, el papá 1 podría aparearse aleatoriamente con el hijo del padre número 10. Después de este apareo aleatorio, la correlación es calculada de nuevo. Si el investigador realiza 100 apareamientos aleatorizados y calcula la correlación cada vez, podrá entonces comparar la correlación original con los 100 apareamientos creados aleatoriamente. Si la correlación original es la mejor (la más alta), el investigador tendrá una mejor idea de que la correlación obtenida pueda ser creíble. Esta pruebas de permutación o de aleatorización han sido muy útiles para ciertas investigaciones y ciertas situaciones de análisis de datos. Se han usado para evaluar los conglomerados obtenidos en un análisis de conglomerados (véase Lee y MacQueen, 1980) y se han propuesto como una solución para el autoanálisis de eficacia de los datos que no son independientes (véase Cervone, 1987).

El azar, la aleatorización y el muestreo aleatorizado están entre las grandes ideas de la ciencia, como se indicó antes. Aunque la investigación puede, por supuesto, realizarse sin usar las ideas de la aleatorización, es difícil concebir cómo podría ser viable y tener validez, al menos en la mayoría de los aspectos de la investigación científica del comportamiento. Los conceptos modernos del diseño de investigación, del muestreo y de la inferencia, por ejemplo, son literalmente inconcebibles sin la idea del azar. Una de las paradojas más relevantes es que a través de la aleatorización o "desorden", se puede tener control sobre las complejidades con frecuencia escandalosas de los fenómenos psicológicos, sociológicos y educativos. Se impone orden al explotar las conductas conocidas de los conjuntos de los eventos azarosos. Uno queda siempre maravillado de lo que se puede llamar la belleza estructural de la probabilidad, del muestreo y de la teoría del diseño y también de su gran utilidad para resolver problemas difíciles del diseño experimental, de la planeación y del análisis e interpretación de datos.

Antes de abandonar este tema, regresemos al punto de vista sobre la aleatorización mencionado antes. Para un sabio no existe lo aleatorio. Por definición, dicho sabio "conocería" la ocurrencia de cualquier evento con completa certeza. Como Poincare (1952/1996) señala, jugar con tal sabio sería una aventura perdedora, de hecho, no sería juego. Si una moneda se lanzara 10 veces, él podría predecir caras o cruces con completa certeza y precisión. Si se lanzaran los dados, este sabio sería infalible respecto de los resultados. ¡Cada número en una tabla de números aleatorios sería predicho correctamente! Obviamente este sabio no necesitaría la investigación y la ciencia. Lo que parece afirmarse con esto es

que aleatorio es un término para la ignorancia. Si nosotros, como el sabio, conociéramos todas las causas que contribuyen a los eventos, entonces no existiría el azar. La belleza de esto, como se dijo antes, es que usamos esta "ignorancia" y la convertimos en conocimiento. Cómo se hace esto será más y más evidente conforme se avance en el estudio.

Algunos libros sobre muestreo

Babbie, E.R. (1990). *Survey research methods* (2a. ed.) Belmont, California: Wadsworth.

Babbie, E.R. (1995). *The practice of social research* (7a. ed). Belmont, California: Wadsworth.

Cowles, M. (1989). *Statistics in psychology: A historical perspective*. Hillsdale, Nueva Jersey: Erlbaum.

Deming, W.E. (1966). *Some theory of samplig*. Nueva York: Dover.

Deming, W.E. (1990). *Sampling design in business research*. Nueva York: Wiley.

Kish, L. (1953). Selection of the sample, en Festinger, L., & Katz, D. (eds.), *Research methods in the behavioral sciences*. Nueva York: Holt, Rinehart and Winston (pp. 175-239).

Kish, L. (1995). *Survey samplig*. Nueva York: Wiley.

Snedecor, G. & Cochran, W. (1989). *Statistical Methods* (8a. ed). Ames, Iowa: Iowa State University Press.

Stephan, F. & McCarthy, P. (1974). *Sampling opinions*. Westport, Connecticut: Greenwood Press.

Sudman, S. (1976). *Applied sampling*. Nueva York: Academic Press.

Warwick, D. & Lininger, D. (1975). *The sample survey: Theory and practice*. Nueva York. McGraw-Hill

Williams, B. (1978). *A sampler on sampling*. Nueva York: Wiley.

RESUMEN DEL CAPÍTULO

1. El muestreo se refiere a tomar una porción de una población o universo, que sea representativa de esa población o universo.
2. Los estudios que usan muestras son económicos, manejables y controlables.
3. Uno de los métodos más populares de muestreo es el muestreo aleatorio.
4. En el muestreo aleatorio se toma una porción (o muestra) de una población o universo de manera que cada miembro de la población o universo tiene la misma probabilidad de ser elegido.
5. El investigador define la población o universo. Una muestra es un subconjunto de la población.
6. Nunca se podrá estar totalmente seguro de que el muestreo aleatorio sea representativo de la población.
7. En el muestreo aleatorio, la probabilidad de seleccionar una muestra con una media cercana a la media poblacional es mayor que la probabilidad de seleccionar una muestra con una media alejada de la media poblacional.
8. El muestreo no aleatorio puede estar sesgado, lo que aumenta las posibilidades de que la media muestral no se acerque a la media poblacional.
9. Se dice que los eventos son aleatorios si no se pueden predecir sus resultados.
10. La asignación aleatoria es otro término para aleatorización. Aquí los participantes son asignados aleatoriamente a grupos de investigación. Se usa para controlar variables indeseables.
11. Hay dos tipos de muestras: las no probabilísticas y las probabilísticas
12. Las muestras no probabilísticas no usan la asignación aleatoria, mientras que las probabilísticas usan el muestreo aleatorio.
13. El muestreo aleatorio simple, el muestreo aleatorio estratificado, el muestreo por racimos y el muestreo sistemático, son cuatro tipos de muestreo probabilístico.

14. El muestreo por cuotas, el muestreo propositivo y el muestreo accidental son tres tipos de muestreo no probabilístico.

SUGERENCIAS DE ESTUDIO

Se recomienda una serie de experimentos con fenómenos donde interviene el azar: juegos usando monedas, dados, barajas, ruleta y tablas de números aleatorios. Estos juegos, utilizados apropiadamente, pueden ayudar a aprender mucho acerca de los conceptos fundamentales de la investigación científica moderna, la estadística, la probabilidad y por supuesto, del azar. Intente resolver los problemas que se sugieren a continuación. No se desanime si encuentra que los ejercicios de este capítulo y los posteriores son laboriosos. Es necesario y útil involucrarse directamente en la rutina de ciertos problemas. Después de trabajar con los problemas que se dan, invente los suyos. Si puede crear problemas inteligentes, de seguro ya está en el camino de entender estos conceptos.

1. De una tabla de números aleatorios, tome 50 números, del 0 al 9. (Si lo desea use los números aleatorios del apéndice C.) Lístelos en columnas de 10 números cada una.
   a) Cuente el total de números impares; cuente el total de números pares. ¿Qué esperaría obtener por el azar? Compare los totales obtenidos con los totales esperados.
   b) Cuente el total de números 0, 1, 2, 4. De forma similar cuente los números 5, 6, 7, 8 y 9. ¿Cuántos del primer grupo obtuvo? ¿Cuántos del segundo grupo? Compare lo que obtuvo con lo esperado debido al azar. ¿Se aleja mucho lo esperado de lo obtenido?
   c) Cuente los números pares y los números nones en cada grupo de 10. Cuente los dos grupos de números 0, 1, 2, 3, 4 y 5, 6, 7, 8, 9 en cada grupo de 10. ¿Difieren mucho los totales de lo esperado por efecto del azar?
   d) Sume cada columna de los 5 grupos de 10 números. Divida cada suma entre 10. (Simplemente mueva el punto decimal un lugar a la izquierda.) ¿Que esperaría obtener como la media de cada grupo, si solamente estuviera "operando" el azar? ¿Qué obtuvo? Sume las cinco adiciones y divida entre 50. La media obtenida, ¿está cercana a la esperada por azar? [Pista: Para obtener lo esperado por el azar, recuerde los límites poblacionales.]
2. Éste es un ejercicio y demostración para el salón de clases. Asigne números arbitrariamente a todos los miembros de la clase, de 1 hasta N, donde N es el total de miembros de la clase. Tome una tabla de números aleatorios e inicie en cualquier parte. Pida a un estudiante, con los ojos cubiertos, que tome un lápiz y apunte hacia la página con la tabla de números aleatorios y que marque algún punto con el lápiz. Escoja n números de dos dígitos entre 1 y N (ignore los números mayores que N y los que se repiten) bajando por las columnas (o de cualquier otra forma específica). n es el numerador de la fracción n/N, que es decidida por el tamaño de la clase. Si $N = 30$, por ejemplo, permita que $n = 10$. Repita el proceso dos veces con diferentes páginas de las tablas de números aleatorios. Ahora tendrá tres grupos iguales. Si N no es divisible entre 3, elimine una o dos personas al azar. Escriba los números aleatorios en el pizarrón para los tres grupos. Pida a cada miembro de la clase que diga su estatura en centímetros y escriba estos valores en el pizarrón, separados de los números pero en los mismos tres grupos. Sume los tres conjuntos de números en cada uno de los conjuntos en el pizarrón; sume también los números aleatorios y las estaturas. Calcule las medias de los seis conjuntos de números y calcule las medias de los conjuntos totales.

a) ¿Qué tan cercanas están las medias de cada uno de los conjuntos de números? ¿Qué tan cercanas están las medias de los grupos a la media total del grupo?
b) Cuente el números de hombres y mujeres en cada uno de los grupos. ¿Se encuentran distribuidos equitativamente ambos sexos entre los tres grupos?
c) Discuta esta demostración. ¿Cuál piensa usted que es su significado para la investigación?

3. En el capítulo 6, se sugirió que el estudiante generara 20 conjuntos de 100 números aleatorios entre 0 y 100 y que calculara las medias y las varianzas. Si usted lo hizo, use los números y estadísticas de ese ejercicio. Si no lo hizo, use los números y estadísticas del apéndice C que está al final del libro.
   a) ¿Qué tan cercanas a la media poblacional están las medias de las 20 muestras? ¿Está "desviada" alguna de las medias? (Usted puede juzgar esto calculando la desviación estándar de las medias y sumando y restando dos desviaciones estándar a la media total.)
   b) Con base en (a), y en su juicio, ¿son "representativas" todas las medias? ¿Qué significa "representativos"?
   c) Tome las medias de los grupos tercero, quinto y noveno. Suponga que 300 sujetos han sido asignados aleatoriamente a los 3 grupos y que éstas son las puntuaciones de alguna medida de importancia en un estudio que usted desea realizar. ¿Qué cree usted que podría concluir a partir de estas tres medias?
4. La mayoría de los estudios publicados en ciencias del comportamiento y educación no usan muestras aleatorias, especialmente muestras aleatorias de grandes poblaciones. En ocasiones, sin embargo, se realizan estudios basados en muestras aleatorias. Uno de tales estudios es el realizado por Osgood, Wilson, O'Malley, Bachman y Johnston (1996). Este estudio merece una lectura cuidadosa, aunque su nivel de sofisticación metodológica incluye un gran número de detalles que van más allá del dominio que usted tiene del tema. Sin embargo, trate de no desanimarse por esta sofisticación. Saque a este documento todo el provecho que sea posible, especialmente en cuanto al muestreo de una gran población de jóvenes. Más adelante en el libro retomaremos este interesante problema. Por ahora quizás la metodología no parezca tan formidable. (Al estudiar investigación, a veces es útil leer más allá de nuestra capacidad presente, pero cuidando de no hacerlo demasiado.)

   Otro estudio de muestras aleatorias de una gran población es el realizado por Voekl (1995). En este estudio el investigador da algunos detalles acerca del uso del muestreo aleatorio estratificado en dos etapas para medir el nivel de cordialidad que el estudiante percibe de su escuela.
5. La asignación aleatoria de sujetos a grupos experimentales es mucho más común que el muestreo aleatorio de sujetos. Un ejemplo de investigación particularmente bueno (excelente, de hecho), en el que los sujetos fueron asignados al azar a dos grupos experimentales, fue realizado por Thompson (1980). De nuevo, no se deje intimidar por los detalles metodológicos de este estudio. Obtenga de él lo que pueda. Note la forma en que se clasificó a los sujetos en grupos de aptitudes y luego su asignación aleatoria a los tratamientos experimentales. También regresaremos a este estudio más adelante. Para entonces, usted será capaz de entender su propósito y diseño, y estará intrigado por el manejo experimental cuidadosamente controlado de un problema educativo muy difícil: los méritos de la llamada instrucción magistral individualizada comparados con la instrucción convencional de lectura-discusión-recitación.
6. Otro ejemplo notable de selección aleatoria es el estudio realizado por Glick, DeMorest y Hotze en 1988. Este estudio es digno de tomarse en cuenta porque

tiene lugar en un escenario real fuera de los laboratorios de la universidad. Los participantes no son por fuerza estudiantes universitarios, sino que son personas de un área pública de venta de alimentos dentro del interior de un gran centro comercial. Los participantes se seleccionaron y después se asignaron a una de seis condiciones experimentales. Este artículo es fácil de leer y el análisis estadístico no va más allá del nivel estadístico elemental.

7. Otro estudio interesante que usa otra variante del muestreo aleatorio es el de Moran y McCullers (1984). En este estudio, los investigadores seleccionaron aleatoriamente fotografías de un anuario. Después agruparon al azar dichas fotografías en 10 grupos de 16 fotos. Se les pidió a estudiantes que no conocían a los estudiantes de las fotos que evaluaran a cada persona en la foto en términos de su atractivo.

Nota especial. En algunos de los estudios sugeridos más arriba y en los del capítulo 6, se dieron instrucciones de seleccionar números de las tablas de números aleatorios o de generar grupos de números aleatorios usando una computadora. Si usted tiene una microcomputadora o tiene acceso a una, puede preferir generar números aleatorios usando la función de generador de números aleatorios de la computadora. Un libro destacado y divertido para leer y aprender cómo hacer esto es el de Walter (1999) "La Guía Secreta de las Computadoras". Walter muestra cómo escribir un sencillo programa de computadora usando lenguaje BASIC, el lenguaje común a la mayoría de las microcomputadoras. ¿Qué tan "buenos" son los números aleatorios generados?("Qué tan buenos" significa "Qué tan aleatorios".) Ya que son producidos en línea con la mejor teoría y práctica contemporánea, deben ser satisfactorios, aunque no cumplan exactamente con los requerimientos de algunos expertos. En nuestra experiencia, son bastante satisfactorios y recomendamos su uso a maestros y estudiantes. Una alternativa es el uso de números aleatorios de la Corporación Rand los cuales se reproducen parcialmente en el apéndice de este libro.

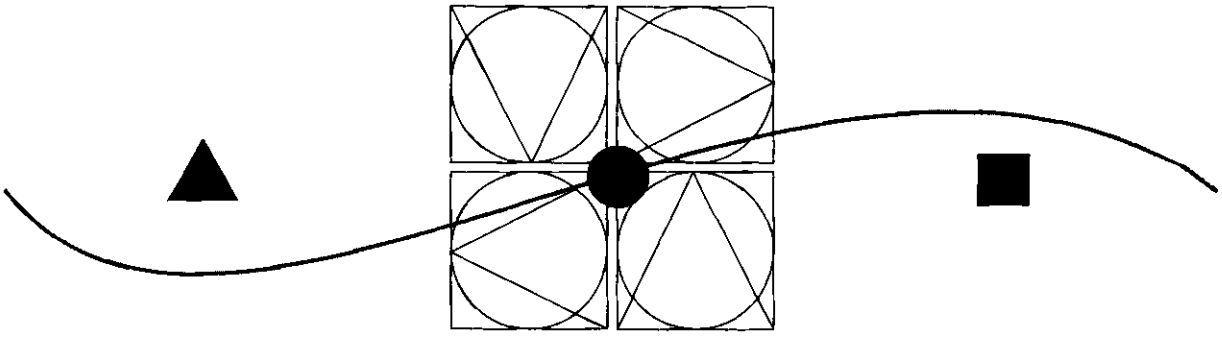
Listado de programa de cómputo para generar la tabla 8.2

```
10 DIM N(100), A$(100)
20 FOR I=1 TO 100: READ A$(I): N(I)=0
30 NEXT I
40 R=0: D=0
50 RANDOMIZE
60 I=1
70 X=RND
80 K=INT(X*100)
90 IF K=0 THEN K=100
100 IF N(K)=1 THEN 70
110 N(K)=1
120 PRINT K,A$(K)
140 Z$=RIGHT$(A$(K),1)
150 IF Z$="d" THEN D=D+1
160 IF Z$="r" THEN R=R+1
170 I=I+1
180 IF I>60 THEN 200
190 GOTO 70
200 FOR I=1 TO 100: PRINT N(I);: NEXT I
220 PRINT " ",D,R
230 DATA heflin-d,shelby-d,murkowski-r,stevens-r,deconcini-d
240 DATA mccain-r,bumpers-d,pryor-d,boxer-d,feinstein-d,campbell-d
250 DATA brown-r,dodd-d,lieberman-d,biden-d,roth-r,graham-d,mack-r
```

```
260 DATA nunn-d,coverdell-r,akaka-d,inouye-d,craig-r,kempthorne-r
270 DATA mosley-brown-d,simon-d,coats-r,lugar-r,harkin-d,grassley-r
280 DATA dole-r,kasselbaum-r,ford-d,mcconnell-r,breaux-d,johnston-d
290 DATA mitchell-d,cohen-r,mikulski-d,sarbanes-d,kennedy-d,kerry-d
300 DATA levin-d,riegel-d,wellstone-d,durenburger-r,cochran-r
310 DATA lott-r,bond-r,danforth-r,bacus-d,burns-r,exon-d,kerrey-d
320 DATA bryan-d,reid-d,gregg-r,smith-r,bradley-d,lautenberg-d
330 DATA bingaman-d,domenici-r,moynihan-d,damato-r,faircloth-r
340 DATA helms-r,conrad-d,dorgan-d,glenn-d,metzenbaum-d,boren-d
350 DATA nickles-r,hatfield-r,packwood-r,wofford-d,specter-r
360 DATA pell-d,chafee-r,hollings-d,thurmond-r,-daschle-d,pressler-r
370 DATA matthews-d,sasser-d,gramm-r,hutchinson-r,bennett-r,hatch-r
380 DATA leahy-d,jeffords-r,robb-d,warner-r,murray-d,gorton-r
390 DATA byrd-d,rockefeller-d,feingold-d,kohl-d,simpson-r,wallop-r
400 END
```

Parte Cuatro

Análisis, interpretación, estadísticas e inferencia

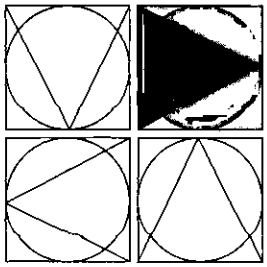


Capítulo 9
Principios del análisis e interpretación

Capítulo 10
El análisis de frecuencias

Capítulo 11
Estadística: propósito, enfoque, método

Capítulo 12
Comprobación de hipótesis y error estándar



CAPÍTULO 9

PRINCIPIOS DEL ANÁLISIS E INTERPRETACIÓN

- MEDIDAS DE FRECUENCIA Y MEDIDAS CONTINUAS
- REGLAS DE CATEGORIZACIÓN
- TIPOS DE ANÁLISIS ESTADÍSTICOS
  - **Distribuciones de frecuencia**
  - **Gráficos y elaboración de gráficos**
  - **Medidas de tendencia central y variabilidad**
  - **Medidas de relaciones**
  - **Análisis de diferencias**
  - **Análisis de varianza y métodos relacionados**
  - **Análisis de perfiles**
  - **Análisis multivariado**
- ÍNDICES
- INDICADORES SOCIALES
- LA INTERPRETACIÓN DE LOS DATOS DE INVESTIGACIÓN
  - **Adecuación de los diseños de investigación, metodología, mediciones y análisis**
  - **Resultados negativos y no concluyentes**
  - **Relaciones no hipotetizadas y hallazgos no anticipados**
  - **Prueba, probabilidad e interpretación**

El analista de la investigación descompone los datos en sus partes constituyentes para obtener respuestas a preguntas de investigación y para probar hipótesis de investigación. Sin embargo, el análisis de los datos de investigación no provee por sí mismo las respuestas a las preguntas de investigación, sino que se requiere la interpretación de los datos. Interpretar es explicar, encontrar significado. Es difícil o imposible explicar datos brutos; se debe primero analizar los datos y entonces se podrán interpretar los resultados del análisis.

Datos, como se usa en investigación del comportamiento, son los resultados de la investigación, a partir de los cuales se hacen las inferencias: generalmente son resultados numéricos, como puntuaciones de pruebas y estadísticas tales como medias, porcentajes y coeficientes de correlación. La palabra *datos* también es usada para representar los resultados de análisis matemáticos y estadísticos; pronto se estudiarán tales análisis y sus resultados. Sin embargo, los datos pueden significar algo más: puede ser información de artículos de un periódico o de una revista, material bibliográfico, diarios, etcétera, es decir, materiales verbales en general. En otras palabras, "datos" es un término general con varios significados. Piense también en los datos de investigación como los resultados de una observación y análisis sistemáticos, utilizados para hacer inferencias y sacar conclusiones. Los científicos observan, asignan símbolos y números a las observaciones y manipulan los símbolos y los números para plantearlos de forma interpretable. Posteriormente, a partir de estos datos, hacen inferencias acerca de las relaciones entre las variables de problemas de investigación.

Análisis significa la categorización, ordenamiento, manipulación y resumen de datos, para responder a las preguntas de investigación. El propósito del análisis es reducir los datos a una forma entendible e interpretable para que las relaciones de los problemas de investigación puedan ser estudiadas y probadas. Un propósito primario de la estadística, por ejemplo, es manipular y resumir los datos numéricos y comparar los resultados obtenidos con lo esperado por el azar. Un investigador hipotetiza que el estilo de liderazgo afecta la participación de los miembros del grupo de ciertas formas. El investigador planea un experimento, ejecuta el plan y recolecta datos de los sujetos. Después por medio del ordenamiento, descomposición y manipulación de los datos determina la respuesta a la pregunta ¿Cómo afectan los estilos de liderazgo a la participación de los miembros del grupo? Debe notarse que esta visión del análisis infiere que la categorización, ordenamiento y resumen de los datos debe ser planeada al inicio de la investigación. Un investigador debe trazar paradigmas de análisis o modelos aun cuando esté trabajando en el problema y en las hipótesis. Solamente de esta forma puede verse, aunque sea débilmente, si los datos y su análisis pueden de hecho contestar las preguntas de investigación.

En la *interpretación*, se toman los resultados del análisis, se hacen las inferencias pertinentes a las relaciones de investigación estudiadas y se sacan conclusiones de estas relaciones. El investigador que interpreta los resultados de investigaciones, los busca por su significado y sus implicaciones. Esto se logra de dos formas: 1) Interpretando las relaciones *dentro* del estudio de investigación y sus datos. Éste es el uso más estrecho y frecuente del término *interpretación*. Aquí, la interpretación y el análisis están estrechamente entrelazados. Casi automáticamente se interpreta cuando se hace el análisis, es decir, que cuando se calcula, digamos, un coeficiente de correlación, casi inmediatamente se infiere la existencia de una relación. 2) La búsqueda del significado más amplio de los datos de investigación. Se comparan los resultados y las inferencias realizadas a partir de los datos, con la teoría y con los resultados de otras investigaciones. Se busca el significado y las implicaciones de los resultados de la investigación dentro de los resultados del estudio y su congruencia o falta de ella, con los resultados de otros investigadores. Más importante aún, se comparan los resultados con las demandas y expectativas de la teoría.

Un ejemplo que puede ilustrar estas ideas es la investigación de cómo la forma en que se proyecta o revela cada persona hacia los demás, influye en la manera como es percibida. La teoría bajo consideración es el interpersonalismo. El interpersonalismo establece que las metas, planes y estrategias propias proveen un significado para entender a las personas y algunas de sus interacciones. Puede involucrar la construcción de modelos mentales de acción. Basados en este marco teórico general, Miller, Cooke, Tsang y Morgan (1992) predijeron que las percepciones o juicios de atribución serían determinados por la estrate-

gia de revelación que una persona decidiera adoptar. Miller *et al.*, estudiaron la diferencia entre tres tipos de revelación: negativa, positiva y jactanciosa. Se desarrollaron escenarios con diferentes métodos de revelación. Se les pidió a los participantes en el estudio que describieran su impresión de la persona en el escenario, en cinco dimensiones de atribución. Cada dimensión fue correlacionada con el tipo de escenario. La correlación (calculada en la forma de η^2[eta^2]) fue muy alta. Éste es el análisis. Los datos se delinearon en una serie de dos conjuntos de medidas, que después fueron comparadas a través de un procedimiento estadístico.[1]

El resultado del análisis —un coeficiente de correlación— debe ser ahora interpretado. ¿Cuál es su significado?, específicamente, ¿cuál es su significado dentro del estudio? ¿Cuál es su significado más amplio a la luz de los hallazgos e interpretaciones de investigaciones relacionadas? Y, ¿qué significado tiene, al confirmar o no confirmar las predicciones teóricas? Si la predicción "interna" se sostiene, entonces se relacionan los hallazgos con los de otras investigaciones, que pueden ser o no consistentes con el hallazgo presente.

La correlación fue alta, por lo que los datos de la correlación son consistentes con las expectativas teóricas. La teoría del interpersonalismo establece que las diferentes estrategias de revelación influyen en la percepción. El fanfarronear es una estrategia de revelación, por lo que ésta debe influir en la percepción. La inferencia específica es que las cosas que uno dice acerca de sí mismo influyen en la opinión que los demás tengan de esa persona. En ciertas situaciones el fanfarronear sobre sí mismo sirve para un propósito útil, la gente lo verá como confiable y exitoso. Mientras que las revelaciones negativas tenderán a hacer que la gente piense que la persona es socialmente sensible, pero no exitosa. Se miden al menos dos variables y se correlacionan. A partir del coeficiente de correlación se da un salto inferencial hacia la hipótesis. Dado que es alta (como se predijo) la hipótesis se apoya. Entonces, se intenta relacionar el hallazgo con otras investigaciones y teorías.

Medidas de frecuencia y medidas continuas

Los datos cuantitativos se presentan en dos formas generales: como medidas de frecuencia y como medidas continuas. Obviamente, las medidas continuas están asociadas con variables continuas (véase la explicación sobre variables continuas y variables categóricas en el capítulo 3). Aunque ambas clases de variables y medidas pueden ser integradas bajo el mismo marco de medición o referencia, en la práctica es necesario distinguirlas.

Frecuencias son los números de objetos en un conjunto o subconjunto. Suponga que U es el conjunto universal con N objetos. Por lo tanto N es el *número* de objetos de U. Permita que U sea fraccionada en $A^1, A^2, \ldots, A_k$. Permita que $n_1, n_2, \ldots, n_k$ sea el número de objetos en $A_1, A_2, \ldots, A_k$. Entonces $n_1, n_2, \ldots, n_k$ son llamadas frecuencias.

Resulta útil ver esto como una función. Suponga que X es cualquier conjunto de objetos con miembros $\{x_1, x_2, \ldots, x_k\}$. Se desea medir un atributo de los miembros del conjunto, el cual será llamado M. Permita que $Y = \{0, 1\}$ y que la medición sea descrita como una función: $f = \{(x, y)$; donde x es un miembro del conjunto X, y y es 1 o 0, dependiendo de que x posea o no a $M\}$. Esto se lee: f, una función o regla de correspondencia es igual al conjunto de pares ordenados (x, y), donde x es un miembro de X, y es 1 o 0, y así

[1] En el estudio de Miller, Cooke, Tsang y Morgan (1992) se empleó más que un análisis de correlación. También realizaron tanto el análisis univariado como el multivariado de varianza. η^2 mide la relación entre la variable independiente: revelación; y la(s) variable(s) dependiente(s): atribución.

sucesivamente. Si x posee M (determinado de alguna manera empírica), se le asigna un 1. Si x no posee M, se le asigna un 0. Para encontrar la frecuencia de objetos con la característica M, se cuenta el número de objetos a los que se les haya asignado el número 1.

Con medidas continuas, la idea básica es la misma. Sólo la regla de correspondencia, f, y los números asignados a los objetos cambian. La regla de correspondencia es más elaborada y los numerales son generalmente 0, 1, 2,... y fracciones de estos numerales. En otras palabras, se escribe una ecuación de medición:

$$f = \{(x, y);\ x \text{ es un objeto, y } y \text{ es cualquier numeral}\}$$

que es una forma generalizada de la función. (Esta ecuación y las ideas que la sustentan se explicarán en detalle en el capítulo 25.) Esta digresión es importante porque ayuda a ver las similitudes básicas del análisis de frecuencias y del análisis de medidas continuas.

Reglas de categorización

El primer paso de cualquier análisis es la categorización. Se dijo anteriormente (capítulo 4) que la partición es el fundamento del análisis. Ahora se explicará por qué. *Categorización* es un sinónimo de partición —es decir, una *categoría* es una partición o una subpartición—. Si un conjunto de objetos es categorizado de alguna forma, es fraccionado de acuerdo con una regla. La regla dice, en efecto, cómo asignar los objetos del conjunto a las particiones o subparticiones. Si esto es así, entonces las reglas de partición que se estudiaron con anterioridad se aplican a los problemas de categorización. Solamente es necesario explicar las reglas, relacionarlas con los propósitos básicos del análisis y ponerlas a trabajar en situaciones analíticas prácticas.

A continuación se describen las cinco reglas de categorización; la (2) y la (3) son las reglas de exhaustividad y de separación discutidas en el capítulo 4. Las otras (4 y 5), en realidad se pueden deducir de las reglas fundamentales (2) y (3). Por razones prácticas, se listan como reglas separadas.

1. Las categorías se establecen de acuerdo con el problema de investigación y sus propósitos.
2. Las categorías son exhaustivas.
3. Las categorías son mutuamente excluyentes e independientes.
4. Cada categoría (variable) se deriva de un principio de clasificación.
5. Cualquier esquema de categorización deberá de estar en un nivel de discurso.

La regla 1 es la más importante. Si las categorizaciones no se establecen de acuerdo a las demandas del problema de investigación, entonces no puede haber respuestas adecuadas a las preguntas de investigación. Hay una pregunta que constantemente debe hacerse: ¿Se ajusta el paradigma del análisis al problema de investigación? Suponga que la pregunta de investigación dice: ¿Cuál es la influencia de la televisión en las habilidades para procesar la comunicación no verbal de los niños? Se ha dicho que demasiada televisión es mala para los niños. ¿Es esto verdad? Cualesquiera que sean los datos obtenidos y el análisis realizado, éstos deben soportar el problema de investigación, que en este caso es la relación entre cantidad de televisión y comprensión de la comunicación no verbal.

La clase más simple de análisis es el análisis de frecuencia. Feldman, Coats y Spielman (1996), en su estudio sobre la cantidad de televisión que se ve y la comprensión de la comunicación no verbal, seleccionaron una muestra de niños y determinaron la frecuencia con que ellos veían la TV. Después, midieron la comprensión que cada niño tenía del

uso estratégico de las manifestaciones emocionales no verbales del personaje principal de un programa de TV. Feldman, Coats y Spielman dividieron a los niños en tres grupos, de acuerdo a la frecuencia de exposición a la TV, en: ligera, moderada y alta. Después contaron cuántos de estos niños eran capaces de ofrecer una respuesta simple o compleja acerca de la presentación visual de las emociones del personaje principal en el programa de TV que veían. El paradigma para el análisis de frecuencia era el siguiente:

Nivel de exposición a la televisión

Regla de Categorización Demostrada	Ligera	Moderada	Alta
Simple		Frecuencia	
Compleja			

Dado que Feldman *et al.* midieron la cantidad de exposición a la TV en forma continua, ellos podrían haber usado este paradigma:

Regla de Categorización Demostrada

Simple	Compleja
Cantidad de exposición a la TV	

Es obvio que los dos paradigmas soportan directamente el problema: en ambos es posible probar la relación entre la comprensión y la exposición a la televisión, si bien es cierto que de diferentes formas. Los autores eligieron el primer método —y encontraron que aquellos niños que veían menos televisión mostraban un mayor nivel de comprensión—. En el grupo de exposición ligera, el 50% de los niños dieron explicaciones complejas y heterogéneas. En el grupo de alta exposición a la TV, 0% mostró un alto nivel de comprensión. El segundo paradigma indudablemente llegaría a la misma conclusión. El punto es que un paradigma analítico es, en efecto, otra forma de formular un problema, una hipótesis y una relación. El hecho de que un paradigma utilice frecuencias mientras que el otro use medidas continuas no afecta de ninguna forma la relación probada. En otras palabras, ambas formas de análisis son lógicamente similares: ambas prueban la misma proposición pero pueden diferir en los datos usados, en pruebas estadísticas, y en sensibilidad y poder.

Hay varias cosas que un investigador podría hacer, que serían irrelevantes para el problema. Si una, dos, o tres variables son incluidas en el estudio sin una razón teórica o práctica para hacerlo, entonces el paradigma analítico sería, al menos parcialmente, irrelevante al problema. Suponga que un investigador estudia la hipótesis de que la educación religiosa aumenta el carácter moral de los niños, y para ello recolecta datos con una prueba de rendimiento en niños de escuelas públicas y de escuelas religiosas. Esto probablemente no tenga relevancia para el problema. El investigador está interesado en las diferencias morales, no en las diferencias de logro, entre los dos tipos de escuelas y entre la instrucción religiosa y la instrucción no religiosa. Podrían incluirse otras variables que tuvieran poco o nada que ver con el problema, como por ejemplo, las diferencias en la experiencia y entrenamiento del profesor o razones alumno-maestro. Si, por otro lado, el investigador piensa que ciertas variables tales como sexo, religión familiar, y quizás variables de personalidad, puedan interactuar con la instrucción religiosa para producir diferencias, enton-

ces sería justificable incorporar tales variables dentro del problema de investigación y, consecuentemente, en el paradigma analítico.[2]

La Regla 2, sobre la exhaustividad, dice que todos los sujetos u objetos de *U*, deben ser usados. Todos los individuos en el universo deben tener la posibilidad de ser asignados a las casillas del paradigma analítico. Con el ejemplo anteriormente considerado, cada niño acude a una escuela religiosa o a una escuela pública. Si, de algún modo, la muestra hubiera incluido niños que asistieran a escuelas privadas, entonces la regla habría sido violada porque habría un número de niños que no se ajustarían al paradigma del problema. (¿Cómo sería un paradigma de análisis de frecuencia? Considere a la honestidad como variable dependiente.) Sin embargo, si el problema de investigación hubiera considerado a los alumnos de escuelas privadas, entonces el paradigma tendría que cambiarse, añadiendo la categoría "privada" a las rúbricas "religiosa" y "pública".

El criterio de exhaustividad no siempre es fácil de satisfacer. Con algunas variables categóricas, no hay problema. Si el género es una de las variables, cualquier individuo ha de ser masculino o femenino. Suponga, sin embargo, que una variable bajo estudio fuera la preferencia religiosa y que se incluyera en el paradigma: Protestante-Católico-Musulmán. Ahora suponga que algunos sujetos fueran ateos o budistas. Claramente se puede observar que el esquema de categorización viola la regla de exhaustividad: algunos sujetos no tendrían casillas a donde ser asignados. Dependiendo del número de casos y del problema de investigación se podría agregar la categoría "Otras", en la cual se incluirán a los sujetos que no son ni protestantes, ni católicos, ni musulmanes. Otra solución, especialmente cuando el número de sujetos en "Otras" es pequeño, es eliminar a estos sujetos del estudio. Otra forma de solucionarlo sería ubicar a estos sujetos, si fuera posible, bajo una categoría ya existente. Algunos ejemplos de otras variables donde se encuentran estos problemas son: la preferencia política, la clase social y los tipos de educación.

La Regla 3 frecuentemente causa preocupación a los investigadores. Esta regla demanda que las categorías sean mutuamente excluyentes, lo que significa, como se aprendió antes, que cada objeto de *U*, cada sujeto de la investigación (es decir, la medición dada a cada sujeto), debe ser asignado a una casilla y solamente a una casilla de un paradigma analítico. Ésta es una función de la definición operacional. Las definiciones de las variables deben de ser claras y sin ambigüedades, de tal manera que sea poco probable que cualquier sujeto se asigne a más de una casilla. Si la preferencia religiosa es la variable que está siendo definida, entonces la definición de los subconjuntos Protestante, Católico y Musulmán debe quedar clara y sin ambigüedades. La definición podría ser: "miembro registrado en una iglesia" o también "nacido en la iglesia" y puede ser tan simple como la identificación que el propio sujeto haga de sí mismo como un protestante, un católico, o un musulmán. Cualquiera que sea la definición, debe permitirle al investigador asignar a cualquier sujeto a ***una y solamente a una*** de las casillas.

La parte de independencia de la regla 3 es a menudo difícil de satisfacer, especialmente con medidas continuas —y algunas veces con frecuencias—. ***Independencia*** significa que la asignación de un objeto a una casilla no afecte, de ninguna forma, la asignación de cualquier otro objeto a esa casilla o a cualquier otra casilla. La asignación aleatoria de un universo infinito o lo suficientemente grande, por supuesto, satisface la regla. Sin la asignación aleatoria, puede haber problemas. Cuando se asignan objetos a las casillas con base

[2] En el capítulo 6, se hicieron algunas consideraciones elementales del análisis de frecuencia con más de una variable independiente. En capítulos posteriores se hará una consideración más detallada tanto del análisis de medidas de frecuencia como del análisis de medidas continuas, con muchas variables independientes. El lector no deberá preocuparse si no logra pleno entendimiento y comprensión de los ejemplos dados anteriormente. Más adelante resultarán más claros.

en que el objeto posea ciertas características, la asignación de un objeto, ahora, puede afectar posteriormente la asignación de otro objeto.

La Regla 4, que dice que cada categoría (variable) se deriva de un principio de clasificación, algunas veces es violada por el neófito. Si se tiene una firme comprensión de la partición, este error puede evitarse fácilmente. La regla implica que, al establecer un diseño analítico, cada variable ha de ser tratada separadamente debido a que cada variable representa una dimensión separada. No se ponen dos o más variables en una categoría o en una dimensión. Si se estuvieran estudiando, por ejemplo, las relaciones entre clase social, sexo y adicción a las drogas, no se pondría a la clase social y al sexo en la misma dimensión.

Para ilustrar esto conviene citar un estudio de Glick, DeMorest y Hotze (1988). Estos investigadores estudiaron las relaciones entre la pertenencia grupal, el espacio personal y la respuesta a una solicitud de ayuda. En este estudio, cómplices de los investigadores buscaron ayuda de una persona con características físicas similares o diferentes (pertenencia a un grupo). Adicionalmente, al pedir ayuda, ellos se encontraban a poca, mediana o larga distancia (espacio personal) del sujeto. La persona a quien se acercaban respondía o no al requerimiento de ayuda. En este estudio, un error en la regla 4 podría verse de la siguiente forma:

	Dentro del grupo	Fuera del grupo	Cerca	Medio	Lejos
Accedió	Frecuencias				
Rehusó					

Claramente este paradigma viola la regla: tiene solamente una categoría derivada de dos variables. Cada variable debe tener su propia categoría. Un paradigma correcto debería verse así:

	Dentro del grupo			*Fuera del grupo*		
	Cerca	Medio	Lejos	Cerca	Medio	Lejos
Accedió	Frecuencias					
Rehusó						

La Regla 5 es la más difícil de explicar porque el término "nivel del discurso" es difícil de definir. Ya fue definido en un capítulo anterior como un conjunto que contiene todos los objetos que entran en una discusión. Si se usa la expresión "universo del discurso", se liga la idea a ideas establecidas. Cuando se habla acerca de U_1, no se trae a colación a U_2 sin una buena razón y sin haber aclarado que se está haciendo esto. Para una discusión sobre los niveles del discurso y su relevancia se puede revisar Kerlinger (1969, pp. 1127-1144, especialmente p. 1131).

El análisis de investigación generalmente mide la variable dependiente: por ejemplo, considere el problema de la pertenencia a un grupo, el espacio personal y la respuesta a una solicitud de ayuda. La pertenencia a un grupo y el espacio personal son las variables independientes; responder a una solicitud de ayuda es la variable dependiente. Los objetos de análisis son las medidas de la respuesta a la solicitud de ayuda. Las variables independientes y sus categorías son realmente usadas para estructurar el análisis de la variable dependiente. El universo del discurso, U, es el conjunto de medidas de la variable dependiente. Las variables independientes pueden percibirse como los principios de partición

usados para fraccionar las medidas de la variable dependiente. Si súbitamente cambiamos a otro tipo de medida de la variable dependiente, entonces habremos cambiado los niveles o universos del discurso.

Tipos de análisis estadísticos

Hay muchos tipos de análisis estadísticos y de presentación que no pueden exponerse en detalle en este libro. Explicaciones posteriores de ciertas formas más avanzadas de análisis estadísticos, tienen como propósito la comprensión básica de la estadística y la inferencia estadística, y la relación de la estadística y la inferencia estadística con la investigación: Aquí, las formas más sofisticadas de análisis estadísticos se exponen brevemente para dar al lector un panorama del tema; sin embargo, se explican solamente en su relación con la investigación. Se asume que el lector ya ha estudiado la estadística descriptiva más simple. Aquellos que no lo han hecho, podrán encontrar buenas explicaciones en libros de texto elementales (Comrey y Lee, 1995; Kirk, 1990; Howell, 1997; Hays, 1994).

Distribuciones de frecuencia

Aunque las distribuciones de frecuencia se usan principalmente para propósitos descriptivos, también pueden ser usadas para otros propósitos de investigación. Por ejemplo, se puede probar si dos o más distribuciones son lo suficientemente similares para garantizar su unión. Suponga que se estudia el aprendizaje verbal de niños y niñas de 6° grado. Después de obtener un gran número de puntuaciones de aprendizaje verbal, pueden compararse y probar las diferencias entre las distribuciones de niños y niñas. Si la prueba muestra que las distribuciones son iguales —y otros criterios se satisfacen— entonces quizás puedan ser combinadas para otros análisis.

Las distribuciones observadas también pueden ser comparadas con distribuciones teóricas. La comparación más conocida de este caso es la llamada *distribución normal*. Puede ser importante saber que las distribuciones obtenidas son normales en forma, o si no son normales, se desvían de la normalidad en ciertas formas específicas. Dicho análisis puede ser útil en otros trabajos teóricos y aplicados, así como en la investigación. En un estudio teórico de habilidades es importante saber si estas habilidades están, de hecho, distribuidas normalmente. Dado que se ha encontrado que muchas de las características humanas se distribuyen normalmente (ver Anastasi, 1958),[3] los investigadores pueden hacer preguntas importantes acerca de características "nuevas" que están siendo investigadas.

La investigación educativa aplicada puede valerse del cuidadoso estudio de la distribución de la inteligencia, las aptitudes y las puntuaciones de aprovechamiento. ¿Es concebible que un programa de aprendizaje innovador pueda cambiar las distribuciones de las puntuaciones de aprovechamiento, digamos, de los niños de 3° y 4° grado? ¿Podría ser que los programas de educación masiva temprana pudieran cambiar los perfiles de las distribuciones, así como los niveles generales de las puntuaciones?

Allport (1947), en su estudio sobre el conformismo social, mostró que aun un fenómeno complejo de comportamiento como el conformismo, podría estudiarse provechosa-

[3] El estudiante de investigación en educación, psicología y sociología debe estudiar la sobresaliente contribución de Anastasi a la comprensión de las diferencias individuales. Su libro también contiene muchos ejemplos de distribuciones de datos empíricos.

mente usando el análisis de distribución. Allport fue capaz de demostrar que muchas de las conductas sociales —parar frente a una luz roja, infracciones de estacionamiento, observancia religiosa, etcétera— se distribuían en forma de una curva *J*, donde la mayoría de las personas son conformistas, pero con un número pequeño predecible de inconformistas, en diferentes grados. Coren, Ward y Enns (1994) presentan numerosas distribuciones con diferentes formas para ciertas percepciones humanas de estímulos físicos basados en la Ley Psicofísica de Steven.

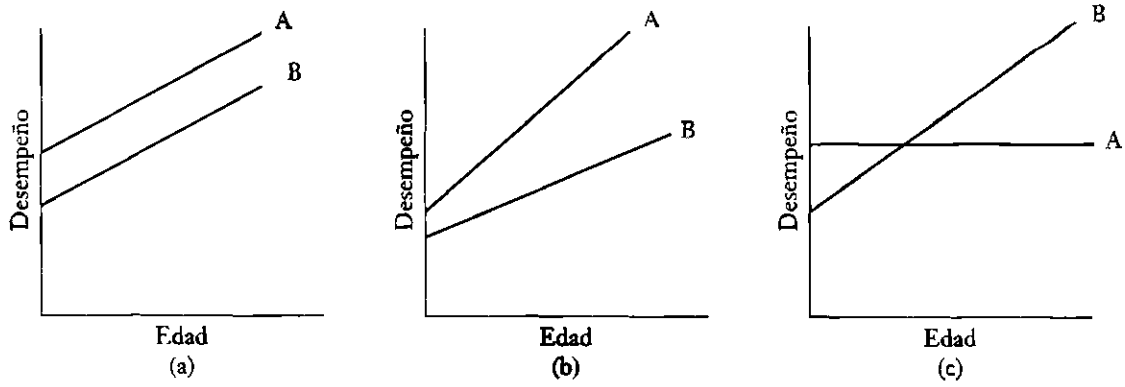
Las distribuciones han sido, probablemente, muy poco usadas en las ciencias del comportamiento y en la educación: el estudio de las relaciones y las pruebas de hipótesis son casi automáticamente asociadas con correlaciones y comparaciones de promedios. El uso de distribuciones es considerado con menos frecuencia. Algunos problemas, sin embargo, pueden ser mejor resueltos usando los análisis de distribución. Estudios de patología y de otras condiciones inusuales son quizás mejor abordados a través de la combinación de los análisis de distribución y los conceptos probabilísticos.

Gráficos y elaboración de gráficos

Una de las más poderosas herramientas del análisis es el gráfico. Un *gráfico* es una representación bidimensional de una relación o relaciones. Exhibe gráficamente conjuntos de pares ordenados en una forma que ningún otro método puede hacerlo. Si existe una relación en un conjunto de datos, un gráfico no sólo la mostrará claramente, sino que también mostrará su naturaleza: positiva, negativa, lineal, cuadrática, etcétera. Aunque los gráficos han sido usados con frecuencia en las ciencias del comportamiento, al igual que las distribuciones, al parecer no han sido lo suficientemente utilizados. Para estar seguros, hay formas objetivas de resumir y probar relaciones, tales como coeficientes de correlación, comparación de medias y otros métodos estadísticos, sin embargo, ninguno de éstos describe tan vívidamente una relación como un gráfico.

Revisando los gráficos del capítulo 5 (figuras 5.1, 5.4, 5.5 y 5.6), se puede notar cómo transmiten la naturaleza de las relaciones. Posteriormente se usarán gráficos de manera más interesante para mostrar la naturaleza de relaciones más complejas entre variables. Para dar al estudiante sólo una muestra de la riqueza de tal análisis, anticiparemos una discusión posterior; de hecho, intentaremos enseñar una idea compleja usando gráficos.

FIGURA 9.1



Los tres gráficos de la figura 9.1 muestran tres relaciones hipotéticas entre la edad, como variable independiente y el desempeño verbal (etiquetado "desempeño") como variable dependiente, de niños de clase media (A) y niños de clase trabajadora (B). Podría llamarse a estos gráficos, gráficos de desarrollo. El eje horizontal es la abscisa y se usa para indicar la variable independiente o *X*. El eje vertical es la ordenada y es usado para indicar la variable dependiente o *Y*. El gráfico (a) muestra la misma relación positiva entre edad y desempeño tanto en la muestra A, como en la muestra B. También muestra que los niños de la muestra A superan a los niños de la muestra B. El gráfico (b) muestra que ambas relaciones son positivas, pero que conforme pasa el tiempo, el desempeño de los niños de la muestra A se incrementó más que el desempeño de los niños de la muestra B. El gráfico (c) es más complejo. Muestra que los niños de la muestra A superaron a los niños de la muestra B en una etapa temprana y que se mantuvieron así hasta una edad mayor, pero los niños de la muestra B, que iniciaron más bajos, avanzaron y continuaron su avance en el tiempo hasta que superaron a los niños de la muestra A. Este tipo de relación es poco probable en el desempeño verbal, pero puede ocurrir con otras variables.

El fenómeno mostrado en los gráficos (b) y (c) es conocido como *interacción*. Brevemente, implica que dos o más variables interactúan en su "efecto" sobre una variable dependiente. En este caso, la edad y el estatus del grupo interactúan en su relación con el desempeño verbal. Expresado de otra forma, la interacción significa que la relación de una variable independiente con una variable dependiente difiere en grupos distintos, como en este caso, o a diferentes niveles de otra variable independiente. El estudio de Behling y Williams (1991) arrojó resultados que podrían ser graficados como en el gráfico (b). En este estudio los investigadores examinaron la percepción del estudiante y del maestro de la inteligencia a partir de diferentes estilos de vestuario de hombres y mujeres. Un gráfico de un estilo de vestuario está en la figura 9.2. Un gráfico similar al gráfico (c) puede construirse a partir de los datos proporcionados por Little, Sterling y Tingstrom (1996). Su estudio involucra la percepción que tenían estudiantes del norte y del sur de Estados Unidos de una persona objetivo descrita, ya sea como norteña o como sureña. Así, los estudiantes del sur dieron puntuaciones similares en el diferencial semántico, tanto a las personas del norte como a las del sur. Sin embargo, los estudiantes del norte dieron a las personas del norte puntuaciones mucho más altas que a aquellas del sur. Esto se ilustra en la figura 9.3. La noción de un efecto de interacción se explicará en detalle y con precisión cuando se estudie el análisis de varianza y el análisis de regresión múltiple.

FIGURA 9.2

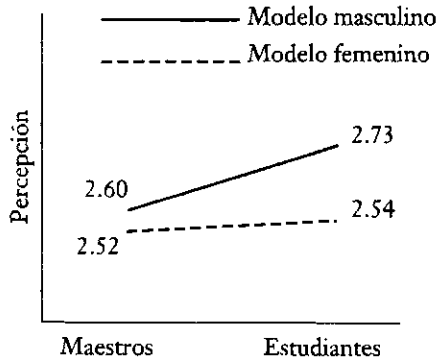
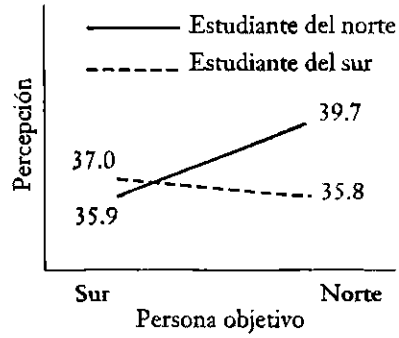


FIGURA 9.3



Mientras que las medias representan una de las mejores formas para reportar datos complejos, confiar completamente en ellos puede ser desafortunado. La mayoría de los casos de diferencias significativas de medias entre grupos, también se acompañan por un considerable traslape de las distribuciones. Anastasi da ejemplos claros (1958) y señala la necesidad de prestar atención a los traslapes y da ejemplos y gráficos de las diferencias de la distribución de sexo, entre otras. En pocas palabras, se les aconseja a los estudiantes de investigación que adquieran el hábito, desde el inicio de su estudio, de prestar atención y comprender las distribuciones de variables y de graficar relaciones de variables.

Medidas de tendencia central y variabilidad

Prácticamente no hay duda de que las medidas de tendencia central y variabilidad son las herramientas más importantes en el análisis de datos conductuales. Dado que gran parte de este libro se ocupará de tales medidas —de hecho, hay toda una sección llamada "El análisis de varianza"— aquí solamente se caracterizarán promedios y varianzas. Los tres promedios principales (o medidas de tendencia central) usadas en investigación —media, mediana y moda— son resúmenes de los conjuntos de las medidas, a partir de los cuales se calculan. Los conjuntos de medidas son demasiado vastos y complejos para poder entenderlos de inmediato. Ellos están "representados" o resumidos por medidas de tendencia central. Éstas indican qué conjuntos de medidas "son parecidos" en promedio, pero también son comparados para probar relaciones. Más aún, las puntuaciones individuales pueden ser útiles, comparadas con aquellas que evalúan el estatus del individuo. Se dice, por ejemplo, que la puntuación individual de A está a tal y tal distancia de la media.

Mientras que la media es el promedio más usado en investigación y sus propiedades son tan deseables que justifican su posición preeminente, por otro lado, la mediana (la medida más medial de un conjunto de medidas) y la moda (la medida más frecuente) pueden algunas veces ser útiles en investigación. Por ejemplo, la mediana, además de ser una importante medida descriptiva, puede ayudarse en pruebas de significancia estadística donde la media es inapropiada (véase Bradley, 1968). El estudio de Allman, Walker, Hart, Laprade, Noel y Smith (1987), donde se comparó la efectividad y los efectos adversos de los colchones de aire y la terapia convencional en pacientes hospitalizados con úlceras de presión, sirve como un buen ejemplo del uso de la mediana como la medida primaria de tendencia central. La moda es usada principalmente para propósitos descriptivos, pero

puede ser útil en investigación para estudiar características de poblaciones y relaciones. Suponga que una prueba de aptitud matemática se aplicó a todos los aspirantes a ingresar a una universidad que apenas había abierto sus admisiones y que la distribución de las puntuaciones resultó ser bimodal. Suponga, además, que solamente se calculó una media y se comparó con las medias de los años anteriores, resultando ser considerablemente menor. La conclusión simple de que la aptitud matemática promedio de los aspirantes fue considerablemente menor que en años previos, oculta el hecho de que debido a la política abierta de admisión muchos aspirantes con antecedentes deficientes en matemáticas fueron admitidos. Aunque éste es un ejemplo obvio y escogido deliberadamente, debe observarse que el hecho de oscurecer fuentes importantes de diferencias puede ser más sutil. A menudo es útil en investigación calcular las medianas y las modas, así como las medias.[4]

Las principales medidas de variabilidad son la varianza y la desviación estándar. Éstas ya se han estudiado y se estudiarán en capítulos posteriores y sólo cabe decir que los reportes de investigación siempre deben incluir medidas de variabilidad. Las medias no deben reportarse sin desviaciones estándar (tampoco sin N, el tamaño de la muestra), ya que una adecuada interpretación de la investigación es virtualmente imposible sin los índices de variabilidad. Otra medida de variabilidad que en años recientes ha adquirido mayor importancia es el *rango*: la diferencia entre la medida más alta y la más baja de un conjunto de medidas. Ahora es posible, especialmente con muestras pequeñas (con N de 20, 15 o menos), usar el rango en pruebas de significancia estadística.

Medidas de relaciones

Hay muchas medidas útiles de relaciones: el coeficiente de correlación producto-momento (r), el coeficiente de correlación de rangos ordenados (*rho*), la razón de correlación (eta: η), la medida de distancia (D), el coeficiente phi (ϕ), el coeficiente de correlación múltiple (R), etcétera. Casi todos los coeficientes de correlación sin importar qué tan diferentes sean en derivación, apariencia, cálculo y uso, hacen en esencia lo mismo: expresan la extensión en que los pares de conjuntos de pares ordenados varían concomitantemente; informan al investigador la magnitud y (generalmente) la dirección de la relación. El valor de algunos varía de –1.00 a +1.00 pasando por 0, donde –1.00 y 1.00 indican una asociación negativa y positiva perfecta respectivamente, y el 0 indica una relación no discernible.

Las medidas de relación son, comparativamente, índices directos de relaciones, en el sentido de que a partir de ellas se adquiere una idea directa del grado de covariación de las variables. El cuadrado del coeficiente de correlación producto-momento, por ejemplo, es un estimado directo de la cantidad de varianza compartida por las variables. Se puede decir, al menos de forma general, qué tan alta o qué tan baja es la relación. Esto contrasta con las medidas de significancia estadística que indican si una relación es o no "significativa" a un nivel específico de significancia. Idealmente, cualquier análisis de datos de investigación debe incluir ambas clases de índices: medidas de significancia de una relación y medidas de la magnitud de la relación.

Las medidas de relación, pero sobre todo los coeficientes de correlación producto-momento, son poco usuales en cuanto a que están sujetos a formas extensas y elaboradas de análisis, principalmente análisis de regresión múltiple y análisis factorial (que se revisarán en capítulos posteriores). Por lo tanto, son herramientas extremadamente útiles y poderosas para el investigador.

---

[4] Diversos tipos de medias y de otras medidas de tendencia central son excepcionalmente bien explicadas en el libro de Tate (1955), un viejo pero valioso documento. Él también da un buen número de ejemplos de distribuciones y gráficos de varios tipos.

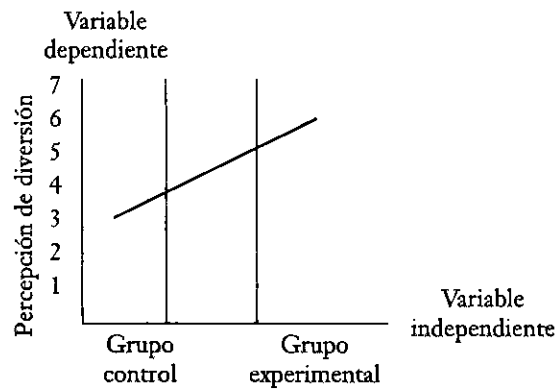
Análisis de diferencias

Los análisis de diferencias, particularmente el análisis de diferencias entre medias, ocupa una parte muy importante del análisis estadístico y de la inferencia. Es importante observar dos cosas acerca de los análisis de diferencias. Primero, por ningún motivo están confinados a las diferencias de medidas de tendencia central. Casi cualquier clase de diferencia puede ser analizada —entre frecuencias, proporciones, porcentajes, rangos, correlaciones y varianzas—. Consideremos las varianzas. Suponga que un psicólogo educativo desea averiguar si cierta forma de instrucción tiene el efecto de hacer a los alumnos más heterogéneos en el aprendizaje de conceptos. La diferencia entre las varianzas de los grupos que recibieron enseñanza por diferentes métodos puede ser probada fácilmente. También puede investigarse si grupos considerados homogéneos, lo son también en variables distintas a las usadas para formar los grupos (véase Comrey y Lee, 1995, pp. 229-234; Mattson, 1986, capítulo 10).

El segundo punto es más importante. Todos los análisis de diferencias son planeados con el propósito de estudiar relaciones. Suponga que alguien cree que el modificar la cantidad de narcisismo tendrá un efecto en las relaciones interpersonales. Carroll, Hoeningmann, Stovall y Whitehead (1996) crearon tres protocolos con diferentes niveles de narcisismo —extremo, moderado y ninguno— y después evaluaron el atractivo de los participantes hacia esa persona. La hipótesis de los investigadores se sustentó, ya que los participantes reportaron un mayor rechazo de la persona con un narcisismo extremo, que de aquellas con otros niveles de narcisismo, sin embargo lo que realmente interesa aquí no son estas diferencias, sino la relación entre el cambio de los niveles narcisismo y su efecto en cómo la gente percibe a la persona; entonces, las diferencias entre las medias realmente reflejan una relación entre la variable independiente y la variable dependiente. Si no hay diferencias significativas entre las medias, la correlación entre la variable independiente y la variable dependiente es 0; y, a la inversa, entre más grandes sean las diferencias, más alta será la correlación, siempre y cuando lo demás permanezca igual.

En el experimento de Strack, Martin y Stepper (1988), se estudió el efecto de la actividad facial de la gente en sus respuestas afectivas. Las personas en el grupo experimental recibieron instrucciones de sostener bolígrafos en la boca, usando solamente sus dientes, mientras veían caricaturas. Las personas del grupo control recibieron instrucciones de sostener los bolígrafos con sus labios, mientras veían el mismo estímulo. El grupo experi-

FIGURA 9.4



mental tuvo una media de 5.09 en una escala de 0 a 9 en la que evaluaban qué tan "divertida" era la caricatura. Sin embargo, el grupo control tuvo una media de 3.90. La diferencia es estadísticamente significativa, por lo que se puede concluir que hay una relación entre los músculos faciales utilizados (y aquellos no utilizados), y la diversión percibida. En capítulos anteriores se graficaron las relaciones entre variables medidas para mostrar la naturaleza de las relaciones. También es posible graficar la presente relación entre la variable independiente experimental (manipulada) y la variable dependiente. Esto se hizo en la figura 9.4, donde las medias se trazaron como se indica. Mientras que el trazo es más o menos arbitrario —por ejemplo, no hay unidades reales de línea base para la variable independiente—, la similitud con los gráficos previos es notoria y la idea básica de una relación ahora es clara.

Si el lector conserva siempre en mente que las relaciones son conjuntos de pares ordenados, la similitud conceptual de la figura 9.4 con gráficos anteriores será evidente. En los gráficos previos, cada miembro de cada par representa una puntuación. En la figura 9.4, un par ordenado consta de un tratamiento experimental y de una puntuación. Si se asigna el valor 1 al grupo experimental y 0 al grupo control, dos pares ordenados pueden ser los siguientes: (1, 5.09), (0, 3.90).

Análisis de varianza y métodos relacionados

Una buena parte de este libro está dedicada al análisis de varianza y los métodos relacionados con éste, por lo que no es necesario discutir mucho de esto por el momento. El lector necesita solamente tener en perspectiva este importante método de análisis. El análisis de varianza es lo que su nombre implica y más: un método para identificar, analizar y probar la significancia estadística de varianzas que provienen de diferentes fuentes de variación; es decir, que una variable dependiente tiene una cantidad total de varianza, parte de la cual es debida al tratamiento experimental, parte al error y parte a otras causas. El papel del análisis de varianza es trabajar con estas diferentes varianzas y fuentes de varianza. Estrictamente hablando, el análisis de varianza es más apropiado para datos experimentales que para datos no experimentales, aunque su inventor Fisher (1950), lo utilizó con ambos. Se considera, entonces, que el análisis de varianza es un método para el análisis de datos recabados en experimentos donde se utiliza, al menos, la aleatorización y manipulación de una variable independiente.

Probablemente no haya mejor forma de estudiar el diseño de investigación que a través del enfoque del análisis de varianza. Aquellos expertos en este enfoque, casi automáticamente piensan en modelos alternativos de análisis de varianza cuando se enfrentan a nuevos problemas de investigación. Tomemos el estudio de Rozin, Nemeroff, Wane y Sherrod (1989) sobre la ley del contagio. Esta ley establece que los objetos que han estado en contacto unos con otros, pueden continuar influenciándose entre sí a través de la transferencia de alguna de sus propiedades. Estos autores construyeron seis objetos (suéter, hamburguesa, manzana, cepillo de cabello [recibido], cepillo de cabello [dado] y un mechón de cabello) y los pusieron en contacto con cuatro diferentes personas (amigo, novio, alguien antipático y alguien indiferente). Se consideró a las cuatro diferentes personas como cuatro niveles de la variable categórica: fuente de origen. Se pidió a los participantes que evaluaran cada objeto en una escala de –100 a +100, donde –100 era "la cosa más desagradable que pueda usted imaginar" y +100 era "la cosa más agradable que usted pueda imaginar". Cero era el punto neutral. Rozin *et al.* (1989) analizaron los datos con un análisis varianza de un factor para cada objeto, usando la fuente de origen (diferentes personas que habían estado en contacto con el objeto) como la variable independiente. El análisis resultaría como el paradigma marcado (a) de la figura 9.5. Si los investigadores

FIGURA 9.5

(a)

Fuente de origen (personas en contacto con el objeto)			
Amigo	**Novio**	**Alguien antipático**	**Alguien indiferente**
Evaluaciones placenteras			

(b)

Objetos	Fuente de origen (personas en contacto con el objeto)			
	Amigo	**Novio**	**Alguien antipático**	**Alguien indiferente**
Suéter Hamburguesa Manzana Cepillo de cabello (recibido) Cepillo de cabello (dado) Mechón de cabello	Evaluaciones placenteras			

FIGURA 9.6

(a)

Proveedores de servicios de emergencia		
Médicos	**Enfermeras**	**Personal prehospitalario**
Conteo de bacteria estafilocóxica		

(b)

Turno laboral	Proveedores de servicios de emergencia		
	Médicos	**Enfermeras**	**Personal prehospitalario**
Día Noche	Conteo de bacteria estafilocóxica		

hubieran usado los objetos como otra variable independiente, pensando que los objetos afectan las evaluaciones de las personas, entonces el paradigma se vería como el marcado (b), que es una análisis de varianza de dos factores. Es claro que el análisis de varianza es un método importante para estudiar diferencias.

De la misma forma, un estudio de Jones, Hoerle y Riekse (1995) comparó la extensión de la presencia de la bacteria estafilococo en los estetoscopios de los proveedores de servicios de emergencia. Se utilizó un análisis de varianza de un factor para hacer esta comparación entre los médicos, enfermeras y personal prehospitalario. La variable independiente era el tipo de proveedor del servicio de emergencia y la variable dependiente era el conteo de bacteria estaflocóxica. La figura 9.6 (a) muestra el paradigma usado en este estudio. Si estos investigadores consideraran que podría haber una diferencia entre los proveedores de servicios de emergencia en los diferentes turnos de trabajo, el paradigma se expresaría como el que se encuentra en la figura 9.6 (b).

Análisis de perfiles

El *análisis de perfiles* es básicamente la evaluación de similitudes en los perfiles de individuos o grupos. Un *perfil* es un conjunto de medidas diferentes de un individuo o grupo, donde

cada una está expresada en la misma unidad de medida. Las puntuaciones de un individuo en un conjunto de diferentes pruebas constituye un perfil, siempre y cuando todas las puntuaciones hayan sido convertidas a un sistema común de medida, como percentiles, rangos y puntuaciones estándar. Los perfiles se han utilizado principalmente con propósitos diagnósticos —por ejemplo los perfiles de puntuaciones de una batería de pruebas son usados para evaluar y asesorar a alumnos de preparatoria—. Sin embargo, el análisis de perfiles ha incrementado su importancia en la investigación sociológica y psicológica, tal como se verá posteriormente cuando se estudie entre otras cosas, la metodología Q.

El análisis de perfiles tiene problemas especiales que requieren consideraciones cuidadosas por parte del investigador. La similitud, por ejemplo, no es una característica general de las personas; solamente hay similitud en características específicas o complejos de características (véase Cronbach y Gleser, 1953). Otra dificultad estriba en qué tipo de información se está dispuesto a sacrificar al calcular los índices de similitud de perfiles. Cuando se utiliza el coeficiente de correlación producto-momento —que es una medida de perfiles—, se pierde nivel, es decir, que se sacrifican las diferencias entre las medias. Esto es una pérdida de *elevación*. Las *r*s producto-momento sólo toman en cuenta la *forma*. Más aún, la *dispersión* (las diferencias en la variabilidad de los perfiles) se pierde al calcular otras clases de medidas de perfiles. En pocas palabras, la información puede perderse, y de hecho, se pierde. El estudiante encontrará una excelente ayuda y guía para el análisis de perfiles en el libro de Nunnally y Bernstein (1993) de psicometría, aunque el tratamiento no es elemental.

Análisis multivariado

Quizás las formas más importantes del análisis estadístico, especialmente en el estado actual del desarrollo de las ciencias del comportamiento, son los análisis multivariados y los análisis factoriales. *Análisis multivariado* es un término general usado para categorizar una familia de métodos analíticos cuya característica principal es el análisis simultáneo de k variables independientes y m variables dependientes. En este libro no nos preocuparemos demasiado acerca de la terminología usada en el análisis multivariado. Para algunos, el análisis multivariado incluye al análisis factorial y otras formas de análisis, como el análisis de regresión múltiple. *Multivariado*, para estas personas infiere más de una variable independiente o más de una variable dependiente, o *ambos*. Otros en el medio usan "análisis multivariado" solamente en el caso de que ambas, la variable dependiente y la variable independiente, sean múltiples. Si un análisis incluye, por ejemplo, cuatro variables independientes y dos variables dependientes, manejadas simultáneamente, entonces es un análisis multivariado.

Puede argumentarse que de todos los métodos de análisis, los métodos multivariados son los más poderosos y apropiados para la investigación científica del comportamiento. El argumento que apoya esta afirmación es muy amplio y complejo y nos apartaría del principal propósito. Básicamente descansa en la idea de que los problemas de investigación del comportamiento son, casi todos, de naturaleza multivariada y que no pueden ser resueltos con un enfoque bivariado (de dos variables), esto es, un enfoque que considere solamente una variable independiente y una variable dependiente a la vez. Esto ha quedado muy claro en mucha de la investigación educativa donde, por ejemplo, los determinantes del aprendizaje y aprovechamiento son complejos: inteligencia, motivación, clase social, instrucción, atmósfera escolar y del salón de clases, la organización escolar, etcétera. Evidentemente variables como éstas interactúan unas con otras y algunas veces unas contra otras, de maneras desconocidas, pero afectan el aprendizaje y el aprovechamiento. En otras palabras, para explicar los complejos fenómenos psicológicos o sociológicos de la

educación, se requiere de herramientas de diseño y análisis que sean capaces de manejar la complejidad que manifiestan, por sí mismas, las múltiples variables independientes y variables dependientes. Un argumento similar puede darse para la investigación psicológica y sociológica.

Este argumento y la realidad que subyace, imponen una pesada carga en aquellos individuos que enseñan y aprenden métodos y enfoques de investigación. Es poco realista e irresponsable estudiar y aprender solamente un enfoque que es básicamente bivariado en su concepción. Los métodos multivariados, sin embargo, son como la realidad conductual que tratan de reflejar: complejos y difíciles de entender. La necesidad pedagógica, en cuanto a este libro concierne, es tratar de expresar los fundamentos del pensamiento de investigación, diseño, métodos y análisis, principalmente a través de un enfoque bivariado modificado. Se extenderá este enfoque, tanto como sea posible, a los conceptos y métodos multivariados, esperando que el estudiante busque más adelante, después de haber recibido los fundamentos adecuados.

La *regresión múltiple*, que es probablemente la forma más útil de los métodos multivariados, analiza las influencias comunes y separadas de dos o más variables independientes sobre una variable dependiente. Esta afirmación tiene limitaciones, especialmente acerca de las contribuciones separadas de las variables independientes, lo que será discutido en el capítulo 33. El aumento del uso de la regresión múltiple como herramienta analítica de las ciencias del comportamiento se debe en su mayor parte, a las computadoras digitales de alta velocidad. Ezekiel y Fox (1959) son dos de los pocos autores cuyos libros sobre regresión múltiple, previa al gran uso de las computadoras, están disponibles. Ezekiel y Fox resumieron los estudios que utilizaron regresión múltiple, antes de la publicación de su libro en 1959. No había muchos de ellos. A partir de la gran disponibilidad de las computadoras y los programas de estadística, el número de estudios que usan la regresión múltiple se ha incrementado exponencialmente. Erlich y Lee (1978) propusieron un uso novedoso del análisis de regresión; lo utilizaron en puntuaciones de pruebas con el propósito de evaluar la responsabilidad educativa. Por otro lado, Griffiths, Bevil, O'Connor y Wieland (1995) usaron la regresión para predecir el nivel de competencia en un examen de anatomía y fisiología; entre las variables predictoras estaban el promedio, así como el tipo de escuela donde habían tomado los cursos propedéuticos de anatomía y fisiología. El método ha sido usado en cientos de estudios, probablemente por su flexibilidad, poder y aplicabilidad general a muchos tipos diferentes de problemas de investigación. (¡También tiene limitaciones!) Por eso, no puede ignorarse en este libro. Afortunadamente, no es tan difícil de entender y aprender a usarlo —dado el suficiente interés y deseo para hacerlo—.

La *correlación canónica* es una extensión lógica de la regresión múltiple. De hecho, es un método de regresión múltiple. Añade más de una variable dependiente al modelo de regresión múltiple; en otras palabras, maneja las relaciones entre conjuntos de variables independientes y conjuntos de variables dependientes por lo que es, teóricamente, un poderoso método de análisis. Sin embargo, tiene limitaciones que pueden restringir su utilidad, tales como la interpretación de los resultados que produce y su capacidad limitada para probar modelos teóricos.

El *análisis discriminante* también está estrechamente relacionado con la regresión múltiple. Como su nombre lo indica, su propósito es discriminar grupos entre sí con base en conjuntos de medidas. También es útil en asignar individuos a grupos, con base en sus puntuaciones en pruebas. Aunque esta explicación no es adecuada, será suficiente por ahora.

En esta etapa es difícil caracterizar, aun en un nivel superficial, la técnica conocida como *análisis multivariado de varianza* porque aún no se ha revisado el análisis de varianza. Por lo anterior se pospone su discusión.

El *análisis factorial* es esencialmente diferente en su clase y en su propósito de otros métodos multivariados. Su propósito fundamental es ayudar al investigador a descubrir e identificar las unidades o dimensiones llamadas *factores* que subyacen a muchas medidas. Por ahora se sabe, por ejemplo, que detrás de muchas medidas de habilidad e inteligencia subyacen algunas dimensiones generales o factores. La aptitud verbal y la aptitud matemática son dos de los factores más conocidos. Al estudiar actitudes sociales se han encontrado factores religiosos, económicos y educativos.

Los métodos multivariados mencionados anteriormente son "estándar" en el sentido de que a ellos nos referimos generalmente al usar el término "métodos multivariados". Sin embargo, hay otros métodos multivariados de igual, e inclusive, mayor importancia. Como se dijo en el prefacio, en un libro de esta naturaleza no es posible dar una explicación técnica adecuada y correcta de todos los métodos multivariados. Por ejemplo aunque tienen una enorme importancia, el análisis estructural de covarianza y el análisis de modelos log-lineales pueden ser demasiado complejos y difíciles de describir y explicar de una forma adecuada y completa. Lo mismo sucede con el método de análisis multidimensional y con el análisis de ruta, que no pueden ser presentados adecuadamente. ¿Entonces, qué se va a hacer? Algunos de estos enfoques y procedimientos son tan poderosos e importantes —de hecho están revolucionando la investigación conductual— que un libro que los ignore será un texto deficiente. La solución al problema fue también expuesto en el prefacio, y vale la pena repetirlo. Los enfoques más comunes y accesibles —el análisis de varianza, la regresión múltiple y al análisis factorial— serán presentados con suficientes detalles técnicos para permitir a un estudiante motivado y entusiasta aplicarlos e interpretar sus resultados. Otros métodos más complicados (como el análisis estructural de covarianza y los modelos log-lineales) serán descritos y explicados "conceptualmente" en sus propósitos y razonamientos con generosas citas y descripciones de investigaciones ficticias y reales. Tal enfoque será usado en capítulos posteriores con las siguientes tres metodologías.

El *análisis de ruta* es un método gráfico del estudio de las supuestas influencias directas e indirectas de las variables independientes entre sí y sobre las variables dependientes. En otras palabras, es un método para describir y probar "teorías" (véase Kerlinger y Pedhazur, 1973; Pedhazur, 1996). Quizás su principal virtud es que requiere que los investigadores expliciten el marco teórico de los problemas de investigación. Para lograr sus objetivos, el análisis de ruta usa los llamados diagramas causales o diagramas de ruta y el análisis de regresión. Los lectores pueden satisfacer un poco su curiosidad examinando, en el capítulo 34, uno o dos de los ejemplos del análisis de ruta que se dan ahí. El análisis de ruta ha sido un marco conceptual útil para explicar las relaciones entre variables. El autor acreditado en desarrollar el análisis de ruta fue Wright (1921). Las aplicaciones de Wright fueron en el campo de la genética. Duncan (1966) y Blalock (1971) popularizaron el trabajo de Wright en las ciencias del comportamiento. Es útil estudiar el análisis de ruta porque ayuda a entender más fácilmente el análisis estructural de covarianza. De hecho, el análisis de ruta forma parte del análisis estructural de covarianza, como se verá en un capítulo posterior.

El *análisis estructural de covarianza* —o modelamiento causal,[5] o modelos de ecuaciones estructurales— es el último enfoque del análisis de estructuras complejas de datos. Este método implica, principalmente, el análisis de variación conjunta de variables que están en una estructura dictada por la teoría. Por ejemplo, se puede estudiar la adecuación de las teorías de inteligencia mencionadas en capítulos anteriores, ajustando las teorías al marco del análisis estructural de covarianza para después evaluar qué tan bien pueden explicar los

[5] El término "modelamiento causal" ya no está en voga, pues ciertos estadistas prominentes señalaron que los análisis basados en la correlación, en las ciencias sociales y conductuales, no podían establecer causa y efecto.

datos reales de una prueba de inteligencia. El método —o más bien, la metodología— es una síntesis matemática y estadística ingeniosa del análisis factorial, la regresión múltiple, el análisis de ruta y la medición psicológica en un único sistema exhaustivo que puede expresar y probar formulaciones teóricas complejas de problemas de investigación. Su creación se atribuye a Joreskog (1970) y a sus asociados, aunque Bentler (1989) ha propuesto el método con fuerza, creando un algoritmo diferente (EQS) al de Joreskog (LISREL).

Los *modelos log-lineales* representan el método multivariado más reciente (o metodología) para analizar datos de frecuencia. Los métodos multivariados mencionados anteriormente están orientados sobre todo a analizar datos obtenidos de medidas continuas: puntuaciones de pruebas, medidas de escalas de aptitudes y personalidad, medidas de variables ecológicas y otros aspectos similares. Como se verá en el siguiente capítulo, los datos de la investigación conductual aparecen ocasionalmente como frecuencias, en especial de individuos, por ejemplo, número de hombres y mujeres, de minorías étnicas y minorías no étnicas, maestros y no-maestros, individuos de clase media y clase trabajadora; y católicos, protestantes y musulmanes. El análisis log-lineal hace posible estudiar combinaciones complejas de dichas variables nominales y, como el análisis estructural de covarianza, probar teorías de las relaciones e influencias de tales variables entre sí. Brevemente se caracterizará la metodología en un capítulo posterior, aunque el reducido espacio y las dificultades técnicas forzarán a limitar la explicación a las ideas básicas involucradas. Se verá, al menos, que como el análisis estructural de covarianza, es uno de los desarrollos metodológicos más poderosos e importantes de la última parte del siglo XX.

Índices

Un *índice* puede definirse de dos formas. Primero, un índice es un fenómeno observable que es sustituido por un fenómeno menos observable. Un termómetro, por ejemplo, da lectura de números que representan grados de temperatura; los números en un velocímetro indican a cuántos kilómetros por hora está avanzando un vehículo. Las puntuaciones de una prueba indican niveles de aprovechamiento, aptitudes verbales, grados de ansiedad, etcétera.

Una segunda, y quizás más útil definición para el investigador dice que *un índice es un número que está compuesto de dos o más números.* Un investigador realiza una serie de observaciones y deriva un solo número de las medidas de las observaciones para resumirlas y expresarlas en forma sucinta. Con esta definición, todas las sumas y promedios son índices ya que incluyen en una sola medida más de una medida. Pero la definición también incluye la idea de los índices como compuestos de diferentes medidas. Los coeficientes de correlación son de este tipo, ya que combinan diferentes medidas en una sola o en un índice.

Existen índices de clase social. Por ejemplo, se puede combinar el ingreso, la ocupación y el lugar de residencia para obtener un buen índice de la clase social. Un índice de cohesión puede obtenerse preguntando a los miembros de un grupo si les gustaría o no seguir formando parte del grupo. Sus respuestas pueden combinarse en un solo número. En negocios y en economía, el poder de compra del dólar americano varía en el tiempo, de manera que es necesario ajustar otros valores para hacer comparaciones significativas. Por ejemplo, tomemos la comparación del costo de un automóvil en 1997 respecto a su costo en 1950. Uno de los primeros pasos es determinar el poder de compra del dólar americano en 1997 y compararlo con el poder de compra que tenía en 1950. El "Bureau of Labor Statistics" (Oficina de estadísticas laborales) registra y publica regularmente el índice de precios al consumidor (IPC). Este índice se utiliza como una medida del costo de la vida.

Los índices son importantes en investigación porque simplifican las comparaciones. De hecho, permiten al investigador hacer comparaciones que, de otra forma, no podrían hacerse o que solamente podrían hacerse con mucha dificultad. Los datos brutos son generalmente demasiado complejos para entenderlos y manejarlos matemática y estadísticamente, por lo que deben reducirse a una forma manipulable. El porcentaje es un buen ejemplo, ya que transforma puntajes brutos a formas comparables.

Los índices generalmente toman la forma de cocientes: un número es dividido entre otro número. Los índices más útiles varían entre 0 y 1.00 o entre –1.00 y +1.00 pasando por 0. Esto los hace independientes del número de casos y permite hacer comparaciones entre muestras y entre estudios. (Éstos generalmente se expresan en forma decimal.) Hay dos tipos de cocientes: las tasas y las proporciones. Un tercer tipo es el porcentaje, que es una variante de la proporción.

Una *tasa* (o *razón*) es un compuesto de dos números que relaciona un número con otro en forma de fracción o de decimal. Cualquier fracción y cualquier cociente son una razón. Tanto el numerador, como el denominador (o ambos) de una razón pueden, por sí mismos, ser tasas. El propósito principal y utilidad de una razón es el relacional, ya que permite la comparación de números. Para hacer esto quizá sea mejor colocar el mayor de los dos números del cociente en el denominador. Esto satisface la condición mencionada anteriormente de tener el rango de los valores de la razón entre 0 y 1, o entre –1.00 y +1.00, pasando por 0. Sin embargo, esto no es absolutamente necesario. Suponga que se desea comparar la tasa de hombres y mujeres graduados de preparatoria con la tasa de hombres y mujeres graduados de secundaria, todo esto en un periodo de varios años. La tasa algunas veces será menor de 1.00 y otras veces será mayor de 1.00, ya que es posible que la preponderancia de un sexo sobre el otro cambie de un año a otro.

Algunas veces las razones dan información más precisa (en cierto sentido) que la que proveen las partes de las que están compuestas. Si se estudiara la relación entre variables educativas y la tasa de impuestos, por ejemplo, y se usaran las tasas de impuestos reales, podría obtenerse una idea errónea de dicha relación. Esto es porque las tasas de impuestos, por característica, son a menudo engañosas. Algunas comunidades con altas *tasas* de impuestos, en realidad tienen niveles relativamente bajos de impuestos. El gravamen de propiedad puede ser bajo. Para evitar las discrepancias entre una comunidad y otra, se puede calcular, para cada comunidad, la tasa de gravamen contra el gravamen real. Entonces una tasa de impuestos ajustada (una "verdadera" tasa de impuestos), puede calcularse multiplicando la tasa de impuestos en uso, por esta fracción. Esto producirá una cifra más precisa para ser usada en los cálculos de relaciones entre la tasa de impuestos y otras variables. La razón de probabilidad es un tipo de índice que es valioso cuando se consideran datos de frecuencia en tablas de contingencia. Se examinará este tipo de índice estadístico en mayor detalle cuando se explique el análisis de datos de frecuencia y el análisis log-lineal.

Una *proporción* es una fracción donde el numerador es una de dos o más frecuencias observadas y el denominador la suma de las frecuencias observadas. La definición de probabilidad dada anteriormente, $p = s/(s+f)$, donde s es igual al número de éxitos y f es igual al número de fracasos, es una proporción. Considere dos números cualesquiera, 20 y 60. La razón de los dos números es 20/60 = .33. (También podría ser 60/20 = 3.) Si estos dos números fueran las frecuencias observadas de la presencia y de la ausencia de un atributo en una muestra total, donde N = 60 + 20 = 80, entonces la proporción sería: 20/(60 + 20) = .25. Otra proporción, por supuesto, es 60/80 = .75.

Un *porcentaje* es simplemente una proporción multiplicada por 100. En el ejemplo anterior sería $20/80 \times 100 = 25\%$. El propósito principal de las proporciones y porcentajes es reducir diferentes conjuntos de números a conjuntos comparables de números con una

base común. Cualquier conjunto de frecuencias puede ser transformado a proporciones o porcentajes para facilitar la manipulación e interpretación estadística y sus manifestaciones.

Es necesario ser precavido ya que es frecuente que haya una mezcla de dos medidas falibles, donde los índices pueden ser peligrosos. El viejo método de calcular el CI (coeficiente intelectual) es un buen ejemplo. El numerador de la fracción es en sí mismo un índice, dado que la edad mental (EM), es un compuesto de varias medidas. Un ejemplo mejor es el llamado coeficiente de desempeño: CD = 100 × EE/EM, donde EE es la edad educacativa y EM es la edad mental. Tanto el numerador como el denominador de la fracción son índices complejos, ya que ambos son la mezcla de mediciones de confiabilidad variable. ¿Qué significa el índice resultante? ¿Cómo interpretarlo prudentemente? Es difícil de decir. En resumen, mientras que los índices son herramientas indispensables para el análisis científico, éstos deben ser usados con cuidado y precaución.

Indicadores sociales

Los indicadores forman una clase especial de variables, aunque están estrechamente relacionados con los índices —de hecho muchas veces son índices— de acuerdo a la definición anterior. Variables tales como el ingreso, expectativa de vida, fertilidad, calidad de vida, nivel de educación (de las personas) y ambiente, pueden ser llamados indicadores sociales. Es evidente que son variables porque es común que se realicen cálculos estadísticos con ellos. Los indicadores sociales son tanto variables como estadísticos. Antes de continuar, es necesario mencionar que Bauer (1966) realizó el primer trabajo con indicadores sociales. Desafortunadamente, es difícil definirlos y no se intentará hacerlo aquí de manera formal. El artículo de Jaeger (1978) documenta las dificultades para definir los indicadores sociales. Sin embargo, los lectores deben saber que la idea de indicadores sociales es importante y lo será más en el futuro. Su uso se está extendiendo a todos los campos y, eventualmente, será estudiado en forma sistemática desde un punto de vista científico, así como desde una perspectiva "pública" y social.

El interés de este libro son los indicadores sociales, entendidos como una clase de variables sociológicas y psicológicas que en el futuro pueden ser útiles para desarrollar y probar teorías científicas sobre las relaciones entre los fenómenos sociales y psicológicos. Ciertos indicadores sociales se usan ahora en los llamados estudios de modelamiento causal de desempeño educativo y ocupacional. En 1972 Duncan, Featherman y Duncan usaron la clase social, la ocupación de los padres y su ingreso, sólo por mencionar algunos. También se han utilizado indicadores psicosociales como la calidad de vida percibida o "felicidad". Un ejemplo de esto puede encontrarse en Campbell, Converse y Rodgers (1976). Sin embargo, en general, parece que se ha hecho poco trabajo metodológico sistemático para categorizar y estudiar los indicadores sociales, la relación entre ellos y sus relaciones con otras variables. La mayoría de los trabajos pueden ser considerados demográficos y estrechamente pragmáticos —en esencia descriptivos—. Sin embargo, una vez que los problemas de confiabilidad y validez son identificados y resueltos, este campo es sumamente prometedor, y deberá ofrecer a los científicos del comportamiento algo más que estadísticas como "en 1956 el 51.2% de la población eran mujeres", o "el 54% de la población mayor de 18 años tenía de 9 a 12 años de educación". Entre los estudios más prometedores se encuentran los realizados por los investigadores Vickie Mays y Susan Cochran acerca de los riesgos de las prácticas sexuales. En un estudio de Cochran DeLeeuw y Mays (1995), usaron dos métodos estadísticos —el análisis de homogeneidad y el análisis de rasgos latentes— para tener una óptima evaluación de los patrones de conducta

sexual. El uso efectivo de estos métodos reduce los indicadores múltiples a una única puntuación que puede ser usada como una variable de resultado en investigaciones relacionadas con el virus de la inmunodeficiencia humana (VIH). Con este tipo de investigación, es posible continuar buscando estudios de análisis factorial de indicadores y estudios de análisis de covarianza, donde los indicadores son variables de las estructuras analizadas. También se puede esperar un aumento general del uso de la idea de indicadores en las áreas sociales y psicológicas de investigación. Esto se ve fácilmente en investigación educativa donde el aprovechamiento de los niños parece estar afectado en formas complejas por diferentes clases de variables, algunas de las cuales son del género de los indicadores sociales. Una de las virtudes del movimiento de indicadores sociales es que dichas influencias sobre el aprovechamiento serán usadas de manera más consciente y sistemática para estudiar y probar teorías de aprovechamiento.

La interpretación de los datos de investigación

Al evaluar la investigación, los científicos pueden disentir en dos temas generales: los datos y la interpretación de los datos. Los desacuerdos sobre los datos se enfocan a problemas tales como la validez y confiabilidad de los instrumentos de medición y la adecuación del diseño de investigación, los métodos de observación y el análisis. Asumiendo competencia, los mayores desacuerdos generalmente se enfocan en la interpretación de los datos. La mayoría de los psicólogos, por ejemplo, estarán de acuerdo en los datos de los experimentos de reforzamiento, pero disentirán vigorosamente en la interpretación de los datos de los experimentos. Tales desacuerdos son, en parte, una función de la teoría. En un libro como éste no se puede profundizar en las interpretaciones de los diferentes puntos de vista teóricos, por lo que nos limitaremos a un objetivo, que es la aclaración de algunos preceptos comunes de la interpretación de los datos *dentro* de un estudio de investigación particular o de una serie de estudios.

Adecuación de los diseños de investigación, metodología, mediciones y análisis

Uno de los temas más importantes en este libro gira alrededor de qué tan apropiada es la metodología para el problema bajo investigación. El investigador generalmente tiene preferencia por ciertos diseños de investigación, métodos de observación, métodos de medición y tipos de análisis. Todos ellos deben ser congruentes y deben encajar unos con otros. Por ejemplo, no es adecuado utilizar un análisis propio de frecuencias con, digamos, medidas continuas tomadas de una escala de actitudes. Es muy importante que el diseño, los métodos de observación, las mediciones y el análisis estadístico sean apropiados para el problema de investigación.

El investigador debe examinar a fondo la adecuación técnica de los métodos, medidas y estadísticas. La adecuación de la interpretación de los datos depende de tal escrutinio. Por ejemplo, una fuente común de debilidad en la interpretación es la negligencia con los problemas de medición. Es una necesidad urgente poner particular atención a la confiabilidad y validez de las medidas de las variables, como se verá en capítulos posteriores. Aun las personas y organizaciones más capaces en investigación titubean en ocasiones. Por muchos años, por ejemplo, la medición de las actitudes sociales comúnmente llamadas "liberalismo" y "conservadurismo" han sido cuestionadas. Por un lado, se ha asumido —aun a la vista de evidencia contraria— que liberalismo y conservadurismo forman parte

de un mismo continuo. Por otro lado, las actitudes sociales han sido medidas con muy pocos reactivos. Incluso algunas organizaciones competentes, instituciones e individuos altamente respetables han cometido estos errores (Barber, 1976). No es un pecado grave equivocarse, pero el pecado real es sacar conclusiones precipitadas, como las características de las personas, con base en mediciones de confiabilidad y validez dudosas (véase Dawes, 1994).

El aceptar sin cuestionamiento la confiabilidad y validez de las mediciones de variables es un error grave. Los investigadores deben de ser especialmente cuidadosos al cuestionar la validez de sus mediciones, dado que todo el marco de la interpretación puede colapsarse sólo en este punto. Si un problema psicológico incluye la variable ansiedad, por ejemplo, y el análisis estadístico muestra una relación positiva entre la ansiedad y el logro, el investigador debe preguntarse a sí mismo y a los datos si la ansiedad medida (o manipulada) es el tipo de ansiedad propia del problema. El investigador puede, por ejemplo, haber medido la ansiedad cuando la variable problema era realmente una ansiedad general. De igual forma, debe preguntarse si la medida elegida de desempeño es válida para los propósitos de la investigación. Si el problema de investigación demanda la aplicación de principios, pero la medida del desempeño es una prueba estandarizada que enfatiza el conocimiento de hechos, entonces la interpretación de los datos puede ser errónea.

En otras palabras, nos enfrentamos aquí al hecho obvio, pero fácilmente ignorado, de que la adecuación de la interpretación depende de cada eslabón en la cadena metodológica, así como de lo apropiado que sea cada eslabón en el problema de investigación y la congruencia de los eslabones entre sí. Esto se ve claramente al revisar resultados negativos o no concluyentes.

Resultados negativos y no concluyentes

Los resultados negativos o no concluyentes son mucho más difíciles de interpretar que los resultados positivos. Cuando los resultados son positivos y cuando apoyan la hipótesis, uno interpreta los datos a través de las líneas de la teoría y del razonamiento que subyacen a las hipótesis. Aunque se formulen cuidadosamente preguntas críticas, las predicciones sostenidas son evidencia para la validez del razonamiento que está detrás del problema enunciado.

Ésta es una de las grandes virtudes de la predicción científica. Cuando se predice algo, y se planea y ejecuta un esquema para probar la predicción, y las cosas resultan como se predijo, lo adecuado del razonamiento y de la ejecución parece sustentarse, aunque nunca se puede estar completamente seguro. Los resultados, aunque predichos, pueden ser como son por razones muy diferentes a las que se creía. Más aún, el hecho de que toda la cadena compleja de teoría, las deducciones a partir de la teoría, el diseño, la metodología, las mediciones y el análisis ha llevado a un resultado presupuesto, es fuerte evidencia de que toda la estructura ha sido adecuada. Aquí, se hace una apuesta compleja, con la suerte en contra. Entonces se lanzan los dados de la investigación o se gira la ruleta de la investigación; si resulta el número predicho, el razonamiento y el procedimiento que llevaron a una predicción exitosa, parecerán ser adecuados. Si se puede repetir la hazaña, entonces la evidencia de lo adecuado de la predicción será aun más convincente.

Pero ahora tomemos el caso negativo. ¿Por qué fueron negativos los resultados? ¿Por qué no salieron como se predijo? Observe que cualquier eslabón débil en la cadena de una investigación puede causar resultados negativos. Esto puede deberse a una, varias o todas las siguientes causas: teoría e hipótesis incorrectas, metodología inapropiada o incorrecta, mediciones inadecuadas o pobres y análisis defectuosos. En 1976, Barber afirmó que in-

cluso podía ser el resultado de un planteamiento incorrecto. Todas estas causas deben ser examinadas y evaluadas minuciosamente para ver si los resultados negativos dependen de una, de varias o de todas ellas. Si se está completamente seguro de que la metodología, la medición y el análisis son adecuados, entonces los resultados negativos podrán ser una contribución definitiva al avance científico. Es con este tipo de resultados, que se puede tener cierta certeza de que las hipótesis son incorrectas.

Relaciones no hipotetizadas y hallazgos no anticipados

Probar relaciones hipotetizadas es algo que se enfatiza en este libro. Sin embargo, esto no significa que otras relaciones en los datos no sean buscadas y probadas; muy al contrario. En la práctica, los investigadores siempre están ansiosos por buscar y estudiar relaciones en sus datos. Las relaciones no predichas pueden ser una clave importante para un entendimiento más profundo de la teoría; pueden resaltar aspectos del problema que no se anticiparon cuando éste se formuló. Por lo tanto, los investigadores —al enfatizar relaciones hipotetizadas— siempre deben estar alertas de relaciones no anticipadas en sus datos.

Suponga que se hipotetiza que un agrupamiento homogéneo de los alumnos será benéfico para los alumnos brillantes pero no para los alumnos con menos habilidades; imagine que esta hipótesis se sustenta, pero se nota una diferencia aparente entre las áreas rurales y suburbanas. La relación parece más fuerte en las áreas suburbanas, ¡pero se encuentra invertida en algunas áreas rurales! Se analizan los datos usando la variable suburbano-rural y se encuentra que el agrupamiento homogéneo parece tener una influencia marcada en los niños brillantes en el área suburbana, pero que hay poca o ninguna influencia en el área rural. Éste sería un hallazgo verdaderamente importante.

Uno de los hallazgos más fuertes y mejor apoyados de la psicología moderna es que el reforzamiento positivo fortalece la tendencia a responder (véase Hergenhahn, 1996). Por ejemplo, se ha creído que para mejorar el aprendizaje de los niños, sus repuestas correctas a problemas deben ser reforzadas positivamente. Sin embargo, sorpresivamente se ha encontrado que la motivación externa a veces tiene efectos perjudiciales. El trabajo de Lepper, Greene y Nisbett (1973) demostró que el reforzamiento positivo extrínseco minaba el interés intrínseco de los niños en un actividad de dibujo, un resultado ciertamente no predecible a partir de la teoría del reforzamiento.[6]

Los hallazgos no predichos e inesperados deben ser tratados con mayor suspicacia que aquellos predichos y esperados. Antes de ser aceptados, deben ser probados en una investigación independientemente, en la que sean predichos y probados de manera específica. Sólo cuando una relación es probada deliberada y sistemáticamente, con los controles necesarios construidos en el diseño, se puede tener fe en ellos. Los hallazgos no anticipados pueden ser fortuitos o espurios.

Tukey (1977) desarrolló métodos para analizar datos en una investigación. El uso de estos métodos se llama *análisis exploratorio de datos*. Tukey, así como Hoaglin, Mosteller y Tukey (1985) han presentado varios diagramas de fácil construcción que resumen y describen los datos. Estos diagramas pueden proveer información útil al investigador para consideraciones adicionales. Uno de los más populares es el *diagrama de tallo y hojas*. Este diagrama es similar al histograma pero tiene la ventaja de no perder los datos originales. El método de tallo y hojas trabaja mejor cuando el tamaño de la muestra es menor de 100. El principio detrás de este método es que un tallo y una hoja se usan para representar

[6] El trabajo Lepper *et al.*, así como el de otros autores, sobre motivación intrínseca y extrínseca se revisa en los artículos de Cameron y Pierce (1994, 1996).

TABLA 9.1 *Datos ficticios usados para demostrar el método de tallo y hojas (N = 46)*

81	54	91	74	88	78	90	77	88	90	69	94	74	76	96	50	93	93	70	77	58	60	75
53	81	73	66	86	81	64	77	56	71	71	56	53	83	85	70	71	76	80	87	62	57	73

cada punto o valor. El tallo se coloca a la izquierda de una línea vertical y la hoja a la derecha de dicha línea.

Tomemos por ejemplo, los datos presentados en la tabla 9.1. La hoja para cada puntuación es el último dígito y el tallo lo constituyen el resto de dígitos de un número. Por ejemplo, el número 81 de la tabla 9.1 se vería como sigue:

Tallo	Hoja
8	1

La figura 9.7 muestra el diagrama final de tallo y hoja, al incluir todos los datos de la tabla 9.1. Con este diagrama se puede tener una idea clara de cómo se ve la distribución, ya que provee una descripción más detallada de los datos que las distribuciones ordinarias de frecuencia o los histogramas. El desarrollo de tales métodos puede ayudar a los investigadores a generar hipótesis para ser probadas.

Prueba, probabilidad e interpretación

La interpretación de los datos culmina en enunciados de probabilidad condicional del tipo "si p, entonces q". Estos enunciados se enriquecen al ser especificados como sigue: si p, entonces q, bajo las condiciones r, s y t. Generalmente se evitan los enunciados causales, por la conciencia de que no pueden ser realizados sin fuerte riesgo de error.

Quizás el problema de la prueba sea de mayor importancia práctica para el investigador que interpreta datos. Hay que aclarar que nada puede ser "probado" científicamente. Todo lo que puede hacerse es buscar evidencia para sostener que determinada proposición es cierta. Una prueba es un asunto deductivo. Los métodos experimentales de investigación no son métodos de prueba, son métodos controlados que brindan evidencia para apoyar la probable verdad o falsedad de las relaciones propuestas. En pocas palabras, ninguna investigación científica puede probar nada, por lo que la interpretación del análisis de los datos de investigación nunca debe usar la palabra *prueba*.

Afortunadamente, para propósitos prácticos de investigación, no es necesario preocuparse mucho acerca de la causalidad y de la prueba. La evidencia con niveles satisfactorios

FIGURA 9.7

Tallo	Hoja
5	0 3 3 4 6 6 7 8
6	0 2 4 6 9
7	0 0 1 1 1 3 3 4 4 5 6 6 7 7 7 8
8	0 1 1 1 3 5 6 7 8 8
9	0 0 1 3 3 4 6

de probabilidad es suficiente para el progreso científico. La causalidad y las pruebas fueron analizadas en este capítulo para sensibilizar al lector del peligro del uso impreciso de los términos. El entendimiento del razonamiento científico, y la práctica y el cuidado razonable en la interpretación de los datos de investigación, son útiles guardianes contra la inferencia inadecuada de datos a conclusiones, aun cuando no garanticen la validez de las interpretaciones.

RESUMEN DEL CAPÍTULO

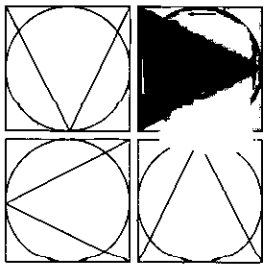
1. El análisis es el proceso de categorizar, ordenar, manipular y resumir datos para responder a las preguntas de investigación.
2. El propósito del análisis es reducir datos a una forma interpretativa, de manera que las relaciones puedan ser estudiadas y probadas.
3. La interpretación toma los resultados del análisis y hace inferencias y discute relaciones.
4. Los datos vienen en forma de medidas de frecuencia y medidas continuas.
5. El primer paso en cualquier análisis es la categorización o partición.
6. Los tipos de análisis estadísticos son:
   a) gráficos
   b) medidas de tendencia central y de variabilidad
   c) medidas de relaciones
   d) análisis de diferencias
   e) análisis de varianza
   f) análisis de perfiles y análisis multivariado
7. Los índices se usan para simplificar las comparaciones. Ejemplo de índices son los porcentajes, los cocientes y las tasas o razones.
8. Los datos y la interpretación de los datos son dos áreas en las que los científicos disienten.
9. Cuando se interpretan los datos de investigación, se debe considerar la adecuación técnica de la metodología de investigación, de los procesos de medición y de la estadística utilizados.
10. Los resultados negativos o no concluyentes son mucho más difíciles de interpretar que los resultados positivos.
11. Al conducir un estudio de investigación, pueden surgir relaciones no hipotetizadas y hallazgos no anticipados.
12. Los hallazgos no predichos e inesperados deben de ser tratados con más suspicacia que los hallazgos predichos y esperados.

SUGERENCIAS DE ESTUDIO

1. Suponga que usted desea estudiar la relación entre la clase social y el resultado de una prueba de ansiedad. ¿Cuáles son las dos principales posibilidades para analizar los datos? (omita la posibilidad de calcular un coeficiente de correlación). Establezca dos estructuras analíticas.
2. Suponga que usted quiere agregar la variable sexo al problema anterior. Establezca las dos clases de paradigmas analíticos.
3. Suponga que un investigador ha probado los efectos de tres métodos de enseñanza de lectura, en la ejecución de la lectura. Tenía 30 sujetos en cada grupo y una pun-

tuación de la ejecución de lectura para cada sujeto. También incluyó el género como una variable independiente: la mitad de los sujetos eran varones y la otra mitad eran mujeres. ¿Cómo se vería su paradigma analítico? ¿Qué va dentro de las casillas?

4. Estudie la figura 9.3. ¿Representan estos diseños o paradigmas de análisis de varianza la partición de las variables? ¿Por qué sí? ¿por qué no? ¿Por qué es importante la partición al establecer diseños de investigación y al analizar los datos? ¿Tienen las reglas de categorización (y partición) algún efecto en la interpretación de los datos? Si es así, ¿qué efectos pueden tener? (Considere los efectos de violar las dos reglas básicas de partición.)



CAPÍTULO 10

EL ANÁLISIS DE FRECUENCIAS

- TERMINOLOGÍA DE DATOS Y VARIABLES
- TABULACIÓN CRUZADA: DEFINICIONES Y PROPÓSITO
- TABULACIÓN CRUZADA SIMPLE Y REGLAS PARA LA CONSTRUCCIÓN DE UNA TABULACIÓN CRUZADA
- CÁLCULO DE PORCENTAJES
- SIGNIFICANCIA ESTADÍSTICA Y LA PRUEBA χ^2
- NIVELES DE SIGNIFICANCIA ESTADÍSTICA
- TIPOS DE TABLAS CRUZADAS Y TABLAS
  - **Tablas unidimensionales**
  - **Tablas bidimensionales**
  - **Tablas bidimensionales, dicotomías "verdaderas" y medidas continuas**
  - **Tablas de tres dimensiones y de *k*-dimensiones**
- ESPECIFICACIÓN
- TABULACIÓN CRUZADA, RELACIONES Y PARES ORDENADOS
  - **La razón de probabilidad**
  - **Análisis multivariado de datos de frecuencia**
  - **Anexo computacional**

Hasta ahora se ha hablado principalmente *acerca* del análisis. Ahora se explicará cómo *hacer* el análisis. La forma más simple de analizar datos para estudiar relaciones es por medio de la partición cruzada de frecuencias. Como se aprendió en el capítulo 4, la partición cruzada es una nueva partición del conjunto U, para formar los subconjuntos de la forma $A \cap B$; es decir, que se forman subconjuntos de la forma $A \cap B$ de los subconjuntos conocidos A y B de U. Se dieron ejemplos en el capítulo 4 y se darán más en breve. La expresión "partición cruzada" se refiere a un proceso abstracto de la teoría de conjuntos. Sin embargo, ahora que la idea de la partición cruzada se aplique al análisis de frecuencias para estudiar relaciones entre variables, se le llamará *tabulación cruzada*, aunque también se le ha llamado algunas veces *fracción cruzada*. El análisis que se mostrará también es conocido como análisis de *contingencia* o análisis de tabla de contingencia.

TABLA 10.1 *Relación entre afiliación al partido político y el voto de conciliación presupuestal, en el senado de Estados Unidos, 1995*

	Republicano	Demócrata	
En contra	1	46	47
	2%	100%	
A favor	52	0	52
	98%	0%	
	53	46	99

Fuente: Datos del *Congressional Quarterly* (1996).

Dado que no es posible seguir adelante sin la estadística, se introducirá una forma de análisis estadístico comúnmente asociado con las frecuencias, la prueba de χ^2 (chi cuadrada), y el concepto de "significancia" estadística. Este estudio de la tabulación cruzada y la χ^2 servirá como introducción a la estadística.

La pugna política entre republicanos y demócratas a menudo se refleja dramáticamente en los votos del Congreso. Una de las recientes votaciones importantes en el senado de Estados Unidos se llevó a cabo en el proyecto de ley fiscal de 1996, referente a la conciliación presupuestal. La lucha republicano-demócrata durante el invierno de 1995 se centró en las propuestas de balance del presupuesto hacia el año 2002: los republicanos se manifestaban, generalmente, a favor de las propuestas y los demócratas en contra de ellas, incluyendo al presidente Clinton. Una de estas propuestas era reducir los gastos en servicios de asistencia social y reducir los impuestos. El proyecto de ley se aprobó por 52 a 47. Esto fue una derrota para el presidente. Lo interesante aquí son los resultados de los votos republicano-demócratas, que se muestran en la tabla 10.1. Por las frecuencias (en este caso) es claro que hay una fuerte relación entre la afiliación al partido político y el voto en la propuesta de ley sobre el presupuesto: los demócratas votaron en contra y los republicanos votaron a favor.

No todas las frecuencias en la tabulación cruzada son así de claras. En la práctica es común calcular porcentajes. Si se hace en una forma que será descrita posteriormente, los porcentajes son los presentados en la esquina inferior derecha de cada casilla. Se nota la fuerza de la relación entre la afiliación al partido político y el voto: 98% de los republicanos votaron a favor y 100% de los demócratas votaron en contra.

Estudios de votaciones similares en la misma época muestran la misma relación general. Por ejemplo, los votos para imponer sanciones a los médicos que realizan abortos tardíos se pueden observar en la tabla 10.2. Nuevamente la relación es fuerte, aunque no tanto como en la votación para la conciliación presupuestal (observe los porcentajes). Un voto "en contra" en este proyecto de ley apoya la posición del presidente.

TABLA 10.2 *Voto del senado de Estados Unidos para imponer sanciones a los médicos que realizan abortos tardíos, 1995*

	Republicano	Demócrata	
En contra	45	9	54
	85%	20%	
A favor	8	36	44
	15%	80%	
	53	45	98

Terminología de datos y variables

En el capítulo 3 se hizo una distinción entre variables activas y variables atributo. Las primeras se refieren a variables experimentales o manipuladas y las segundas a variables medidas. El término "atributo" se usó porque es general y puede abarcar las propiedades de un objeto animado o inanimado. Desafortunadamente "atributo" algunas veces se ha usado para significar las llamadas variables categóricas en este libro. Con esta acepción, por ejemplo, sexo, raza, religión y otras variables categóricas similares han sido llamadas atributos, también "variables cualitativas"; ambos usos parecen equivocados. Un atributo es cualquier propiedad de cualquier objeto, ya sea que el objeto sea medido en términos de todo o nada, o con un conjunto de medidas continuas. Esta definición será utilizada en este libro, no para contravenir cualquier uso convencional, si esto fuera posible, sino para aclarar la distinción entre variables experimentales y variables medidas.

Las llamadas variables categóricas son también conocidas, quizás en forma más precisa, como "variables nominales", porque corresponden al nivel de medición "nominal", el cual se aprenderá más adelante. Dado que en este capítulo y en los siguientes debe quedar muy clara la diferencia entre las variables continuas y las variables categóricas, se anticipará brevemente una discusión posterior y se definirá *medición*. Cuando los números o símbolos asignados a los objetos no tienen un significado numérico más allá de la presencia o ausencia de la propiedad o atributo que están midiendo, esta medida es llamada "nominal". Una variable nominal, es la que se ha estado llamando "categórica". Nombrar a algo ("nominal") es colocarlo en una categoría ("categórica"). Algunos datos categóricos se dan naturalmente, como el género (femenino-masculino) o el color de ojos (azul, café, gris, avellana). Otros datos categóricos son creados al categorizar los datos medidos en una escala continua.

Todo esto quizás sea más claro con la siguiente ecuación de conjuntos, que es una definición general de medición:

$$f = \{(x, y): x = \text{cualquier objeto, y } y = \text{cualquier numeral}\}$$

que se lee: f es una regla de correspondencia que es definida como un conjunto de pares ordenados (x, y), donde x es algún objeto y y es algún número asignado a x. Ésta es una definición general que cubre todos los casos en medición. Obviamente, y puede ser un conjunto de medidas continuas o simplemente el conjunto $\{0, 1\}$. Las variables categóricas o nominales son aquellas variables donde $y = \{0, 1\}$, donde 0 y 1 son asignados con base en que el objeto x posea o no alguna propiedad o atributo definido. Las variables continuas son aquellas variables donde $y = \{0, 1, 2..., k\}$, o algún sistema numérico donde los números indican más o menos el atributo en cuestión. (Matemáticamente es difícil definir *medidas continuas*, y la definición dada anteriormente no es satisfactoria. Sin embargo, el lector sabrá lo que significa.)

El nivel de medición de este capítulo es en su mayoría nominal. Aun cuando se usan variables continuas, éstas son convertidas a variables nominales. Si de esta conversión resultan categorías que pueden ser ordenadas en términos de "importancia", "cantidad" o atributos jerárquicos similares, estos datos son llamados *ordinales*. Una categoría puede poseer más de algún atributo que las otras categorías. En general, la conversión de datos continuos a nominales o a ordinales no debería hacerse porque desperdicia (descarta) información (varianza). Sin embargo hay ocasiones en las que, a juicio del investigador, es necesario o deseable tratar a una variable continua como variable nominal. Por ejemplo, es posible medir una variable potencialmente continua sólo de manera burda por un observador que juzgue si un objeto posee o no un atributo. Mientras que hay grados de con-

ducta agresiva, podría ser posible sólo decir si un individuo exhibió o no una conducta agresiva.

Tabulación cruzada: definiciones y propósito

Una tabulación cruzada es una presentación tabular numérica de los datos, generalmente en forma de frecuencias o de porcentajes en la que las variables se dividen de forma cruzada. Una forma común de la fracción cruzada o tabulación cruzada es la partición cruzada usada para estudiar las relaciones entre las variables. Es una forma común de análisis que puede utilizarse con casi cualquier tipo de datos, aunque se usa principalmente con datos categóricos o nominales. Además de su uso real en la investigación, la tabulación cruzada es una herramienta pedagógica muy útil. Su claridad y simplicidad la hacen una herramienta útil para aprender cómo estructurar los problemas de investigación y cómo analizar los datos. La tabulación cruzada son particiones cruzadas, como se indicó antes, por lo que las reglas de la partición y los conceptos de conjuntos ya aprendidos pueden aplicarse fácilmente a este análisis.

La tabulación cruzada también se usa de forma descriptiva. El investigador puede no estar interesado en las relaciones, sino solamente en describir una situación existente. Por ejemplo, considere el caso en que una tabla fracciona la clase social contra la posesión de aparatos de televisión, refrigeradores, etcétera. Ésta es una comparación descriptiva más que una tabulación cruzada de variables, aunque la posesión del televisor pudiera ser de algún tipo de variable. El interés aquí es exclusivamente el análisis de los datos obtenidos para probar o explorar relaciones.

La tabulación cruzada permite al investigador determinar la naturaleza de las relaciones entre las variables, pero tiene también otros propósitos adicionales: puede ser usada para organizar datos de una forma conveniente en un análisis estadístico, para luego aplicar una prueba estadística a esos datos. También es posible calcular los índices de asociación.

Otro propósito de la tabulación cruzada es el control de las variables. Como se verá posteriormente, la tabulación cruzada permite estudiar y probar una relación entre dos variables mientras se controla una tercera variable. De esta forma, las relaciones "espurias" pueden ser desenmascaradas y las relaciones entre variables pueden ser "especificadas", es decir, que las diferencias en el grado de relación en diferentes niveles de una variable control, pueden ser determinadas.

Otro propósito de la tabulación cruzada, referido anteriormente, fue que su uso y su estudio sensibiliza al estudiante y al que practica la investigación, en el diseño y estructura de los problemas de investigación. Existen beneficios al reducir un problema de investigación a una tabulación cruzada, de hecho, si no es posible crear un diagrama del paradigma del problema de investigación, ya sea como análisis de varianza o como tabulación cruzada, entonces el problema no está claro en la mente, o bien, no se tiene realmente un problema de investigación.

Tabulación cruzada simple y reglas para la construcción de una tabulación cruzada

La forma más simple de una tabulación cruzada es una tabla de 2 por 2 (o 2×2). Ya se dieron dos ejemplos anteriormente. Un tercer ejemplo se presenta en la tabla 10.3. Los datos son de un estudio de Payette y Clarizio (1994), donde se examinó la influencia de las características del estudiante en su clasificación errónea como poseedor, o no, de un

Tabla 10.3 *Frecuencias de estudiantes que no mostraron una discrepancia severa bajo los indicadores de género y decisión de elegibilidad (estudio de Payette y Clarizio)*[a]

	Género		
Elegibilidad	Mujeres	Hombres	
Elegible	17 (.40)	16 (.21)	33
No elegible	26 (.60)	60 (.79)	86
	43	76	119

[a] Los números en el centro de cada casilla son frecuencias. Los números en paréntesis de cada casilla son los porcentajes calculados para el género de acuerdo a la elegibilidad, por ejemplo, 17/43 = .40, y 60/76 = .79. Estos últimos están escritos como proporciones: al multiplicar por 100, las proporciones se transforman en porcentajes. En lo sucesivo, se sigue la convención de escribir proporciones.

problema de aprendizaje (PA). Las características del estudiante incluidas fueron raza, género y estatus intelectual, de aprovechamiento y de nivel-grado. Cada estudiante en el proyecto fue clasificado como elegible o no elegible para su ubicación en el grupo de problemas de aprendizaje. Bajo los principios reales mencionados por Payette y Clarizio, una discrepancia severa era definida como bajo aprovechamiento. Los datos en la tabla representan el número de hombres y mujeres que no mostraron una discrepancia severa, pero que fueron clasificados como elegibles o no elegibles. Payette y Clarizio encontraron que el número de hombres y mujeres clasificados como elegibles era muy similar. Sin embargo, las mujeres tenían mayor probabilidad de ser clasificadas como "elegibles" que los hombres (.40 contra .21). Aunque podrían discutirse las razones de esta diferencia, el principal propósito aquí es mostrar cómo se construyó esta tabla.

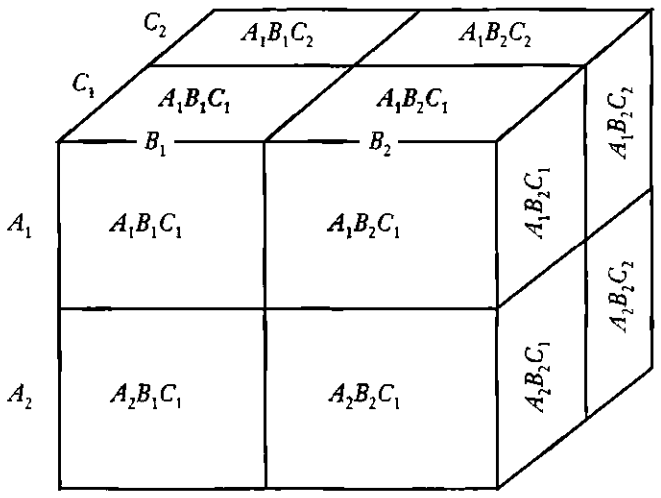
Parecen no existir reglas aceptadas de forma general respecto a cómo construir tabulaciones cruzadas. Se sabe, sin embargo, que son particiones cruzadas y que deben seguir las reglas de la partición o categorización discutidas anteriormente. Esas reglas eran: 1) las categorías se establecen de acuerdo a la hipótesis de investigación; 2) las categorías son independientes y mutuamente excluyentes; 3) las categorías son exhaustivas; 4) cada categoría es derivada de un solo principio de clasificación, y 5) todas las categorías están en un nivel de discurso. En los estudios donde hay una clara distinción sobre cuál es la variable independiente y cuál es la variable dependiente, se reportan los niveles de la variable independiente en las columnas de la tabla de contingencia y los resultados de la variable dependiente en los renglones.

En la figura 10.1 se muestra una tabulación cruzada de 2×2, con los símbolos de las variables. A_1 y A_2 son las particiones de la variable A; B_1 y B_2 son las particiones de la variable B. Las celdas A_1B_1, A_1B_2, A_2B_1, A_2B_2, son simplemente las intersecciones de los subconjuntos de A y B: A_1B_1, A_1B_2, A_2B_1, A_2B_2. Cualquier objeto en U, el universo de objetos, puede ser categorizado como A_1B_1, A_1B_2, A_2B_1, A_2B_2. Si U es una muestra de niños, B es el género y A es delincuencia, entonces un miembro de A_1B_1 es un delincuente masculino, mientras que un miembro A_2B_2 es una niña no delincuente.

Figura 10.1

	B_1	B_2
A_1	A_1B_1	A_1B_2
A_2	A_2B_1	A_2B_2

FIGURA 10.2



En la tabla 10.3, *B* sería el género; *A* sería la elegibilidad; A_1 es igual a elegible; A_2 es igual a no elegible; B_1 es igual a mujer; B_2 es igual a hombre. Entonces A_1B_1 es una mujer elegible para *PA* y A_2B_2 es un hombre clasificado como no elegible. Tablas más grandes, de 2×3, 2×4, 3×2, etcétera, son solamente extensiones de esta idea.

En el caso de tres variables, estrictamente hablando, se requiere de un cubo. Suponga que hay tres variables dicotomizadas *A*, *B*, *C*. La situación real se parecería a la que se muestra en la figura 10.2. Cada casilla es un cubo con una triple etiqueta. Todos los cubos visibles han sido etiquetados apropiadamente. Si las variables *A*, *B* y *C* fueran sexo, clase social y delincuencia, respectivamente, entonces por ejemplo un miembro de la casilla $A_2B_2C_1$ sería una mujer de clase trabajadora que es delincuente. Dado que manejar cubos es incómodo, se utilizará un sistema más simple. La tabla de la tabulación cruzada de tres variables puede semejarse a la que se muestra en la figura 10.3. Se retomará la tabulación cruzada de tres variables posteriormente.

Cálculo de porcentajes

Los porcentajes se calculan de la variable independiente hacia la variable dependiente. En los estudios donde no es posible etiquetar las variables como independientes y dependientes, la regla, por supuesto no aplica; pero en la mayoría de los casos es aplicable. En la tabla

FIGURA 10.3

	B_1		B_2	
	C_1	C_2	C_1	C_2
A_1	$A_1B_1C_1$	$A_1B_1C_2$	$A_1B_2C_1$	$A_1B_2C_2$
A_2	$A_2B_1C_1$	$A_2B_1C_2$	$A_2B_2C_1$	$A_2B_2C_2$

10.1 y en la tabla 10.2, se calculan los porcentajes desde republicanos y demócratas hacia los votos a favor y en contra, por ejemplo, 51/52 = .98 y 1/52 = .02 en la tabla 10.1, y 9/45 = .20 y 36/45 = .80 en la tabla 10.2. En las tres tablas anteriores, la convención usada fue colocar las variables independientes en la parte superior de la tabla y las variables dependientes al lado de la tabla. Se pudo haber hecho también en forma invertida, pero cuando hay más de una variable independiente, las tablas de contingencia publicadas son frecuentemente impresas de arriba a abajo. En la figura 10.3, por ejemplo, *B* y *C* serían las variables independientes y *A* la variable dependiente.

Si se observa nuevamente la tabla 10.3, que contiene los datos del estudio de Payette y Clarizio, ¿indica esta tabla una relación mayor a la esperada por el azar, entre género y elegibilidad por problemas de aprendizaje? Las proporciones en las cuatro casillas de la tabla, ¿se alejan significantemente de las proporciones esperadas por el azar? Si así sucede, se dice que hay una relación entre las variables. Suponga que se ha realizado una prueba estadística y que sus resultados indican un alejamiento de las proporciones mayor que el esperado por el azar. (Se mostrará cómo realizar esta prueba en breve.) Entonces, se afirma que hay una relación estadísticamente significativa entre el género y la elegibilidad para problemas del aprendizaje.

Pero ¿cuál es la naturaleza de la relación? Esto se determina al estudiar la tabla, especialmente los porcentajes (proporciones). La parte más pesada de la relación parece ser la columna "elegible": 40% de las mujeres son elegibles aun cuando no muestren una discrepancia severa, mientras que solamente 21% de los hombres fueron colocados aquí. Como resultado, pocas mujeres fueron consideradas no elegibles al compararlas con los hombres.

Las tablas cruzadas con frecuencias pueden ser interpretadas sin convertirlas en porcentajes, pero es aconsejable convertirlas siguiendo la regla dada anteriormente: calcular una columna (o renglón) a la vez, de la variable independiente hacia la variable dependiente. Para hacer esto, primero se suman las frecuencias en los renglones y en las columnas y luego se colocan las sumas resultantes en la parte inferior y al lado de la tabla. En la tabla 10.3 se incluyeron dichas sumas y son llamadas "frecuencias marginales" o "marginales". (En realidad, para calcular los porcentajes, solamente las sumas de las columnas de la tabla 10.3 necesitan ser calculadas. Tanto las sumas de los renglones como de las columnas se necesitarán posteriormente.) En las relaciones de la tabla 10.1 y de la tabla 10.2, la variable independiente es, claramente, la afiliación al partido político y la variable dependiente es el voto en el asunto. En la tabla 10.3, la variable independiente es el género y la variable dependiente la elegibilidad. A veces, determinar qué variable es cuál no es tan simple. De cualquier manera, en las tres tablas se calcularon los porcentajes por columnas, o de la variable independiente (columnas) hacia la variable dependiente (renglones).

Para estar seguros de saber lo que se está haciendo, hay que calcular los porcentajes de la tabla 10.3. Tomemos los renglones separadamente: el renglón de las mujeres: 17 ÷ 43 = .40 y 26 ÷ 43 = .60. Éstas son las proporciones. Si se multiplican por 100 (solamente moviendo el punto decimal dos lugares a la derecha) resulta, por supuesto, 40% y 60%. Ahora la columna de los hombres: 16 ÷ 76 = .21 y 60 ÷ 76 = .79, o 21% y 79%. (Observe que cada columna debe dar un total de 1.00, o 100%). La relación ahora es clara. Las mujeres son (proporcionalmente) más tendientes a ser clasificadas como elegibles, que los hombres. Note cómo el porcentaje de la tabulación cruzada resalta la relación, que no era tan clara en las frecuencias debido al número desigual de mujeres (43) y hombres (76). En otras palabras, el cálculo del porcentaje transforma ambos renglones a una base común y fortalece la comparación y la relación.

Aquí pueden surgir dos preguntas: 1) ¿Por qué no calcular los porcentajes de otra forma: de la variable dependiente a la variable independiente? 2) ¿Por qué no calcular los

porcentajes con base en la tabla completa? No hay nada propiamente equivocado en estas preguntas. En el primer caso, sin embargo, se estaría haciendo a los datos una pregunta diferente. En el segundo caso, se estarían transformando los datos de frecuencia a porcentajes o proporciones sin cambiar el patrón de las frecuencias.

El problema de Payette-Clarizio se enfocó hacia la clasificación equivocada de niños elegibles o no elegibles para un tratamiento por problemas de aprendizaje. Una hipótesis implicada en el problema es: quienes toman la decisión están sesgados en su decisión respecto a las niñas. Éste es un enunciado de la clase "si p entonces q": si se es niña, entonces se tiene mayor probabilidad de ser elegida como poseedor de problemas de aprendizaje. No puede haber duda respecto a las variables independiente y dependiente, por lo tanto el cálculo de los porcentajes está determinado ya que debemos preguntar: si se trata de una niña ¿qué proporción de ellas será clasificada como elegible? La pregunta es contestada en la primera columna de la tabla 10.3: .40, o 40%. (Por supuesto que la segunda columna es también importante para la relación total.)

El cálculo de los porcentajes a través de los renglones es equivalente a la hipótesis: si se es elegible para presentar problemas de aprendizaje, entonces el género es femenino; pero no se está tratando de explicar el género, ya que el género no es la variable dependiente. Si aún así se calculan los porcentajes, éstos resultarían erróneos (véase sugerencia de estudio 3). El razonamiento teórico para calcular los porcentajes partiendo de la variable independiente hacia una variable dependiente está basado en la consideración de que los porcentajes calculados de esta forma son probabilidades condicionales (véase capítulo 7), cuyos enunciados correctos se derivan del problema de investigación. Por ejemplo, para la tabla 10.1 podemos decir: "si es republicano, entonces vota en contra", que es un enunciado condicional. En lenguaje de teoría de conjuntos y de probabilidad, esto es: la probabilidad de B_1, un voto en contra, dado A_1, republicano, o:

$$p(B_1|A_1) = \frac{p(A_1 \cap B_1)}{P(A_1)} = \frac{1/99}{53/99} = .02$$

y ésta es la probabilidad condicional: la probabilidad de B_1, dado A_1. También es el porcentaje de la casilla $A_1 B_1$ [republicano-en contra] de la tabla 10.1.

Significancia estadística y la prueba χ^2

Es necesario interrumpir el estudio de la tabulación cruzada para aprender un poco acerca de estadística y así anticipar el trabajo y estudio del siguiente capítulo. Aunque es posible discutir acerca de la tabulación cruzada y cómo se construye sin usar estadísticas, en realidad no es posible avanzar hacia el análisis y la interpretación de los datos de frecuencia sin usar al menos algo de estadística. Así que se examinará una de las pruebas estadísticas más simples, pero más útiles, la prueba χ^2 (chi cuadrada).

Observe las frecuencias de la tabla 10.3. ¿Realmente expresan una relación entre género y elegibilidad para problemas de aprendizaje? ¿O podrían haberse dado por el azar? ¿Son estas frecuencias un patrón entre muchos patrones de frecuencias, que se podrían haber obtenido por medio de una tabla de números aleatorios (selección limitada solamente por las frecuencias marginales dadas)? Tales preguntas deben hacerse para cada conjunto de resultados de frecuencias obtenidos de muestras. Hasta ser contestadas, no tiene caso avanzar en la interpretación de los datos. Si los resultados pudieran haber sucedido por el azar, ¿qué caso tiene intentar interpretarlos?

¿Qué quiere decir que un resultado obtenido es "estadísticamente significativo"? ¿Que se aparta "significativamente" de lo esperado por el azar? Suponga que se realiza un experimento real, 100 veces (lanzar una moneda 100 veces). Cada experimento es como un lanzamiento de moneda o como un lanzamiento de dados. El resultado de cada experimento puede ser considerado como un punto muestral. El espacio muestral propiamente concebido, es un número infinito de tales experimentos o puntos muestrales. Por conveniencia, se considera a las 100 réplicas del experimento como el espacio muestral *U*. Esto no es nada nuevo. Es lo que se hizo con las monedas y los dados.

Tomemos un ejemplo simple, la administración universitaria está considerando cambiar su sistema de calificación, pero desea conocer las actitudes de los catedráticos hacia el cambio propuesto. La administración ha encontrado, por experiencias anteriores, que si la mayoría de los catedráticos no aprueba un cambio, el nuevo sistema puede tener serios problemas. Por medio de un procedimiento conveniente, se les pregunta a 100 catedráticos seleccionados al azar, su opinión hacia el cambio propuesto. Sesenta de ellos aprueban el cambio y 40 lo desaprueban. La administración debe preguntar ahora: ¿Es ésta una mayoría "significativa"? Los administradores razonan como sigue: si los catedráticos fueran completamente indiferentes al respecto, sus respuestas serían como dadas al azar —ahora de esta forma, ahora de otra—. La frecuencia esperada en una hipótesis de indiferencias sería, por supuesto 50/50, el resultado esperado por el azar.

Para contestar la pregunta sobre si 60/40 difiere significativamente de la indiferencia o del azar, se realiza una prueba estadística χ^2. Se estructura una tabla (tabla 10.4) para obtener los términos necesarios para el cálculo de la χ^2. El término f_o representa "frecuencia obtenida" y f_e representa "frecuencia esperada". La función de la prueba estadística es comparar los resultados obtenidos con aquellos esperados con base en el azar. Entonces, se comparan f_o con f_e. En el supuesto de la indiferencia o del azar, se escribe 50/50; pero se obtuvieron 60 y 40. La diferencia es 10 con respecto al 50. ¿Podría una diferencia tan grande de 10 haber ocurrido por azar? Otra forma de plantear la pregunta es: si se realizara el mismo experimento 100 veces y solamente estuviera operando el azar (esto es, que los catedráticos contestaran las preguntas indiferentemente o, en efecto, al azar) ¿cuántas de las 100 veces podría esperarse una desviación tan grande como 60/40? Si se lanza una moneda 100 veces, sabemos que a veces se obtendrán 60 caras y 40 cruces, y 40 caras y 60 cruces; ¿cuántas veces ocurriría tal discrepancia (si en realidad es tan grande) por azar? La prueba χ^2 es una forma conveniente para obtener una respuesta.

He aquí la fórmula de la χ^2:

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right]$$

que dice simplemente: "reste cada frecuencia esperada, f_e, de la frecuencia obtenida f_o, eleve esta diferencia al cuadrado, divida esta diferencia cuadrada entre la frecuencia esperada f_e, y después sume estos cocientes". Esto se hizo en la tabla 10.4. Para estar seguro de que el lector conoce lo que se ha hecho escribiremos ahora:

$$\chi^2 = \frac{(60-50)^2}{50} + \frac{(40-50)^2}{50} = \frac{100}{50} + \frac{100}{50} = 4$$

Pero ¿qué significa $\chi^2 = 4$? χ^2 es una medida de qué tanto se apartan las frecuencias obtenidas de las frecuencias esperadas por el azar. Dado que se tiene forma de conocer lo que es esperado por el azar y dado que las observaciones son independientes, entonces se puede

TABLA 10.4 *Cálculo de la χ^2: aprobación y desaprobación de los catedráticos hacia los cambios propuestos en el sistema de calificación*

	Aprueba	Desaprueba
f_o	60	40
f_e	50	50
$f_o - f_e$	10	−10
$(f_o - f_e)^2$	100	100
$(f_o - f_e)^2/f_e$	100/50 = 2	100/50 = 2

calcular la χ^2. Entre más grande sea la χ^2, mayor es la desviación de las frecuencias obtenidas respecto a las frecuencias esperadas por el azar. El valor de la χ^2 puede variar desde 0, lo que indica ninguna desviación de las frecuencias obtenidas respecto a las esperadas, hasta un gran número de valores crecientes.

Además de la fórmula anterior, es necesario conocer los *grados de libertad* (*gl*) del problema, y tener una tabla de valores críticos de χ^2. Las tablas de chi cuadrada se encuentran en casi todos los textos de estadística, junto con las instrucciones de cómo usarlas. La tabla 10.5a presenta una tabla abreviada de χ^2. Diversas explicaciones sobre los grados de libertad también se dan en los libros de texto de estadística (véase Walker, 1951; Graciano y Raulin, 1993). Se puede decir que los "grados de libertad" definen la amplitud de variación contenida en un problema estadístico. En el problema anterior hay un grado de libertad porque el número total de casos está fijado en 100, y porque tan pronto como se da una de las frecuencias, la otra queda determinada inmediatamente. Es decir, que no hay grados de libertad cuando dos números deben sumar 100, y uno de ellos, por ejemplo, 40, es dado. Una vez que 40 o 45, o cualquier otro número es dado no hay donde ir. El número remanente no tiene libertad para variar.

Para entender más acerca de lo que está sucediendo aquí, suponga que se calculan todas las χ^2 para todas la posibilidades: 40/60, 41/59, 42/58, ..., 50/50, ..., 60/40. Haciendo esto se tiene el conjunto de valores que se muestra en la tabla 10.5b. (Al leer la tabla, es útil considerar a la primera frecuencia de cada par como "cara", o "de acuerdo con", o "masculino", o cualquier otra variable.) Solamente dos de estas χ^2, los valores de 4.00 asociados con 40/60 y 60/40, son estadísticamente significativos. Son estadísticamente significativos porque al verificar la tabla de χ^2 (tabla 10.5a) para un grado de libertad se encuentra una entrada de 3.84 lo que es llamado el nivel de significancia .05. Todos los

TABLA 10.5a *Distribución de probabilidades de χ^2*

gl	nivel .25	nivel .10	nivel .05	nivel .01
1	1.32	2.71	3.84	6.63
2	2.77	4.61	5.99	9.21
3	4.11	6.25	7.81	11.3
4	5.39	7.78	9.49	13.3
5	6.63	9.24	11.1	15.1
6	7.84	10.6	12.6	16.8
7	9.04	12.0	14.1	18.5
8	10.2	13.4	15.5	20.1
9	11.4	14.7	16.9	21.7
10	12.5	16.0	18.3	23.2
11	13.7	17.3	19.7	24.7

TABLA 10.5b *Frecuencias y χ^2[a] correspondientes*

Frecuencias	χ^2
40/60	4.00
41/59	3.24
42/58	2.56
43/57	1.96
44/56	1.44
45/55	1.00
46/54	.64
47/53	.36
48/52	.16
49/51	.04
50/50	0

[a] Los valores de χ^2 para 51/49, ..., 60/40 son, por supuesto, los mismos que se muestran en la tabla, pero en orden inverso.

otros valores de χ^2 en el tabla 10.5b son menores a 3.84. Tomemos, por ejemplo, la χ^2 para 42/58, que es 2.56, y al consultar la tabla, 2.56 cae entre los valores de χ^2 con probabilidades de .10 y .25 o 2.71 y 1.32, respectivamente. Esto representa realmente una probabilidad cercana a .14. En la mayoría de los casos no es necesario buscar dónde caen, sino que solamente se requiere observar que no alcanza el nivel de .05 para 3.84. Si no lo hace, se concluye que no es estadísticamente significativo al nivel de .05. El lector puede ahora preguntar: "¿qué es el nivel .05?" y "¿por qué el nivel .05?" "¿Por qué no .10 o aun .15?" Para contestar estas preguntas es necesario desviarse un poco del tema.

Niveles de significancia estadística

El nivel .05 quiere decir que un resultado que es significativo al nivel .05 puede ocurrir por azar no más de 5 veces en 100 ensayos. En el ejemplo de las respuestas a la pregunta de la administración de 60 acuerdos y 40 desacuerdos, se puede decir que una discrepancia tan grande como ésta *ocurriría por azar* cerca de 5 veces o menos en 100 ensayos.

Un nivel de significancia estadística es elegido en forma algo arbitraria. Se atribuye esta elección a Fisher (1950), pero ciertamente no es por completo arbitraria. Otro nivel de significancia frecuentemente usado es el nivel .01. Los niveles .05 y .01 corresponden claramente a dos y tres desviaciones estándar de la media de una distribución normal de probabilidad. (Una distribución normal de probabilidad es la curva simétrica con forma de campana que el lector probablemente ya ha visto. Se hablará de ella más tarde.)

Regresemos al experimento de lanzar una moneda 100 veces. Resultó cara 52 veces y cruz 48 veces (consulte la tabla 10.5b, $\chi^2 = .16$, un resultado claramente no significativo). Suponga que la moneda no fue lanzada un conjunto de 100 veces sino 100 conjuntos de 100 lanzamientos, lo que equivaldría a 100 experimentos. De estos 100 experimentos se podrían obtener una variedad de resultados: 58 + 42, 46 + 54, 51 + 49, etcétera. Cerca de 95 o 96 de estos experimentos producirían caras con márgenes de 40 y 60. Esto es, que solamente 4 o 5 de estos experimentos producirían menos de 40 o más de 60 caras. De forma similar, si se realiza un experimento y se encuentra una diferencia entre dos medias, después de una prueba estadística apropiada, con nivel de significancia de .05, entonces habrá una razón para creer que la diferencia de medias obtenida no es meramente una

diferencia por azar, aunque *podría* serlo. Si el experimento se hiciera 100 veces y realmente no hubiera diferencia entre las medias, cuando mucho 5 de estas 100 réplicas podrían mostrar diferencias entre las medias lo suficientemente grandes para ser consideradas "significativas".

Aunque esta discusión puede ayudar a aclarar lo que es la significancia estadística, aún no se responden todas las preguntas realizadas anteriormente. El nivel .05 fue elegido en un principio —y ha persistido entre los investigadores— porque es considerado una forma de especulación razonablemente buena. No es ni demasiado alto ni demasiado bajo para la mayoría de la investigación científica social. Muchos investigadores prefieren el nivel de significancia de .01, el cual es un nivel muy alto de certeza; de hecho es "certeza práctica". Algunos investigadores dicen que el nivel de .10 puede usarse algunas veces, aunque otros dicen que 10 resultados debidos al azar en 100 son demasiados y que no estarían dispuestos a arriesgar una decisión con tales probabilidades. Otros dicen que el nivel de .01, o una probabilidad en 100, es demasiado riguroso, y que resultados "realmente" significativos pueden ser descartados por esta rigidez.

¿Debe elegirse un determinado nivel de significancia y ajustarse totalmente a él? Ésta es una pregunta difícil. Los niveles .05 y .01 han sido recomendados ampliamente. Hay una tendencia nueva que recomienda reportar los niveles de significancia de todos los resultados. Esto es, si un resultado es significativo al nivel .12, por ejemplo, debería ser reportado de esa forma. Algunos investigadores objetan esta práctica, ya que dicen que se debe hacer una apuesta y adherirse a ella. Otra escuela de pensamiento recomienda trabajar con los llamados "intervalos de confianza". Muchos investigadores dicen que los resultados no son significativos si no alcanzan el nivel .05 o .01. Rozeboom (1960) recomienda el uso de los intervalos de confianza y reportar los valores precisos de probabilidad de los resultados experimentales. Sin embargo, Brady (1988) establece que tal precisión generalmente carece de significado en las ciencias sociales y conductuales por la imprecisión de las mediciones. La idea básica es que en lugar de rechazar categóricamente las hipótesis si el grado .05 no se alcanza, se puede decir que la probabilidad de que el valor desconocido caiga entre .30 y .50 es de .95. Ahora bien, si la proporción empírica obtenida es de .60, por ejemplo, entonces ésta es una evidencia para que el investigador corrija su hipótesis sustantiva, o en lenguaje de hipótesis nula, la hipótesis nula se rechaza. Una excelente revisión de este tipo de problemas se encuentra en el libro de Kirk (1972), el cual contiene muchos ensayos importantes respecto a estos temas. Cohen (1994), Simon (1976, 1987), y Simon y Roscoe (1984) han argumentado contra el uso de estas pruebas de significancia. Estos temas son profundos y complejos y no pueden ser discutidos adecuadamente aquí.

En este libro el enfoque de los niveles estadísticos se usará por su simpleza. Para el estudiante que no tiene en mente hacer ninguna investigación, el asunto no es muy serio pero aquellos que se involucren en investigación deberán estudiar otros procedimientos, tales como métodos de estimación estadística, intervalos de confianza y métodos exactos de probabilidad. Un resultado estadísticamente significativo no implica significancia personal o práctica. Babbie (1990) ha mencionado cuatro puntos importantes respecto a su rechazo del uso de pruebas de significancia en la investigación de ciencias sociales. Él establece que los supuestos que subyacen a las pruebas estadísticas generalmente no se encuentran en ciertos tipos de estudios de investigación social. Estos supuestos se centran alrededor de los métodos de muestreo usados en investigación. Babbie también considera que hay una tendencia de los investigadores a interpretar las pruebas de significancia estadística como la fuerza de asociación o como significancia sustantiva.

Para ilustrar el cálculo y el uso de la prueba χ^2 con la tabulación cruzada, ahora se aplicarán a los datos de frecuencia de la tabla 10.1. La fórmula dada previamente se usa, pero con la tabulación cruzada su aplicación es más complicada que la que se hizo en la

TABLA 10.6 *Cálculo de χ^2, datos de la tabla 10.1*

26.1616[a] 1 –24.1616[b]	27.8384 52 24.1616	53
21.8384 46 –24.1616	24.1616 0 24.1616	46
47	52	99

[a] $f_e = (53 \times 47) / 99 = 25.616$; $(53 \times 52) / 99 = 27.8384$; etcétera.

[b] $f_o - f_e = 1 - 25.1616 = -24.1616$; etcétera.

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(1 - 25.1616)^2}{25.1616} + \frac{(52 - 27.8384)^2}{27.8384} + \frac{(46 - 21.8384)^2}{21.8384} + \frac{(0 - 24.1616)^2}{24.1616}$$

$$= 23.2013 + 20.9704 + 26.7319 + 24.1616 = 95.0653$$

tabla 10.4. La principal diferencia es el cálculo de las frecuencias esperadas. Los cálculos necesarios se muestran en la tabla 10.6. Las frecuencias esperadas, f_e, se ubican en la esquina superior izquierda de cada celda y son calculadas como se muestra en la nota de pie *a* de la tabla. Las frecuencias obtenidas, f_o, se dan en el centro de cada casilla. Los términos $f_o - f_e$, requeridos por la fórmula, se pueden ver en la esquina inferior izquierda de cada casilla, y son los mismos en todas las casillas, excepto por el signo. Esto es para las tablas de 2×2. La fórmula de la χ^2 simplemente requiere elevar al cuadrado estas diferencias, dividiendo los cuadrados por las frecuencias esperadas, y sumando los resultados. Estos cálculos se indican más abajo: $\chi^2 = 95.0653$, con un grado de libertad. (¿Por qué un grado de libertad?) Al observar la tabla de los valores de χ^2, un grado de libertad en el nivel .01, se lee 6.635. Dado que el valor excede esto sustancialmente, puede decirse que la χ^2 es estadísticamente significativa, que los resultados obtenidos probablemente no son debidos al azar y que la relación expresada en la tabla es "real" en el sentido de que probablemente no se deba al azar. Observe que χ^2 necesita una corrección si N es pequeña. La regla implica el uso de la llamada corrección por continuidad, que consiste en restar .5 de la diferencia absoluta entre f_o y f_e en la fórmula de χ^2 antes de elevar al cuadrado, cuando las frecuencias *esperadas* son menores que 5 en tablas de 2×2. Esta corrección es llamada "corrección de Yates" (véase Comrey y Lee, 1995).

La χ^2, como cualquier otro estadístico que indique significancia estadística, no nos dice nada acerca de la magnitud de la relación. Es una prueba de la independencia de las variables, entendiendo independencia en el sentido en que se expuso en el capítulo 9. No es, estrictamente hablando, una medida de asociación. Uno de los más viejos problemas de la estadística es indexar la fuerza o magnitud de la asociación o relación entre variables categóricas. Su complejidad impide su explicación aquí, pero un estadístico que es fácilmente aplicable y que puede ser usado con una tabla de contingencia de cualquier tamaño es la V de Cramer, una medida de asociación basada en el valor de la chi cuadrada. La fórmula es:

$$V = \sqrt{\frac{\chi^2}{N(k-1)}}$$

El valor de k es determinado por el número de renglones o por el número de columnas en la tabla de contingencia. El que tenga el valor más pequeño, el número de renglones o el número de columnas, es usado para el valor de k. N es la frecuencia total. En este caso es 99, dado que solamente 99 senadores votaron. Si se sustituye el valor de χ^2 calculado anteriormente, N y k en la ecuación, se obtiene:

$$V = \sqrt{\frac{95.0653}{99(1)}} = \sqrt{.9602} = .9799 \approx .98$$

que es un índice de la fuerza de la relación.

La V de Cramer es la generalización del coeficiente phi (ϕ). En las tablas de 2×2, la V de Cramer y phi son idénticos. Ocasionalmente el coeficiente de contingencia, C, aparece en la literatura. El consenso general es que este valor C no es tan adecuado como la V de Cramer. Por un lado, no es realmente comparable entre tablas de contingencia de diferentes tamaños. Por otro lado, nunca puede alcanzar el valor de 1.00, que es el valor de una asociación perfecta. Las mismas críticas no son aplicables a la V de Cramer o a phi (ϕ). Sin embargo, como lo señalaron Comrey y Lee (1992), estas medidas de asociación, especialmente el coeficiente phi, son materia de otros problemas. Hays (1994) recomienda firmemente el uso de medidas de asociación junto con pruebas de significancia estadística. En general, el mejor consejo para manejar datos categóricos es calcular χ^2 (para determinar la significancia estadística), calcular V, calcular los porcentajes como se explicó anteriormente y después interpretar los datos usando toda la información.

Tipos de tablas cruzadas y tablas

En general hay tres tipos de tablas: unidimensional, bidimensional y k-dimensional. El número de variables determina el número de dimensiones de una tabla: una tabla unidimensional tiene una variable, una tabla bidimensional tiene dos variables, y así sucesivamente. No importa cuántas categorías tenga cada variable, el número de variables siempre determina la dimensión de la tabla. Ya se ha considerado la tabla bidimensional, donde dos variables —una independiente y una dependiente— se constrastan entre sí. A menudo es fructífero y necesario considerar más de dos variables de manera simultánea. Teóricamente no hay límite para el número de variables que pueden ser consideradas en un mismo tiempo. Las únicas limitaciones son de tipo práctico: el tamaño insuficiente de la muestra y la dificultad para comprender las relaciones contenidas en una tabla multidimensional.

Tablas unidimensionales

Hay dos clases de tablas unidimensionales. Una es una "verdadera" tabla unidimensional, que es de poco interés aquí porque no expresa una relación. Tales tablas se presentan frecuentemente en las revistas o periódicos, publicaciones del gobierno, etcétera. Al reportar el número o proporción de hombres y mujeres en San Francisco, el número de automóviles de diferentes marcas producidos en 1992, el número de niños en cada uno de los grados

TABLA 10.7 *Réplica del estudio de conformismo de Asch (datos de Walker y Andrade)*

	Grupos de edad (años)				
	3-5	6-8	9-11	12-14	15-17
% conformismo	85	42	38	9	0
% no conformismo	15	58	62	91	100

de X sistema escolar, tenemos "verdaderas" tablas unidimensionales. Solamente una variables es usada en la tabla.

Los científicos sociales a veces escogen reportar sus datos en tablas que parecen unidimensionales pero que realmente son bidimensionales. Consideremos la tabla reportada por Walker y Andrade en 1996. Este estudio extrajo una muestra de niños en edad escolar quienes participaron en una réplica del estudio de Asch sobre conformismo, de 1956. En el estudio de Asch el participante era ubicado en un grupo de "cómplices" del experimentador, que se comportaban como si también fueran participantes en el estudio. La tarea implicaba elegir una de tres líneas que fuera del mismo largo que la línea de prueba. En el ensayo clave, un cómplice elegía a propósito la línea incorrecta. El interés era ver si el participante se conformaba y escogía la misma línea incorrecta cuando la elección era apoyada por los otros cómplices. La tabla 10.7 muestra el porcentaje de veces que el participante se conformaba, en cada grupo de edad. (En la tabla original solamente se incluyeron los porcentajes por renglón del primer renglón.) La tabla parece unidimensional, pero realmente expresa una relación entre dos variables: edad y conformismo.

El punto clave es que las tablas de este tipo no son realmente unidimensionales. En la tabla 10.7, una de las variables, el conformismo, está expresada en forma incompleta. Para aclarar esto solamente se suma otro renglón de porcentajes al lado de los que ya están en la tabla original (esto ya se ha hecho en la tabla 10.7). Este renglón puede ser etiquetado como "no conformismo". Ahora se tiene una tabla bidimensional completa, y las relaciones se hacen obvias. (A veces esto no se puede hacer porque los datos para "completar" la tabla no se tienen.)

Como otro ejemplo, consideremos los datos presentados en la tabla 10.8. Los datos fueron tomados de un estudio de Child, Potter y Levine (1946). En este estudio los valores expresados en los libros de texto de los niños de tercer año fueron analizados en su contenido. La tabla 10.8 muestra los porcentajes de casos en que se dio reforzamiento por varios modos de adquisición. Como en el estudio de Walker y Andrade, solamente se dio un nivel de respuesta. Se agregó el otro nivel de respuesta en la tabla 10.8, en el último renglón y los valores correspondientes están en paréntesis.

TABLA 10.8 *Datos incompletos (presentados en el estudio de Child, Potter y Levine)*

	Esfuerzo	Comprando, vendiendo, negociando	Pidiendo, deseando, tomando lo que se oferta	Dominancia, agresión, robo, trampa
% en que se premió	93	80	68	41
(% en que no se premió)	(7)	(20)	(32)	(59)

TABLA 10.9 *Efecto de la cantidad solicitada en el tamaño de la donación realizada (estudio de Doob y McLaughlin)*[a]

	Cantidad solicitada		
	Cantidad no especificada	**Solicitudes más pequeñas ($5, $10, $25)**	**Solicitudes más grandes ($59, $100, $250)**
Tamaño de la donación realizada			
<$30	52%	36%	44%
$30-$49	19%	38%	8%
$50-$74	21%	16%	29%
$75-$99	1%	1%	4%
$100	7%	8%	12%
>100	0%	1%	3%

[a] $\chi^2 = 111.3$ ($p < .01$); $V = .26$.

Tablas bidimensionales

Las tablas bidimensionales o tabulación cruzada tienen dos variables, cada una con dos o más subclases. La forma más simple de una tabla bidimensional, como se ha visto, es llamada 2-por-2 o simplemente 2×2. Las tablas bidimensionales no están limitadas a la forma de 2×2; de hecho no hay una limitante lógica en el número de subclases que cada variable pueda tener. A continuación se presentan algunos ejemplos de tablas $m \times n$.

Doob y McLaughlin en 1989 estudiaron la relación entre la dimensión donación–no donación y la cantidad solicitada para la donación. Se pidió a los participantes en este estudio hacer una donación en dinero. La cantidad de dinero solicitada fue manipulada para examinar su efecto en la gente: si donaba o no donaba. En este artículo se presenta una tabla que relaciona el tamaño de la donación y la cantidad solicitada. Reportaron la tabulación cruzada de 6×3 de la tabla 10.9. Los resultados mostraron que el tamaño de la donación está relacionado con la cantidad solicitada. El valor obtenido de la chi cuadrada fue $\chi^2 = 111.3$, que es altamente significativo y el de $V = .26$, una relación media. (Los autores no calcularon una medida de asociación.) Aquí se muestra un método simple pero efectivo para probar la hipótesis y analizar los datos. Los investigadores encontraron que las solicitudes mayores eran más efectivas. Este estudio también es notorio por yuxtaponer una variable continua (cantidad de la donación) con una variable ordinal (cantidad solicitada). Esta tabla también ilustra un punto que parece confundir a los estudiantes: que los números de m y n de una tabulación cruzada $m \times n$ indican el número de subclases o subcategorías, y no el número de variables (m representa el número de categorías de la primer variable y n el número de categorías de la segunda variable).

Otro ejemplo de una tabla bidimensional que aborda datos de estudio interesantes proviene de la investigación clásica de Stouffer (1995)[1] sobre el conformismo y la toleran-

[1] Este libro contiene un análisis exhaustivo de la tabulación cruzada y casi puede ser considerado un modelo de cómo analizar las relaciones a través de la tabulación cruzada. Todas las especificaciones de Stouffer sobre sus datos son especialmente valiosas. Como ejemplo, véase el capítulo 4 donde Stouffer yuxtapone edad, educación, tolerancia y otras variables.

TABLA 10.10 *Relación entre educación y tolerancia (estudio de Stouffer)*

Porcentaje de distribución de las puntuaciones en la escala de tolerancia	Graduados de universidad	Universidad incompleta	Graduados de preparatoria	Preparatoria incompleta	Graduados de primaria
Menos tolerante	5	9	12	17	22
Entre ambos	29	38	46	54	62
Más tolerante	66	53	42	29	16
N	308	319	768	576	792

cia. Stouffer estudió la relación entre tolerancia, por un lado, y muchas otras variables sociológicas por otro lado. Una de estas últimas fue la educación. Stouffer buscó una respuesta a la pregunta: ¿Cuál es la relación entre la cantidad de educación y el grado de tolerancia? La tabulación cruzada que se muestra en la tabla 10.10 es ilustrativa. Un estudio de esta tabla muestra que la relación entre las dos variables existe: evidentemente, a mayor educación, mayor tolerancia.

Observemos brevemente un análisis similar de una clase diferente de problema de investigación. Shaw, Borough y Fink (1994) estudiaron la relación entre la orientación sexual percibida y la conducta de ayuda. Estos investigadores esencialmente preguntaron: ¿Hay una relación entre recibir ayuda y la orientación sexual de la persona que la solicita? Usando la "técnica del número equivocado" los investigadores obtuvieron una medida no reactiva de homofobia. La tabla 10.11 presenta un hallazgo parcial. Los números principales en las casillas son frecuencias y los porcentajes (o proporciones) se dan en paréntesis. Viendo los porcentajes, es evidente que la gente tiende más a ayudar a una persona que es heterosexual que a una persona que es homosexual. La χ^2 resultante fue 18.34 y fue estadísticamente significativa a un nivel $\alpha = .01$. La V de Cramer = .48. Sin embargo, es interesante observar que no existe una relación significativa entre el sexo de quien responde y la orientación sexual percibida de quien solicita la ayuda (esta tabla no se muestra en el texto). El resultado de la prueba χ^2 fue .33.

Tablas bidimensionales, dicotomías "verdaderas" y medidas continuas

Muchas tablas bidimensionales reportan datos nominales "verdaderos", datos de variables que son realmente dicotomías: sexo, vivo-muerto, y otros similares. Sin embargo, muchas

TABLA 10.11 *Relación entre la orientación sexual percibida y el comportamiento de prestar ayuda (estudio de Shaw, Borough y Fink)*

	Orientación del solicitante		
	Heterosexual	Homosexual	Respuestas totales
Ayuda	32(80)	13(33)	45
No Ayuda	8(20)	27(67)	35
Total	40	40	80

TABLA 10.12 *Relación entre autoestima y raza en los niños de escuelas de Baltimore (estudio de Rosenberg y Simmons)*

Autoestima	Raza	
	Afroamericanos (%)	Americanos blancos (%)
Baja	19	37
Media	35	30
Alta	46	33
	100	100
N	1 213	682

de estas tablas tienen una o ambas variables presumiblemente continuas y dicotomizadas o tricotomizadas de forma artificial. En un estudio sobre la autoestima de niños afroamericanos en escuelas públicas de Baltimore, Rosenberg y Simmons (1971) mostraron que la autoestima de los niños afroamericanos no era, como se pensaba, más baja que la de los niños americanos blancos. La variable independiente, raza, está en la parte superior de la tabla 10.12 y la variable dependiente, la autoestima, a un lado. (Así, los porcentajes son calculados hacia abajo en las columnas.) Observe también que una variable continua, autoestima, se ha convertido en una variable ordinal.

Tablas de tres dimensiones y de *k*-dimensiones

Es teóricamente posible realizar análisis cruzados con cualquier número de variables, pero en la práctica el límite es de tres o cuatro, con más frecuencia de tres. Las razones para tal limitación son obvias: se necesitan *N* muy grandes y, lo más importante, la interpretación de los datos se hace considerablemente más difícil. Otro punto que hay que tener en mente es: nunca usar un análisis complejo cuando un análisis más simple puede lograr el trabajo analítico. Aun así, las tablas de tres y de cuatro dimensiones pueden ser útiles y brindar información indispensable.

El análisis de tres o más variables simultáneamente tiene dos propósitos importantes. Primero, estudiar las relaciones entre tres o más variables. Tomemos un ejemplo de tres dimensiones con las variables *A*, *B* y *C*. Se pueden estudiar las relaciones entre *A* y *B*, *A* y *C*, *B* y *C*, y entre *A*, *B* y *C*. El segundo propósito es controlar una variable al estudiar la relación entre las otras dos variables. Por ejemplo, se puede estudiar la relación entre *B* y *C* mientras se controla *A*. Un uso importante de este concepto es ayudar a detectar las relaciones espurias, otro uso es para "especificar" una relación, indicado cuándo o bajo qué condiciones, una relación es más o menos pronunciada.

Especificación

La *especificación* es el proceso de describir las condiciones bajo las cuales una relación existe o no existe, o existe en mayor o menor grado. Un ejemplo ayudará a aclarar este enunciado y también da la oportunidad de introducir las tablas de contingencia *k*-dimensionales así como el análisis multivariado de los datos de frecuencia.

Suponga que un investigador está interesado en la hipótesis de que el nivel de aspiración está relacionado positivamente con el éxito en la universidad. Específicamente, la

TABLA 10.13 *Relación entre el nivel de aspiración y el logro escolar, datos hipotéticos*

	ANA	BNA	
EU	140	60	200
NEU	60	140	200
	200	200	(400)

hipótesis dice que a mayor grado de aspiración, mayor es la probabilidad de graduarse. Suponga además, que el investigador tiene una medida dicotómica relativamente cruda del nivel de aspiración, así como una medida del éxito en la universidad. Esta medida sería si el estudiante se graduó o no. Las variables y categorías, entonces, son ANA (alto nivel de aspiración), BNA (bajo nivel de aspiración), EU (éxito en la universidad) y NEU (no éxito en la universidad). El investigador toma una muestra aleatoria de 400 estudiantes de segundo año de una universidad y obtiene su grado de aspiración midiéndolo directamente en ellos. Los 400 estudiantes se dividen en dos mitades con base en la medida del grado de aspiración. Al final de los tres años, se categoriza a los estudiantes en función a que se hayan graduado o no se hayan graduado. Suponga que los resultados son los que se muestran en la tabla 10.13.[2] Hay evidentemente una relación entre las variables: $\chi^2 = 64$, es significativa al nivel de .001, y la $V = .40$.

El investigador muestra estos resultados a un colega varón, un individuo amargo, que dice que éstos son cuestionables y que si se hubiera considerado la clase social, la relación podría ser muy diferente. Él razona que la clase social y el nivel de aspiración están fuertemente relacionados, y que la relación original podría sostenerse para los estudiantes de la clase media, pero no para los estudiantes de clase trabajadora. Afortunadamente, cuando revisa los datos recolectados descubre que tiene los índices de la clase social de todos los sujetos. El resultado de usar una tabulación cruzada de tres variables se muestra en la tabla 10.14. La inspección de los datos muestra que su colega estaba en lo cierto, y que la relación entre el grado de aspiración y el éxito en la universidad es considerablemente más pronunciado para los estudiantes de la clase media (CM), que para los estudiantes de clase trabajadora (CT).

El investigador puede estudiar las relaciones con mayor profundidad mediante el cálculo de los porcentajes en forma separada para la clase media y la clase trabajadora como se muestra en la tabla 10.14. En este caso, dado que las frecuencias en cada renglón de las mitades de la tabla totalizan 100, las frecuencias son, en efecto, porcentajes. Puede verse que la relación entre el nivel de aspiración y el éxito universitario es más fuerte en los estudiantes de clase media que en los estudiantes de clase trabajadora.

En el análisis anterior, los datos fueron especificados: se mostró, al introducir la variable clase social, que la relación entre el nivel de aspiración y el éxito en la universidad era mayor en un grupo (clase media) que en otro (clase trabajadora). Esto es similar al fenómeno de interacción discutido en el capítulo 9, donde se estableció que la interacción infiere que una variable independiente afecta de forma diferente a una variable dependiente en diferentes niveles o facetas de otra variable independiente. Estrictamente ha-

[2] Los totales marginales de la tabla 10.13 (y también los de la tabla 10.14) se han hecho iguales para simplificar la discusión y resaltar ciertos puntos que se observarán ahora y posteriormente. Esto es, por supuesto, no realista: las tablas de frecuencia pocas veces son así de complacientes.

TABLA 10.14 *Relaciones entre nivel de aspiración, clase social y logro escolar (datos hipotéticos)*

	CM		CT		
	ANA	BNA	ANA	BNA	
EU	80	20	60	40	200
NEU	20	80	40	60	200
	100	100	100	100	(400)
		(200)		(200)	

blando, "interacción" es un término usado en la investigación experimental y en el análisis de varianza, como se verá en capítulos subsecuentes. Existe la duda sobre si el término puede ser aplicado en la investigación no experimental y en la clase de análisis que ahora se examina. La posición tomada en este libro es que la interacción es un fenómeno general de gran importancia que ocurre tanto en investigación experimental como no experimental. La "validez" de la interacción en la investigación no experimental, sin embargo, es mucho más difícil de establecer que en la investigación experimental. De hecho, esto sucede en la "validez" de todas las relaciones en la investigación no experimental, como se verá en forma detallada en los capítulos 22 y 23. En resumen, las relaciones especificadas de la tabla 10.13 pueden ser vistas como una interacción o simplemente como una especificación de relaciones. Lo principal, por supuesto, es entender lo que está sucediendo: las relaciones son fuertes, débiles o aun de cero en diferentes niveles de otras variables independientes. En el ejemplo anterior, la relación entre el nivel de aspiración y el éxito universitario es diferente en las dos clases sociales. Con tales enunciados multivariados, se logra un acercamiento al corazón y espíritu de la investigación científica, el análisis y la interpretación.

Tabulación cruzada, relaciones y pares ordenados

Una relación es un conjunto de pares ordenados. Dos de las formas en las que se puede expresar un conjunto de pares ordenados son: 1) por un listado de pares y 2) graficándolos. Un coeficiente de correlación es un índice que expresa la magnitud de una relación. Una tabulación cruzada expresa los pares ordenados en una tabla de frecuencias.

Para mostrar cómo estas ideas están relacionadas, tomemos los datos ficticios de la tabla 10.15. El estudio consiste en la relación entre el control estatal de un sistema económico y la democracia política. En una investigación de democracia política en países modernos, Bollen (1979) hipotetizó que a mayor control del sistema económico de un país,

TABLA 10.15 *Relación entre control estatal del sistema económico y desarrollo político (datos ficticios)*

	B_1 bajo CE		B_2 alto CE		
A_1 bajo DP	(0, 0)	2	(0, 1)	8	10
A_2 alto DP	(1, 0)	10	(1, 1)	3	13
	12		11		

menor es su nivel de democracia política. Suponga que de una muestra de 23 países, se cuentan 12 de ellos con un bajo control económico (bajo CE) y 11 países con alto control económico (alto CE). También hay 13 países con un desarrollo político elevado (alto DP) y 10 países con un bajo desarrollo político (bajo DP). Esto da los totales marginales de una tabulación cruzada de 2 × 2, aunque no indica cuántos países hay en cada casilla.

Ahora se cuenta el número de países con bajo CE que tienen un alto DP y el número de países con alto CE que tienen bajo DP. Estos conteos se anotan en las casillas apropiadas de una tabulación cruzada de 2 × 2 como en la tabla 10.15. Se encuentra que las frecuencias de las casillas se apartan significativamente de lo esperado por el azar, por lo tanto, existe una relación significativa entre el control económico estatal y el desarrollo político.

Para las tablas de 2 × 2 donde las frecuencias esperadas son pequeñas (<10), se debe usar la prueba exacta de significancia desarrollada por Fisher (1950). Otras alternativas serían usar la corrección de Yates para una prueba χ^2, o usar las tablas de Finney (véase Pearson y Hartley, 1954; Ferguson, 1971; Comrey y Lee, 1995).

Para poder ver los pares ordenados claramente, se cambian las etiquetas de las variables, así B_1 es igual a bajo CE, B_2 es igual a alto CE, A_1 es igual bajo DP y A_2 es igual a alto DP. Las etiquetas A y B han sido insertadas apropiadamente en la tabla 10.15. Ahora, ¿cómo se establecen los pares ordenados de la tabulación cruzada? Esto se hace asignando cada uno de los 23 países a una de las siguientes combinaciones de subgrupos: (1, 1), (0, 1), (1, 0), (0, 0) (véase las designaciones en la tabla 10.15). En otras palabras, a A_1 y

TABLA 10.16 *Arreglo de pares ordenados de la tabla 10.15*

Países	A	B	Intersecciones de la tabulación cruzada
1	1	0	
2	1	0	
3	1	0	
5	1	0	A_2B_1
6	1	0	
7	1	0	
8	1	0	
9	1	0	
10	1	0	
11	1	1	
12	1	1	A_2B_2
13	1	1	
14	0	0	A_1B_1
15	0	0	
16	0	1	
17	0	1	
18	0	1	
19	0	1	A_1B_2
20	0	1	
21	0	1	
22	0	1	
23	0	1	

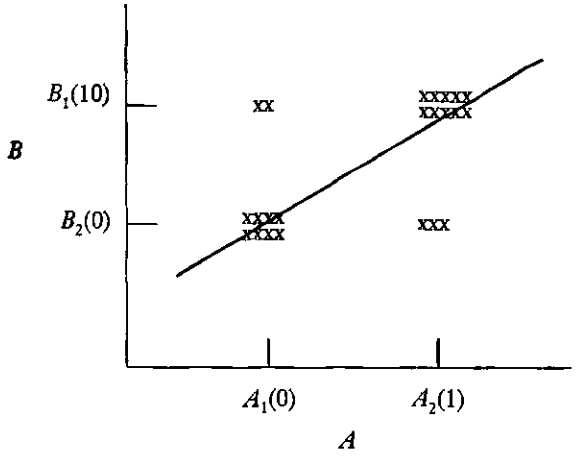
B_1 se les asignan ceros y a A_2 y B_2 se les asignan unos. Si un país tiene bajo CE y alto DP, entonces es un A_2 B_1; consecuentemente el par ordenado asignado para éste es (1, 0). Los primeros 10 países de la tabla 10.16 pertenecen a la categoría A_2 B_1, por lo tanto se les asignan (1, 0). De la misma forma, a los países restantes se les asignan los pares ordenados de números de acuerdo con su correspondiente subconjunto. La lista completa de 23 pares ordenados se presenta en la tabla 10.16. Las categorías o intersecciones de las tabulaciones cruzadas (conjunto) han sido indicadas.

La relación es el conjunto de pares ordenados de 1 y 0. La tabla 10.16 es solamente una forma diferente de expresar la misma relación mostrada en la tabla 10.15. Es posible calcular un coeficiente de correlación para ambas tablas. Si, por ejemplo, calculamos un coeficiente de correlación, una r producto-momento de los datos de la tabla 10.16, se obtiene .56. (La r producto-momento calculada con 1 y 0 es llamada coeficiente phi (ϕ).

Grafiquemos la relación en 2 ejes, A y B, en ángulo recto, donde A y B representen las dos variables contenidas en las tablas 10.15 y 10.16. Se busca estudiar la relación entre A y B. La figura 10.4 muestra los pares ordenados graficados y también muestra una línea de "relación" que corre a través de los racimos más grandes de pares. ¿Dónde está la relación? ¿Hay un conjunto de pares ordenados que defina una relación significativa entre A y B? Se apareó la puntuación de cada país en A con la puntuación de cada país en B y se graficaron los pares en los ejes A y B. Regresando a lo sustancial de la relación, ahora, para cada país, se aparea la puntuación individual de control económico con su correspondiente puntuación de desarrollo político; de esta manera se obtiene un conjunto de pares ordenados el cual representa una relación. Sin embargo, la pregunta real no es si existe una relación entre A y B sino cuál es la naturaleza de dicha relación.

Se puede ver en la figura 10.4 que la relación entre A y B es bastante fuerte. Esto está determinado por los pares ordenados que son en su mayoría (A_1B_2) y (A_2B_1). Hay, comparativamente, pocos pares (A_1B_1) (A_2B_2). Diciéndolo en palabras, las puntuaciones de bajo CE se aparean con puntuaciones de alto DP (1); y las puntuaciones de alto CE se aparean con bajo DP (0) con pocas excepciones (5 casos de 23). No se puede nombrar a esta relación de forma sucinta, como en una relación de "matrimonio" o "hermandad". Sin embargo, se le puede llamar "control económico estatal-desarrollo político", lo que significa que hay una relación de estas variables en el sentido de pares ordenados.

FIGURA 10.4



La razón de probabilidad

Un estadístico muy útil que puede ser calculado partiendo de las tablas de contingencia de 2×2 es la razón de probabilidad. Este estadístico es difícil de definir verbalmente, pero muy fácil de ilustrar. Por definición, es la razón o tasa de dos probabilidades. Las probabilidades son calculadas como la razón de la probabilidad de que el evento ocurra con la probabilidad de que dicho evento no ocurra. Por ejemplo, tomemos un mazo de 52 barajas; si se desea conocer la probabilidad de que salga una reina, se establece la siguiente razón:

$$\text{Probabilidad } (\textit{reina}) = \frac{4}{52} = \frac{1}{13} = 0.077$$

La probabilidad de que no salga una reina es

$$\text{Probabilidad } (\textit{no reina}) = \frac{48}{52} = \frac{12}{13} = 0.923$$

La razón de probabilidad de que salga una reina sería

$$\text{Probabilidad } (\textit{reina}) = \frac{4/52}{48/52} = \frac{1}{12} = 0.083$$

Si se utilizan los datos presentados en la tabla 10.13, puede verse cómo funciona la razón de probabilidad y por qué es útil en muchas situaciones. Para ser consistentes con el ejemplo anterior, se cambiaron las frecuencias a probabilidades o proporciones. La tabla que sigue lo refleja.

	ANA	BNA
EU	.7	.3
NEU	.3	.7

Las probabilidades de éxito, si el estudiante está en el grupo de alto nivel de aspiración, son

$$\text{Probabilidad } (\textit{éxito} \mid \textit{alto}) = \frac{.7}{.3} = 2.33$$

Esto indica que los estudiantes en el grupo de alto nivel de aspiración tienen 2.33 veces más probabilidad de ser exitosos en la universidad. Las probabilidades de éxito, si el estudiante está en el grupo de bajo nivel de aspiración son

$$\text{Probabilidad } (\textit{éxito} \mid \textit{bajo}) = \frac{.3}{.7} \approx .43$$

Esto podría llevar a la interpretación de que hay menos de media probabilidad de que un estudiante del grupo de bajo nivel de aspiración termine con éxito la universidad. Si se calcula la tasa entre estas dos probabilidades, se obtiene la razón de probabilidad:

$$\textit{Razón de probabilidad} = \frac{2.333}{0.429} = 5.444$$

La razón de probabilidad indica que los estudiantes en el grupo de alto nivel de aspiración tienen 5.444 veces más probabilidad de tener éxito o terminar con éxito la universidad que los estudiantes de bajo nivel de aspiración.

La razón de probabilidad nos da información útil. Ayuda a tratar de explicar qué está sucediendo. El estadístico chi cuadrada sigue siendo el método preferido; sin embargo, es incapaz de dar el tipo de información que la razón de probabilidad proporciona. El concepto que subyace a la razón de probabilidad es más difícil para los estudiantes; sin embargo, aprender acerca de este estadístico es importante cuando se trabaja con datos categóricos; es especialmente útil cuando se consideran las tablas de contingencia multifactoriales o análisis en los que se usan funciones logísticas. Se revisará más de este estadístico en el capítulo 35, donde se aprenderá también un estadístico chi cuadrada diferente. Howell (1997) presenta un ejemplo interesante acerca de la efectividad de las aspirinas en la disminución de la incidencia de ataques cardiacos. Las probabilidades individuales fueron muy pequeñas; sin embargo, la razón de probabilidad fue muy grande. Una persona en el grupo que no toma aspirina tiene 1.83 veces más probabilidad de sufrir un ataque cardiaco que una persona que toma dosis bajas de aspirina.

Análisis multivariado de datos de frecuencia

La mayor parte de la discusión anterior se limitó a dos variables: una variable independiente y una variable dependiente. Sin embargo, muchos análisis de datos de frecuencia son de tres y más variables. Un ejemplo ficticio con tres variables se dio anteriormente en la tabla 10.14. Mientras que la mayoría de los casos de más de tres variables pueden ser analizados e interpretados usando porcentajes, los estudios de datos con cuatro o más variables no son sujetos de análisis e interpretación de esta naturaleza y se necesita otro enfoque. Aún con tres variables a veces es necesario otro enfoque porque los datos son demasiado complejos y sutiles para una interpretación simple. Con una tabulación cruzada de dos variables hay solamente una relación: aquella entre *A* y *B*. Con tres variables, sin embargo, hay cuatro relaciones de posible interés: *AB*, *AC*, *BC* y *ABC*. Hasta aquí se han estudiado las tabulaciones cruzadas de tres por dos variables. La tabulación cruzada de tres variables, *ABC*, es como la que se mostró en la tabla 10.14, y en este caso puede ser más útil si se analiza el estudio de la relación entre el nivel de aspiración y el éxito en la universidad en dos muestras: clase media y clase trabajadora. Esto es, se estudia si la relación entre el nivel de aspiración y el éxito en la universidad es el mismo en la clase media que en la clase trabajadora. Si es el mismo, se "establece" una *no varianza*; si es diferente, entonces se tiene una *interacción*: la relación es tal en la clase media pero es tal otra en la clase trabajadora.

Desde el inicio de los años 70 ha habido cambios importantes en la conceptualización de los problemas de investigación y en el análisis de datos. Algunos de los trabajos notables que han contribuido en el área de las tablas de contingencia multivariadas con datos de frecuencias son los de Grizzle, Starmer y Koch (1969); Bishop, Fienberg y Holland (1976); Goodman (1971) y Clogg (1979). Antes del desarrollo del análisis multivariado de medidas continuas y de frecuencias, el análisis —y su conceptualización— era en su mayoría bivariado. Los investigadores estudiaron las relaciones entre pares de variables, como se ha hecho en este capítulo. Mientras que la idea de estudiar la operación de muchas variables simultáneamente era bien conocida, el significado práctico de hacerlo tuvo que esperar hasta el surgimiento de la computadora y de otras formas diferentes de pensamiento. Más adelante en este libro se examinará la naturaleza de la computadora y su importante papel en la investigación. También se dará una descripción más completa del análisis multivariado de los datos de frecuencia. En la edición previa de este libro se introdujo una breve discusión en este capítulo, acerca de los modelos log-lineales para tablas multivariadas de frecuencia/contingencia. Desde entonces el campo se ha expandido lo suficiente para merecer una sección más larga, que será presentada en los capítulos que tratan sobre estadística multivariada.

Anexo computacional

La tabulación cruzada de dos dimensiones puede realizarse usando el programa de cómputo SPSS. Hay dos diferentes disposiciones que el usuario debe saber. La primera involucra un conjunto de datos de valores brutos. Un ejemplo de tal conjunto de valores brutos se muestra en la tabla 10.16. Con los datos brutos, se necesita dar una instrucción al SPSS para procesar los datos creando primero una tabla de contingencia seguida por el análisis. La segunda disposición es usada cuando el investigador ya ha construido la tabla de contingencia y necesita solamente obtener el análisis estadístico para esa tabla. Un ejemplo de esto se muestra en las tablas 10.14 y 10.15.

Para ilustrar la primera disposición se usan los datos ofrecidos en la tabla 10.16. Es necesario partir de que el lector ya haya leído el anexo computacional del capítulo 6 y conoce cómo usar el programa SPSS para Windows. Esto incluye el saber cómo definir las variables y cómo ingresar los datos a la hoja de cálculo del SPSS. La figura 10.5 muestra la pantalla del SPSS después de que los datos se han ingresado y el análisis estadístico apropiado está cerca de ser seleccionado. Note que la tabla 10.16 tiene 23 observaciones, pero en la figura 10.5 se muestran sólo los primeros 14 casos debido a las restricciones de espacio. Note también las similitudes entre la tabla 10.16 y la figura 10.5, respecto a la disposición de los datos.

FIGURA 10.5

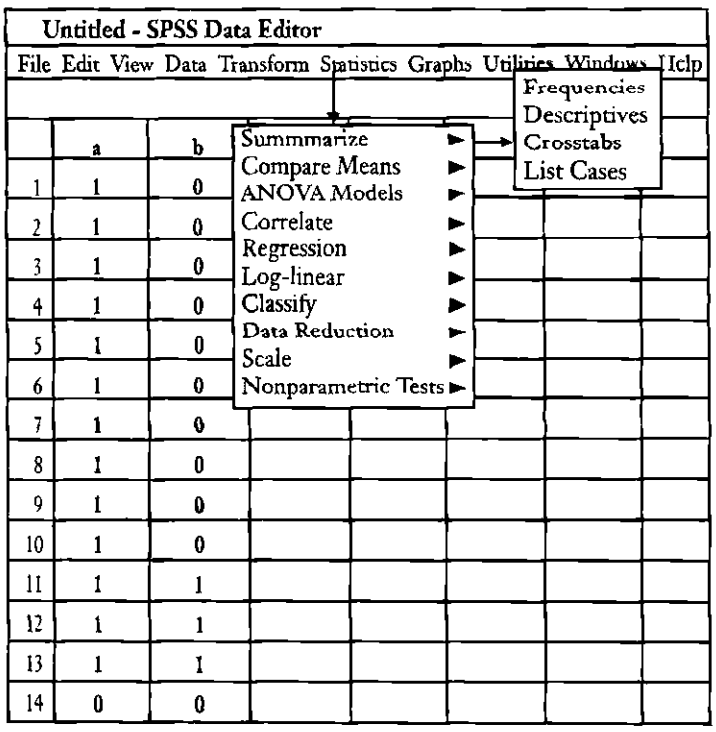
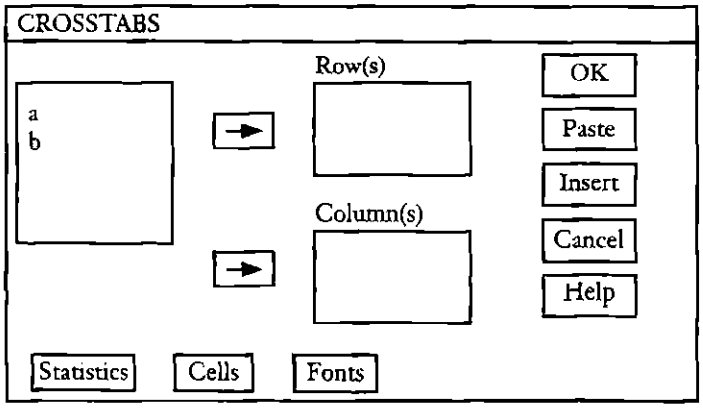


FIGURA 10.6



Después, se selecciona "Statistics" con un clic; en el siguiente menú seleccione "Crosstabs" para llegar a la pantalla que se muestra en la figura 10.6. Esta pantalla permite seleccionar qué variable estará en las filas o renglones (variable dependiente) en la tabla de contingencia y cuál estará en las columnas (variable independiente). También se necesita hacer clic en el botón "Statistics" para seleccionar los estadísticos que quiere desplegar en sus resultados. Para seleccionar la variable de las filas se resalta la variable "a" en la caja que está más a la izquierda y se hace clic en la flecha de arriba. Esto moverá la variable "a" a la caja de "Row(s)". En seguida, se resalta la variable "b" y se hace clic en la flecha inferior y la variable "b" se moverá de la caja de la izquierda a la caja de la de "Column(s)". La figura 10.7 muestra el resultado final de estas operaciones.

A continuación, hacer clic en "Statistics", lo cual producirá otra pantalla. De esta pantalla y para los propósitos, seleccionar "Chi-square" y los estadísticos "Phi & Cramer's V". Éstos son seleccionados al hacer clic en la caja que está junto a tales estadísticos. Una vez hecho esto, hacer clic en el botón "Continue" y regresará a la pantalla previa, que se

FIGURA 10.7

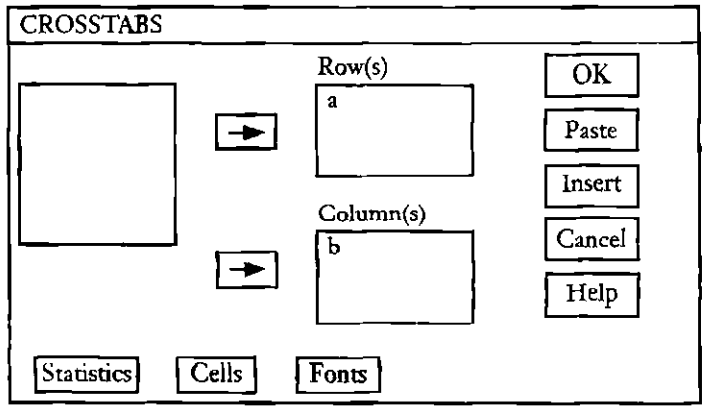
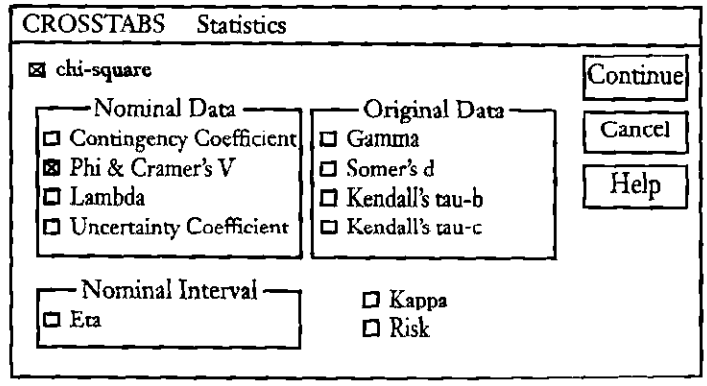


FIGURA 10.8



muestra en la figura 10.7. Una vez que se ha regresado a esta pantalla, se hace clic en el botón "OK". Entonces se verá que el SPSS cambia a la pantalla de resultados y despliega los resultados del análisis estadístico solicitado. Estos resultados se presentan en la figura 10.9.

La segunda disposición incluye que se haga el análisis usando solamente la tabla de contingencia en lugar de los valores de datos brutos. Se definirán las variables *A* y *B* en el SPSS, nuevamente, aunque en esta ocasión solamente se ingresarán las identificaciones para cada casilla. Recuerde que en la tabla 10.15 se dio la combinación (0, 0) para bajo DP-bajo CE. También se dio tal designación a las otras casillas de la tabla de contingencia, esto es bajo DP-alto CE tenían (0, 1), alto DP-bajo CE, tenían (1, 0) y alto DP-alto CE, tenían(1, 1). La figura 10.10 muestra la hoja de cálculo del SPSS donde se realizó esto en

FIGURA 10.9

		B	Low EC	High EC	
	Count		0	1	Row Totals
A					
Low PD	0		2	8	10 43.5
High PD	1		10	3	13 56.5
	Column Total		12 52.2	11 47.8	23

Chi-Square	Value	DF	Significance
Pearson	7.33963	1	.00675
Continuity Correction	5.23565	1	.02213

Phi −.56490
Cramer's V .56490

FIGURA 10.10

Untitled - SPSS Data Editor

File Edit View Data Transform Statistics Graphs Utilities Windows Help

	a	b	count	var	var	var	
1	0	0	2				
2	0	1	8				
3	1	0	10				
4	1	1	3				
5							

las primeras dos columnas. Note que hay una columna etiquetada "Count". En esta columna se ingresan los conteos de frecuencia para cada casilla. Por ejemplo, (0, 0) o bajo DP-bajo CE tuvo una frecuencia de dos. Junto a la designación (0, 0) en la hoja de cálculo, bajo la columna "Count", ingresar 2. Para (0, 1) ingresar un 8, un 10 para (1, 0) y un 3 para (1, 1).

Después de ingresar los datos apropiados al SPSS, es necesario indicarle que se tiene una disposición especial. El SPSS generalmente espera que la disposición esté en la forma mostrada en la figura 10.5. Para informar al SPSS, se selecciona "Data", en la barra superior, lo que llevará a otro menú. De este menú se elige "Weight Cases" (figura 10.11).

Note que "Weight Cases" está en negritas, para indicar que se va a elegir esa opción. Después de elegir esa opción aparecerá una nueva pantalla donde se indicará al SPSS cómo ponderar los casos. Esta pantalla se muestra en la figura 10.12. Observe que en la

FIGURA 10.11

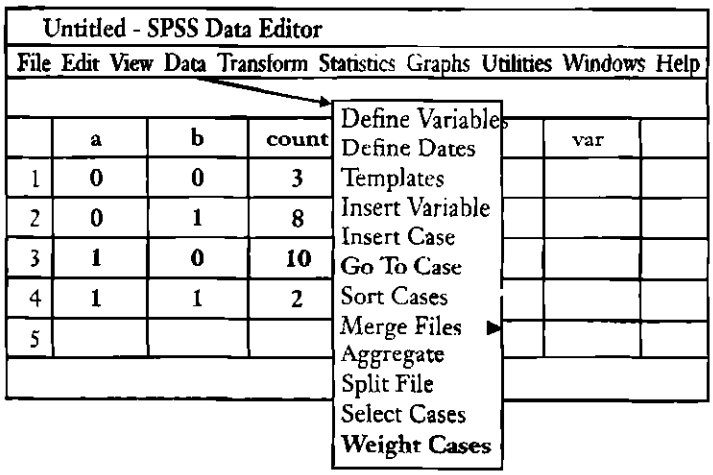
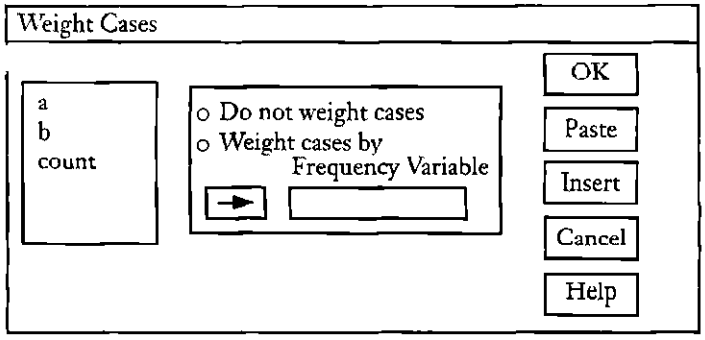


FIGURA 10.12



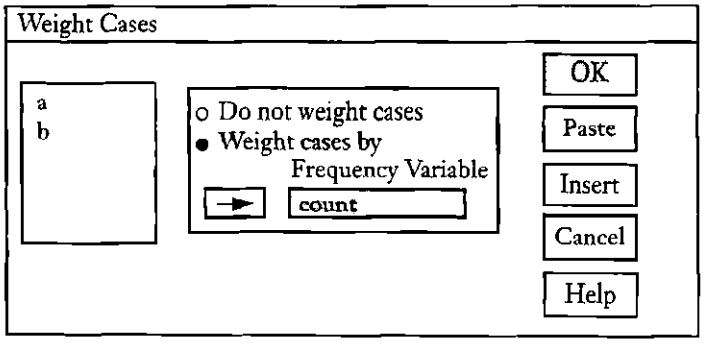
caja de la extrema izquierda están las tres variables: "a", "b" y "count". Primero, hacer clic en el botón etiquetado "Weight cases by" (figura 10.13). Entonces, con el ratón se resalta la variable "count" en la caja de la extrema izquierda. Haciendo clic en el botón de la flecha derecha del panel central, se verá que la variable "count" se mueve de la caja de la izquierda a la caja de la derecha. Una vez completado este movimiento, se hace clic en el botón "OK". Con esto se regresará a la hoja de cálculo del SPSS.

Para realizar el análisis estadístico se necesita seguir los pasos descritos en la primera disposición. Éstos se muestran en las figuras 10.6, 10.7 y 10.8. El resultado será idéntico al obtenido usando la primera disposición descrita en la figura 10.9. La clave para realizar el análisis de la tabla de contingencia por esta segunda disposición depende de cómo fueron designadas las casillas de la tabla de contingencia. Si se tiene una tabla de contingencia de 2×3, la designación sería (0, 0), (0, 1), (0, 2), (1, 0), (1, 1) y (1, 2).

RESUMEN DEL CAPÍTULO

1. Se introdujo a los fundamentos de cómo realizar el análisis con datos de frecuencia de partición cruzada.
2. La partición cruzada también es llamada tabulación cruzada, análisis de contingencia o análisis de tabla de contingencia.

FIGURA 10.13



3. Las variables categóricas también son llamadas variables nominales.
4. La tabulación cruzada es una presentación tabular numérica de los datos.
5. La tabulación cruzada se usa para determinar la naturaleza de las relaciones entre variables.
6. La forma más simple de una tabulación cruzada es una tabla de 2 por 2 o una tabla de cuatro casillas.
7. La regla generalmente aceptada en la construcción de las tablas de tabulación cruzada usa las columnas para los niveles de la variable independiente y los renglones para los resultados de la variable dependiente.
8. Los porcentajes en la tabulación cruzada son calculados de la variable independiente hacia la variable dependiente.
9. El estadístico chi cuadrada (χ^2) se usa para determinar la significancia estadística en una tabulación cruzada.
10. La significancia estadística se define como un resultado empírico que difiere significativamente de lo esperado por el azar.
11. El nivel de significancia estadística es elegido arbitrariamente; .05 y .01 son por lo general los niveles aceptados en las ciencias del comportamiento.
12. Si un resultado observado es significativo al nivel .05, se dice que el resultado pudo ocurrir por azar en no más de cinco de cada 100 ensayos del mismo experimento.
13. La V de Cramer o el coeficiente phi (ϕ) son medidas de asociación entre dos variables en una tabulación cruzada. El coeficiente de phi es usado en tablas de 2×2 y la V de Cramer es útil para tablas más grandes.
14. Tipos de tabulaciones cruzadas:
    a) Unidimensional
    b) Bidimensional
    c) Tri y *k*-dimensionales
15. La especificación es el proceso de describir las condiciones bajo las cuales una relación existe o no existe.
16. Una relación es un conjunto de pares ordenados. La tabulación cruzada expresa pares ordenados en una tabla de frecuencias.
17. El análisis de tablas multidimensionales es también llamado análisis log-lineal. Estas tablas son más complejas para analizar y requieren cálculos mucho más complejos.

SUGERENCIAS DE ESTUDIO

1. Freedman, Wallington y Bless (1967) presentan un estudio clásico que probó la hipótesis de que el sentirse culpable lleva a las personas a ser complacientes. Estos investigadores indujeron la culpa en los sujetos experimentales haciéndolos mentir acerca de una prueba que iban a tomar. A los sujetos control no se les hizo mentir. A los sujetos se les preguntó si estaban o no dispuestos a participar en un estudio no relacionado (variable dependiente: complacencia). Los autores reportaron la siguiente tabla de frecuencias:

	Experimental (mentir)	Control (no mentir)
Complace	20	11
No complace	11	20

Calcular χ^2, V y los porcentajes. Interprete los resultados. ¿Se acepta la hipótesis? ¿La relación es débil, moderada o fuerte?
(Respuestas: $\chi^2 = 5.23$ ($p < .05$); $V = .29$. Sí, la hipótesis se acepta. La relación es débil a moderada.)

2. El *Congressional Quarterly* (1993) reportó que el 3 de agosto de 1993, el senado de Estados Unidos votó para autorizar 1 500 millones de dólares para el Programa de Servicio Nacional. Esto proporcionaría a la gente de 17 años de edad o mayores $4 725.00 por año a lo largo de dos años en premios de educación por trabajo en programas de servicio a la comunidad. La votación fue como sigue:

	Republicano	Demócrata
A favor	7	51
En contra	37	4

Calcular χ^2, V y los porcentajes. Interprete los resultados.
(Respuestas: $\chi^2 = 59.45$; $V = .78$.)

3. Zavala, Barnett, Smedi, Istvan y Matarazzo (1990) investigaron la relación entre el consumo de cigarros, alcohol y café entre el personal de la armada de Estados Unidos. Una de sus tablas se reproduce parcialmente en seguida.

	Fumadores	Ex fumadores	No fumadores
Consumo de café			
0 tazas	24	12	66
1 a 2 tazas	10	3	8
3 o + tazas	16	3	6

   a) Examine los datos cuidadosamente, luego interprete la tabla.
   b) Calcule los porcentajes, primero por columnas y luego por renglones. ¿Cambia la interpretación? Si lo hace, ¿cómo cambia?

4. Si es posible, consiga un programa que calcule χ^2 (muchos están disponibles comercialmente y algunos otros pueden ser bajados de Internet). Usando ese programa, analice los ejemplos y los problemas en esta sección. Verifique sus respuestas.

5. ¿Han cambiado las ocupaciones de las mujeres bajo el impacto del movimiento de igualdad de derechos? Aquí se presentan datos del reporte del censo de EUA (en miles). Estos datos fueron obtenidos de la página Web de la Oficina de Censos de Estados Unidos: http://www.census.gov[3]

	1983		1995	
	Hombres	Mujeres	Hombres	Mujeres
Profesional, gerencial, administrativo	13 943	9 649	18 365	16 953
Contadores, ventas, servicio	11 068	20 198	13 320	24 097

(Nota: Los datos anteriores fueron obtenidos sumando las categorías profesional + gerencial + administrativos; contadores + ventas + servicio.)

[3] Las direcciones para ser usadas en Internet, se conservarán en inglés, pues es como se deben escribir para poder tener acceso al sitio o a la página de la dirección correspondiente.

a) Calcule los porcentajes, teniendo cuidado de calcularlos partiendo de la variable independiente hacia la variable dependiente, como es usual.
b) Calcule χ^2 y V para 1983 y 1995, de forma separada. (Use los datos anteriores; es decir, olvide el hecho de que las cifras indican miles. Esto afecta a la χ^2 pero no a la V.)
c) Interprete los resultados de sus cálculos. (Sea circunspecto, el método de sumar los números de las categorías puede haber originado un sesgo, e incluso ser incorrecto.)
d) En *b*, arriba, usted calculó χ^2 y V usando las frecuencias tabuladas como están. Ahora haga los mismos cálculos usando los números en miles (es decir, en lugar de 13 943, use 13 943 000). Observe el enorme incremento en χ^2, pero V es la misma. He aquí una generalización: con números muy grandes, virtualmente todo es estadísticamente significativo. Ésta es una ventaja de las medidas de asociación, que permanecen sin ser afectadas por la magnitud de los números.

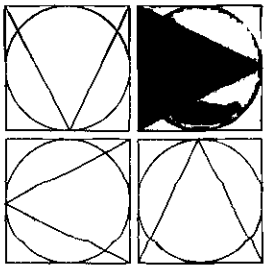
6. Los siguientes datos fueron recolectados por Glick, DeMorest y Hotze (1988) en su estudio acerca de la pertenencia al grupo, el espacio personal y la solicitud de un pequeño favor. Este estudio fue descrito brevemente en un capítulo anterior. Los investigadores querían determinar si la similitud de las características personales entre quien solicita el favor y el solicitado, influyen o no en el hecho de que el solicitado acceda a la petición. También fue de interés ver si la distancia entre el solicitante y el solicitado influye en la complacencia.

Tipo de cómplice

	Externo al grupo			Dentro del grupo		
	Distancia			Distancia		
	Cerca	Medio	Lejos	Cerca	Medio	Lejos
Respuesta a la solicitud						
Aceptó	1	6	12	10	12	9
Rechazó	14	9	3	5	3	6

a) Calcule los porcentajes e interprete. Considere cada tipo de cómplice en forma separada.
b) ¿Cómo influye la distancia en la complacencia?, ¿fuertemente?, ¿moderadamente? ¿Hay la misma relación con el solicitante dentro del grupo que con el solicitante externo al grupo?
c) Este estudio deberá ser analizado usando una tabla de contingencia multidimensional. Explique por qué.

7. Si usted tiene disponible una versión del SPSS para Windows, trate de analizar los datos de las sugerencias de estudio 1, 2 y 3.



CAPÍTULO 11

ESTADÍSTICA: PROPÓSITO, ENFOQUE Y MÉTODO

- EL ENFOQUE BÁSICO
- DEFINICIÓN Y PROPÓSITO DE LA ESTADÍSTICA
- ESTADÍSTICA BINOMIAL
- LA VARIANZA
- LA LEY DE LOS NÚMEROS GRANDES
- LA CURVA NORMAL DE PROBABILIDAD Y LA DESVIACIÓN ESTÁNDAR
- INTERPRETACIÓN DE DATOS USANDO LA CURVA NORMAL DE PROBABILIDAD CON DATOS DE FRECUENCIA
- INTERPRETACIÓN DE DATOS UTILIZANDO LA CURVA NORMAL DE PROBABILIDAD CON DATOS CONTINUOS

El enfoque básico

El principio básico detrás del uso de las pruebas estadísticas de significancia puede enunciarse de la siguiente forma: comparar los resultados obtenidos con lo esperado por el efecto del azar; dicho de otra forma, ¿se obtuvo lo que se esperaba por efecto del azar? Cuando se realiza una investigación y se obtienen resultados estadísticos, éstos se comparan con los resultados esperados por el azar. En el capítulo 7 se dieron ejemplos donde se comparaban los resultados empíricos del lanzamiento de dados y monedas con las expectativas teóricas. Por ejemplo, si un dado se lanza un gran número de veces, la proporción esperada de que resulte un cuatro es un sexto del número total de lanzamientos. En el capítulo 10 se aprendió que el fundamento de la prueba χ^2 es la comparación de las frecuencias observadas de eventos, con las frecuencias esperadas por el azar. En realidad, las nociones estadísticas del capítulo 10 se presentaron previamente al presente capítulo, en parte para ofrecer al estudiante experiencia preliminar respecto a los resultados obtenidos y a los esperados.

En el capítulo 7 se describió una demostración donde un par de dados fueron lanzados 72 veces; teóricamente, el 7 caería $1/6 \times 72 = 12$ veces. No obstante, la tabla 7.2 indica que el 7 cayó en 15 de los 72 lanzamientos, en lugar de 12. Entonces surgen diversas interrogantes: ¿el resultado obtenido difiere en forma significativa del resultado esperado teóricamente? ¿Este resultado obtenido difiere de lo esperado por el azar, lo suficiente como para garantizar la creencia de que es efecto de algo distinto al azar? ¿Los resultados pueden ser explicados únicamente por el azar?

Preguntas como ésas constituyen la esencia del enfoque estadístico. Los estadistas son escépticos, ya que no creen en la "realidad" de los resultados empíricos hasta que han sido sometidos al análisis estadístico. Ellos suponen que los resultados se deben al azar, hasta que se compruebe lo contrario. Son probabilistas rigurosos; la esencia de su enfoque a los datos empíricos consiste en establecer expectativas basadas en probabilidades como sus hipótesis y tratar de ajustar los datos empíricos al modelo de probabilidad. Si los datos empíricos se "ajustan" al modelo de probabilidad, entonces se dice que no son "estadísticamente significativos"; si no se ajustan, y se apartan "lo suficiente" del modelo del azar, entonces se les considera "estadísticamente significativos".

Este capítulo, y muchos de los sucesivos, están dedicados al enfoque estadístico de los problemas de investigación. En el presente capítulo se extiende la discusión del capítulo 7 sobre probabilidad a conceptos básicos de la media, varianza y desviación estándar. También se explica e interpreta la llamada ley de los números grandes y la curva normal de probabilidad, así como algunos aspectos de su amplia utilidad en estadística. En el próximo capítulo se aborda la idea de la comprobación estadística en sí misma. Estos dos capítulos constituyen los fundamentos.

Definición y propósito de la estadística

La estadística es la teoría y el método de analizar datos cuantitativos obtenidos de muestras de observaciones para estudiar y comparar fuentes de varianza de los fenómenos, para ayudar en la toma de decisiones para aceptar o rechazar relaciones hipotetizadas entre los fenómenos, y para contribuir en la extracción de inferencias confiables a partir de observaciones empíricas.

En esta definición se plantean cuatro propósitos de la estadística. El primero es el más común y tradicional: reducir grandes cantidades de datos de manera que puedan manejarse y comprenderse. Por ejemplo, es imposible analizar 100 puntuaciones, pero si se calculan una media y una desviación estándar, una persona capacitada puede interpretarlas fácilmente. La definición de *estadístico* se deriva de este uso y propósito tradicionales de la estadística. Un estadístico es una medida calculada a partir de una muestra. El estadístico se contrasta con un parámetro, que es un valor poblacional. Si se calcula la media de U (una población o universo), ésta es un parámetro. Tome un subconjunto (muestra) A de U. La media de A es un estadístico. Para los propósitos de este libro, los parámetros representan un interés teórico ya que generalmente son desconocidos y son estimados con los estadísticos. Por ello, la mayoría de las veces se manejan muestras o subconjuntos estadísticos, los cuales se consideran como representativos de U. Por lo tanto, los estadísticos son resúmenes de las muestras —y quizás, con frecuencia, de las poblaciones— a partir de las cuales fueron calculados. Las medias, medianas, varianzas, desviaciones estándar, percentiles, porcentajes, etcétera, que se calculan a partir de muestras, son estadísticos.

Un segundo propósito de la estadística consiste en ayudar al estudio de poblaciones y muestras. Este uso no será discutido aquí, ya que es bien conocido; además de que ya se estudió algo al respecto de muestras y poblaciones en capítulos previos.

Un tercer propósito de la estadística es ayudar en la toma de decisiones. Si un psicólogo educativo necesita saber cuál de tres métodos de instrucción promueve mayor aprendizaje al menor costo, la estadística puede ayudar a obtener este conocimiento. Este uso de la estadística es comparativamente más reciente.

Aunque la mayoría de las situaciones de decisión resultan más complejas, se utilizará un ejemplo que ya es bastante familiar: suponga que usted es quien toma las decisiones en un juego de dados. Su primera tarea es determinar los resultados de los lanzamientos de dados, los cuales son, obviamente, del 2 al 12. Usted observa las diferentes frecuencias de los números; por ejemplo, el 2 y el 12 caerán probablemente con menos frecuencia que el 7 o el 6. Después, usted calcula las probabilidades de los diferentes resultados. Finalmente con base en la cantidad de dinero que espera ganar, diseña un sistema de apuestas. Usted decide, por ejemplo, que como la probabilidad de obtener un 7 es de 1/6, usted pedirá a su oponente que apueste 5 a 1, y no cantidades iguales en el primer lanzamiento. Para hacer más dramática la situación, suponga que dos jugadores operan con diferentes sistemas de toma de decisiones (este ejemplo fue sugerido por Bross en 1953). Usted es el jugador *A*, y propone el siguiente juego: *A* ganará si resulta 2, 3 o 4. El oponente *B*, ganará con 5, 6 o 7 (los resultados del 8 al 12 se descartarán). Es obvio que su sistema de toma de decisiones es defectuoso, ya que se basa en la suposición de que los resultados 2, 3, 4, 5, 6 y 7 son equiprobables. El jugador *B* la pasará bien en este juego.

El cuarto y último propósito de la estadística (ayudar a realizar inferencias confiables a partir de los datos observados) está muy relacionado y, de hecho, forma parte del propósito de ayudar a tomar decisiones acerca de las hipótesis. Una inferencia constituye una proposición o generalización derivada por medio del razonamiento a partir de otras proposiciones, o de la evidencia. En otras palabras, una inferencia es una conclusión a la que se llega por medio del razonamiento. En estadística diversas inferencias se pueden extraer de las pruebas de hipótesis estadísticas. Se "concluyó" previamente que los métodos *A* y *B* difieren realmente. A partir de la evidencia se concluye que si, por ejemplo, $r = .67$, las dos variables realmente están relacionadas.

Las inferencias estadísticas tienen dos características: 1) Las inferencias se hacen usualmente de muestras a poblaciones. Cuando se dice que las variables *A* y *B* están relacionadas, porque la evidencia estadística es $r = .67$, esto se infiere porque $r = .67$ en esta muestra es $r = .67$, o cercano a esto, en la población de la cual se extrajo la muestra. 2) Las inferencias se utilizan cuando los investigadores no están interesados en las poblaciones, o solamente tienen un interés secundario en éstas. Un investigador educativo estudia el supuesto efecto de las relaciones entre los miembros del consejo escolar y los administradores educativos en jefe, por un lado, y el estado de ánimo de los maestros, por el otro. La hipótesis afirma que cuando las relaciones entre los consejeros y los administradores se tensan, el estado de ánimo de los maestros se encontrará más afectado que cuando no es así. El investigador tiene interés en probar esta hipótesis únicamente en el condado *Y*. Después de realizar el estudio y obtener los resultados estadísticos, comprueba la hipótesis, por ejemplo, de que el estado de ánimo es más bajo entre los profesores del sistema *A* que entre aquellos de los sistemas *B* y *C*. El investigador infiere que la proposición hipotética inicial es correcta, a partir de la evidencia estadística de la diferencia entre el sistema *A*, por un lado, y los sistemas *B* y *C*, por el otro, en el condado *Y*. En realidad es posible que el interés del investigador se limite estrictamente al condado *Y*.

Para resumir lo anterior, los propósitos de la estadística pueden reducirse a un propósito principal: ayudar a realizar inferencias. Éste es uno de los propósitos básicos del diseño, metodología y estadística de la investigación. Los científicos buscan realizar inferencias a partir de datos. La ciencia de la estadística, con su poder para reducir datos a formas más manejables (estadísticos), y para estudiar y analizar varianzas, permite a los científicos unir

estimados de probabilidad a las inferencias que extraen de los datos. La estadística dice, en efecto, "la inferencia que extrajo es correcta a tal o cual nivel de significancia. Puede actuar como si su hipótesis fuera verdadera, recordando que existe tal o cual probabilidad de que sea falsa". Debe quedar razonablemente claro por qué algunos estadísticos contemporáneos llaman a la estadística la disciplina de la toma de decisiones en la incertidumbre. También debe quedar razonablemente claro que, sabiéndolo o no, las personas realizan inferencias de manera continua, calculando las probabilidades de varios resultados o hipótesis, y tomando decisiones con base en el razonamiento estadístico. La estadística, al usar la teoría de la probabilidad y las matemáticas, vuelve el proceso más sistemático y objetivo.

Estadística binomial

Al contar objetos, el sistema numérico resulta simple y útil. Siempre que se cuentan cosas, se hace con base en algún criterio, alguna variable o atributo, en el lenguaje de investigación. Ya se han dado muchos ejemplos: caras, cruces, números de dados, sexo, actos agresivos, preferencia política, etcétera. Si una persona o cosa posee el atributo, se dice que esta persona o cosa está "incluida". Cuando algo se "incluye" porque posee el atributo en cuestión, se le asigna el número 1. Si no posee el atributo, se le asigna el 0. Éste es un sistema binomial.

Con anterioridad se definió a la media como $M = \Sigma X/n$. La varianza es $V = \Sigma\chi^2/n$, donde $\chi = X - M$ (cada χ es una desviación de la puntuación en bruto X con respecto a la media). La desviación estándar es $DE = \sqrt{V}$. Obviamente estas fórmulas funcionan para cualquier puntuación; aquí se utilizan tan sólo con 1 y 0, y resulta útil modificar la fórmula para la media, ya que $\Sigma X/n$ no es lo suficientemente general debido a que en ella se asume que todas las puntuaciones son equiprobables. Una fórmula más general y que puede utilizarse cuando no se asume equiprobabilidad, es:

$$M = \Sigma[X \cdot w(X)] \qquad \textbf{(11.1)}$$

donde $w(X)$ representa el peso (*weight*, en inglés) asignado a una X; $w(X)$ simplemente significa la probabilidad que cada X tiene de ocurrir. La fórmula dice: multiplique cada X, cada puntuación, por su peso (probabilidad), y luego sume todos. Considere que si todas las X tienen la misma probabilidad, esta fórmula es la misma que $\Sigma X/n$.

La media del conjunto {1, 2, 3, 4, 5} es:

$$M = \frac{1 + 2 + 3 + 4 + 5}{5} = \frac{15}{5} = 3$$

Realizándolo a través de la ecuación 11.1, es lo mismo obviamente, pero el cálculo se ve diferente:

$$M = 1 \cdot \frac{1}{5} + 2 \cdot \frac{1}{5} + 3 \cdot \frac{1}{5} + 4 \cdot \frac{1}{5} + 5 \cdot \frac{1}{5} = 3$$

¿Y por qué tantas sutilezas? Véase el siguiente ejemplo. Si se lanza una moneda al aire, $U = \{C, X\}$. La media del número de caras sería, de acuerdo a la ecuación 11.1

$$M = 1 \cdot \frac{1}{2} + 0 \cdot \frac{1}{2} = \frac{1}{2}$$

Si se lanzan dos monedas al aire nuevamente, $U = \{CC, CX, XC, XX\}$. La media del número de caras, o el número de caras esperadas, es

$$M = 2 \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} + 0 \cdot \frac{1}{4} = \frac{4}{4} = 1$$

Esto significa que si se lanzan dos monedas muchas veces, el número promedio de caras por lanzamiento de las dos monedas es 1. Si se muestrea una persona de un grupo de 30 hombres y 70 mujeres, la media de hombres sería $M = 3/10 \cdot 1 + 7/10 \cdot 0 = 0.3$. La media de mujeres sería $M = 3/10 \cdot 0 + 7/10 \cdot 1 = 0.7$. Éstas son las medias para un resultado. (Esto sería parecido a decir "un promedio de 2.5 hijos por familia".)

Lo que se ha afirmado en estos ejemplos es que la media de cualquier experimento (un solo lanzamiento de una moneda, el muestreo de una persona) es la probabilidad de ocurrencia de uno de dos posibles resultados (caras, un hombre). Si se da el resultado, se le asigna un 1, y si no se da, se le asigna un 0. Esto equivale a decir $p(1) = p$ y $p(0) = 1 - p$. Si en el experimento de un solo lanzamiento de la moneda se asigna 1 a cara y 0 a cruz, entonces $p(1) = 1/2$ y $p(0) = 1 - 1/2 = 1/2$. Al lanzar una moneda dos veces, se asigna 1 a cada cara resultante y 0 a cada cruz. Suponga que el resultado de interés es "caras", por lo que $U = \{CC, CX, XC, XX\}$. La media sería:

$$M = \frac{1}{4} \cdot 2 + \frac{1}{4} \cdot 1 + \frac{1}{4} \cdot 1 + \frac{1}{4} \cdot 0 = 1$$

¿Podrá llegarse al mismo resultado de manera más sencilla? Sí. Solamente es necesario sumar las medias para cada resultado. La media del resultado del lanzamiento de una moneda es 1/2. Para dos lanzamientos es 1/2 + 1/2 = 1. Para determinar las probabilidades en el lanzamiento de una moneda, ponderamos, 1 (caras) con su probabilidad y 0 (cruces) con su probabilidad. Esto da $M = p \cdot 1 + (1 - p) \cdot 0 = p$. Ahora tome el ejemplo del muestreo de hombres y mujeres, suponiendo que p es igual a la probabilidad de que un hombre sea elegido para la muestra en un solo resultado, y $1 - p = q$ corresponde a la probabilidad de que sea una mujer. Entonces $p = 3/10$ y $q = 7/10$. Si el interés radica en conocer la media de que un hombre sea muestreado, entonces $M = p \cdot 1 + q \cdot 0 = p$, $M = 3/10 \cdot 1 + 7/10 \cdot 0 = 3/10 = p$, por lo que la media es 3/10 y la probabilidad es de 3/10. Evidentemente $M = p$, o la media es igual a la probabilidad.

¿Qué sucede en el caso de una serie de resultados? Se utiliza S para la suma de n resultados. El ejemplo del lanzamiento de monedas ya se consideró anteriormente. Tome nuevamente el ejemplo del muestreo de hombres y mujeres. La media de la ocurrencia de un hombre es de 3/10 y la media de la ocurrencia de una mujer es de 7/10. Si se muestrearan 10 personas, ¿cuál sería la media de los hombres? O, de otra forma, ¿cuál es la expectativa de los hombres? Si se suman las 10 medias de los resultados individuales, se obtiene la respuesta:

$$M(m_{10}) = M_1 + M_2 + \cdots + M_{10} \qquad \textbf{(11.2)}$$

$$= 3/10 + 3/10 + \cdots + 3/10 = 30/10 = 3$$

En una muestra de 10 sujetos, se esperaría obtener 3 hombres. El mismo resultado podría haberse obtenido con $3/10 \cdot 10 = 3$; pero $3/10 \cdot 10$ es pn, o

$$M(m_n) = pn \qquad \textbf{(11.3)}$$

En n ensayos la media de ocurrencias del resultado asociado con p es pn.

La varianza

En el capítulo 6 la varianza se definió como $V = \Sigma\chi^2/n$. En el presente capítulo se seguirá dicha definición, pero con el cambio de algunos símbolos (por la misma razón que en la fórmula de la media):

$$V = \Sigma[w(X)(X - M)^2] \tag{11.4}$$

Para dejar claro qué es una varianza —y una desviación estándar— en la teoría de la probabilidad, se darán dos ejemplos. Recuerde que en un binomio sólo existen dos resultados posibles, 1 y 0. Por lo tanto, X es igual a 1 o 0. Se preparó una tabla para ayudar a calcular la varianza del resultado cara al lanzar una moneda:

Resultado	X	$w(X) = p$	$(X - M)^2$	$(11/2)^2$
C	1	1/2	$(1 - 1/2)^2 = 1/4$	
X	0	1/2	$(0 - 1/2)^2 = 1/4$	

Entonces, la varianza es:

$$V = 1/2(1 - 1/2)^2 + 1/2(0 - 1/2)^2 = 1/2 \cdot 1/4 + 1/2 \cdot 1/4 = 1/4$$

La media es 1/2 y la varianza es 1/4. La desviación estándar es la raíz cuadrada de la varianza, o:

$$\sqrt{1/4} = 1/2$$

Sin embargo, la varianza de un resultado individual no tiene mucho significado. En realidad se busca la varianza de la suma de un número de resultados. Si los resultados son independientes, la varianza de la suma de los resultados es la suma de la varianza de los resultados:

$$V(m_n) = V_1 + V_2 + \cdots + V_n \tag{11.5}$$

Para 10 lanzamientos de una moneda, la varianza de caras es $V(H_{10}) = 10 \cdot 1/4 = 10/4 = 2.5$. Antes se mostró que $M(S_n) = np$; pero ahora se busca una fórmula para la varianza, es decir, en lugar de la ecuación 11.5 se requiere de una fórmula simple y directa. Con un poco de manipulación algebraica se puede llegar a dicha fórmula:

$$V = p(1 - p) = pq \tag{11.6}$$

Ésta es la varianza de un resultado. La varianza del número de veces que ocurre un resultado es, como en las ecuaciones 11.2, 11.3 y 11.5, la suma de las varianzas de los resultados individuales, o:

$$V(m_n) = npq \tag{11.7}$$

La desviación estándar es:

$$DE(m_n) = \sqrt{npq} \tag{11.8}$$

Las ecuaciones 11.3, 11.7 y 11.8 son importantes y útiles. Pueden aplicarse en muchas situaciones estadísticas. A continuación se mostrarán dos o tres aplicaciones. Primero considere un ejemplo en el cual, de una muestra de 100 sujetos ($n = 100$), 60 se mostraron a favor de un asunto político, y los 40 restantes se mostraron en contra. Suponiendo equiprobabilidad, $p = 1/2$ y $q = 1/2$, $M(m_{100}) = np = 100 \cdot 1/2 = 50$, $V(m_{100}) = npq = 100 \cdot 1/2 \cdot 1/2 = 25$, y $DE(m_{100}) = \sqrt{25} = 5$. Se encontró que había 60 acuerdos, por lo que ésta es una desviación de dos desviaciones estándar con respecto a la media de 50, $60 - 50 = 10$, y $10/5 = 2$. Para el segundo ejemplo se usará el experimento de lanzamiento de monedas del capítulo sobre probabilidad. En él, se obtuvieron 52 caras en 100 lanzamientos. Los cálculos son los mismos que los realizados previamente; puesto que hubo 52 caras, la desviación con respecto a la media, o frecuencia esperada, es $52 - 50 = 2$. En términos o unidades de desviación estándar, es $2/5 = .4$ unidades de desviación estándar con respecto a la media. Ahora se retoma una de las preguntas originales: ¿Estas diferencias son "estadísticamente significativas"? Por medio de la chi cuadrada se encontró que el resultado de 60 sujetos a favor es estadísticamente significativo y que el resultado de 52 caras no lo fue. ¿Se podrá hacer lo mismo con la presente fórmula? Sí se puede. Además, la belleza de este método radica en que puede aplicarse a todo tipo de números, no únicamente a los números binomiales. Sin embargo, antes de demostrarlo, se debe estudiar brevemente la llamada ley de los números grandes y las propiedades de la desviación estándar y de la curva normal de probabilidad.

La ley de los números grandes

La ley de los números grandes le tomó a Jacob Bernoulli (alias Jacques o James) 20 años para desarrollarla. En esencia es tan simple que uno se pregunta por qué le llevó tanto tiempo desarrollarla. Bernoulli, quien desarrolló esta ley en 1713, la llamó el "teorema de oro". Poisson le dio el nombre de "la ley de los números grandes" en 1837. Newman (1988) hace una detallada e interesante descripción de los alcances y controversias que existen respecto de este teorema. De manera general, esta ley sostiene que al incrementarse el tamaño de la muestra, n, existe una disminución en la probabilidad de que el valor observado de un evento, A, se desvíe del "verdadero" valor de A por no más de una cantidad fija, k. Siempre que los miembros de las muestras se elijan de forma independiente, mientras mayor sea el tamaño de la muestra, más cerca se estará del "verdadero" valor de la proporción de la población. Suponga que se lanza una moneda recién acuñada 100 veces y se registra el número de caras obtenidas; después se lanza la misma moneda 1 000 veces y también se registra el número de caras. De acuerdo con la ley de los números grandes, existe una mayor probabilidad de que los 1 000 lanzamientos produzcan 510 caras (una diferencia de 10 caras de las 500 esperadas), que el evento de 100 lanzamientos resulte en 60 caras (también una diferencia de 10 caras de las 50 esperadas). Lo que esto indica esencialmente es que los errores son menores en el experimento de 1 000 ensayos, que en el de 100 ensayos. El teorema también es un camino para la comprobación de hipótesis estadísticas, como se verá más adelante; además juega un papel particularmente importante en el teorema de Tchebysheff, el cual establece que si se tiene un número k mayor o igual a 1, y un conjunto de n mediciones, se garantiza (sin importar la forma de la distribución) que por lo menos $(1 - 1/k^2)$ de las mediciones caerán dentro de k unidades de desviación estándar hacia cualquier lado de la media.

Suponga que se lanza una moneda 1, 10, 50, 100, 400 y 1 000 veces, y que se desea conocer los resultados de las caras. Se calculan medias, varianzas, desviaciones estándar y

TABLA 11.1 *Medias, varianzas, desviaciones estándar y probabilidades esperadas del resultado de caras con diferentes tamaños de muestra**

n	$M(m_n) = np$	$V(m_n) = npq$	$DE(m_n)$	$M(C_n) = p$	$V(C_n) = pq/n$
1	1/2	.25	.50	1/2	14
10	5	2.50	1.58	1/2	1/40
50	25	12.50	3.54	1/2	1/200
100	50	25.00	5.00	1/2	1/400
400	200	100.00	10.00	1/2	1/1 600
1 000	500	250.00	15.81	1/2	1/4 000

* Véase el texto para la explicación de los símbolos en esta tabla.

dos nuevas medidas. La primera de ellas es la proporción de resultados favorables (caras en este caso) en la muestra total. A esta medida se le llamará m_n y se define como $C_n = m_n/n$ (recuerde que m_n es el número total de veces que ocurre el resultado favorable en n ensayos). Entonces la fracción de tiempo en que ocurre el resultado favorable es C_n. La media de C_n es p, o $M(C_n) = p$ [esto se deduce de la ecuación 11.3, donde $M(m_n) = pn$, y como $C_n = m_n/n$, entonces $M(C_n) = M(m_n/n = np/n = p$]. En pocas palabras, $M(C_n)$ es igual a la probabilidad esperada. La segunda medida es la varianza de C_n, que se define: $V(C_n) = pq/n$. La varianza, $V(C_n)$, es una medida de la variabilidad de la media, $M(C_n)$. Posteriormente se profundizará sobre la raíz cuadrada de $V(C_n)$, llamado el error estándar de la media. Los resultados de los cálculos se presentan en la tabla 11.1.

Observe que, aunque las medias, varianzas y desviaciones estándar de las sumas aumentan con el tamaño de las muestras, las $M(C_n)$ o p permanecen igual; esto es que el número promedio de caras, $M(C_n)$, siempre es 1/2. Pero la varianza del número promedio de caras, $V(C_n)$, disminuye conforme el tamaño de las muestras aumenta. De nuevo, $V(C_n)$ es una medida de la variabilidad de los promedios. Como la tabla 11.1 claramente indica, el número promedio de resultados debe acercarse cada vez más al valor "verdadero", que en este caso es 1/2. (El estudiante debe reflexionar cuidadosamente sobre este ejemplo antes de continuar.)

La curva normal de probabilidad y la desviación estándar

La curva normal de probabilidad es la curva en forma de campana que a menudo se encuentra en los libros de texto de estadística y psicología. Su importancia proviene del hecho de que grandes cantidades de eventos azarosos tienden a distribuirse en la forma de la curva. La llamada teoría de los errores utiliza esta curva. Se considera que muchos fenómenos —físicos y psicológicos— se distribuyen en forma aproximadamente normal. La estatura, la inteligencia, las aptitudes y el desempeño son ejemplos conocidos. Las medias de las muestras se distribuyen normalmente. El lector debe evitar la creencia no probada de que todos o casi todos los fenómenos se distribuyen de forma normal. Siempre que sea posible, los datos deben ser verificados con métodos apropiados, sobre todo por medio de diagramas o gráficos, ya que los datos frecuentemente son engañosos. Considere como ejemplo la aptitud, que en la población total puede estar distribuida de forma normal, pero suponga, por ejemplo, que se estudia si las puntuaciones del *Graduate Record Examination* (GRE) predicen éxito en la escuela de posgrado. Las correlaciones reportadas

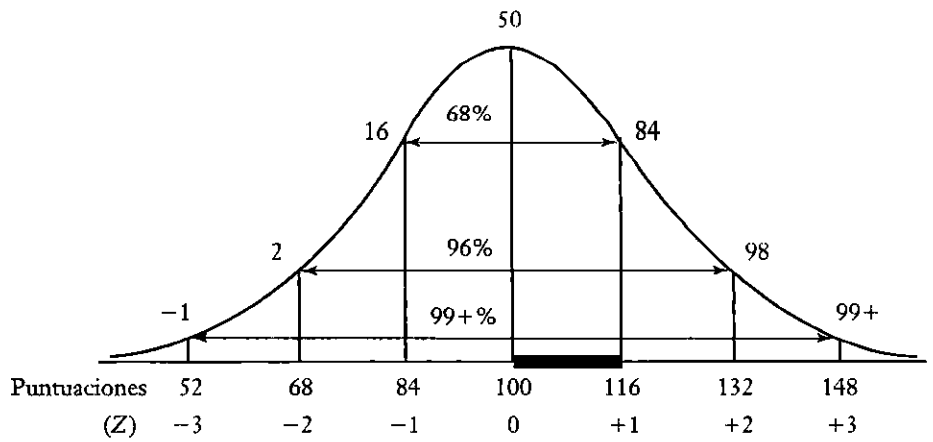
entre el éxito y las calificaciones del GRE no tienen valores muy altos (Morrison y Morrison, 1995). Se considera que las puntuaciones del GRE se distribuyen normalmente; sin embargo, esto no sucede con las personas que han sido admitidas en una escuela de posgrado *de alto nivel, donde las calificaciones en esta prueba se toman seriamente. Debido a que* sólo se admite a aquellos que obtienen altas calificaciones, y no a quienes obtienen bajas calificaciones, resulta que a estos últimos no se les mide su nivel de éxito. Esto trae como consecuencia que no sean incluidos en el cálculo de la relación entre las calificaciones del GRE y el éxito posterior. Una distribución truncada (que ya no es normal) conlleva un valor de correlación bajo (Kirk, 1990; House, 1983). Es difícil concebir a la estadística moderna sin esta curva. Todo texto sobre estadística tiene una tabla llamada "tabla de la desviación normal" o "tabla de la curva normal".

La razón estadística más importante para utilizar la curva normal consiste en poder interpretar fácilmente las probabilidades de los estadísticos que se calculan. Si los datos son, como se dice, "normales" o aproximadamente normales, se tiene una clara interpretación de lo que se hace.

Existen dos tipos de gráficos que generalmente se usan en la investigación del comportamiento. En uno de ellos, como ya se ha visto, los valores de una variable dependiente se grafican contra los valores de una variable independiente. El segundo gran grupo de gráficos tiene un propósito distinto: mostrar la distribución de una sola variable. En el eje horizontal los valores se ubican de forma similar a los del primer tipo de gráfico; pero en el eje vertical se ubican frecuencias o intervalos de frecuencia, o probabilidades.

Se dibuja una curva normal y se especifican dos conjuntos de valores sobre el eje horizontal. En uno de los conjuntos se utilizan puntuaciones de una prueba de inteligencia, con una media de 100 y una desviación estándar de 16. Suponga que la muestra es de 400 sujetos y que los datos (las puntuaciones) están distribuidos de forma aproximadamente normal (se dice que los datos están "distribuidos normalmente"). La curva se parece a la presentada en la figura 11.1. Imagine un eje Y (vertical) con frecuencias (o proporciones) marcadas sobre el eje. Las principales características de las curvas normales son la unimodalidad (una curva), la simetría (un lado similar al otro) y ciertas propiedades matemáticas, las cuales son de principal interés ya que permiten realizar inferencias estadísticas de poder considerable.

FIGURA 11.1



Una desviación estándar puede concebirse como una extensión a lo largo de la línea base de la curva, que va de la media o mitad de la línea base, hacia la izquierda o derecha, hasta el punto donde la curva se inflexiona. También puede visualizarse como un punto en la línea base a cierta distancia de la media. Una desviación estándar a partir de la media de esta distribución en particular es 100 + 16 = 116. La línea gruesa en la figura 11.1 indica la distancia de 100 a 116. De forma similar, una desviación estándar debajo de la media es 100 – 16 = 84. Dos desviaciones estándar se representan por 100 + (2)(16) = 132, y 100 – (2)(16) = 68. Si se tiene la suficiente confianza en que los datos en cuestión se distribuyen normalmente, entonces puede dibujarse una curva como la anterior, marcar la media y marcar las desviaciones estándar; esto también se hizo en la figura 11.1. La línea base también se graduó en unidades de desviación estándar (marcadas con Z en la figura). En lugar de utilizar puntuaciones de 100, 116 y 68, por ejemplo, se pueden usar puntuaciones de desviación estándar, que son 0, +1, –2, etcétera; se pueden señalar puntos entre éstos, por ejemplo, media desviación estándar arriba de la media es, en puntuaciones brutas, 100 + (1/2)(16) = 108; en puntuaciones de desviación estándar es 0 + .5 = .5. Estas puntuaciones de desviación estándar se denominan puntuaciones estándar o puntuaciones Z. Hablando en términos prácticos, las puntuaciones Z varían entre aproximadamente –3 y +3, pasando por el 0. Para transformar cualquier puntuación en bruto a una puntuación Z, se utiliza la fórmula $Z = \chi/DE$, donde $\chi = X - M$ y DE es la desviación estándar de la muestra. Las χ se llaman puntuaciones de desviación. Ahora puede dividirse la desviación estándar entre cualquier χ para convertir la X (puntuaciones en bruto) en una puntuación Z. Como ejemplo, suponga que $X = 120$; entonces $Z = (120 - 100)/16 = 20/16 = 1.25$, lo que significa que una puntuación en bruto de 120 equivale a una puntuación Z de 1.25, o que se encuentra una desviación estándar y cuarto arriba de la media.

Si se utilizan puntuaciones Z y el área total bajo la curva es igual a 1.00, entonces se habla de una curva de forma estándar. Esto de inmediato sugiere probabilidad. Las porciones del área de la curva se conciben e interpretan como probabilidades. Si el área total bajo la curva completa es igual a 1.00, y se dibuja una línea vertical de la línea base hacia arriba sobre la media ($Z = 0$) hasta la parte superior de la campana, las áreas a ambos lados de dicha línea vertical son iguales a 1/2 o 50%. Sin embargo, también pueden dibujarse líneas verticales en cualquier otro punto, partiendo de la línea base, a una desviación estándar arriba de la media ($Z = 1$) o a dos desviaciones estándar debajo de la media ($Z = -2$). Para interpretar tales puntos en términos de área —y en términos de probabilidad— se deben conocer las propiedades del área de la curva.

Los porcentajes aproximados de las áreas correspondientes a una, dos y tres desviaciones estándar arriba y debajo de la media, están indicadas en la figura 11.1. Para los propósitos presentes no es necesario utilizar los porcentajes exactos. El área entre $Z = -1$ y $Z = +1$ es aproximadamente 68%. El área entre $Z = -2$ y $Z = +2$ es aproximadamente 96% (la cifra exacta es .9544 pero se utiliza .96 porque facilita la interpretación). El área entre $Z = -3$ y $Z = +3$ es 99%. De la misma forma, todas las otras posibles distancias de la línea base, y sus áreas asociadas, pueden convertirse en porcentajes de la curva completa. Es importante recordar que, puesto que el área de la curva completa es igual a 1.00 o 100% y que, por lo tanto, es equivalente a U en la teoría de la probabilidad, los porcentajes de área pueden ser interpretados como probabilidades. De hecho, los valores de la tabla de probabilidad normal se dan en porcentajes de áreas correspondientes a puntuaciones Z.

Estos porcentajes aplican únicamente para una distribución normal. Si la forma de la distribución no es normal, estos porcentajes no aplican. Para encontrar los porcentajes para una curva de distribución no normal, se puede aplicar el teorema de Tchebysheff antes mencionado. Con este teorema se garantiza un 75% entre $Z = -2$ y $Z = +2$, y un 89.9% entre $Z = -3$ y $Z = +3$.

Interpretación de datos usando la curva normal de probabilidad con datos de frecuencia

Para formular preguntas acerca de las probabilidades de eventos, es necesario regresar al lanzamiento de monedas. Estrictamente hablando, las frecuencias de caras y cruces son eventos discontinuos, mientras que la curva normal de probabilidad es continua. Pero esto no debe causar preocupación, ya que las aproximaciones son cercanas. Es posible especificar con gran precisión y facilidad las probabilidades de la ocurrencia de eventos azarosos. En lugar de calcular probabilidades exactas, como se hizo previamente, las probabilidades se pueden estimar a partir del conocimiento de las propiedades de la curva normal. Esta aproximación a la curva normal de la distribución binomial resulta más precisa y útil cuando N es grande y el valor de p (la probabilidad de uno de los dos eventos) es cercana a .5. Comrey y Lee (1995, pp. 186-187) muestran cuánto cambia la aproximación para diferentes valores de p y N.

Suponga que nuevamente se lanzan 100 monedas, y se calcula que el número promedio de veces que probablemente resultarán caras es $M(m_{100}) = np = 100 \cdot 1/2 = 50$, y que la desviación estándar es:

$$DE(m_{100}) = \sqrt{V(m_{100})} = \sqrt{npq} = \sqrt{100 \cdot 1/2 \cdot 1/2} = \sqrt{25} = 5$$

Utilizando los porcentajes de la curva (probabilidades), se pueden hacer enunciados de probabilidad. Por ejemplo, se puede decir que en 100 lanzamientos la probabilidad de obtener caras entre una desviación estándar debajo de la media ($Z = -1$) y una desviación estándar arriba de la media ($Z = +1$), es aproximadamente .68. Existen, entonces, dos de tres posibilidades de que el número de caras sea entre 45 y 55 (50 ± 5). Hay una posibilidad de tres, aproximadamente, de que el número de caras sea menor que 45 o mayor que 55; es decir, $q = 1 - p = 1 - .68 = .32$.

Considere dos desviaciones estándar encima y debajo de la media. Estos puntos serían $50 - (2)(5) = 40$ y $50 + (2)(5) = 60$. Sabiendo que cerca del 95-96% de los casos probablemente caerán dentro de este rango, es decir, entre $Z = -2$ y $Z = +2$, o entre 40 y 60, puede decirse que la probabilidad de que el número de caras no será menor que 40 o mayor que 60, es aproximadamente .95 o .96. En otras palabras, existen solamente cuatro o cinco posibilidades en 100 de que resulten caras menos de 40 o más de 60 veces. Puede suceder, pero es poco probable.

Si se desea o necesita tener plena certeza (como en ciertos casos de investigación médica o de ingeniería), se puede recurrir hasta tres desviaciones estándar, $Z = -3$ y $Z = +3$, o quizá un poco menos de tres desviaciones estándar (el nivel .01 está aproximadamente a 2.58 desviaciones estándar). Tres desviaciones indican que el número de caras está entre 35 y 65. Puesto que tres desviaciones estándar arriba y debajo de la media, en la figura 11.1, cubren más del 99% del área de la curva, puede afirmarse que prácticamente se tiene la certeza de que el número de caras resultantes en 100 lanzamientos de una moneda recién acuñada no será menos de 35 ni más de 65. La probabilidad es mayor de .99. Si se lanzara una moneda 100 veces y se obtuvieran, por ejemplo, 68 caras, se podría concluir que probablemente existe un defecto en la moneda. Por supuesto que podrían resultar 68 caras; pero es muy poco probable que esto suceda con una moneda nueva.

El problema anterior respecto a acuerdos y desacuerdos se maneja exactamente de la misma forma que el de las monedas. El resultado de 60 acuerdos y 40 desacuerdos es poco probable de ocurrir; de hecho, existen solamente unas cuatro posibilidades en 100 de que se dé tal resultado por el azar. Esto ya se sabía a partir de la prueba de chi cuadrada y de la

prueba de probabilidad exacta, y ahora se cuenta con un tercer procedimiento que por lo común es aplicable a todo tipo de datos, cuando éstos se distribuyen normalmente o casi normalmente.

Interpretación de datos utilizando la curva normal de probabilidad con datos continuos

Suponga que se tienen las puntuaciones de una prueba de matemáticas de una muestra de 100 alumnos del quinto año. La media de las calificaciones es 70 y la desviación estándar es 10. Por conocimiento previo se sabe que la distribución de las puntuaciones de esta prueba es aproximadamente normal. En efecto, los datos pueden ser interpretados usando la curva normal; aunque aquí resulta importante la confiabilidad de la media. ¿Qué tanto se puede depender de esta media? ¿Se obtendrá la misma media con futuras muestras de alumnos similares de quinto grado? Si la media es poco confiable, es decir, que fluctúa ampliamente de una muestra a otra, cualquier interpretación de las puntuaciones de la prueba de alumnos en particular sería arriesgada. Una puntuación de 75 podría ser promedio en un momento, pero si la media no es confiable, este 75 podría ser una puntuación superior en futuras pruebas. En otras palabras, se requiere de una media confiable, de la que se pueda depender.

Considere que se aplica la misma prueba al mismo grupo de alumnos una y otra vez; yendo todavía más lejos, suponga que la prueba se aplica 100 000 veces, con todos los aspectos en las mismas condiciones: los niños no aprenden nada nuevo en todas estas repeticiones, no se cansan, las condiciones ambientales son iguales, etcétera.

Si se calculara una media y una desviación estándar para cada aplicación, se obtendría una gigantesca distribución de medias (y de desviaciones estándar). ¿Cómo se vería esta distribución? Primero, formaría una curva normal en forma de campana. Las medias tienen la propiedad de distribuirse adecuadamente en una curva normal, aun cuando las distribuciones originales de donde fueron obtenidas no sean normales. Esto se debe a que se asumió que "todos los aspectos permanecieron en las mismas condiciones", por lo que no existen fuentes de fluctuación de las medias, excepto por el azar. Las medias fluctuarán, pero estas fluctuaciones se deberán solamente al azar. La mayoría de las fluctuaciones se agruparán alrededor de la llamada media "verdadera", es decir, el "verdadero" valor de la gigantesca población de medias; unas pocas tendrán valores extremos. Si se repitiera el experimento de 100 lanzamientos de una moneda muchas veces, se encontraría que las caras se agruparían alrededor del valor "verdadero": 50. Algunas estarían ligeramente más arriba y otras ligeramente más abajo; unas pocas estarían muy arriba y otras pocas muy abajo. En resumen, las caras y las medias obedecen a la misma "ley". Puesto que se supone que no influyen otros factores, se debe concluir que las fluctuaciones se deben al azar. Respecto a los errores por azar, si hubiera suficientes, también se distribuirían de forma normal. Ésta es la llamada *teoría de los errores*.

Continuando con el tema de las medias, si se tuvieran los datos de las múltiples aplicaciones de la prueba de matemáticas al mismo grupo, se calcularían una media y una desviación estándar. Tal media calculada estaría cercana al valor de la media "verdadera". Si se tuviera un número infinito de medias de un número infinito de aplicaciones de la prueba y se calculara la media de las medias, entonces se obtendría la media "verdadera". Esto sería similar para la desviación estándar de las medias. En efecto, ello no puede hacerse ya que no se tiene un número infinito, ni siquiera lo bastante grande, de aplicaciones de la prueba.

Por fortuna existe una forma más simple para resolver el problema. Consiste en aceptar la media calculada para la muestra como la media "verdadera", y después estimar qué tan precisa es esta decisión (o suposición). Para hacerlo, se calcula un estadístico conocido como el error estándar de la media. Se define de la siguiente manera:

$$EE_M = \frac{\sigma_{pob}}{\sqrt{n}} \tag{11.9}$$

donde el error estándar de la media es EE_M; la desviación estándar de la población (σ se lee "sigma"), σ_{pob}; y el número de casos en la muestra, n.

Hay un pequeño obstáculo aquí: no se conoce o no se puede conocer la desviación estándar de la población. Recuerde que tampoco se conocía la media de la población, pero se estimó con la media de la muestra. De forma similar, se estima la desviación estándar de la población con la desviación estándar de la muestra. Entonces, se utiliza la siguiente fórmula:

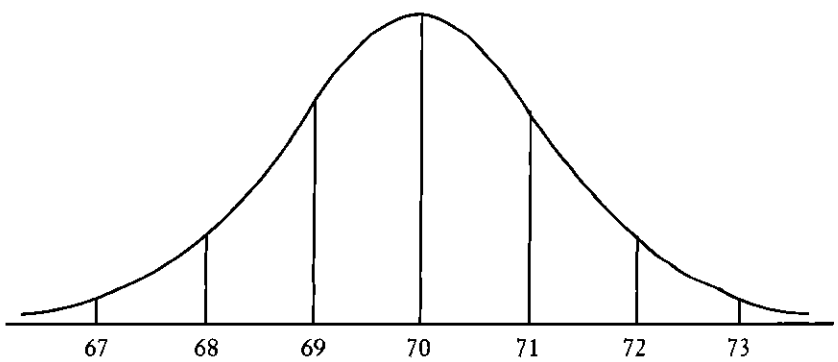
$$EE_M = \frac{DE}{\sqrt{n}} \tag{11.10}$$

Ahora puede estudiarse la confiabilidad de la media de la prueba de matemáticas. Se calcula:

$$EE_M = \frac{10}{\sqrt{100}} = \frac{10}{10} = 1$$

Nuevamente, considere una población grande de medias de esta prueba. Si se integran en una distribución y se grafica la curva de dicha distribución, ésta se observará como la curva mostrada en la figura 11.2. Es importante recordar que se trata de una distribución imaginaria de *medias de muestras* y *no* de una distribución de puntuaciones. Resulta sencillo notar que las medias de esta distribución no son muy variables. Si se duplica el error estándar de la media, se obtiene 2. Se resta y se suma esta cifra a la media de 70: 68 a 72. Existe una probabilidad aproximada de .95 de que la media ("verdadera") de la población se encuentre dentro del intervalo 68 a 72, es decir, aproximadamente el 5% de las veces las medias de muestras aleatorias de este tamaño caerán fuera de este intervalo.

Figura 11.2



Si se realizan los mismos cálculos con los datos de la prueba de inteligencia de la figura 11.1, se obtendría:

$$EE_M = \frac{16}{\sqrt{400}} = \frac{16}{20} = .80$$

Tres errores estándar arriba y debajo de la media de 100 dan el rango 97.60 a 102.40, es decir, que la media "verdadera" muy probablemente (con menos de 1% de probabilidad de equivocarse) se encuentra dentro del intervalo de 97.60 a 102.40. Las medias son confiables con muestras de tamaño razonable. Aun con muestras relativamente pequeñas, la media resulta muy estable (véase los datos de la prueba de inteligencia del capítulo 8). De una población se extrajeron cinco muestras de 20 puntuaciones de inteligencia cada una. La media poblacional era 95. Se calcularon las medias de las cinco muestras, así como los errores estándar de las medias de las primeras dos muestras, y después se interpretaron. Más adelante se hicieron comparaciones con el valor "verdadero" de 95. La media de la primera muestra fue 93.55, la desviación estándar fue 12.22 y el error estándar de la media, $EE_M = 2.73$. El rango de las medias al nivel .05 fue: 88.09 a 99.01. En efecto el valor 95 cae dentro de este rango. La media de la segunda muestra estaba más desviada: 90.20, la desviación estándar fue 9.44 y el error estándar de la media, $EE_M = 2.11$. El rango al nivel .05 fue: 85.98 a 94.42; el valor 95 no cae dentro de este rango. El rango al nivel .01 fue: de 83.87 a 96.53. Ahora el valor 95 sí está incluido. Esto no está nada mal para muestras de tan sólo 20 sujetos. En muestras de 50 o 100 sujetos resultaría aún mejor. La media de las cinco medias fue 93.31; la desviación estándar de estas medias fue 2.73. Compare ésta con los errores estándar calculados para las dos muestras: 2.73 y 2.11. En el capítulo 12 se dará una demostración más convincente respecto de la estabilidad de las medias.

Entonces, el error estándar de la media es una desviación estándar. Es una desviación estándar de un número infinito de medias. Sólo el error debido al azar hace fluctuar las medias, por lo que el error estándar de la media (o, si se prefiere, la desviación estándar de las medias) es una medida de azar o error en sus efectos sobre una medida de tendencia central.

Resulta necesaria una advertencia: toda la teoría estudiada aquí está basada en el supuesto de que se trata de muestras aleatorias y de observaciones independientes. Si se infringen estos supuestos, el razonamiento, aunque no se invalida totalmente, puede ser cuestionado. Las estimaciones del error pueden estar sesgadas en menor o mayor grado; el problema es que no se puede decir qué tan sesgado está un error estándar. Hace algunos años, Guilford y Fruchter (1977) dieron ejemplos interesantes de los sesgos encontrados cuando se infringen los supuestos. Con un gran número de pilotos de la fuerza aérea, encontraron que algunas veces las estimaciones de los errores estándar estaban considerablemente sesgadas. Nadie puede dar reglas rápidas y exactas. La máxima probablemente afirmaría: siempre que sea posible, debe usarse el muestreo aleatorio y mantener las observaciones independientes. Simon (1987) apoyaría esta regla.

Si no puede usarse el muestreo aleatorio, y existen dudas respecto a la independencia de las observaciones, deben calcularse e interpretarse los estadísticos, pero es necesario ser muy cauteloso con las interpretaciones y las conclusiones, ya que pueden resultar erróneas. Debido a dichas posibilidades de error, se ha dicho que las estadísticas son engañosas, e incluso inútiles. Como cualquier otro método —consultar una autoridad, utilizar la intuición, etcétera— la estadística puede ser engañosa; pero aun cuando las medidas estadísticas estén sesgadas, en general lo están menos que los juicios de autoridad y de intuición. No es que los números mientan; los números no saben lo que están haciendo. Son los seres humanos que usan los números quienes pueden estar informados o mal informa-

dos, sesgados o no, con conocimiento o ignorancia, inteligentes o necios. Los números y la estadística no deben ser tratadas con demasiado respeto ni con demasiado desprecio. Al calcular estadísticos debe actuarse como si fueran "verdaderos", pero siempre manteniendo cierta reserva hacia ellos; se requiere estar dispuesto a no creer en ellos si la evidencia indica su descrédito.

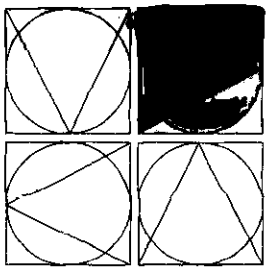
RESUMEN DEL CAPÍTULO

1. El principio básico que subyace al uso de las pruebas estadísticas de significancia consiste en comparar los resultados obtenidos (observados, empíricos) con lo esperado por el azar.
2. Cuatro propósitos de la estadística son:
   a) reducir los datos a formas manejables y entendibles;
   b) ayudar en el estudio de poblaciones y muestras;
   c) ayudar en la toma de decisiones, y
   d) auxiliar para realizar inferencias confiables de muestras a poblaciones.
3. Los datos binomiales consisten de dos posibles resultados.
4. Bajo ciertas condiciones, la curva normal puede usarse como una aproximación de la distribución binomial.
5. La ley de los números grandes establece que cuanto más grande sea la muestra, más se acercará el valor de la muestra al valor verdadero (de la población).
6. Los eventos azarosos tienden a distribuirse en forma de una curva normal.
7. El uso de la curva normal simplifica la interpretación del análisis de los datos.
8. La curva normal posee ciertas propiedades matemáticas que hacen atractiva su aplicación en el análisis e interpretación estadísticos.
9. Las puntuaciones estándar Z son transformaciones lineales (reexpresiones) de puntuaciones en bruto.
10. El uso de puntuaciones Z incrementa el poder de interpretación de los datos ya que están expresadas en "unidades de desviación estándar".
11. Las puntuaciones Z de diferentes distribuciones pueden compararse significativamente entre sí.
12. La conversión de puntuaciones en bruto que se encuentran distribuidas normalmente en puntuaciones Z permite el empleo de la tabla de la curva normal para determinar porcentajes, áreas y probabilidades.

SUGERENCIAS DE ESTUDIO

1. La estadística sirve para resumir conjuntos grandes de datos. Dé un ejemplo donde la estadística pueda resultar engañosa al utilizarla para evaluar a una sola persona, compañía o grupo.
2. Explique cómo difieren los estadísticos y los ciudadanos comunes en su concepto del término *error*.
3. ¿Cuál es el principal propósito de la estadística?
4. Al usar la curva normal de probabilidad, aproximadamente el .68 del área bajo la curva se ubica entre ± 1 desviación estándar de la media. Para ± 2 desviaciones estándar es .96. ¿Cuáles serían los porcentajes aproximados si la curva no fuera normal?

5. Una amiga lanza una moneda al aire 1 000 veces, y obtiene 505 caras y 495 cruces. Ella afirma que su resultado apoya su idea de que la moneda está en buenas condiciones. Sin embargo, se sabe que una moneda en buen estado debe generar 500 caras. Digamos que ella tiene razón. ¿Cómo puede explicarse la diferencia de 5 caras (o cruces)?
6. Mencione la distinción entre un parámetro y un estadístico.



CAPÍTULO 12

COMPROBACIÓN DE HIPÓTESIS Y ERROR ESTÁNDAR

- EJEMPLOS: DIFERENCIAS ENTRE MEDIAS
- DIFERENCIAS ABSOLUTAS Y RELATIVAS
- COEFICIENTES DE CORRELACIÓN
- PRUEBA DE HIPÓTESIS: HIPÓTESIS SUSTANTIVAS Y NULAS
- NATURALEZA GENERAL DE UN ERROR ESTÁNDAR
- UNA DEMOSTRACIÓN MONTE CARLO
  - **Procedimiento**
  - **Generalizaciones**
  - **Teorema del límite central**
  - **Error estándar de las diferencias entre medias**
- INFERENCIA ESTADÍSTICA
  - **Comprobación de hipótesis y los dos tipos de errores**
- LOS CINCO PASOS DE LA COMPROBACIÓN DE HIPÓTESIS
  - **Determinación del tamaño de la muestra**

El error estándar,[1] como estimado de la fluctuación debida al azar, es la medida contra la cual se verifican los resultados de los experimentos. ¿Existe una diferencia entre las medias de dos grupos experimentales? Si es así, ¿la diferencia es una diferencia "real" o sólo una consecuencia de las muchas diferencias relativamente pequeñas que pudieron haber surgido por el azar? Para contestar esta pregunta, se calcula el error estándar de las diferencias entre medias, y la diferencia obtenida se compara con tal error estándar. Si es suficientemente mayor que el error estándar, se dice que se trata de una diferencia "significativa". Un razonamiento similar puede aplicarse a cualquier estadístico; por lo tanto, existen muchos errores estándar: de coeficientes de correlación, de diferencias entre medias, de medias,

[1] El término "error" aquí se refiere a la fluctuación encontrada entre diferentes muestras del mismo tamaño, tomadas de la misma población. No debe entenderse como "equivocaciones".

de medianas, de proporciones, etcétera. Los propósitos de este capítulo son: 1) examinar la noción general del error estándar, 2) aprender cómo se prueban las hipótesis utilizando el error estándar, y 3) conocer el importante papel que éste juega en la estimación del tamaño de la muestra.

Ejemplos: diferencias entre medias

Un problema particularmente difícil en la psicología contemporánea se centra en la pregunta de si el comportamiento está controlado más por factores situacionales o ambientales, o por la predisposición de los individuos. McGee y Snyder (1975), utilizando una supuesta diferencia entre los individuos que salan su comida antes de probarla y aquellos que la prueban antes de salarla, hipotetizaron que los individuos que conforman su comportamiento por predisposición salan su comida antes de probarla; mientras que los individuos que conforman su comportamiento según la situación prueban su comida antes de salarla. Ellos concluyeron además que los primeros atribuirían más rasgos a sí mismos que los últimos. Encontraron que los del primer grupo, los "saladores", atribuyeron a sí mismos una media de 14.87 rasgos; mientras que el segundo grupo, los "probadores", atribuyeron a sí mismos una media de 6.90 rasgos. La dirección de la diferencia fue como los autores predijeron. ¿El tamaño de la diferencia entre las medias, 7.97, es suficiente para garantizar la afirmación de los autores de que su hipótesis fue apoyada? Una prueba de la significancia estadística de esta diferencia mostró que era altamente significativa. (Esta afirmación es una generalización de la original.)

Un problema psicológico creciente, donde cerca del 75% de los afectados no buscan ayuda, es el trastorno de pánico. Con el incremento de las regulaciones impuestas por organizaciones de administración de salud (OAS), es posible que aún menos individuos afectados busquen tratamiento. El estudio de Gould y Clum (1995) proporciona datos que parecen muy prometedores para aliviar parcialmente este problema. Gould y Clum estudiaron el beneficio de un programa de autoayuda para tratar a las víctimas del trastorno de pánico. En un gran esfuerzo para reclutar sujetos para su estudio, lograron formar dos grupos de participantes y ambos consistieron de enfermos con trastorno de pánico. Uno de los grupos recibió instrucciones y algunas sesiones de asesoría sobre autoayuda. La autoayuda incluyó la lectura del libro *Coping with Panic* (*Manejo del pánico*). El otro grupo, denominado como lista de espera, no recibió tratamiento (se les dijo que estaban en lista de espera para la terapia). Cada paciente fue evaluado durante un periodo de 14 semanas, cubriendo tres etapas importantes: pretratamiento, postratamiento y seguimiento. Una de las medidas fue el número de ataques de pánico por semana. Antes del tratamiento, el grupo de autoayuda tuvo una media de 2.6 ataques por semana; mientras que el grupo en lista de espera reportó una media de 1.8 ataques. Después del tratamiento, el grupo de autoayuda tuvo una media de 0.9 (un cambio de −1.7 en la media) y el grupo en lista de espera reportó una media de 2.1 (un cambio de +0.3 en la media). En el periodo de seguimiento, el grupo de autoayuda tuvo un promedio de 0.5 ataques; por su parte, el grupo en lista de espera reportó 2.5. Las hipótesis de estos investigadores fue sustentada. Una prueba de la significancia estadística de esta diferencia mostró que resultaba altamente significativa.

El punto importante de estos dos ejemplos en el contexto presente es que la significancia estadística de la diferencia entre las medias fue probada mediante un error estándar. El error estándar, en este caso, fue el error estándar de la diferencia entre las medias. Se encontró que la diferencia en ambos estudios fue significativa. El estudio de McGee y Snyder (1975) señala que aquellos individuos que perciben que el comportamiento es

influido por rasgos individuales tienden a salar su comida antes de probarla; mientras que aquellos individuos cuya percepción está más orientada al ambiente prueban su comida antes de salarla. En el estudio de Gould y Clum (1995), el programa de autoayuda es una forma más prometedora de tratamiento para el trastorno de pánico. Mientras que el grupo en lista de espera experimentó un cambio no significativo en términos del promedio de ataques de pánico, el grupo de autoayuda mostró una considerable mejoría. Gould y Clum (1995) utilizaron otras mediciones dependientes, tales como los síntomas del pánico y el manejo de la ansiedad del pánico, y encontraron un patrón similar de significancia. Ahora se estudiará un ejemplo donde la diferencia entre las medias no resultó significativa.

Gates y Taylor (1925), en un estudio antiguo y bien conocido sobre la transferencia del aprendizaje, formaron dos grupos apareados de 16 alumnos cada uno. Al grupo experimental se le dio práctica en la memorización de dígitos, y al grupo control no. La mejoría promedio del grupo experimental, inmediatamente después del periodo de práctica, fue de 2.00. La mejoría promedio del grupo control fue de 0.67, una diferencia media de 1.33. De cuatro a cinco meses después, los niños de ambos grupos fueron evaluados nuevamente. La mejoría promedio del grupo experimental fue de 0.35; y la del grupo control, de 0.36. Este resultado fue sorpresivo ya que se esperaba que el grupo experimental tuviera mejor desempeño que el grupo control, como había sucedido al principio del estudio. En este caso, el desempeño del grupo control fue igual al desempeño de los sujetos del grupo experimental. Difícilmente se requieren pruebas estadísticas para datos como éstos.

Diferencias absolutas y relativas

Puesto que las diferencias entre estadísticos (especialmente entre medias) se prueban y se reportan mucho en la literatura, es necesario obtener cierta perspectiva sobre los tamaños absolutos y relativos de tales estadísticos. Aunque el análisis utiliza diferencias entre medias como ejemplos, los mismos puntos se aplican a las diferencias entre proporciones, coeficientes de correlación, etcétera. En un estudio de Scattone y Saetermoe (1997) se encontró que las personas de origen asiático nacidas en Estados Unidos eran más receptivas hacia las personas con discapacidades que los asiáticos nacidos en otros países. Con el empleo de una escala de distancia social con valores que van del 1 al 5, donde el 5 indica una alta aceptación, los asiáticos nacidos en Estados Unidos tuvieron una media de 4.17; mientras que los asiáticos nacidos en otros países tuvieron una media de 3.71. La diferencia entre las medias fue de 0.46 y resultó estadísticamente significativa. ¿Tendrá algún significado una diferencia tan pequeña como ésta? Contraste esta pequeña diferencia con la diferencia de medias en el consumo de cerveza entre hombres y mujeres, obtenida por Zavela, Barrett, Smedi, Istvan y Matarazzo (1990). Zavela y sus colaboradores estudiaron las diferencias entre géneros respecto al consumo de alcohol, cigarrillos y café. En lo que se refiere al consumo mensual de cerveza, los hombres tuvieron una media de 18.68; y las mujeres, de 9.14. La diferencia entre estas medias fue de 9.54 y resultó estadísticamente significativa.

El problema aquí en realidad lo constituyen dos problemas: uno sobre el tamaño absoluto y relativo de las diferencias, y otro sobre la significancia práctica o "real" *versus* la significancia estadística. La que aparentemente es una pequeña diferencia puede, al examinarse de cerca, no resultar tan pequeña. En un estudio de Evans, Turner, Ghee y Getz (1990) sobre la relación entre el papel andrógino y el tabaquismo, se encontró una diferencia de medias de 0.164 entre los sujetos andróginos, y los no andróginos respecto a la frecuencia con que fumaban. La diferencia de 0.164 probablemente es trivial, aunque estadísticamente significativa. El 0.164 se derivó de una escala de 7 puntos sobre la fre-

cuencia de la conducta de fumar y, por lo tanto, es realmente muy pequeño. Ahora, tome un ejemplo completamente diferente de un importante estudio de Miller y DiCara (1968) sobre el condicionamiento instrumental de la secreción de orina. Las medias de un grupo de ratas, antes y después de entrenarlas para secretar orina, fueron 0.017 y 0.028, y la diferencia tuvo una significancia estadística muy alta. Pero la diferencia fue de sólo 0.011. ¿Será demasiado pequeña para considerarla seriamente? Ahora tiene que considerarse la naturaleza de las medidas. Las pequeñas medias de 0.017 y 0.028 se obtuvieron de mediciones de la secreción de orina de las ratas. Cuando se considera el tamaño de las vejigas de las ratas y que el condicionamiento instrumental (recompensa por secretar orina) produjo una diferencia de medias de 0.011, el significado de esta diferencia se vuelve dramático: ¡incluso es bastante grande! (Los datos se analizarán en un capítulo posterior y quizá esto resulte más claro.)

Por lo común no se debe ser demasiado entusiasta respecto a diferencias de medias de 0.20, 0.15, 0.08, etcétera; pero se debe ser cauteloso y hábil al analizarlas. Suponga que se reporta como estadísticamente significativa una diferencia muy pequeña y se piensa que esto es ridículo. También suponga que se trata de la diferencia de medias entre la longitud de las dendritas de grupos de ratas bajo experiencias enriquecedoras y de privación, en los primeros días de sus vidas (Camel, Withers y Greenough, 1986). Obtener cualquier diferencia en la ramificación de las dendritas neuronales a causa de la experiencia es un logro destacado y, obviamente, un descubrimiento científico importante.

Coeficientes de correlación

Los coeficientes de correlación se reportan en grandes cantidades en las revistas científicas. Deben formularse preguntas respecto a la significancia de los coeficientes y a la "realidad" de las relaciones que expresan. Por ejemplo, para resultar estadísticamente significativo, un coeficiente de correlación calculado entre 30 pares de mediciones debe ser de aproximadamente 0.31 al nivel de 0.05, y 0.42 al nivel de 0.01. Con 100 pares de mediciones, el problema es menos severo (de nuevo la ley de los números grandes); al nivel de 0.05, una r de 0.16 es suficiente; al nivel de 0.01, una r de 0.23 lo logra. Si las r son menores que estos valores, se considera que no son significativamente diferentes de cero.

Si se extraen, por ejemplo, 30 pares de números de una tabla de números aleatorios y se les correlaciona, teóricamente la r debería estar cerca de cero. Con claridad, deben existir relaciones cercanas a cero entre conjuntos de números aleatorios; sin embargo, en ocasiones, los conjuntos de pares pueden resultar estadísticamente significativos y con r razonablemente altas, "debido al azar". A cualquier costo, los coeficientes de correlación, así como las medias y las diferencias, deben ser elevados respecto a la significancia estadística, comparándolos contra sus errores estándar. Por fortuna, esto es fácil de hacer, ya que las r, para los diferentes niveles de significancia y para diferentes tamaños de muestras, se ofrecen en tablas en la mayoría de los textos de estadística. Por ello, al utilizar r no es necesario calcular ni utilizar el error estándar de una r. Sin embargo, los cálculos que originan estas tablas deben ser comprendidos.

De los miles de coeficientes de correlación reportados en la literatura de investigación, muchos son de baja magnitud. ¿Qué tan bajo es bajo? ¿En qué punto un coeficiente de correlación es demasiado bajo como para tomarlo en serio? Generalmente una r menor a 0.10 no puede tomarse con mucha seriedad; una r de 0.10 significa que tan sólo el 1% ($0.10^2 = 0.01$) de la varianza de y se comparte o explica con x. Por otro lado, si una r de 0.30 resulta estadísticamente significativa, puede ser relevante porque quizá señale una relación importante. El problema se complica con r comprendidas entre 0.20 y 0.30. (Recuer-

de que con N grandes, las r entre 0.20 y 0.30 son estadísticamente significativas.) Para estar seguros, una r de, por ejemplo, 0.20 indica que las dos variables comparten tan sólo el 4% de su varianza. Pero una r de 0.26 (7% de la varianza compartida), o incluso una de 0.20, pueden ser relevantes, ya que tal vez provean de un avance importante a la teoría y a las investigaciones subsecuentes. El problema se vuelve complejo. En investigación básica, las correlaciones bajas (que deben ser estadísticamente significativas, por supuesto) enriquecen la teoría y la investigación. Es en la investigación aplicada donde la predicción resulta importante, y donde han crecido los juicios de valor respecto a las correlaciones bajas y a las cantidades triviales de varianza compartida. No obstante, en la investigación básica el panorama se complica más. Una conclusión es segura: los coeficientes de correlación, como otros estadísticos, deben probarse respecto a su significancia estadística.

Prueba de hipótesis: hipótesis sustantivas y nulas

El principal propósito de investigación de la estadística inferencial consiste en poner a prueba hipótesis de investigación por medio de la comprobación de hipótesis estadísticas. De forma general, los científicos utilizan dos tipos de hipótesis: sustantivas y estadísticas. Una *hipótesis sustantiva* es el tipo común de hipótesis analizadas en el capítulo 2, donde se expresa una afirmación conjetural de la relación entre dos o más variables. Por ejemplo, la hipótesis "a mayor cohesión de un grupo, mayor será su influencia sobre sus miembros" es una hipótesis sustantiva expresada por Schacter, Ellertson, McBride y Gregory (1951). La teoría de un investigador afirma que esta variable se relaciona con la otra variable. La afirmación de la relación constituye una hipótesis sustantiva.

Estrictamente hablando, una hipótesis sustantiva no puede someterse a prueba, sin antes traducirse a términos operacionales. Una forma muy útil para probar hipótesis sustantivas es a través de hipótesis estadísticas. Una *hipótesis estadística* es un enunciado conjetural, en términos estadísticos, de relaciones estadísticas deducidas a partir de relaciones de hipótesis sustantivas. Este burdo enunciado requiere de traducción: una hipótesis estadística expresa un aspecto de la hipótesis sustantiva original, en términos cuantitativos y estadísticos, es decir, $\mu_A > \mu_B$, la media A es mayor que la media B; $r > +0.20$, el coeficiente de correlación es mayor que +0.20; $\mu_A > \mu_B > \mu_C$, al nivel de 0.01: χ^2 es significativa al nivel de 0.05; etcétera. Una hipótesis estadística constituye una predicción sobre cómo resultarán los estadísticos utilizados al analizar los datos cuantitativos de un problema de investigación. Para el análisis sobre la comprobación de hipótesis se utilizará μ para indicar la media poblacional; y M para la media muestral. Las hipótesis estadísticas se expresan en términos de valores de la población. Después de recolectar los datos, la media calculada a partir de la muestra se expresará como M.

Las hipótesis estadísticas deben probarse en contra de algo. No es posible probar tan sólo una hipótesis estadística aislada; es decir, no se prueba directamente la proposición estadística $\mu_A > \mu_B$, por sí misma. Se prueba *contra* una proposición alternativa. De hecho, pueden existir varias alternativas para $\mu_A > \mu_B$ y la alternativa que generalmente se selecciona es la hipótesis nula, que fue inventada por Sir Ronald Fisher. La *hipótesis nula* es una proposición estadística que esencialmente enuncia que no existe relación entre las variables (del problema). La hipótesis nula señala: "Estás equivocado, no existe relación; contradíceme si puedes." Dice lo anterior en términos estadísticos tales como $\mu_A = \mu_B$; o $\mu_A - \mu_B = 0$; $r_{xy} = 0$; χ^2 no es significativa; t no es significativa, etcétera.

Algunas veces, los investigadores utilizan inconscientemente hipótesis nulas como hipótesis sustantivas; por ejemplo, en vez de afirmar que un método de presentación de materiales escritos tiene un mayor efecto en el recuerdo en relación con otro método,

pueden decir que no existe diferencia alguna entre ambos métodos. Esto refleja falta de experiencia, ya que, en efecto, utilizan la hipótesis nula estadística como una hipótesis sustantiva, confundiendo los dos tipos de hipótesis. Estrictamente hablando, cualquier resultado significativo, ya sea positivo o negativo, apoya la hipótesis; pero ésta no es ciertamente la intención; la intención es obtener evidencia estadística para apoyar la hipótesis sustantiva, por ejemplo, en $\mu_A > \mu_B$. Si el resultado es estadísticamente significativo $\mu_A > \mu_B$, entonces se acepta la hipótesis sustantiva (se rechaza la hipótesis nula de que $\mu_A = \mu_B$). Al utilizar de manera sustantiva la hipótesis nula, se pierde el poder de la hipótesis sustantiva, lo cual equivale al hecho de que el investigador hiciera una predicción específica sin oportunidades.

Por supuesto que siempre existe la rara posibilidad de que una hipótesis nula sea la hipótesis sustantiva. Si, por ejemplo, un investigador desea demostrar que dos métodos de enseñanza no producen diferencias en el rendimiento, entonces quizá la hipótesis nula sea apropiada. El problema con esto es que lógicamente coloca al investigador en una posición complicada, ya que es bastante difícil —quizás imposible— demostrar la "validez" empírica de una hipótesis nula. Después de todo, si se apoya la hipótesis $\mu_A = \mu_B$, bien puede ser uno de los muchos posibles resultados debidos al azar, en lugar de una no diferencia significativa. Es factible encontrar buenas discusiones sobre la comprobación de hipótesis en Giere (1979), en los capítulos 6, 8, 11 y 12, especialmente en el capítulo 11.

Fisher (1950) afirma: "Puede decirse que cada experimento existe solamente para dar a los hechos la oportunidad de desprobar la hipótesis nula." Es una atinada afirmación, pero ¿qué significa? Suponga que se tiene una hipótesis de que el efecto del método *A* es superior al del método *B*. Si se resuelve satisfactoriamente el problema de definir lo que se quiere significar con "superior" (estableciendo un experimento y cosas así), ahora debe especificarse una hipótesis estadística. En este caso se puede afirmar que $\mu_A > \mu_B$ (la media del método *A* es o será mayor que la media del método *B*, con base en determinada medida criterio). Suponga que después del experimento, las dos medias son 68 y 61, respectivamente. Parecería que se apoya la hipótesis sustantiva, ya que 68 > 61, o μ_A es mayor que μ_B. Sin embargo, como se aprendió antes, esto no es suficiente, ya que la diferencia puede ser una de las muchas posibles diferencias similares debidas al azar.

En efecto, se estableció lo que puede llamarse la hipótesis del azar: $\mu_A = \mu_B$, o $\mu_A - \mu_B = 0$. Éstas son hipótesis nulas. Entonces se deben anotar las hipótesis; primero se escribe la hipótesis estadística que refleja el significado operacional-experimental de la hipótesis sustantiva; después se escribe la hipótesis nula contra la cual se prueba el primer tipo de hipótesis. A continuación se observan los dos tipos de hipótesis convenientemente expresadas:

$$H_0: \mu_A = \mu_B$$
$$H_1: \mu_A > \mu_B$$

H_1 representa la "hipótesis 1". Con frecuencia existe más de una de dichas hipótesis, por lo que se etiquetan como H_1, H_2, H_3, etcétera. H_0 representa la "hipótesis nula". Note que, en este caso, la hipótesis nula se pudo haber anotado H_0: $\mu_A - \mu_B = 0$. Esta forma muestra de dónde obtuvo su nombre la hipótesis nula: la diferencia entre μ_A y μ_B es cero. Pero resulta poco manejable de esta manera, en especial cuando se prueban tres o cuatro medias u otros estadísticos. $\mu_A = \mu_B$ es general y, por supuesto, significa lo mismo que $\mu_A - \mu_B = 0$ y $\mu_B - \mu_A = 0$. Considere que es fácil escribir $\mu_A = \mu_B = \mu_C = \ldots = \mu_N$.

Aunque como investigador se desea demostrar que H_1 es verdadera, no puede hacerse fácilmente en forma directa. Suponga que la hipótesis sustantiva lleva al investigador a escribir la hipótesis estadística H_1: $\mu_A \neq \mu_B$. Esta hipótesis podría volver a escribirse H_1: μ_A

$- \mu_B \neq 0$. Para probar esta hipótesis de manera directa se necesitaría probar un número infinito de valores, es decir, que se requeriría probar todas y cada una de las situaciones donde $\mu_A - \mu_B$ *no* es igual a cero. En la comprobación de hipótesis, el procedimiento dicta que se pruebe la hipótesis nula. La hipótesis nula se expresa como H_0: $\mu_A - \mu_B = 0$. Observe que apunta directamente a un valor, en específico al cero. Es necesario reunir datos empíricos para demostrar que la hipótesis nula es insostenible. En términos estadísticos, "se rechazaría H_0". Al hacerlo se indica que se tiene un resultado significativo, lo cual lleva a apoyar H_1. Si se apoya H_1, a su vez, conlleva a la sustentación de la hipótesis sustantiva. Si no existen datos empíricos suficientes para refutar la hipótesis nula, no puede rechazarse la hipótesis nula. Estadísticamente se diría que "no se logró rechazar H_0" o "no rechazar H_0". Considere que no se "acepta" H_0 porque los resultados fueron "no significativos". Sin importar los resultados, tan sólo es posible "no lograr rechazar" H_0 o "no rechazar" H_0; nunca es posible "aceptar" H_0. Para "aceptar" H_0 se requeriría repetir el estudio un número infinito de veces, y obtener exactamente cero cada vez. Por otro lado, es posible "no lograr rechazar" H_0 puesto que los resultados no son lo suficientemente diferentes de lo que se podría predecir (bajo el supuesto de que H_0 sea verdadera) para garantizar la conclusión de que es falsa.

La condición de la H_0 resulta similar a la de un acusado en un juicio, en el cual es considerado "inocente" hasta que se pruebe que es "culpable". Si el juicio resulta en un veredicto de "no culpable", ello no quiere decir que el acusado sea "inocente", tan sólo significa que no pudo demostrarse la culpa más allá de la duda razonable. Cuando el investigador no logra rechazar H_0, eso no significa que H_0 sea verdadera, sino que no pudo demostrarse su falsedad más allá de la duda "razonable". Propst (1988) y Kenney (1985) realizan una interesante analogía de la comprobación de hipótesis en el sistema judicial.

Naturaleza general de un error estándar

Si éste fuera el mejor de todos los posibles mundos de investigación, no habría error aleatorio, y si no hubiese error aleatorio, no habría necesidad de pruebas estadísticas de significancia. De hecho, el término *significancia* no tendría ningún significado. Cualquier diferencia sería una diferencia "real"; pero esto tampoco sucede. Siempre existen errores debidos al azar (y también errores de sesgo), y en la investigación del comportamiento con frecuencia contribuyen sustancialmente a la varianza total. Los errores estándar son medidas de este error y se utilizan, como se ha indicado una y otra vez, como un tipo de patrón contra el cual se verifica la varianza experimental.

El *error estándar* es la desviación estándar de la distribución muestral de cualquier medición —la media o el coeficiente de correlación, por ejemplo—. En la mayoría de los casos, no es factible conocer los valores de la población o universo (parámetros); deben estimarse a partir de medidas de la muestra, por lo común de muestras únicas.

Suponga que se extrae una muestra aleatoria de 100 niños de las aulas de octavo grado en determinado sistema educativo. Resulta difícil o imposible medir al universo completo de alumnos de octavo grado. Se calculan la media y la desviación estándar de una prueba aplicada a los niños, resultando los estadísticos $M = 110$; $DE = 10$. Las preguntas importantes a plantearse son: ¿Qué tan precisa es esta media? O, si se extranjera un número grande de muestras aleatorias de 100 alumnos de octavo grado de esta misma población, ¿serían las medias de estas muestras 110 o cercanas a 110?; y si fuesen cercanas a 110, ¿qué tan cerca estarían? Lo que se hace, en efecto, es crear una *distribución hipotética de medias muestrales*, todas calculadas a partir de muestras de 100 alumnos, cada una obtenida de la población original de alumnos de octavo grado. Si se pudiera calcular la media de esta

población de medias, o si se supiera cuál es ésta, todo resultaría simple. Pero no se conoce dicho valor ni es posible conocerlo, ya que las posibilidades de extraer muestras diferentes son bastante numerosas. Lo mejor que puede hacerse es *estimarlo con el valor muestral o la media de la muestra.* En este caso, simplemente se dice "sea la media muestral igual a la media poblacional" y espere estar en lo correcto. Después debe probarse la ecuación con el error estándar.

Un argumento similar se aplica a la desviación estándar de la población total (de las puntuaciones originales). No se conoce y tal vez nunca se conocerá; pero puede ser estimada con la desviación estándar, calculada a partir de la muestra. De nuevo se dice "sea la desviación estándar de la muestra igual a la de la población". Se sabe que probablemente no tengan el mismo valor; aunque también se sabe que, si el muestreo ha sido aleatorio, probablemente sean cercanas.

En el capítulo 11 se utilizó la desviación estándar de la muestra como un sustituto de la desviación estándar de la población en la fórmula para el *error estándar de la media*:

$$EE_M = \frac{DE}{\sqrt{n}} \qquad (12.1)$$

Éste también se llama el *error de muestreo*; así como la desviación estándar es una medida de la dispersión de las puntuaciones originales, el error estándar de la media es una medida de la dispersión de la distribución de medias muestrales. *No* es la desviación estándar de la población de puntuaciones individuales. *No* es lo mismo que probar a cada miembro de la población y, después, calcular la media y la desviación estándar de dicha población.

Una demostración Monte Carlo

Para tener material de trabajo, ahora se recurre a la computadora y a los denominados métodos Monte Carlo, que son métodos de simulación asistidos por computadora diseñados para obtener soluciones a problemas matemáticos, estadísticos, numéricos y aun verbales, por medio del uso de procedimientos alcatorios y muestras de números alcatorios. En general asociados con problemas matemáticos cuyas soluciones son imposibles, los métodos Monte Carlo han extendido su uso para "probar" las características estadísticas de muestras de poblaciones grandes. Por ejemplo, las consecuencias de violar los supuestos subyacentes a las pruebas estadísticas de significancia pueden estudiarse efectivamente al simular distribuciones estadísticas con números aleatorios, e introduciendo violaciones de los supuestos en el procedimiento para estudiar las consecuencias. En las ciencias del comportamiento, los procedimientos Monte Carlo por lo común son estudios empíricos de modelos estadísticos y de otros tipos, que utilizan los números aleatorios generados por computadora para ayudar a simular los procesos aleatorios necesarios para estudiar los modelos. De cualquier manera, ahora se utilizará una forma elemental del método Monte Carlo para probar un teorema estadístico de lo más importante y para explorar la variabilidad de medias y el uso del error estándar de la media. También se busca establecer un fundamento para entender el uso de la computadora en el estudio de los procesos aleatorios.

Procedimiento

Un programa de computadora está diseñado para generar 4 000 números aleatorios distribuidos uniformemente entre el 0 y el 100 (de tal manera que cada número tiene la

TABLA 12.1 *Medias, desviaciones estándar y errores estándar de las medias, cinco muestras de 100 números aleatorios (de 0 a 100)*[a]

	Muestras				
	1	2	3	4	5
M	52.21	49.64	51.37	49.02	55.51
DE	29.62	27.91	29.83	26.72	29.23
EE_M	2.96	2.79	2.98	2.67	2.92

[a] Estadísticos de la población: $M = 50.33$; $DE = 29.1653$; $N = 4\,000$.

misma probabilidad de ser "extraído") en 40 conjuntos de 100 números cada uno, y para calcular diversos estadísticos con los números. Considere este conjunto de 4 000 números como una población, o *U*. La media de *U* es 50.33 (de acuerdo al cálculo real de la computadora) y la desviación estándar es 29.17. Se desea estimar esta media a partir de muestras extraídas aleatoriamente de *U*. Por supuesto, que en una situación real por lo común no se conoce la media de la población. Una de las virtudes de los procedimientos Monte Carlo consiste en que se puede conocer lo que usualmente no se conoce.

Cinco de los 40 conjuntos de 100 números se extraen aleatoriamente. (Los conjuntos extraídos son los conjuntos número 5, 7, 8, 16 y 36 [véase apéndice C].) Se calculan las medias y las desviaciones estándar así como los errores estándar de la media de los cinco conjuntos. Estos estadísticos se reportan en la tabla 12.1. Se quiere presentar una idea intuitiva de lo que es el error estándar de la media y cómo se utiliza.

Primero se calcula la desviación estándar (*DE*) de esta muestra de medias. Si tan sólo se trata a las cinco medias (53.21, 49.64, 51.37, 49.02 y 55.51) como puntuaciones ordinarias y se calcula la media de estas medias y su desviación estándar, se obtiene: $M = 51.75$; $DE = 2.38$. La media de las 4 000 puntuaciones es 50.33. Cada una de las cinco medias es un estimado muestral de esta media poblacional. Note que tres de ellas, 49.64, 51.37 y 49.02, están bastante cercanas a la media poblacional; y dos de ellas, 53.21 y 55.51, están más alejadas de ella. Parece ser que tres de las muestras proveen buenos estimados de la media poblacional y que dos no lo hacen, ¿o sí?

La desviación estándar de 2.48 es *similar* al error estándar de la media (por supuesto que no es el error estándar de la media, ya que ha sido calculada a partir de sólo cinco medias). Supóngase que sólo se hubiera extraído una muestra con $M = 53.21$ y $DE = 29.62$, lo cual es la situación común de investigación, y que se calculó el error estándar de la media:

$$EE_M = \frac{DE}{\sqrt{n}} = \frac{29.62}{\sqrt{100}} = 2.96$$

Este valor es un estimado de la desviación estándar de las *medias* poblacionales de muchísimas muestras con 100 casos, cada una extraída aleatoriamente de la población. La población del ejemplo tiene 40 grupos y, por lo tanto, 40 medias (que, por supuesto, no son muchísimas medias). La desviación estándar de estas medias es en realidad 3.10. El EE_M calculado para la primera muestra es cercano a este valor poblacional: 2.96, como un estimado de 3.10.

Los cinco errores estándar de las medias se muestran en la tercera línea de datos de la tabla 12.1; fluctúan muy poco —de 2.67 a 2.98— aunque las medias de los conjuntos de 100 puntuaciones varíen considerablemente. La desviación estándar de 2.48, calculada

TABLA 12.2 *Medias de 20 conjuntos de 4 000 números aleatorios generados por computadora (0 a 100)*[a]

50.3322	49.9447	50.1615	50.0995
50.1170	49.5960	51.0585	51.1450
49.8200	49.3175	49.5822	50.6440
49.8227	49.9022	49.7505	49.8437
49.5875	50.6180	50.0990	49.3605

[a] La media de las medias es igual a 50.0401; la desviación de las medias es igual a 0.4956; el error estándar de la media de la primera muestra es igual a 0.4611.

para las cinco medias, es solamente una estimación razonable de la desviación estándar de la población de medias; aun así es un estimado. El punto importante e interesante es que el error estándar de la media, el cual es un estimado "teórico", calculado de los datos de cualquiera de los cinco grupos, es un estimado preciso de la variabilidad de las medias de las muestras de la población.

Para reforzar estas ideas, ahora se expondrá otra demostración Monte Carlo de mucho mayor magnitud. El mismo programa de cómputo utilizado para producir los 4 000 números aleatorios de los que se habló anteriormente, se utiliza ahora para producir otros 15 conjuntos de 4 000 números aleatorios cada uno, distribuidos uniformemente entre 0 y 100, es decir, que se generaron un total de 80 000 números aleatorios, en 20 conjuntos de 4 000 cada uno. De nuevo, la media teórica de los números entre 0 y 100 es 50. Considere a cada uno de los 20 conjuntos como una muestra de 4 000 números. Las medias de los 20 conjuntos se presentan en la tabla 12.2.

Las 20 medias se agrupan cerca y alrededor del 50: la más baja es 49.3175; la más alta, 51.1450, y la mayoría están cerca de 50. La media de las 20 medias es 50.0401, muy cerca en realidad de la expectativa teórica de 50. La desviación estándar de las 20 medias es 0.4956; la desviación estándar de la primera muestra de 4 000 casos (véase la nota *a* de la tabla 12.1) es 29.1653. Si se utiliza dicha desviación estándar para calcular el error estándar de la media, se obtiene:

$$EE_M = \frac{29.1653}{\sqrt{4\,000}} = .4611$$

Observe que este estimado del error estándar de la media es cercano a la desviación estándar calculada de las 20 medias. No sería un error utilizarlo para evaluar la variabilidad de las medias de muestras de 4 000 números aleatorios. Claramente, las medias de muestras grandes son estadísticos muy estables, y los errores estándar resultan buenos estimados de su variabilidad.

Generalizaciones

Ahora se pueden realizar diversas generalizaciones de gran utilidad para la investigación. Por ejemplo, las medias muestrales son estables en el sentido de que son mucho menos variables que las medidas a partir de las cuales calcularon. Esto, por supuesto, es verdadero por definición. Las varianzas, las desviaciones estándar y los errores estándar de la media son aún más estables, ya que fluctúan dentro de rangos relativamente estrechos. Aun cuando

las medias muestrales del ejemplo variaron tanto como cuatro o cinco puntos, los errores estándar fluctuaron en no más de un punto y medio. Lo anterior significa que se puede tener bastante confianza en que los estimados de las medias muestrales estarán muy cerca de la media poblacional de dichas medias. Además, la ley de los números grandes afirma que a mayor tamaño de la muestra, más cercanos estarán probablemente los estadísticos a los valores poblacionales.

Una pregunta difícil para los investigadores es: ¿Se mantienen siempre estas generalizaciones, especialmente con muestras no aleatorias? La validez de las generalizaciones depende del muestreo aleatorio. Si el muestreo no es aleatorio, no se puede saber en realidad si se mantienen las generalizaciones. No obstante, con frecuencia se debe actuar como si de hecho se mantuvieran, aun con muestras no aleatorias. Por fortuna, si se es cuidadoso al estudiar los datos para detectar idiosincrasia muestral sustancial, es posible utilizar ventajosamente la teoría. Por ejemplo, las muestras pueden examinarse para expectativas fáciles de verificar: si se espera un número aproximadamente igual de hombres y mujeres en una muestra, o proporciones conocidas de republicanos y demócratas o de jóvenes y viejos, se vuelve sencillo contar estos números. Hay expertos que insisten en el muestreo aleatorio como una condición de la validez de la teoría —y esto es correcto hasta cierto punto—. Sin embargo, si no se debe aplicar la teoría con muestras no aleatorias, se debe abandonar el uso de los estadísticos y de las inferencias que de ellos se derivan. La realidad es que los estadísticos parecen funcionar muy bien aun con muestras no aleatorias, siempre y cuando el investigador conozca las limitaciones de dichas muestras. El investigador necesita ser todavía más cuidadoso con muestras no aleatorias que con muestras aleatorias. La réplica de estudios no aleatorios es una obligación.

Teorema del límite central

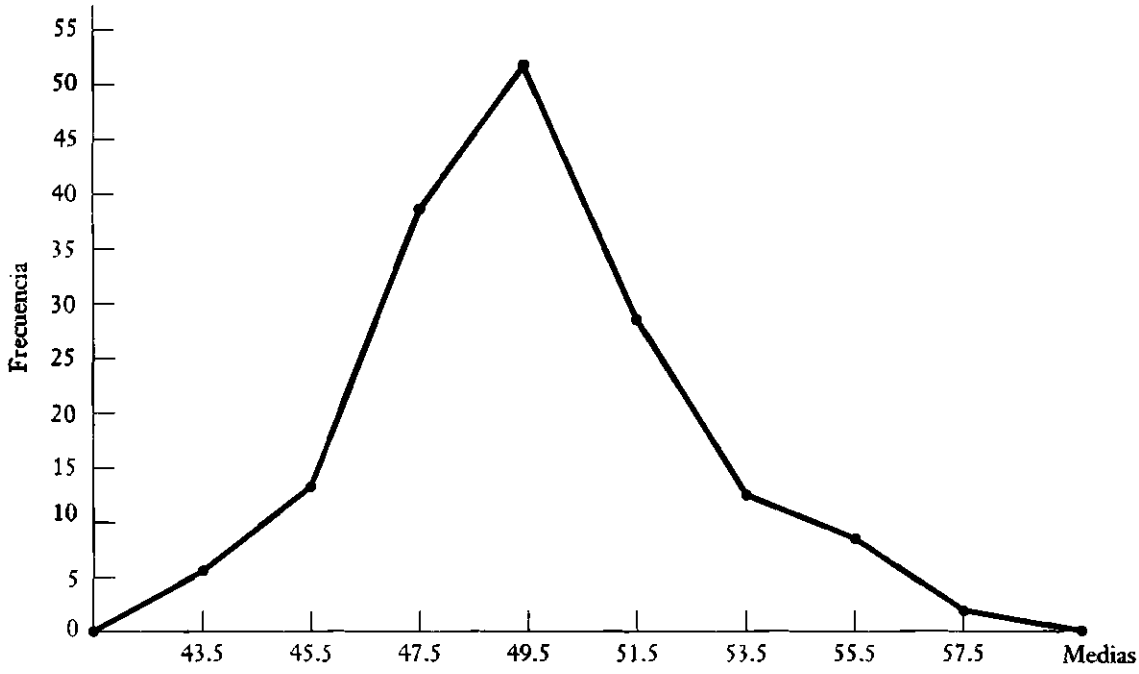
Antes de estudiar el uso real del error estándar de la media, se requiere conocer un poco acerca de una generalización extremadamente importante sobre las medias: *si las muestras son extraídas aleatoriamente de una población, las medias de las muestras tenderán a distribuirse normalmente*. Mientras mayor sea el tamaño de las *N*, más resulta así. La forma y el tipo de distribución de la población original no provoca ninguna diferencia; es decir, la distribución de la población no tiene que estar distribuida normalmente (véase Hays, 1994, pp. 251-254 para un buen ejemplo sobre la forma en que funciona el teorema).

Por ejemplo, la distribución de los 4 000 números aleatorios en el apéndice C es rectangular, ya que los números están distribuidos de manera uniforme. Si el teorema del límite central es válido desde el punto de vista empírico, entonces las medias de los 40 conjuntos de 100 puntuaciones cada uno deberían distribuirse de forma aproximadamente normal; si es así, esto es notable. Y así sucede, aunque una muestra de 40 medias apenas es suficiente para demostrar la tendencia; por lo tanto, se generan por computadora otras tres poblaciones de 4 000 números aleatorios diferentes, distribuidos uniformemente, separados en 40 subconjuntos de 100 números cada uno.

Las medias para los $4 \times 40 = 160$ subconjuntos de 100 números cada uno se calcularon y se incorporaron en una distribución. Un polígono de frecuencias de las medias se presenta en la figura 12.1, donde puede verse que las 160 medias se observan casi como la curva normal en forma de campana. En apariencia el teorema del límite central "funciona", y es necesario recordar que esta distribución de medias se obtuvo de distribuciones rectangulares de números.

¿Por qué molestarse con todo esto? ¿Por qué es importante demostrar que las distribuciones de medias se aproximan a la normalidad? Se trabajó bastante con medias en el análisis de datos, y si están normalmente distribuidas entonces se pueden utilizar las pro-

FIGURA 12.1



piedades conocidas de la curva normal para interpretar los datos de investigación obtenidos. Saber que aproximadamente el 96% de las medias se ubicará entre dos desviaciones estándar (errores estándar), por arriba y por debajo de la media, es información valiosa, pues un resultado obtenido puede ser evaluado contra las propiedades conocidas de la curva normal. En el capítulo 11 se estudió el uso de la curva normal para interpretar medias; ahora se estudiará lo que quizás es un uso más interesante de la curva para evaluar las diferencias entre medias.

Error estándar de las diferencias entre medias

Una de las estrategias más frecuentes y útiles en investigación consiste en comparar medias de muestras. A partir de las diferencias entre medias se infieren efectos de la variable independiente. Cualquier combinación lineal de medias también está gobernada por el teorema del límite central; es decir, que las diferencias entre medias se distribuirán normalmente, si se tienen muestras suficientemente grandes. (Una combinación lineal es cualquier ecuación de primer grado, por ejemplo, $Y = M_1 - M_2$. $Y = M_1^2 - M_2$ no es lineal.) Por lo tanto, es posible utilizar la misma teoría con las diferencias entre medias que aquella que se usa con medias.

Suponga que se asignan 200 sujetos a dos grupos aleatoriamente, 100 a cada grupo. A un grupo se le muestra una película sobre relaciones intergrupales (grupo *A*), por ejemplo, y al otro grupo no se le muestra ninguna película (grupo *B*); después, se les aplica a ambos grupos una medida de actitud. La puntuación media del grupo *A* es 110, y la del grupo *B*

es 100. El problema es: ¿la diferencia de 10 unidades es una diferencia "real", una diferencia estadísticamente significativa? ¿O es una diferencia que pudo haber surgido por azar (más de 5 veces en 100, por ejemplo, o alguna otra cantidad) cuando, de hecho, no existe una diferencia?

Si, de manera similar, se crea otro par muestras de 100 elementos cada una y se calculan las diferencias entre las medias de estas muestras y se sigue el mismo procedimiento experimental, ¿se obtendrá consistentemente esta diferencia de 10? De nuevo se utiliza el error estándar para evaluar las diferencias, pero ahora se tiene una *distribución muestral de diferencias entre medias.* Es como si se calculara cada $M_i - M_j$ y se considerara como una X. Entonces, las diversas diferencias entre las medias muestrales son consideradas como las X de una nueva distribución; de todos modos, la desviación estándar de esta distribución muestral de diferencias *es similar* al error estándar. Sin embargo, este procedimiento es sólo una ilustración, porque en realidad esto no se hace. Aquí, de nuevo se estima el error estándar de los dos primeros grupos, *A* y *B*, utilizando la siguiente fórmula:

$$EE_{M_A - M_B} = \sqrt{EE^2{}_{M_A} + EE^2{}_{M_B}} \qquad (12.2)$$

donde $EE^2{}_{M_A}$ y $EE^2{}_{M_B}$ son los errores estándar elevados al cuadrado, del grupo *A* y grupo *B*, respectivamente.

Suponga que el experimento se realiza con cinco pares de grupos, es decir, 10 grupos, dos a la vez. Las cinco diferencias entre las medias son 10, 11, 12, 8, 9. La media de estas diferencias es 10; la desviación estándar es 1.414; este 1.414 es, de nuevo, *similar* al error estándar de la distribución muestral de las diferencias entre medias, en el mismo sentido que el error estándar de la media en el análisis anterior. Si ahora se calcula el error estándar de la media para cada grupo (las desviaciones estándar son inventadas para los dos grupos, $DE_A = 8$ y $DE_B = 9$), se obtiene:

$$EE_{M_A} = \frac{DE_A}{\sqrt{n_A}} = \frac{8}{\sqrt{100}} = .8, \qquad EE_{M_B} = \frac{DE_B}{\sqrt{n_B}} = \frac{9}{100} = .9$$

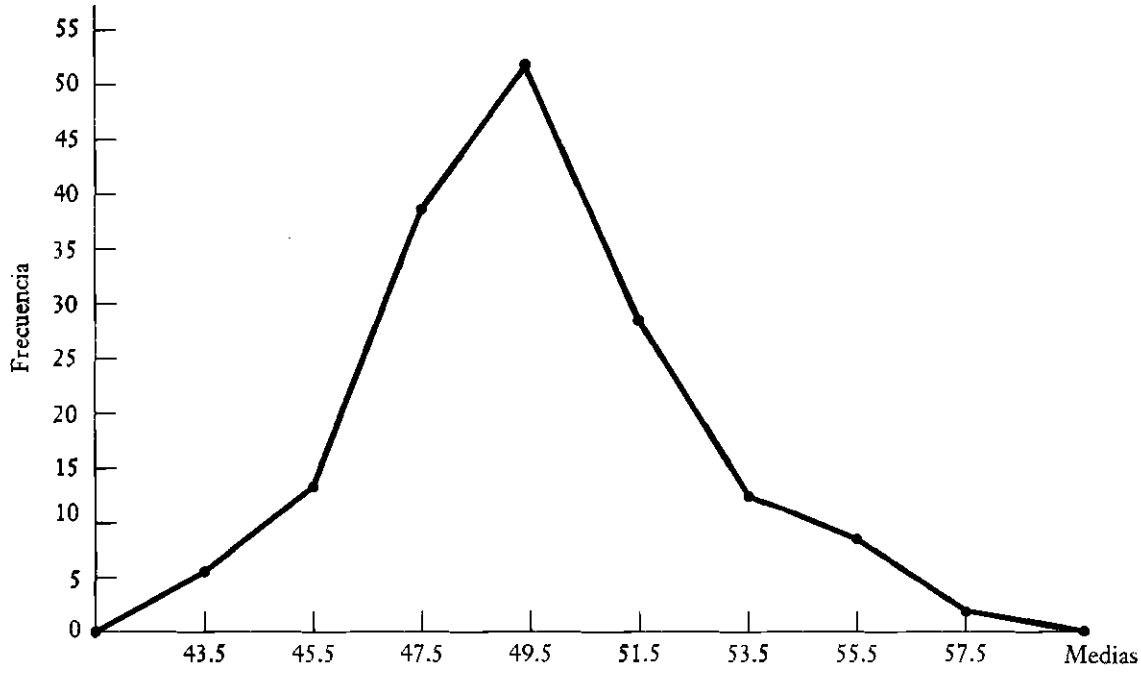
Mediante la ecuación 12.2 se calcula el error estándar de las diferencias entre las medias:

$$EE_{M_A - M_B} = \sqrt{EE^2{}_{M_A} + EE^2{}_{M_B}} = \sqrt{(.8)^2 + (.9)^2} = \sqrt{.64 + .81} = \sqrt{1.45} = 1.20$$

¿Qué se hace con el 1.20 resultante, ahora que se tiene? Si las puntuaciones de los dos grupos hubieran sido escogidas de una tabla de números aleatorios y no existieran condiciones experimentales, no se esperarían diferencias entre las medias; sin embargo, como ya se mencionó, siempre existen diferencias relativamente pequeñas debidas a factores del azar; estas diferencias son aleatorias. *El error estándar de las diferencias entre las medias es un estimado de la dispersión de estas diferencias.* Pero una medida de estas diferencias es en sí lo que es un estimado de dichas diferencias para la población entera. Por ejemplo, el error estándar de las diferencias entre las medias es 1.20, lo cual indica que, debido sólo al azar, la diferencia entre M_A y M_B fluctuará aleatoriamente alrededor de 10. Esto es, ahora puede ser 10 y después quizá 10.2 o 9.8, etcétera. Sólo en raras ocasiones las diferencias excederán, digamos, 13 o 7 (aproximadamente tres veces el *EE*). Otra forma de expresarlo es decir que el error estándar de 1.20 indica los límites (si se multiplica el 1.20 por el factor apropiado) que probablemente no excederán las diferencias muestrales.

¿Qué tiene que ver todo esto con el experimento? Es precisamente aquí donde se evalúan los resultados experimentales. El error estándar de 1.20 estima las fluctuaciones

FIGURA 12.1



piedades conocidas de la curva normal para interpretar los datos de investigación obtenidos. Saber que aproximadamente el 96% de las medias se ubicará entre dos desviaciones estándar (errores estándar), por arriba y por debajo de la media, es información valiosa, pues un resultado obtenido puede ser evaluado contra las propiedades conocidas de la curva normal. En el capítulo 11 se estudió el uso de la curva normal para interpretar medias; ahora se estudiará lo que quizás es un uso más interesante de la curva para evaluar las diferencias entre medias.

Error estándar de las diferencias entre medias

Una de las estrategias más frecuentes y útiles en investigación consiste en comparar medias de muestras. A partir de las diferencias entre medias se infieren efectos de la variable independiente. Cualquier combinación lineal de medias también está gobernada por el teorema del límite central; es decir, que las diferencias entre medias se distribuirán normalmente, si se tienen muestras suficientemente grandes. (Una combinación lineal es cualquier ecuación de primer grado, por ejemplo, $Y = M_1 - M_2$. $Y = M_1^2 - M_2$ no es lineal.) Por lo tanto, es posible utilizar la misma teoría con las diferencias entre medias que aquella que se usa con medias.

Suponga que se asignan 200 sujetos a dos grupos aleatoriamente, 100 a cada grupo. A un grupo se le muestra una película sobre relaciones intergrupales (grupo *A*), por ejemplo, y al otro grupo no se le muestra ninguna película (grupo *B*); después, se les aplica a ambos grupos una medida de actitud. La puntuación media del grupo *A* es 110, y la del grupo *B*

es 100. El problema es: ¿la diferencia de 10 unidades es una diferencia "real", una diferencia estadísticamente significativa? ¿O es una diferencia que pudo haber surgido por azar (más de 5 veces en 100, por ejemplo, o alguna otra cantidad) cuando, de hecho, no existe una diferencia?

Si, de manera similar, se crea otro par muestras de 100 elementos cada una y se calculan las diferencias entre las medias de estas muestras y se sigue el mismo procedimiento experimental, ¿se obtendrá consistentemente esta diferencia de 10? De nuevo se utiliza el error estándar para evaluar las diferencias, pero ahora se tiene una *distribución muestral de diferencias entre medias.* Es como si se calculara cada $M_i - M_j$ y se considerara como una X. Entonces, las diversas diferencias entre las medias muestrales son consideradas como las X de una nueva distribución; de todos modos, la desviación estándar de esta distribución muestral de diferencias *es similar* al error estándar. Sin embargo, este procedimiento es sólo una ilustración, porque en realidad esto no se hace. Aquí, de nuevo se estima el error estándar de los dos primeros grupos, A y B, utilizando la siguiente fórmula:

$$EE_{M_A - M_B} = \sqrt{EE^2{}_{M_A} + EE^2{}_{M_B}} \qquad (12.2)$$

donde $EE^2{}_{M_A}$ y $EE^2{}_{M_B}$ son los errores estándar elevados al cuadrado, del grupo A y grupo B, respectivamente.

Suponga que el experimento se realiza con cinco pares de grupos, es decir, 10 grupos, dos a la vez. Las cinco diferencias entre las medias son 10, 11, 12, 8, 9. La media de estas diferencias es 10; la desviación estándar es 1.414; este 1.414 es, de nuevo, *similar* al error estándar de la distribución muestral de las diferencias entre medias, en el mismo sentido que el error estándar de la media en el análisis anterior. Si ahora se calcula el error estándar de la media para cada grupo (las desviaciones estándar son inventadas para los dos grupos, $DE_A = 8$ y $DE_B = 9$), se obtiene:

$$EE_{M_A} = \frac{DE_A}{\sqrt{n_A}} = \frac{8}{\sqrt{100}} = .8, \qquad EE_{M_B} = \frac{DE_B}{\sqrt{n_B}} = \frac{9}{100} = .9$$

Mediante la ecuación 12.2 se calcula el error estándar de las diferencias entre las medias:

$$EE_{M_A - M_B} = \sqrt{EE^2{}_{M_A} + EE^2{}_{M_B}} = \sqrt{(.8)^2 + (.9)^2} = \sqrt{.64 + .81} = \sqrt{1.45} = 1.20$$

¿Qué se hace con el 1.20 resultante, ahora que se tiene? Si las puntuaciones de los dos grupos hubieran sido escogidas de una tabla de números aleatorios y no existieran condiciones experimentales, no se esperarían diferencias entre las medias; sin embargo, como ya se mencionó, siempre existen diferencias relativamente pequeñas debidas a factores del azar; estas diferencias son aleatorias. *El error estándar de las diferencias entre las medias es un estimado de la dispersión de estas diferencias.* Pero una medida de estas diferencias es en sí lo que es un estimado de dichas diferencias para la población entera. Por ejemplo, el error estándar de las diferencias entre las medias es 1.20, lo cual indica que, debido sólo al azar, la diferencia entre M_A y M_B fluctuará aleatoriamente alrededor de 10. Esto es, ahora puede ser 10 y después quizá 10.2 o 9.8, etcétera. Sólo en raras ocasiones las diferencias excederán, digamos, 13 o 7 (aproximadamente tres veces el *EE*). Otra forma de expresarlo es decir que el error estándar de 1.20 indica los límites (si se multiplica el 1.20 por el factor apropiado) que probablemente no excederán las diferencias muestrales.

¿Qué tiene que ver todo esto con el experimento? Es precisamente aquí donde se evalúan los resultados experimentales. El error estándar de 1.20 estima las fluctuaciones

aleatorias. ¿Pudo $M_A - M_B = 10$ haber surgido por el azar, como un resultado de fluctuaciones aleatorias, como se describió? Debe ir quedando claro que esto no es posible, excepto bajo circunstancias muy inusuales. Se evalúa esta diferencia de 10 comparándola con la estimación de las fluctuaciones aleatorias. *¿Se trata de una de ellas?* Se hacen ahora las comparaciones por medio de la razón *t* o prueba *t*:

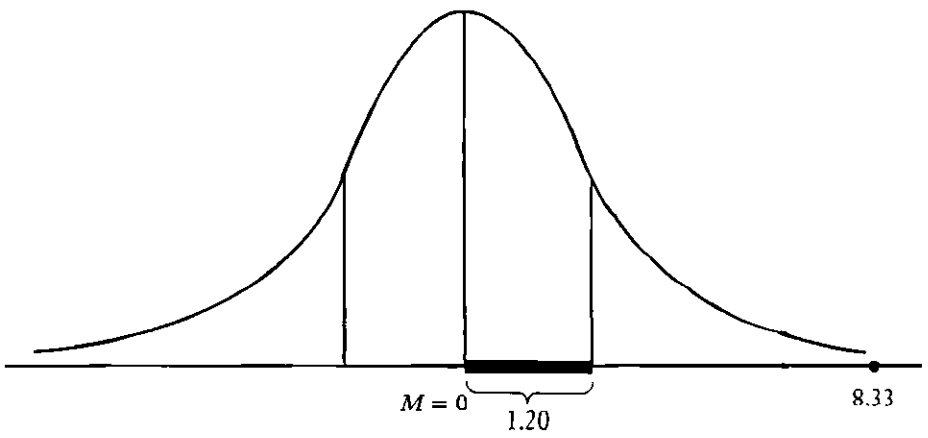
$$t = \frac{M_A - M_B}{EE_{M_A - M_B}} = \frac{110 - 100}{1.20} = \frac{10}{1.20} = 8.33$$

La ecuación establece que la diferencia medida entre M_A y M_B estaría a 8.33 desviaciones estándar (unidades de error) de distancia de una media hipotética de cero (diferencia de cero, no diferencia entre las dos medias).

En teoría no habría ninguna diferencia si los sujetos estuvieran aleatorizados y si no hubiese habido manipulación experimental. Se tendrían, en efecto, dos distribuciones de números aleatorios de los que se esperarían tan sólo fluctuaciones debidas al azar; sin embargo, aquí se tiene una diferencia relativamente enorme de 10, comparada con un insignificante 1.20 (el estimado de las desviaciones aleatorias). Decididamente algo debe estar ocurriendo aquí, además del azar, y ese algo es aquello que se está buscando. Presumiblemente es el efecto de la película o el efecto de la condición experimental, siempre y cuando otras condiciones hayan sido controladas lo suficiente, por supuesto.

Observe la figura 12.2, que representa una *población de diferencias entre* medias con una media de cero y una desviación estándar de 1.20. (La media se establece en cero porque se asume que la media de todas las diferencias de medias es cero.) ¿En qué parte de la línea basal del diagrama se colocaría la diferencia de 10? Para contestar esta pregunta, primero debe convertirse el 10 en unidades de desviación estándar (o error estándar). (Recuerde las puntuaciones estándar del capítulo 11.) Esto se hace dividiendo el 10 entre la desviación estándar (error estándar), que es 1.20: 10/1.2 = 8.33. Sin embargo, esto es lo que se obtuvo al calcular la razón *t*; es, entonces, simplemente la diferencia entre M_A y M_B, 10, expresada en unidades de desviación estándar (error estándar). Ahora puede incluirse en la línea basal del diagrama; observe el punto que se encuentra hacia la derecha: claramente la diferencia de 10 constituye una desviación, que se ubica tan alejada que probablemente no

FIGURA 12.2



pertenece a la población en cuestión. En resumen, la diferencia entre M_A y M_B es estadísticamente significativa, tanto que alcanza lo que Bernoulli llamó "certeza moral". Una diferencia tan grande o desviación de las expectativas por azar, difícilmente, puede atribuirse sólo al azar; las probabilidades son, en realidad, mayores de cien mil millones a una. Puede suceder por azar; pero es poco probable que suceda. Una pregunta importante es: ¿qué tan grande debe ser una diferencia, o en lenguaje estadístico, qué tan lejos de la media hipotética de cero debe estar una desviación para ser significativa? Esta pregunta no puede ser contestada de manera definitiva en este libro. Con muestras grandes, el nivel 0.05 representa 1.96 desviaciones estándar de la media; y el nivel 0.01, 2.58 desviaciones estándar de la media. Pero existen complicaciones, especialmente con muestras pequeñas; el estudiante, como siempre, necesita estudiar un buen texto de estadística. Una regla simple es: 2 desviaciones estándar son significativas (aproximadamente al nivel 0.05); 2.5 desviaciones estándar son muy significativas (aproximadamente al nivel 0.01) y 3 desviaciones estándar son altamente significativas (un poco menos del nivel 0.001).

Tal es el error estándar y sus usos. Los errores estándar de otros estadísticos se utilizan de la misma manera. Es una herramienta útil e importante, pues constituye un instrumento básico en la investigación contemporánea. De hecho, sería difícil imaginar la metodología moderna e imposible imaginar la estadística actual sin el error estándar; como elemento clave para la inferencia estadística, su importancia no puede sobrestimarse. Mucha de la inferencia estadística se reduce a una familia de fracciones resumida en la siguiente fracción:

$$\frac{\text{Estadístico}}{\text{Error estándar del estadístico}}$$

Inferencia estadística

Inferir significa derivar una conclusión a partir de premisas o de la evidencia. *Inferir estadísticamente* quiere decir derivar conclusiones probabilísticas a partir de premisas probabilísticas. Se concluye probabilísticamente, es decir, a un nivel especificado de significancia. Se infiere, probabilísticamente, si un resultado experimental se desvía de las expectativas por el azar y si la hipótesis nula no es "verdadera", que una influencia "real" está operando. Si, en el experimento de los métodos, $M_A > M_B$ y $M_A \neq M_B$, o H_1 es "verdadera" y H_0 no es "verdadera", se infiere que el método A es "superior" al método B, entendiéndose "superior" en el sentido definido en el experimento.

Otra forma de inferencia, discutida profundamente en el capítulo sobre muestreo, es aquella que establece que la inferencia se realiza de una muestra hacia una población. Puesto que, por ejemplo, el 55% de una muestra aleatoria de 2 000 personas en Estados Unidos dice que votará por cierto candidato presidencial, se infiere que si se le preguntara a toda la población estadunidense, respondería de manera similar. Esta es una inferencia bastante arriesgada. Uno de los peligros más graves de la investigación (o quizás deba decirse de cualquier razonamiento humano) consiste en el salto inferencial de los datos muestrales a los hechos de la población. Con frecuencia se realizan saltos inferenciales en política, economía, educación y otras áreas de gran importancia. Por ejemplo, si el gobierno recorta los gastos, la inflación decrecerá; si se utilizan máquinas de enseñanza, los niños aprenderán más. Los científicos también dan saltos inferenciales —en ocasiones muy grandes— con una diferencia importante: el científico está (o debería estar) consciente de dichos saltos y de que siempre son riesgosos.

Puede afirmarse, en resumen, que la estadística permite a los científicos probar hipótesis sustantivas indirectamente al permitirles probar hipótesis estadísticas directamente (si

es que es posible probar algo directamente). En este proceso, ellos utilizan hipótesis nulas, hipótesis escritas por el azar. Prueban la "verdad" de hipótesis sustantivas al someter hipótesis nulas a pruebas estadísticas basadas en razonamientos probabilísticos; después hacen inferencias apropiadas. De hecho, el objetivo de todas las pruebas estadísticas consiste en probar qué tanto se justifican las inferencias. Un revisor de este capítulo ha cuestionado el mensaje implícito del capítulo, es decir, que todas las pruebas estadísticas de hipótesis incluyen errores estándar. Esta implicación sería desafortunada, ya que, como se verá en capítulos posteriores, existen otros medios que se utilizan con frecuencia para evaluar la significancia estadística. Por ejemplo, las pruebas de análisis de varianza no paramétrico presentadas en el capítulo 16 dependen de los rangos, y las complejas pruebas del análisis estructural de covarianza del capítulo 37 dependen de comparaciones de covarianzas (correlaciones) y de la comparación de estructuras latentes con datos empíricos.

Comprobación de hipótesis y los dos tipos de errores

En un experimento de lanzamiento de monedas se pueden probar las hipótesis de que la moneda está o no equilibrada. Las hipótesis se expresan de la siguiente manera:

$$H_0: p = 1/2$$
$$H_1: p \neq 1/2$$

donde H_0 es igual a la hipótesis sometida a prueba, y p es igual a la probabilidad verdadera de que resulte cara. La hipótesis sometida a prueba, H_0, establece que p, la probabilidad verdadera de obtener cara en cualquier ensayo, es 1/2. Si esto es verdadero, entonces la moneda está equilibrada. Por supuesto que en la práctica no puede garantizarse que el número de caras obtenido con una moneda equilibrada sea exactamente 1/2, a menos que la moneda sea lanzada un número infinito de veces —algo que es imposible—. Con una moneda recién acuñada, el número de caras se aproxima a 50% conforme se incrementa el número de ensayos.

En un experimento de lanzamiento de monedas donde 12 de 16 lanzamientos resultan caras, se sospecha que la moneda está dando demasiadas caras; la probabilidad de tal evento se puede obtener utilizando la fórmula binomial (véase Comrey y Lee, 1995, capítulo 7) o consultando una tabla de valores binomiales (véase Beyer, 1971, p. 44). La probabilidad o valor p para el resultado obtenido es de 0.038. Si se eligiera de antemano el nivel 0.05 de significancia, el resultado sería declarado "significativo", pues $0.038 < 0.05$. Sin embargo, no sería significativo si se hubiera escogido el nivel de significancia de 0.01, ya que $0.038 > 0.01$.

Si se condujera otro experimento con la misma moneda y resultaran 15 caras en 19 lanzamientos, la probabilidad de que esto suceda, si se asume que la moneda está equilibrada, es de 0.0096. En este caso, los resultados serían significativos no sólo al nivel 0.05 ($0.0096 < 0.05$), sino que también lo serían al nivel de 0.01 ($0.0096 < 0.01$).

En el ejemplo donde se obtuvieron 12 caras en 16 lanzamientos de una moneda, se rechaza la hipótesis de que la moneda está equilibrada, a causa de que la probabilidad de ocurrencia de dicho evento (dado que la moneda está equilibrada) es de 0.038, y este valor es menor a la cantidad tolerable de 0.05. Rechazar la H_0 sería un error si, de hecho, la moneda está equilibrada; a este error se le llama *error tipo I*. Una moneda equilibrada podría generar 12 o más caras en 16 lanzamientos; la posibilidad de esta ocurrencia es 0.038 o 38 de 1 000 repeticiones del mismo experimento de 16 lanzamientos. No se sabe de antemano si este experimento en particular es uno de los 38 posibles cuando una moneda equilibrada origina 12 caras en 16 lanzamientos o si la moneda en realidad está desequi-

librada. No obstante se rechaza H_0 con el conocimiento de que se pudo haber cometido un error; aunque la probabilidad de que eso ocurra es menor a 0.05. La conclusión de rechazar H_0 es correcta, en promedio, más del 95% de las veces. Para el nivel de 1% de significancia, rechazar una hipótesis nula verdadera ocurre un promedio de una vez en cada 100 experimentos. Para el nivel del 5%, ocurre un promedio de cinco veces en cada 100 experimentos. Por lo tanto, rechazar una hipótesis nula verdadera constituye un error tipo I. El símbolo utilizado para representar la probabilidad de un error tipo I es la letra griega α (alfa). El término "nivel de confianza" se intercambia frecuentemente con "nivel de significancia" y "nivel alfa".

Un segundo tipo de error, denominado un *error tipo II*, se comete cuando la H_0 es falsa, pero a partir del análisis se concluye que la H_0 es verdadera. Esto es, aceptar una hipótesis nula falsa es un error tipo II. En general, observar 8 caras en 16 lanzamientos de una moneda es evidencia de que la moneda está equilibrada. Sin embargo, una moneda desequilibrada (una donde la probabilidad de obtener caras seas 0.25 en lugar de 0.5) puede generar 8 caras en 16 lanzamientos; la facilidad de que ello suceda no es tan alta con una moneda equilibrada, aunque una moneda desequilibrada puede hacerlo. El experimento puede repetirse muchas veces antes de formular un juicio; no obstante, en algunos experimentos del mundo real, como los encontrados en estudios de ingeniería sobre factores humanos, no resulta financieramente posible repetirlos. Por lo común se tiene un resultado experimental único a partir del cual se toma una decisión. En el ejemplo anterior, si la moneda está desequilibrada y la conclusión del experimento es que está equilibrada, entonces se ha cometido un error tipo II. La letra griega utilizada para representar la probabilidad de un error tipo II es β (beta).

La mayoría de los investigadores novatos tienden a establecer un criterio muy riguroso del error tipo I, con lo cual existe menos probabilidad de cometerlo. Sin embargo, existe una relación entre los errores tipo I y tipo II que debe considerarse antes de hacer que la decisión de cometer cualquiera de los errores sea demasiado rigurosa. Si se reduce la probabilidad de un error tipo I, aumenta la probabilidad del error tipo II en una muestra de tamaño fijo. A su vez, al reducir la probabilidad del error tipo II se incrementa la probabilidad del error tipo I. Como regla, al seleccionar un nivel de significancia debe decidirse qué tipo de error es más importante evitar o minimizar. Para tener la certeza de que un evento de cierta importancia ha sido identificado antes de reportarlo, se requiere usar un criterio de significancia bastante riguroso, como 0.01. Por otro lado, si existe mayor preocupación por no perder algo, debe usarse un nivel menos riguroso, como 0.05. La tabla 12.3 y la figura 12.3 muestran la relación entre los errores tipo I y tipo II. Probst (1988) presenta una absorbente discusión respecto a la relación entre los errores tipo I y tipo II en algunas situaciones reales.

Al examinar la figura 12.3, las áreas sombreadas indican la probabilidad de un error tipo I. El punto crítico de decisión es el punto que divide la distribución que sostiene que "H_0 es verdadera" de manera que el 0.05 o 0.01 del área se ubica a la derecha del punto. Al determinar la probabilidad de un error tipo I, queda determinada la probabilidad de un error tipo II. Al mover el punto crítico de decisión, el error tipo I se vuelve más pequeño o más grande y, a cambio, el error tipo II se hace más grande o más pequeño.

El tamaño de la muestra se relaciona con ambos tipos de error. Con un valor fijo de α y un tamaño muestral n fijo, se predetermina el valor de β. Si β es demasiado grande, puede reducirse al incrementar el nivel de α para una n fija, o al incrementar n para un nivel fijo de α. Aunque β rara vez se determina en un experimento, los investigadores pueden asegurarse de que es razonablemente pequeño recolectando una muestra grande.

El concepto del poder de una prueba surge del error tipo II, β; de hecho, el poder de una prueba se define como $1 - \beta$. El poder de una prueba es la probabilidad de rechazar

FIGURA 12.3

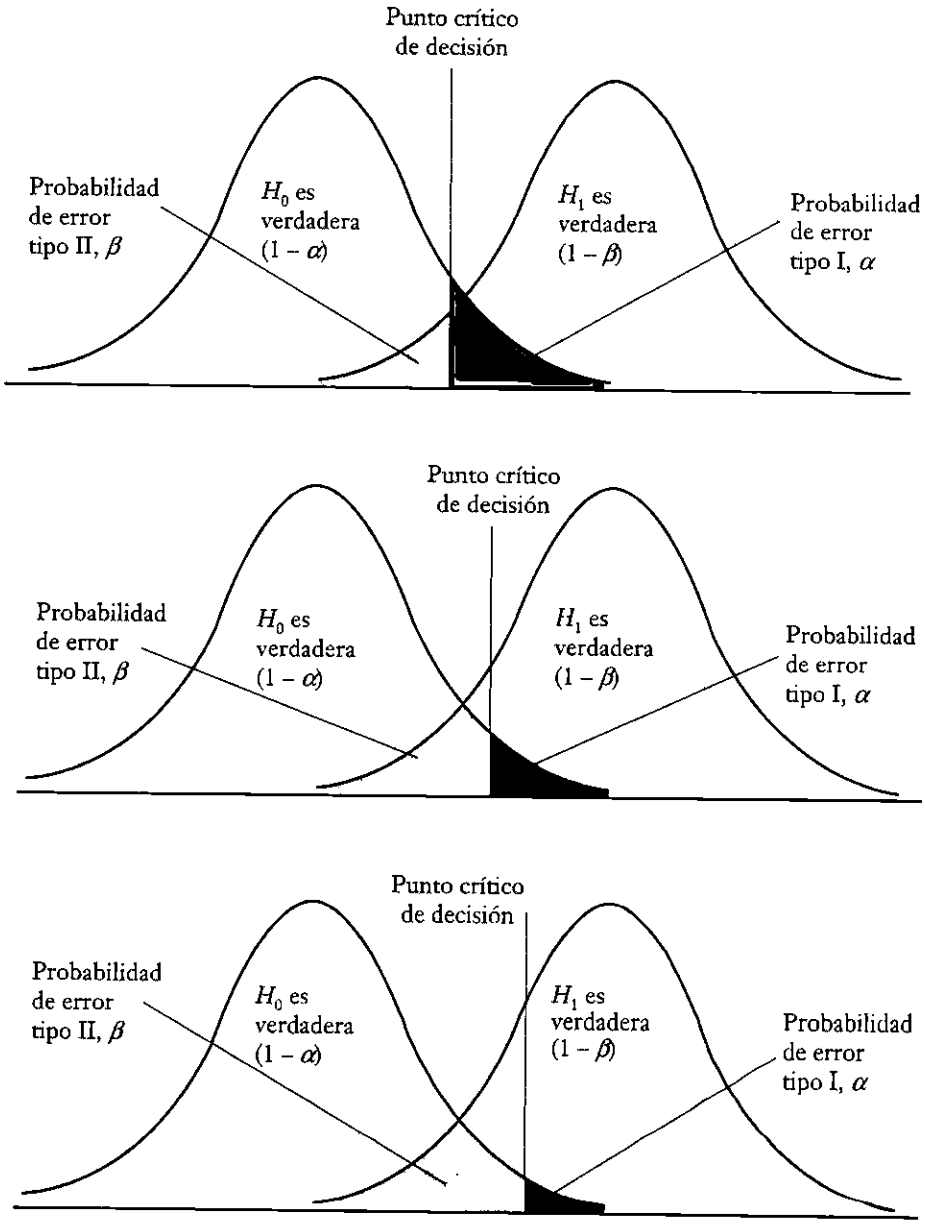

Punto crítico de decisión
Probabilidad de error tipo II, β
H_0 es verdadera (1 – α)
H_1 es verdadera (1 – β)
Probabilidad de error tipo I, α
Punto crítico de decisión
Probabilidad de error tipo II, β
H_0 es verdadera (1 – α)
H_1 es verdadera (1 – β)
Probabilidad de error tipo I, α
Punto crítico de decisión
Probabilidad de error tipo II, β
H_0 es verdadera (1 – α)
H_1 es verdadera (1 – β)
Probabilidad de error tipo I, α

TABLA 12.3 *Errores de decisión tipo I y tipo II*

		Verdadero estado del asunto	
		La hipótesis nula es correcta	**La hipótesis experimental es correcta**
La decisión	No se rechaza H_0	Decisión correcta, $1 - \alpha$	Error tipo II, β
	Se rechaza H_0	Error tipo I, α	Decisión correcta, $1 - \beta$

una hipótesis nula falsa. Se dice que una prueba es más poderosa que otra cuando tiene más posibilidades de descubrir diferencias significativas que esa otra. Dichas pruebas con diferentes niveles de poder pueden, además, compararse con un índice de eficiencia de poder, que generalmente va de 0.63 a 1.00. Cuando una prueba tiene una eficiencia de poder de 0.75, en comparación con otra prueba, ello indica que la prueba más débil requiere un tamaño de muestra de 100 para conseguir el mismo nivel de poder que la prueba más fuerte tiene con un tamaño de muestra de 75. Por lo común el *poder de prueba* no se calcula, ya que existen tablas disponibles para estimarlo. Un tratamiento más completo del tema se encuentra en Cohen (1988). La noción de poder se emplea con frecuencia al estimar el tamaño de la muestra. Diversos programas de cómputo para guiar el análisis sobre el poder se encuentran disponibles. Uno de los más prestigiados y mejor conocidos es el elaborado por Borenstein, Cohen y Rothstein (1997), llamado "Power Precision!" Otros son N & Nsurv y PASS (Power Analysis and Sample Size). Estos programas son costosos y, al momento de escribir estas líneas, sólo PASS está diseñado para correr en Windows. Los otros dos son programas compatibles con el sistema operativo DOS. Aunque la información de Internet se torna obsoleta rápidamente, en este momento hay un sitio donde el investigador puede bajar una lista y una revisión de los programas para calcular el poder. La dirección del sitio Web es http://www.interchg.ubc.ca/cacb/power/.[2] Un programa con base en DOS para el análisis de poder está disponible con el libro de Woodward, Bonett y Brecht (1990).

Los cinco pasos de la comprobación de hipótesis

Después de la discusión en las secciones previas, es el momento de poner en su lugar los cinco principales pasos utilizados en la comprobación de hipótesis. Al utilizar una hipótesis sustantiva, puede establecerse de forma estadística; aunque se ha hecho referencia a ella como hipótesis estadística, muchos estadísticos le llaman la hipótesis de investigación, experimental o alterativa. El paso 1 consiste en establecer dicha hipótesis estadística; generalmente se expresa en términos de valores poblacionales y contiene tanto el signo de no igual que ($\neq$), como el de mayor que ($>$) o el de menor que ($<$). Por ejemplo, la hipótesis estadística podría ser H_1: $\mu_A > \mu_B$ o $\mu_A - \mu_B = 0$. El paso 2 implica enunciar la hipótesis nula, H_0, la cual contiene el signo igual ($=$). Por ejemplo, ésta podría ser H_0: $\mu_A = \mu_B$ o $\mu_A - \mu_B = 0$. El paso 3 incluye calcular el estadístico de la prueba utilizando datos empíricos. El estadístico de la prueba generalmente es un tipo de puntuación estándar que expresa una diferencia en términos de unidades de error estándar (desviación). El paso 4 consiste en la definición

[2] Esta valiosa información fue suministrada por uno de los revisores anónimos de este libro de texto.

de una regla de decisión, la cual provee los lineamientos para evaluar el estadístico de la prueba. La probabilidad de un error tipo I, es decir α, está considerada en la determinación del valor crítico que se utiliza en la regla. Encontrar el valor crítico también implica la determinación (cálculo) de los grados de libertad y el uso de una tabla de valores críticos. La regla de decisión indica si debe o no rechazarse la hipótesis nula. El paso 5 da el salto de la inferencia: de la decisión tomada en el paso 4, regresa al problema en cuestión. Relaciona los resultados de la prueba estadística con la hipótesis sustantiva. La tabla 12.4 muestra un resumen de estos cinco pasos.

Determinación del tamaño de la muestra

Al iniciar un estudio surge la pregunta respecto a qué tan grande debe ser la muestra que se obtendrá. Esta pregunta resulta importante porque el interés radica en conseguir la mejor información al menor costo. Para aquellos investigadores que llevan a cabo grandes investigaciones, donde el costo de la recolección de datos es alto, la determinación del tamaño de la muestra resulta crítica. Cuando un investigador solicita financiamiento para el estudio, la determinación del tamaño de la muestra como parte de la propuesta de investigación es importante porque informa cuál será el costo del proceso de recolección de datos, en términos de tiempo y esfuerzo. Un tamaño de muestra demasiado grande representa un desperdicio de recursos; un tamaño de muestra demasiado pequeño es también un desperdicio de esfuerzo, pues no será lo suficientemente grande para detectar un efecto (diferencia) significativo. La forma en que se extraen las muestras y su tamaño determinan la cantidad total de información relevante contenida en una muestra. En el capítulo 8 se analizaron muchos procedimientos de muestreo. Aquí, después de la introducción de algunos estadísticos y en particular, del error estándar, se verá cómo se determinan los tamaños de las muestras. Con un poco de manipulación algebraica e información adicional, el error estándar posibilita la determinación del tamaño de la muestra.

Al incrementar el tamaño de la muestra, la distribución muestral se vuelve más estrecha y el error estándar se vuelve más pequeño. Como consecuencia, una muestra grande incrementa la probabilidad de detectar una diferencia. Sin embargo, una muestra demasiado grande hará que una diferencia muy pequeña resulte estadísticamente significativa, sin tener necesariamente una significancia práctica. Aunque se tratará de simplificar los conceptos y procedimientos implicados, el proceso de determinación del tamaño de mues-

TABLA 12.4 *Resumen de los cinco pasos de la comprobación de hipótesis*

Pasos para la comprobación de hipótesis	Notas
1. Establecer la hipótesis nula	H_0: $\mu_1 = \mu_2$ (note que la hipótesis nula contiene el signo =).
2. Establecer la hipótesis alternativa	H_1: $\mu_1 \neq \mu_2 (\mu_1 > \mu_2$ o $\mu_1 < \mu_2)$.
3. Calcular de los estadísticos de la prueba	Los estadísticos pueden ser z, t, F, χ^2 Calculados de datos observados.
4. Regla de decisión	Use α, gl y la tabla para determinar el valor crítico.
5. Relacionar la decisión con el problema original	Ésta es la parte inferencial.

tras para estudios de investigación no resulta trivial ni sencillo. De hecho Williams (1978) afirma que es uno de los problemas más difíciles en la estadística aplicada. La respuesta dada por estos métodos no es completamente precisa y sólo debe utilizarse como una guía para ayudar a tomar decisiones inteligentes acerca de la conducta del estudio. Aun así, dicho uso implica una mejoría respecto a otros métodos con reglas intuitivas, que los científicos utilizan sin justificación. Una de estas reglas es la decisión de seleccionar n número de participantes con base en una proporción del tamaño de la población. Aunque el segundo autor de este libro (HBL) ha oído sobre dichas reglas, no las ha encontrado escritas y con justificación en ningún lado.

Primero es necesario introducir cómo se determina el tamaño de las muestras para muestras aleatorias simples. Aquí el investigador debe conocer el valor real de la desviación estándar poblacional σ, o un estimado de él; los estimados provienen de datos o estudios previos. No obstante, si no están disponibles, el investigador puede usar el rango, lo cual requerirá un estimado del valor más grande y del valor más pequeño en las mediciones. Mendenhall y Beaver (1994) recomiendan dividir el rango entre 4 para obtener un estimado de σ. Williams (1978) recomienda dividir el rango entre 6. Segundo, el investigador necesita especificar el nivel de precisión (qué tan cercana está la media muestral de la media poblacional [verdadera]). Algunos se refieren a esto como la cantidad de error que el investigador está dispuesto a tolerar, entre la media muestral y la media verdadera. El tercer ingrediente es la cantidad de riesgo (en términos de probabilidad) o certeza que es aceptable para el investigador, lo cual se conoce tradicionalmente como la probabilidad del error tipo I, α.

La fórmula para calcular el tamaño de la muestra para una muestra aleatoria simple es:

$$n = \frac{Z^2\sigma^2}{d^2} \tag{12.3}$$

donde

Z^2 = puntuación estándar correspondiente a la probabilidad de riesgo especificada. Si el riesgo es 0.10 (es decir, $\alpha = 0.10$), $Z = 1.645$. Para un riesgo de 0.05, $Z = 1.96$, y para 0.01 la Z es 2.575.
σ = la desviación estándar de la población.
d = desviación especificada.

Ésta es la precisión deseada de la media muestral. ¿Qué tan cercana debe estar la media muestral a la media verdadera?

Ejemplo

Una investigadora está diseñando un estudio respecto a los estudiantes universitarios. Ella seleccionará dos grupos de estudiantes y desea determinar el número apropiado de estudiantes que debe muestrear para el estudio. La variable dependiente en este estudio es el promedio de las calificaciones de los estudiantes. Ella siente que puede tolerar 0.2 desviaciones entre la media muestral y la media verdadera: está dispuesta a tomar un riesgo de 0.05. Investigaciones previas que han utilizado el promedio de las calificaciones han reportado una desviación estándar de aproximadamente 0.6.

Para un riesgo con probabilidad de 0.05, el valor Z correspondiente es 1.96. La desviación estándar es 0.6 y la desviación es 0.2. Utilizando la fórmula dada anteriormente, se estima que el tamaño de muestra requerido es:

$$n = \frac{1.96^2(0.6^2)}{.2^2} = \frac{3.842(.36)}{.04} = \frac{1.383}{.04} \approx 34.6 \approx 35$$

Esto es, 35 sujetos por grupo. Por lo tanto, la investigadora necesitará 70 sujetos.

Si el muestreo proviene de una población finita de tamaño N, y el muestreo se realiza sin remplazamiento, Williams (1978) sugiere el siguiente ajuste a la fórmula expresada con anterioridad:

$$n' = \frac{n}{1 + n/N}$$

n' es el tamaño estimado de la muestra, n es el tamaño estimado de la muestra al utilizar la fórmula 12.3 y N es el tamaño de la población. Utilizando el ejemplo anterior, si se hubiera determinado que el tamaño de la población era $N = 1\,000$, entonces n' sería:

$$n' = \frac{70}{1 + 70/1\,000} = 65.421 \approx 66, \text{ o } 33 \text{ en cada grupo}$$

Este método requiere conocer tan sólo la desviación estándar de las poblaciones o su estimado y α, la probabilidad de un error tipo I. Guilford y Fruchter (1978) presentan un método que también utiliza β, la probabilidad del error tipo II. Al especificar β, como se mencionó antes, también se especifica el poder de la prueba estadística mediante $1 - \beta$. Los investigadores que desean protegerse de α y β pueden usar la fórmula de Guilford y Fruchter para encontrar un tamaño de muestra que les brinde el riesgo deseado.

La fórmula es:

$$n = \frac{(Z_\beta - Z_\alpha)^2 \sigma^2}{d^2}$$

donde α = la desviación estándar de la población.

d = desviación especificada. Ésta es la precisión deseada de la media muestral, es decir, ¿qué tan cerca debe estar la media muestral de la media verdadera?

Z_α = distancia del valor crítico a la media en H_0 (en unidades de desviación estándar, con el signo apropiado).

Z_β = distancia del valor crítico a la media en H_1 (en unidades de desviación estándar, con el signo apropiado).

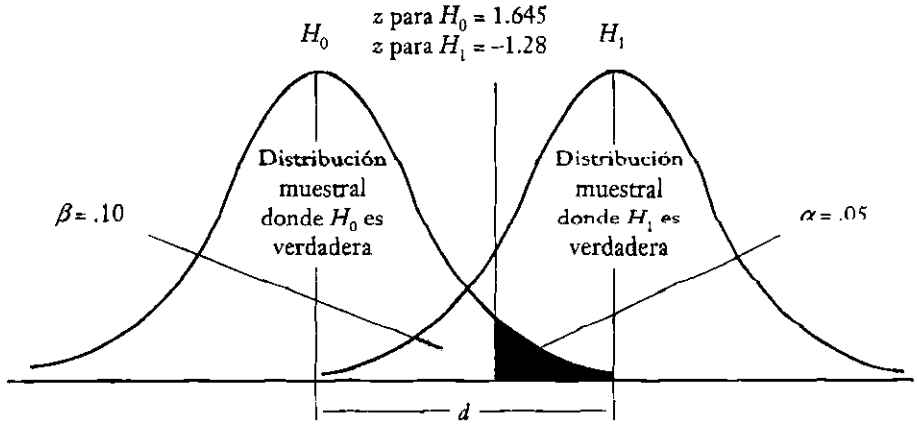
Para demostrar cómo funciona esta fórmula, es necesario referirse a la figura 12.3, que muestra la relación entre α y β. El tamaño de la muestra puede determinarse especificando tanto α como β, junto con la desviación estándar. Suponiendo que la desviación estándar se midió con precisión, el número de puntos de datos recolectado en un estudio de investigación cumpliría con la especificación establecida por los niveles α y β. Con valores específicos de α y β, las dos distribuciones muestrales pueden mostrarse, de tal manera que pueda encontrarse el valor crítico apropiado. Por ejemplo, si para el estudio se establece que $\alpha = 0.05$ y $\beta = 0.10$, los valores Z que satisfacen esto serían -1.28 para la distribución H_1 y 1.645 para la distribución H_0. La figura 12.4 muestra que esta $Z = -1.28$ sería el valor que cortaría $\beta = 0.10$ en la distribución "H_1 es verdadera". Para ese mismo punto marcado como "valor crítico" en la figura 12.4, correspondería a $Z = 1.645$ en la distribución "H_0 es verdadera".

Si se utilizan los datos del ejemplo anterior, el tamaño de la muestra estimado sería:

$$n = \frac{(-1.28 - 1.645)^2(0.6^2)}{.2^2} = \frac{18.656(.36)}{.04} = \frac{6.716}{.04} = 167.9 = 168$$

Esto es, 168 participantes por grupo.

FIGURA 12.4



El procedimiento descrito antes aplica para la prueba de una cola. Para la prueba de dos colas, sólo cambiará la Z_α Si la prueba es de dos colas, entonces en vez de utilizar toda la α en una sola cola, se usaría $\alpha/2$ en su lugar. Para el ejemplo anterior, el valor Z adecuado sería 1.96.

El empleo, esencialmente, de los mismos datos para los dos ejemplos resultó con valores diferentes, ¿por qué? Recuerde que la probabilidad de un error tipo II es no rechazar la hipótesis nula cuando existe una diferencia verdadera. En el ejemplo, sería necesario usar 35 sujetos por grupo para rechazar la hipótesis nula; así no preocuparía la posibilidad de perder ninguna oportunidad. Si se maneja este ejemplo con la fórmula de Guilford y Fruchter, entonces se considera β, la probabilidad del error tipo II, convirtiéndose en una prueba más sensible para detectar una diferencia verdadera. Por lo tanto, con una $n = 168$, los investigadores no solamente tendrían suficientes sujetos para rechazar H_0 a un nivel $\alpha = 0.05$, sino que también serían suficientes para darles un poder $(1 - \beta)$ de 0.90.

RESUMEN DEL CAPÍTULO

1. El error estándar es la desviación estándar de la distribución muestral de los estadísticos de la muestra.
2. Los errores estándar sirven para evaluar
   a) diferencias entre medias.
   b) diferencias entre la correlación de la muestra y cero.
3. Diferencias pequeñas pueden resultar estadísticamente significativas si el error estándar es proporcionalmente más pequeño.
4. Los errores estándar sirven como un instrumento de medición contra el que se examina la varianza experimental.
5. Monte Carlo es un método utilizado para crear datos simulados, para numerosas situaciones donde la recolección de datos puede resultar costosa o no factible.
6. El método Monte Carlo puede ser utilizado para demostrar el comportamiento y significado del error estándar.

7. El teorema del límite central es uno de los teoremas más importantes en la estadística.
8. Con el teorema del límite central, la distribución muestral de las medias muestrales es aproximadamente normal en su forma, aunque la distribución de la cual se extrajeron las muestras no fuera normal.
9. Una hipótesis sustantiva consiste en un enunciado conjetural de la relación entre dos variables.
10. Las hipótesis estadísticas son un nuevo planteamiento de hipótesis sustantivas, en términos estadísticos.
11. Las pruebas de hipótesis involucran a las hipótesis nula y estadística.
12. Existen cinco pasos básicos para la comprobación de hipótesis.
13. El error estándar es una parte importante en la determinación del tamaño de la muestra.

SUGERENCIAS DE ESTUDIO

1. Por fortuna abundan las buenas referencias en el tema de la estadística. Los libros que se mencionan a continuación pueden ser de utilidad. Escoja uno o dos para complementar su estudio. Al consultar un libro sobre estadística, no se desanime si no comprende cabalmente todo lo que lee. De hecho, algunas veces se sentirá desconcertado por completo. Conforme adquiera entendimiento del lenguaje y métodos de la estadística, la mayoría de las dificultades desaparecerán.

Comrey, A. L. y Lee, H. B. (1995). *Elementary statistics: A problem-solving approach* (3a. ed.). Dubuque, Iowa: Kendall-Hunt. Es un buen libro para el estudiante principiante. Los temas están organizados en la forma de 50 problemas.

Freedman, D., Pisani, R. y Purves, R. (1997). *Statistics* (3a. ed.). Nueva York: Norton. Accesible para el estudiante principiante. Excelentes análisis de estudios y problemas interesantes. Orientado a las aplicaciones. Evita el uso de símbolos y de la notación estadística.

Glass, G. y Hopkins (1996). *Statistical methods in education and psychology* (3a. ed.). Boston: Allyn & Bacon. Un libro bien escrito, con un buen tratamiento de conceptos difíciles. Ofrece una interesante demostración por computadora del teorema del límite central.

Hays, W. L. (1994). *Statistics* (5a. ed.). Fort Worth, Texas: Harcourt Brace. Excelente libro, exhaustivo, una autoridad en la materia, orientado hacia la investigación, pero no resulta elemental. Su cuidadoso estudio debería constituir una meta de estudiantes e investigadores serios.

Kirk, R. E. (1990). *Statistics: An introduction* (3a. ed.). Fort Worth, Texas: Holt, Rinehart y Winston. Un tratamiento de la estadística bien escrito e informativo; una buena referencia para principiantes.

Mattson, D. E. (1984). *Statistics: Difficult concepts, understandable explanations.* Oak Park, Illinois: Bolchazy-Carducci Publishers. Cada capítulo está dividido en lecciones. Da un buen tratamiento a datos sobre salud pública.

Natrella, M. G. (1966). *Experimental Statistics. National Bureau of Standards Handbook 91.* Washington, DC: U.S. Government Printing Office. Un libro antiguo pero bien presentado, producido por el gobierno de Estados Unidos. Contiene una serie de tablas que resultan útiles para estimar tamaños de muestras para diferentes pruebas estadísticas.

Snedecor, G., Cochran, W. y Cox, D. R. (1989). *Statistical method* (8a. ed.). Ames: Iowa State University Press. Sólido, una autoridad en la materia, útil, pero no es elemental. Excelente libro de referencia.

2. Las proporciones de hombres y mujeres votantes en cierto condado son 0.70 y 0.30, respectivamente. En un distrito electoral de 400 personas, hay 300 hombres y 100 mujeres. ¿Podría decirse que la proporción distrital de hombres y mujeres votantes difiere significativamente de la del condado?
   [Respuesta: Sí. $\chi^2 = 4.76$. El valor de entrada en la tabla de χ^2, al nivel 0.05, para $gl = 1$, es 3.84.]
3. Un investigador en el área del prejuicio experimentó con varios métodos sobre cómo responder a los comentarios de las personas con prejuicios acerca de los miembros de grupos minoritarios. El investigador asignó aleatoriamente a 32 sujetos a dos grupos, 16 en cada grupo. Con el primer grupo se utilizó el método A; y con el segundo, el método B. Las medias de ambos grupos en una prueba de actitud, administrada después de aplicar los métodos, fueron A: 27 y B: 25. Ambos grupos tuvieron una desviación estándar de 4. ¿Difieren significativamente las dos medias de los grupos?
   [Respuesta: No. (27 – 25)/1.414 = 1.414.]
4. Los 4 000 números aleatorios distribuidos uniformemente discutidos en el texto y los estadísticos calculados a partir de los números aleatorios se presentan en el apéndice C al final del libro. Utilice una tabla de números aleatorios —los 4 000 números aleatorios servirán— y mueva un lápiz en el aire con los ojos cerrados, para después bajarlo en cualquier punto de la tabla. Descienda por las columnas a partir del punto señalado por el lápiz y copie 10 números dentro del rango de 1 a 40. Permita que éstos sean los números de 10 de los 40 grupos. Las medias, varianzas y desviaciones estándar se proporcionan inmediatamente después de la tabla de los 4 000 números aleatorios. Copie las medias de los grupos seleccionados aleatoriamente. *Redondee las medias; esto es, 54.33 se convierte en 54, 47.87 se convierte en 48,* etcétera.
   a) Calcule la media de las medias y compárela con la media poblacional de 50 (50.33, en realidad). ¿Se aproximó?
   b) Calcule la desviación estándar de las 10 medias.
   c) Tome el primer grupo seleccionado y calcule el error estándar de la media, usando $N = 100$ y la desviación estándar reportada. Haga lo mismo para el cuarto y quinto grupos. ¿Los EE_M son similares? Interprete el primer EE_M. Compare los resultados de los incisos ***b***) y ***c***).
   d) Calcule las diferencias entre la primera y la sexta medias, y entre la cuarta y décima medias. Pruebe la significancia estadística de ambas diferencias. ¿Deberían ser estadísticamente significativas? Justifique su respuesta. Diseñe una situación experimental e imagine que la cuarta y décima medias son sus resultados. Interprete.
   e) Discuta el teorema del límite central en relación al inciso ***d***).
5. Hasta ahora, la varianza y la desviación estándar han sido calculadas con N en el denominador. En los libros de estadística, el estudiante encontrará la fórmula de la varianza como: $V = \Sigma\chi^2/N$, o $V = \Sigma\chi^2/(N-1)$. La primera fórmula se utiliza cuando sólo se describe una muestra o población. La segunda se utiliza cuando se estima la varianza de una población a partir de la varianza de la muestra (o desviación estándar). Con una N grande existe una diferencia práctica mínima. En capítulos posteriores se verá que los denominadores de los estimados de la varianza siempre tienen

$N - 1$, $k - 1$, etcétera. Éstos en realidad son grados de libertad. La mayoría de los programas computacionales utilizan $N - 1$ para calcular desviaciones estándar. Quizás el mejor consejo sea utilizar siempre $N - 1$; aun cuando no sea apropiado, no provocará mucha diferencia.

6. Los estadísticos no siempre son bien vistos. A los marxistas, por ejemplo, no les agradan (¿por qué supone usted que así es?). En un interesante estudio sobre educación se utilizó un diseño con un grupo control; sin embargo, no se emplearon pruebas estadísticas de significancia ni medidas de la magnitud de relaciones: véase DeCorte y Verschaffel (1981). El estudiante encontrará interesante la lectura de este estudio.
7. En educación se ha discutido mucho respecto a las supuestas virtudes de un ambiente educativo "abierto". En un estudio de Wright (1975) sobre la diferencia entre ambientes escolares "abiertos" y "tradicionales", se reportan varias diferencias de medias interesantes; entre estas diferencias de medias, aquellas del significado de palabras y creatividad verbal (p. 453) resultaron como sigue:

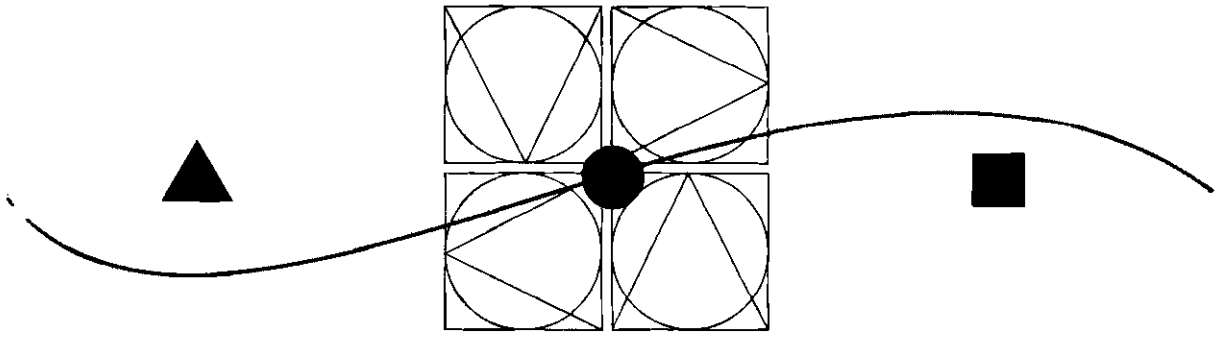
	Significado de palabras		*Creatividad verbal*	
	Tradicional	*Abierto*	*Tradicional*	*Abierto*
N	50	50	50	50
M	4.84	4.35	135.38	129.60
DE	1.19	.78	23.5	19.2

Calcule las dos razones t e interprete los resultados. (Utilice la ecuación 12.1 y sustituya en la ecuación 12.2.)
[Respuestas: significado de palabras: $t = 2.43$ (p, .05); creatividad verbal: $t = 1.35$ (n.s.).]

8. Revise el estudio de Scattone y Saetermoe (1997). Note que los autores realizaron una prueba t de las medias. Sin embargo, estas medias reflejan la variable independiente. En general el análisis de datos con pruebas t y otros estadísticos similares se realiza con medidas de la variable dependiente. ¿Se equivocaron los autores? Si es así, ¿por qué? ¿Podría no ser significativa una prueba t de la variable dependiente o las medidas de "discapacidad"? De ser así, ¿qué pasa con la hipótesis de los autores? (Aquí se ignoran otros posibles tipos de análisis.)
[Consejo: ¿qué se predice en problemas de este tipo? Piense en las hipótesis como enunciados del tipo "si p, entonces q".]
9. Se le pide a una investigadora que realice un estudio sobre las puntuaciones de una prueba de inteligencia. Un distrito escolar específico afirma que los estudiantes presentan una puntuación promedio de 90 en la prueba. La investigadora necesita obtener una muestra de tamaño n que sea suficientemente grande para obtener una media muestral que no difiera de 90 por más de 2 puntos, con un 99% de confianza. El distrito también reporta una desviación estándar de 10.2. ¿Qué tan grande debería ser n?
10. Utilizando los datos del problema 9, si se sabe que el distrito tiene 1 500 estudiantes, ¿qué tan grande debería ser n?

Parte Cinco

Análisis de varianza

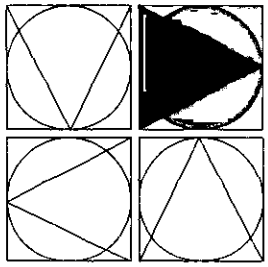


Capítulo 13
Análisis de varianza: fundamentos

Capítulo 14
Análisis factorial de varianza

Capítulo 15
Análisis de varianza: grupos correlacionados

Capítulo 16
Análisis de varianza no paramétricos y estadísticos relacionados



CAPÍTULO 13

ANÁLISIS DE VARIANZA: FUNDAMENTOS

El análisis de varianza no constituye simplemente un método estadístico, es un enfoque y una forma de pensamiento. También es una de las muchas expresiones de lo que se conoce como el modelo lineal general. Este modelo es, en realidad, una ecuación lincal (*lineal* significa que ningún término de la ecuación tiene poderes mayores a 1) que expresa las fuentes de la varianza de un conjunto de medidas. De forma adecuada para el análisis de varianza, podría escribirse de la siguiente manera:

$$y = b_0x_0 + b_1x_1 + b_2x_2 \ldots + b_kx_k + e$$

(Observe que ninguna de las *x*s tiene potencia mayor que 1, es decir, que no hay x^2 o x^3.) Si se concibe que la puntuación de un individuo, y, tiene una o más fuentes de varianza, x_1,

x_2,..., entonces se capta, de manera general, la idea del modelo. Las b son pesos que expresan los grados relativos de influencia de las x que explican a y. La letra e representa el error; expresa los factores desconocidos que ejercen influencia sobre y, junto con el omnipresente error aleatorio. La ecuación es general: se ajusta a la mayoría de las situaciones analíticas, en las que se desea explicar la variación de un conjunto de medidas de una variable dependiente, y. Para los modelos de análisis de varianza, la ecuación se simplifica a una de varias formas específicas, que no se requiere examinar por ahora. El punto es que las medidas de la variable dependiente se conciben como constituidas por dos o más componentes, y el papel del análisis de varianza consiste en determinar las contribuciones relativas de estos componentes a la variación de la variable dependiente. Como se verá hacia el final del libro, ésta es una de las metas de la regresión múltiple, así como de otros métodos analíticos. Resulta necesario intentar hacer estas abstracciones más concretas y comprensibles. Sin embargo, por ahora, sólo se requiere tener en mente lo siguiente: la varianza total de la variable dependiente de cualquier situación estadística se descompone en las fuentes que originan o componen la varianza.

En el presente capítulo y en los capítulos 14 y 15 se explora el análisis de varianza. Se pondrá énfasis en algunas nociones fundamentales y generales que subyacen al método. El propósito de los capítulos no es tan sólo enseñar el análisis de varianza y los métodos relacionados como estadísticos, su intención es transmitir las ideas básicas de los métodos en relación con la investigación y los problemas de investigación. Para lograr este propósito pedagógico, se utilizarán ejemplos simples. No hay mucha diferencia si se emplean 5 o 500 puntuaciones, o 2 o 20 variables; las ideas fundamentales y las concepciones teóricas son las mismas. En este capítulo se estudia el *análisis de varianza de un factor*. En los próximos dos capítulos se consideran el análisis de varianza factorial y el análisis de varianza de grupos o sujetos correlacionados. Para entonces, el estudiante deberá tener una buena base sobre diseño de investigación.

Descomposición de la varianza: un ejemplo simple

En el capítulo 6 se analizaron dos conjuntos de puntuaciones en la modalidad de varianza. La *varianza total* de todas las puntuaciones se descompuso en una *varianza entre los grupos* y una *varianza dentro de los grupos*. Ahora se retomará el tema del capítulo 6 utilizando, con modificaciones, el ejemplo de los dos grupos que ahí se explicó, corrigiendo también el método para su cálculo. Después se extienden considerablemente las ideas respecto al análisis de varianza.

Suponga que un investigador está interesado en la eficacia relativa de los métodos A_1 y A_2. Se utiliza el término *métodos* aquí y en otras partes ya que la palabra es general y fácil de entender. Los estudiantes pueden asir la sustancia de diferentes métodos de su propio campo. Por ejemplo, en educación podrían ser métodos de enseñanza; en psicología, métodos de reforzamiento o de activación de la atención; en ciencias políticas, métodos de participación en procesos políticos, etcétera. El investigador toma a diez estudiantes como la muestra, y los divide aleatoriamente en dos grupos. Cada uno de los grupos se asigna al azar a los tratamientos experimentales. Después de un tiempo razonable, mide el aprendizaje de los alumnos en ambos grupos, utilizando una prueba de rendimiento. Los resultados, junto con ciertos cálculos, se presentan en la tabla 13.1.

Nuestro trabajo consiste en localizar y calcular las diferentes varianzas que conforman la varianza total. La varianza total y las otras varianzas se calculan como se hizo antes, pero con una importante diferencia: en lugar de utilizar N o n en el denominador de las fracciones de varianza, se usan los grados de libertad, los cuales generalmente se definen

TABLA 13.1 *Dos conjuntos de datos experimentales hipotéticos con sumas, medias y sumas de cuadrados*

	A_1	x	x^2	A_2	x	x^2	
	4	0	0	3	0	0	
	5	1	1	1	−2	4	
	3	−1	1	5	2	4	
	2	−2	4	2	−1	1	
	6	2	4	4	1	1	
ΣX	20			15			$\Sigma X_t = 35$
M	4			3			$M_t = 3.5$
Σx^2			10			10	

como un caso menos que N o n; es decir, $N - 1$ y $n - 1$. En el caso de grupos, en lugar de k (el número de grupos), se utiliza $k - 1$. Mientras que este método representa una gran ventaja desde el punto de vista estadístico, desde una perspectiva conceptual-matemática, hace el trabajo un poco más difícil. Primero realizamos los cálculos, y después regresamos a la dificultad que esto supone.

Para calcular la varianza total, se utiliza la siguiente fórmula:

$$V_t = \frac{\Sigma x^2}{N - 1} \tag{13.1}$$

donde Σx^2 es igual a la suma de cuadrados, como antes, $x = X - M$, o la desviación de cualquier puntuación con respecto a la media; y N es igual al número de casos en la muestra total. Para calcular V_t simplemente se toman todas las puntuaciones, sin importar su agrupación, y se calculan los términos necesarios de la ecuación 13.1, como en la tabla 13.2. Puesto que $N - 1 = 10 - 1 = 9$, $V_t = 22.50/9 = 2.5$. Por lo tanto, si se acomodan los datos de la tabla 13.1 pasando por alto el hecho de que pertenezcan a un grupo o al otro, $V_t = 2.5$.

Existe la varianza entre los grupos, la cual se debe, presumiblemente, a la manipulación experimental; esto quiere decir que el experimentador hizo algo a un grupo y otra cosa diferente al otro grupo. Estos distintos tratamientos deberían hacer diferentes a los grupos y a sus medias, lo que provocaría una *varianza entre grupos*. Considere las dos medias como cualquier otra puntuación (X) y calcule su varianza (véase tabla 13.3).

Aún existe otra fuente de varianza: el siempre presente error aleatorio. En el capítulo 6 se aprendió que ésta puede obtenerse calculando la varianza *dentro* de cada grupo de forma separada, y después promediar estas varianzas separadas. Hacemos esto utilizando las cantidades dadas en la tabla 13.1. Cada grupo tiene $\Sigma x^2 = 10$. Al dividir cada una de estas sumas de cuadrados entre sus grados de libertad, se obtiene:

$$\frac{\Sigma x_{A1}^2}{n_{A1} - 1} = \frac{10}{4} = 2.5$$

y

$$\frac{\Sigma x_{A2}^2}{n_{A2} - 1} = \frac{10}{4} = 2.5$$

El promedio resulta, por supuesto, 2.5. Por lo tanto, la varianza *dentro de los grupos*, V_d, es 2.5. Ya se han calculado tres varianzas: $V_t = 2.5$, $V_e = 0.50$, $V_d = 2.5$. La ecuación teórica

TABLA 13.2 *Cálculo de V_t a partir de los datos de la tabla 13.1*

	X	x	x^2
	4	.5	.25
	5	1.5	2.25
	3	−.5	.25
	2	−1.5	2.25
	6	2.5	6.25
	3	−.5	.25
	1	−2.5	6.25
	5	1.5	2.25
	2	−1.5	2.25
	4	.5	.25
Σx	35		
M	3.5		
Σx^2			22.50

presentada en el capítulo 6 indica que la varianza total se compone de fuentes separadas de varianza: la varianza entre los grupos y la varianza dentro de los grupos; lógicamente ambas deben sumar la varianza total. La ecuación teórica es la siguiente:

$$V_t = V_e + V_d \tag{13.2}$$

Puesto que 2.5 no es igual a 0.50 y 2.5, algo debe estar mal. El problema es que los grados de libertad se utilizaron en los denominadores de la fórmula de la varianza, en lugar de N, n y k. Si se hubieran utilizado N, n y k, las relaciones de la ecuación 13.2 se habrían mantenido (véase capítulo 6). Si N, n y k se hubieran utilizado, los valores habrían sido $V_t = 2.25$, $V_e = 0.25$ y $V_d = 2$.

El estudiante podría preguntar: ¿por qué no seguir el procedimiento con N, n y k? Y si no se puede seguir, ¿para qué molestarse en hacer todo esto? La respuesta es que el cálculo de las varianzas con N, n y k resulta matemáticamente correcto, pero estadísticamente "insatisfactorio". Otro aspecto importante del análisis de varianza es el estimado de los valores de la población. Es posible mostrar que el uso de los grados de libertad en el denominador de la fórmula de la varianza produce estimados no sesgados de los valores de la población, un aspecto de gran importancia estadística. El valor de lidiar con el presente procedimiento es para mostrar claramente al lector la base matemática de este razonamiento. Sin embargo, debe recordarse que las varianzas, como se utilizan en el análisis de varianza, no son necesariamente aditivas.

TABLA 13.3 *Cálculo de V_e a partir de los datos de la tabla 13.1*

	X	x	x^2
	4	.5	.25
	3	.5	.25
ΣX	7		
M	3.5		
Σx^2			.50

$$V_e = \frac{\Sigma x_b^2}{k-1} = \frac{.50}{2-1} = .50$$

Por otro lado, las sumas de cuadrados son siempre aditivas (se calculan a partir de las puntuaciones, y no se dividen entre ningún otro valor) y son también medidas de variabilidad. Las sumas de cuadrados se calculan, estudian y analizan excepto en la etapa final del análisis de varianza. Para convencerse sobre la propiedad aditiva de las sumas de cuadrados, note que la suma de las sumas de cuadrados entre y dentro de los grupos dan como resultado la suma de cuadrados total. Si se multiplica la suma de cuadrados entre grupos por el número de casos en cada grupo, es decir n:

$$\Sigma x_t^2 = n\Sigma x_c^2 + \Sigma x_d^2$$

O numéricamente, $22.50 = (5)(0.50) + 20$.

El razonamiento que subyace a la expresión $n\Sigma\chi_c^2$ en esta ecuación es el siguiente: la definición de un estimado sin sesgo de la varianza de la población de medias es $V_M = \Sigma x^2/(n-1)$. Pero a partir de nuestro razonamiento respecto al error estándar y a la varianza estándar, se sabe que $V_M = VE_M = V/n$. Sustituyendo en la primera ecuación, se obtiene $V/n = \Sigma x^2/(k-1)$, por lo tanto $V = n\Sigma x^2/(k-1)$. Debe notarse aquí que la expresión $n\Sigma x_c^2$, indicada en el capítulo 6, en realidad es la suma de cuadrados *entre*, y no Σx_c^2, como se indicó en ese capítulo y otros subsecuentes. Es decir, en lugar de escribir Σx_c^2, los estadísticos escriben sc_c, que en realidad es $n\Sigma x_c^2$.

El enfoque de la razón *t*

Con los datos de la tabla 13.1, calculamos varios estadísticos para los datos de A_1 y A_2 de manera separada: las varianzas, las desviaciones estándar, los errores estándar de las medias y las varianzas estándar de las medias. Los métodos de análisis utilizados en la primera parte de este capítulo no se utilizan en los cálculos reales, a causa de que son demasiado engorrosos; se presentan aquí sólo por razones pedagógicas. Por desgracia, el método de cálculo acostumbrado tiende a oscurecer las relaciones y operaciones importantes que subyacen al análisis de varianza. Estos cálculos se presentan en la tabla 13.4 (note que ahora V se calcula con $n-1$ en lugar de n).

Ahora considere la idea estadística central detrás del análisis de varianza. La pregunta que el investigador se hace es: ¿Las medidas difieren entre sí significativamente? Resulta obvio que 4 no es igual a 3; pero la pregunta debe hacerse estadísticamente. Se sabe que si se extraen conjuntos de números aleatorios, las medias de los conjuntos no serán iguales; sin embargo, no deberían ser demasiado diferentes, sino sólo dentro de los márgenes de las fluctuaciones debidas al azar. De esta manera, la pregunta se convierte en: ¿4 difiere

TABLA 13.4 *Diferentes estadísticos calculados a partir de los datos de la tabla 13.1*

	A_1	A_2
V:	$\frac{\Sigma x^2}{n-1} = \frac{10}{4} = 2.5$	$\frac{10}{4} = 2.5$
DE:	$\sqrt{2.5} = 1.58$	$\sqrt{2.5} = 1.58$
EE_M:	$\frac{DE}{\sqrt{n}} = \frac{1.58}{\sqrt{5}} = .705$	$\frac{1.58}{\sqrt{5}} = .705$
VE_M:	$\frac{V}{n} = \frac{2.5}{5} = .50$	$\frac{2.5}{5} = .50$

significativamente de 3? De nuevo, se establece la hipótesis nula: H_0: $\mu_{A1} - \mu_{A2} = 0$, o $\mu_{A1} = \mu_{A2}$. La hipótesis sustantiva era: H_1: $\mu_{A1} > \mu_{A2}$. ¿A cuál hipótesis apoya la evidencia? En otras palabras, no se trata simplemente de preguntar si 4 es absolutamente mayor que 3, sino más bien de preguntar si el 4 difiere del 3 más allá de las diferencias esperadas por el azar.

Esta pregunta puede ser rápidamente contestada utilizando los métodos del capítulo anterior. Primero se calcula el error estándar de las diferencias entre las medias:

$$EE_{M_{A1} - M_{A2}} = \sqrt{EE^2_{M_{A1}} + EE^2_{M_{A2}}}$$

$$= \sqrt{(.705)^2 + (.705)^2} = \sqrt{.994} = .997 = 1.00 \text{ (redondeado)}$$

Ahora, la razón *t*:

$$t = \frac{M_{A1} - M_{A2}}{EE_{M_{A1} - M_{A2}}} = \frac{4 - 3}{1.00} = \frac{1}{1} = 1$$

Puesto que la diferencia que se evalúa no es mayor que la medida del error, resulta obvio que no es significativa. El numerador y el denominador de la razón *t* son iguales. La diferencia 4 – 3 = 1 constituye claramente una de las diferencias que pudieron haber ocurrido con números aleatorios. Recuerde que una diferencia "real" se refleja en la razón *t* por un numerador considerablemente mayor que el denominador.

El enfoque del análisis de varianza

En el análisis de varianza el enfoque es conceptualmente similar, aunque el método difiere. El método es general: se pueden probar las diferencias entre más de dos grupos con respecto a su significancia estadística; mientras que la prueba *t* únicamente aplica a dos grupos (con dos grupos, como se verá en breve, los resultados de los dos métodos son realmente idénticos). El método del análisis de varianza usa enteramente varianzas, en lugar de usar las diferencias y errores estándar, aunque el razonamiento sobre las diferencias y el error estándar subyace al método. Dos varianzas se confrontan siempre una contra otra. Una varianza, presumiblemente debida a la variable o variables experimentales (independientes) se confronta contra otra varianza, la debida probablemente al error o al azar. Para comprender esta idea es necesario regresar al problema.

Encontramos que la varianza entre los grupos fue 0.50. Ahora debemos encontrar una varianza que refleje el error: la varianza dentro de los grupos. Después de todo, ya que calculamos la varianza dentro de los grupos, al calcular la varianza de cada grupo en forma separada y luego promediando las dos (o más) varianzas, este estimado del error no se ve afectado por las diferencias entre las medias. Por lo tanto, *si ninguna otra cuestión está causando la variación en las puntuaciones*, es razonable considerar a la varianza dentro de los grupos como una medida de la fluctuación aleatoria; si esto es así, entonces se puede comparar *la varianza debida al efecto experimental*, es decir, *la varianza entre los grupos, contra esta medida del error aleatorio: la varianza dentro de los grupos*. La única pregunta sería: ¿Cómo se calcula la varianza dentro de los grupos?

Recuerde que la varianza de una población de medias puede estimarse con la varianza estándar de la media (el error estándar elevado al cuadrado). Una manera de obtener la varianza dentro de los grupos es calcular la varianza estándar de cada uno de los grupos y, después, promediarlas. Esto deberá producir un estimado del error que puede ser utilizado para evaluar la varianza de las medias de los grupos. El razonamiento aquí resulta

básico: para evaluar las diferencias entre las medias es necesario referirse a una población de medias teórica que se obtendría del muestreo aleatorio de grupos de puntuaciones, como los grupos de puntuaciones que aquí se tienen. En el presente caso se tienen dos medias muestrales con cinco puntuaciones en cada grupo. (Conviene recordar que se podrían tener tres, cuatro o más medias de tres, cuatro o más grupos; el razonamiento es el mismo.) Si los participantes fueran asignados aleatoriamente a los grupos y nada hubiera operado (es decir, que no existieron manipulaciones experimentales ni otras influencias sistemáticas), entonces es posible estimar la varianza de las medias de la población de medias a partir de la varianza estándar de las medias (EE_M^2 o VE_M). Cada grupo proporciona un estimado de este tipo. Dichos estimados variarán en cierto grado entre ellos, y pueden unirse haciendo un promedio para formar un estimado general de la varianza de las medias de la población.

Como se aprendió antes, la fórmula del error estándar de la media es: $EE_M = DE/\sqrt{n}$. Para obtener la varianza estándar de la media tan sólo se eleva al cuadrado esta expresión: $EE_M^2 = (DE)^2/n = VE_M = V/n$. La varianza de cada grupo fue 2.5. Al calcular las varianzas estándar para cada grupo se obtiene: $VE_M = V/n = 2.50/5 = 0.50$. Si se les promedia, obviamente resultará 0.50. Observe que cada varianza estándar fue calculada a partir de cada grupo *de manera separada y luego promediada*. Por lo tanto, esta varianza estándar promedio no se ve influenciada por las diferencias entre las medias, como se analizó anteriormente. La varianza estándar promedio es, entonces, una *varianza dentro de los grupos*; es un estimado de los errores aleatorios.

No obstante, si se hubieran utilizado números aleatorios, el mismo razonamiento aplicaría para la varianza entre grupos, la varianza calculada a partir de las medias en cuestión. Se calculó una varianza de las medias de 4 y 3: resultó 0.50. Si los números fueran aleatorios sería posible estimar la varianza de la población de medias, calculando la varianza de las medias obtenidas.

Sin embargo, note cuidadosamente que si operara cualquier influencia extraña, si existiera la influencia de algún efecto experimental, entonces la varianza calculada a partir de las medias obtenidas ya no sería un buen estimado de la varianza de la población de medias. Si en realidad hubiera operado cualquier influencia experimental (o cualquier influencia distinta al azar), el efecto podría ser el incremento de la varianza de las medias obtenidas. En cierto sentido, éste es el propósito de la manipulación experimental: incrementar la varianza entre las medias, para hacer que las medias sean diferentes entre sí. Esto es el punto esencial en el análisis de varianza. Si una manipulación experimental ha ejercido influencia, entonces debería manifestarse en las diferencias entre medias encima y más allá de las diferencias que surgen únicamente por el azar; y la varianza entre grupos debería mostrar esta influencia al hacerse más grande que lo esperado por el azar. Resulta claro que puede utilizarse la V_e, como una medida de la influencia experimental. También debe ser claro, como se mostró antes, que puede emplearse V_d como una medida de la variación aleatoria. Por lo tanto, ya casi llegamos al final de un viaje largo pero productivo: es factible evaluar la varianza entre grupos V_e por medio de la varianza dentro de grupos, V_d; dicho de otra manera, se puede sopesar la información experimental contra el error o el azar.

Se podría evaluar V_e al restarle V_d, sin embargo, en el análisis de varianza se divide la V_e entre la V_d. La razón así formada se denomina la razón F. Snedecor nombró a la razón F en honor a Ronald Fisher, el inventor del análisis de varianza. Snedecor fue quien desarrolló las tablas F utilizadas para evaluar las razones F. Primero se calcula la razón F a partir de los datos observados y luego se verifica el resultado contra un valor de la tabla de la razón F (la tabla de la razón F con las instrucciones para su uso, puede encontrarse en cualquier libro de texto sobre estadística). Si la razón F obtenida resulta tan grande o más grande

que la especificada en la tabla, entonces las diferencias reflejadas por la V_e son estadísticamente significativas. En tal caso, la hipótesis nula de no diferencia entre las medias se rechaza al nivel de significancia determinado. En este caso:

$$F = \frac{V_e}{V_d} = \frac{.50}{.50} = 1$$

Obviamente no se requiere la tabla de la razón F para saber que la razón F no es significativa. Evidentemente las dos medias de 4 y 3 no difieren entre sí de manera significativa; en otras palabras, de las muchas muestras aleatorias posibles de pares de grupos de cinco casos cada uno, este caso en particular podría ser fácilmente uno de ellos. Si la diferencia hubiese sido bastante mayor, lo suficientemente grande para equilibrar la balanza de la razón F, entonces la conclusión hubiera sido bastante diferente, como se verá a continuación. Note que la prueba t y el análisis de varianza producen el mismo resultado. Con dos grupos solamente, o un grado de libertad $(k - 1)$, $F = t^2$, o $t = \sqrt{F}$. Esta igualdad demuestra que en el caso de dos grupos, no importa si se calcula t o F. (Pero en la mayoría de los casos es más fácil calcular el análisis de varianza que la t.) Sin embargo, con tres o más grupos, no se cumple la igualdad y siempre debe calcularse F. Por lo tanto, F es la prueba general de la cual t es un caso especial.

Ejemplo de una diferencia estadísticamente significativa

Suponga que el investigador hubiera obtenido resultados bastante diferentes, digamos que las medias hubieran sido 6 y 3, en lugar de 4 y 3. Ahora tomamos el ejemplo anterior y añadimos una constante de 2 *a cada puntuación de* A_1. Esta operación, por supuesto, tan sólo restituye las puntuaciones utilizadas en el capítulo 6. Antes se indicó que el añadir (o restar) una constante a un conjunto de puntuaciones cambia la media por la constante, *pero no tiene ningún efecto en la* varianza. Las cifras se presentan en la tabla 13.5.

Resulta importante notar cuidadosamente que los valores de Σx^2 son los mismos de antes, 10. También debe notarse que las varianzas, V, son las mismas, 2.5, y lo mismo sucede con las varianzas estándar, pues cada una es de 0.50. En lo que respecta a dichos estadísticos, no hay una diferencia entre este ejemplo y el ejemplo previo. Pero al calcular la varianza entre los grupos V_e (tabla 13.6), se observa que ésta es nueve veces más grande que antes: 4.50 contra 0.50. Sin embargo, la V_d es exactamente igual a la anterior. Esto representa un aspecto importante. Se reitera: añadir una constante a un conjunto de puntuaciones [que es equivalente a una manipulación experimental, ya que uno de los propósitos de un experimento de esta clase es aumentar o disminuir un conjunto de medidas (las medidas del grupo experimental), mientras el otro conjunto no cambia (las medidas del grupo control)] no tiene un efecto sobre la varianza dentro de grupos, mientras que la varianza entre grupos cambia drásticamente. Considere que los *estimados* de V_e y V_d son independientes entre sí (si no lo son, por cierto, se violan las reglas y supuestos de la prueba F).

La razón F es $F = V_e/V_d = 4.50/.50 = 9$. Evidentemente la información obtenida acerca de las medias es mucho mayor que el error. ¿Querrá esto decir que la diferencia $6 - 3 = 3$ es una diferencia estadísticamente significativa? Si revisamos una tabla de la razón F, encontramos que, en este caso, una razón F de 5.32 o mayor es significativa al nivel .05 (posteriormente en este capítulo se explican los detalles para leer una tabla de la razón F). Para ser significativa al nivel .01, en este caso la razón F tendría que ser 11.26 o mayor. La razón F aquí es 9, que es mayor que 5.32, pero menor que 11.26. Parece ser que la diferen-

TABLA 13.5 *Datos de un experimento hipotético con dos grupos: datos alterados de la tabla 13.1*

	A_1	x	x^2	A_2	x	x^2
	4 + 2 = 6	0	0	3	0	0
	5 + 2 = 7	1	1	1	−2	4
	3 + 2 = 5	−1	1	5	2	4
	2 + 2 = 4	−2	4	2	−1	1
	6 + 2 = 8	2	4	4	1	1
ΣX	30			15		
M	6			3		
ΣX^2			10			10

$$V: \frac{10}{4} = 2.5 \qquad \frac{10}{4} = 2.5$$

$$VE: = \frac{V}{n} = \frac{2.5}{5} = .50 \qquad \frac{2.5}{5} = .50$$

cia de 3 es una diferencia estadísticamente significativa al nivel de .05. **Por lo tanto, $6 \neq 3$ y se rechaza la hipótesis nula.**

Cálculo del análisis de varianza de un factor

Con el auge que se vive gracias al desarrollo de las computadoras, un investigador interesado en realizar un análisis de varianza seguramente tendrá una computadora y un programa adecuado para realizar análisis estadísticos. El uso de una computadora será la primera opción en lo que a cálculos se refiere. No obstante si se está en una situación en la que no se cuenta con una computadora, sino que sólo se tiene una calculadora, hacer los cálculos del ANOVA de un factor no resulta difícil ni complejo. La presente sección se preparó para quienes sientan que necesitan saber cómo calcular un ANOVA de un factor, o quienes deseen hacerlo con una calculadora.

El análisis de varianza de un factor es más fácil de hacer que el procedimiento descrito en la sección previa. Para mostrar el método se utilizará el ejemplo que acaba de explicarse. El lector ya debe estar preparado para seguir el procedimiento sin dificultad. Note que

TABLA 13.6 *Cálculo de la varianza entre grupos, datos de la tabla 13.5*

	X	x	x^2
	6	1.5	2.25
	3	−1.5	2.25
ΣX	9		
M	4.5		
Σx^2			4.50

$$V_b = \frac{\Sigma x_b^2}{k-1} = \frac{4.50}{2-1} = 4.50$$

las puntuaciones de desviación (x) no se utilizan en lo absoluto; puede realizarse el cálculo completo con puntuaciones en bruto. Habrá ciertas diferencias en las varianzas. En el ejemplo previo, se utilizaron las varianzas estándar para demostrar la lógica subyacente del análisis de varianza. Sin embargo, en el siguiente método, aunque se utilizará la misma metodología, se omiten ciertos pasos debido a que es posible efectuar los cálculos de una manera más sencilla.

Los cálculos de la tabla 13.7 pueden seguirse fácilmente. Primero, en el cuerpo de la tabla se puede notar que las puntuaciones en bruto, las X, están todas elevadas al cuadrado; después se sumaron para dar las $\sum X^2$ en la parte inferior de la tabla (190 y 55). El propósito de realizarlo consiste en obtener $\sum X_t^2 = 245$ (190 + 55), en la parte inferior derecha de la tabla; $\sum X^2$ se lee: "la sumatoria total de todas las X elevadas al cuadrado". Las $\sum X$ y M se calculan de la forma usual (aunque en realidad no se necesitan las M, excepto para interpretaciones posteriores). Después, se eleva al cuadrado la suma de cada grupo y se escribe $(\sum X)^2$. Son $(30)^2 = 900$ y $(15)^2 = 225$. Se requiere tener cuidado aquí, ya que un error frecuente consiste en confundir $\sum X^2$ y $(\sum X)^2$. En la parte inferior derecha de la tabla se anotan los correspondientes X_t, $(\sum X)^2$, M_t y $\sum X_t^2$. Éstos son estadísticos de todas las puntuaciones como un solo conjunto y se calculan de la misma manera que los estadísticos de un grupo individual.

Se continúa con los cálculos de las sumas de cuadrados (de aquí en adelante *sc*). En el análisis de varianza se calculan y utilizan de manera casi exclusiva sumas de cuadrados. Las varianzas o cuadrados medios se reservan para el análisis final de la tabla del análisis de varianza (en la parte inferior de la tabla 13.7). Lo que se busca con este procedimiento son las sumas de cuadrados total, entre y dentro, o sc_t, sc_e y sc_d. Primero, el cálculo de C, el término de corrección: puesto que se están utilizando puntuaciones en bruto y como se están calculando sumas de cuadrados, que son las sumas de las *desviaciones* elevadas al cuadrado, se deben reducir las puntuaciones en bruto a puntuaciones de desviación. Para lograrlo, se resta C de cada uno de los cálculos, lo cual realiza la reducción: cambia, en efecto, las X a x. El cálculo de C resulta obvio; aquí es de 202.50.

Ahora se calcula la suma de cuadrados total, sc_t: 42.50. La suma de cuadrados entre grupos, o entre medias, no es tan obvia. La suma de las puntuaciones de cada grupo se eleva al cuadrado y luego se divide entre el número de puntuaciones en el grupo; después se suman estos promedios, y al resultado se le resta C. El resultado es la suma de cuadrados entre grupos o sc_e. Éste es todo el proceso del análisis de varianza de un factor. La suma de cuadrados dentro, sc_d, se calcula con una resta. La siguiente ecuación es importante y es menester recordarla:

$$sc_t = sc_e + sc_d \qquad \textbf{(13.3)}$$

Hoy casi todas las calculadoras cuyo precio actual es de alrededor de 10 dólares incluyen las teclas de las funciones estadísticas. Por lo común hay una tecla que permite al usuario obtener la media, y otra para determinar la desviación estándar. En muchos casos, dichas calculadoras contienen una función para calcular la desviación estándar usando N, y otra que utiliza $N - 1$. Aprender a utilizar estas teclas de funciones simplifica bastante los cálculos y disminuye los errores; además, estas funciones también ayudan a calcular el ANOVA de un factor. Por ejemplo, el término C puede calcularse como $M^2 \times N$, lo cual implica alimentar los datos en la calculadora, independientemente del grupo al que pertenezcan y, luego, se obtiene C presionando la tecla para la media, elevándolo al cuadrado y multiplicando el resultado por el número de datos (puntuaciones). De la misma forma, se puede obtener la suma de cuadrados total, sc_t, por medio de la fórmula $DE^2 \times (N)$ o $x_t^2 \times (N - 1)$. [Nota: la s minúscula se calculó utilizando los grados de libertad, es decir, $N - 1$ en lugar de N.] Esto se hace presionando la tecla de la desviación estándar, elevándola al

TABLA 13.7 *Cálculo del análisis de varianza: datos ficticios*

	X_{A1}	X_{A1}^2	X_{A2}	X_{A2}^2	
	6	36	3	9	N = 10
	7	49	1	1	n = 5
	5	25	5	25	k = 2
	4	16	2	4	
	8	64	4	16	
ΣX:	30		15		$X_t = 45$
$(\Sigma X)^2$:	900		225		$(\Sigma X_t)^2 = 2\,025$
M	6		3		$M_t = 9$
ΣX^2		190		55	$\Sigma X_t^2 = 245$

$$C = \frac{(\Sigma X_t)^2}{N} = \frac{(45)^2}{10} = \frac{2\,025}{10} = 202.50$$

$$sc\ total = \Sigma X_t^2 - C = 245 - 202.50 = 42.50$$

$$sc\ entre = \left[\frac{(\Sigma X_{A1})^2}{n_{A1}} + \frac{(\Sigma X_{A2})^2}{n_{A2}}\right] - C$$

$$= \left[\frac{(30)^2}{5} + \frac{(15)^2}{5}\right] - 202.50 = (180 + 45) - 202.50 = 22.50$$

cuadrado y multiplicándola por N o por $N - 1$. Se multiplicaría por N si la tecla de función utilizada fuera para la desviación estándar calculada con N; se multiplicaría por $N - 1$ si la función fuera para calcular la desviación estándar con $N - 1$. Por lo tanto, los datos se introducen en la calculadora una sola vez y se pueden obtener C y sc_t presionando unas cuantas teclas.

Recuerde la ecuación 13.2: $V_t = V_e + V_d$. La ecuación 13.3 es la misma ecuación pero en la forma de sumas de cuadrados. La ecuación 13.2 no puede utilizarse debido a que, como se señaló antes, es una formulación teórica que sólo funciona bajo las condiciones especificadas. La ecuación 13.3 siempre funciona, es decir, que las sumas de cuadrados en el análisis de varianza siempre son aditivas. Así, con una pequeña manipulación algebraica se puede observar que $sc_d = sc_t - sc_e$; en otras palabras, para obtener la suma de cuadrados dentro, simplemente se resta la suma de cuadrados entre grupos de la suma de cuadrados total. En la tabla del análisis de varianza vemos que 42.50 – 22.50 = 20. (También es posible calcular de forma directa la suma de cuadrados dentro de grupos.)

Después de completar los cálculos anteriores, se calculan los grados de libertad (*gl* en la tabla final). Aunque ya se han dado las fórmulas, no resultan necesarias para la operación. Para conocer el total de grados de libertad, sólo se toma el número total de partici-

Fuente de la variación	*gl*	*sc*	*CM*	*F*
Entre grupos	k – 1 = 1	22.50	22.50	9.0(0.05)
Dentro de grupos	N – k = 8	20.00	2.50	
Total	N – 1 = 9	42.50		

pantes, menos uno. Si, por ejemplo, hubiera tres grupos experimentales con 30 sujetos cada uno, los grados de libertad totales serían $N - 1 = 90 - 1 = 89$. Los grados de libertad entre grupos son el número de grupos experimentales, menos uno; con tres grupos experimentales, $k - 1 = 3 - 1 = 2$. Con el ejemplo de la tabla 13.7, $k - 1 = 2 - 1 = 1$. Los grados de libertad dentro de los grupos, al igual que la suma de cuadrados dentro de los grupos, se obtienen por medio de una resta; en este caso, $9 - 1 = 8$. Después se dividen las sumas de cuadrados entre sus respectivos grados de libertad (*sc/gl*), para obtener los cuadrados medios entre y dentro de grupos, denominados como *CM* en la tabla. En el análisis de varianza, se les llama "cuadrados medios" o "medias cuadráticas". Por último, se obtiene la razón *F* dividiendo la varianza entre los grupos o cuadrado medio entre, entre la varianza dentro de los grupos o el cuadrado medio dentro: $F = CM/CM_d = 22.50/2.50 = 9$. Esta razón *F* final (también llamada razón de varianza) se compara contra los valores correspondientes en la tabla de la razón *F*, para determinar su significancia, como se mencionó antes.

Una tabla abreviada de la razón *F* se presenta en la tabla 13.8; para utilizarla, primero debemos decidir el nivel de significancia (ya sea .05 o .01), después se buscan en el primer renglón los grados de libertad para la varianza entre grupos. En el ejemplo previo es $k - 1 = 1$. Ahora se buscan hacia abajo en la primera columna los grados de libertad para la varianza dentro de grupos, que es $N - k = 8$. El valor que buscamos (también llamado valor crítico) se encuentra en la intersección del renglón y la columna correspondientes a los

TABLA 13.8 *Valores críticos de F*

	gl entre grupos			
gl dentro de grupos	1	2	3	4
1	161 **4 052**	200 **4 999**	216 **5 403**	225 **5 625**
2	18.51 **98.49**	19.00 **99.00**	19.16 **99.17**	19.25 **99.25**
3	**10.13** **34.12**	9.55 **30.82**	9.28 **29.46**	9.12 **28.71**
4	7.71 **21.20**	6.94 **18.00**	6.39 **15.98**	6.26 15.52
5	6.61 **16.26**	5.14 **10.92**	4.76 **9.78**	4.53 **9.15**
6	5.99 **13.74**	5.14 **10.92**	4.76 **9.78**	4.53 **9.15**
7	5.59 **12.25**	4.74 **9.55**	4.35 **8.45**	4.12 **7.85**
8	5.32 **11.26**	4.46 **8.65**	4.07 **7.59**	3.84 **7.01**
9	5.12 **10.56**	4.26 **8.02**	3.86 **6.99**	3.63 **6.42**
10	4.96 **10.04**	4.10 7.56	3.71 **6.55**	3.48 **5.99**

grados de libertad que acabamos de ubicar. Al hacerlo, encontramos dos valores: 5.32 y **11.26**. El valor en negritas es para $\alpha = .01$ y el otro es para $\alpha = .05$.

Un ejemplo de investigación

Para ilustrar el uso del análisis de varianza de un factor en investigación, se utilizarán los datos de un antiguo estudio experimental de Hurlock (1925), ya mencionado antes en este libro. Los resultados se muestran en la tabla 13.9. Los datos no fueron analizados de esta manera por Hurlock, ya que en aquel entonces todavía no estaba disponible el análisis de varianza. Las primeras tres líneas de la tabla 13.9 fueron reportadas por Hurlock, y el resto de las cifras fueron calculadas por los autores, a partir de estas cifras (véase el anexo del capítulo). Hurlock dividió a 106 alumnos de cuarto y sexto grados en cuatro grupos: E_1, E_2, E_3 y C. Utilizó cinco formas de una prueba de sumas: A, B, C, D y E. El primer día aplicó la forma A a todos los participantes. En los siguientes cuatro días se les aplicaron las diferentes formas de la prueba a los grupos experimentales E_1, E_2 y E_3. El grupo C (grupo control) fue separado de los otros grupos y se le aplicaron diferentes formas de la prueba en cuatro días distintos. A los participantes del grupo C se les pidió que trabajaran como acostumbraban. Sin embargo, un día antes de la aplicación de la prueba, se pasaba al grupo E_1 al frente del salón y se le *felicitaba* por su buen desempeño; luego pasaba el grupo E_2 al frente y se le *reprendía* por su pobre desempeño. A los miembros del grupo E_3 se les *ignoraba*. Al quinto día del experimento se administró la forma E a todos los grupos. Las puntuaciones consistían en el número de respuestas correctas en esta forma de la prueba. En la tabla 13.9 se presenta un resumen de los datos, junto con la tabla del análisis de varianza final.

Puesto que $F = 10.08$, que es significativa al nivel .001, tiene que rechazarse la hipótesis nula de no diferencias entre las medias. Evidentemente la manipulación experimental fue efectiva, sin embargo no existe una diferencia grande entre los grupos ignorado y control, lo que constituye un descubrimiento interesante. El grupo felicitado presenta la media más alta y la media del grupo reprendido se ubica entre la del grupo felicitado y la de los otros dos grupos. El estudiante puede completar la interpretación de los datos. Después de un análisis de varianza de este tipo, algunos investigadores prueban pares de medias con pruebas *t*. Tal procedimiento es cuestionable, a menos que antes del análisis se hayan realizado predicciones sobre diferencias específicas entre medias, o grupos de medias. Dicho problema se retomará posteriormente en este capítulo (véase la sugerencia de estudio número 6).

TABLA 13.9 *Resumen de datos y análisis de varianza de los datos (del estudio de Hurlock)*

	E1: felicitado	E2: reprendido	E3: ignorado	C: control
n:	27	27	26	26
M:	20.22	14.19	12.38	11.35
DE:	7.68	6.78	6.06	4.21

Fuente de la variación	*gl*	*SC*	*CM*	*F*
Entre grupos	3	1 260.06	420.02	10.08(0.001)
Dentro de grupos	102	4 249.29	41.66	
Total	105	5 509.35		

TABLA 13.10 *Fuerte relación entre métodos de instrucción y rendimiento*

Variable independiente (métodos de instrucción)	Variable dependiente (rendimiento)	Medias
Método A_1	10 9 9 8	9
Método A_2	7 7 7 7	7
Método A_3	5 4 4 3	4

Fuerza de las relaciones: correlación y análisis de varianza

Las pruebas de significancia estadística, como la *t* y la *F*, por desgracia no indican la magnitud o la fuerza de las relaciones. Si una prueba *t* de la diferencia entre dos medias es significativa, tan sólo le indica al investigador que existe una relación. De la misma forma, si una prueba *F* es significativa, solamente señala que existe una relación. En ambos casos la relación se infiere a partir de las diferencias significativas entre dos, tres o más medias. Una prueba estadística como la razón *F* indica, indirectamente, si existe o no una relación entre la variable (o variables) independientes y la variable dependiente.

En contraste con las pruebas de significancia estadística como la *t* y la *F*, los coeficientes de correlación son medidas relativamente directas de las relaciones. Poseen un mensaje intuitivo, directo y fácil de percibir, ya que la unión de dos conjuntos de puntuaciones tiene una apariencia más obvia de relación y cumple la definición dada previamente de una relación como un conjunto de pares ordenados. Si por ejemplo, $r = .90$, es fácil ver que el orden de los rangos de las medidas de ambas variables es muy similar. Sin embargo, las razones *t* y *F* se alejan uno o dos pasos de la relación real. Entonces, una importante pregunta técnica de investigación es cómo se relacionan *t* y *F* por un lado, con medidas tales como *r*, por el otro.

En un análisis de varianza, la variable al margen de la tabla de datos (métodos de incentivar en el ejemplo de Hurlock) es la variable independiente. Las medidas que se encuentran en el cuerpo de la tabla reflejan la variable dependiente (es decir, el rendimiento matemático en el ejemplo de Hurlock). El análisis de varianza funciona con la relación entre estos dos tipos de variables. Si la variable independiente tiene un efecto sobre la variable dependiente, entonces esto alterará la "igualdad" de las medias de los grupos experimentales que se esperaría si los números analizados fueran aleatorios. El efecto de una variable independiente realmente influyente consiste en volver desiguales las medias. Puede decirse, entonces, que cualquier relación existente entre las variables independiente y dependiente se refleja en la desigualdad de las medias. Mientras más desiguales sean

TABLA 13.11 *Relación de cero entre métodos de instrucción y rendimiento*

Variable independiente (métodos de instrucción)	Variable dependiente (rendimiento)	Medias
Método A_1	4 8 10 7	7.25
Método A_2	3 5 4 9	5.25
Método A_3	7 7 7 9	7.50

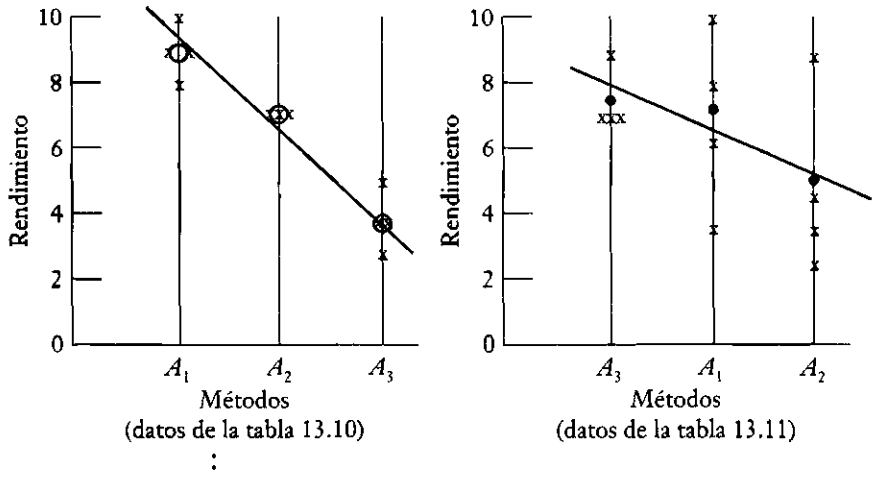
las medias, cuanto más lejos están una de la otra y mayor será la relación, siempre y cuando los demás aspectos se mantengan constantes.

Si no existe relación entre la variable independiente y la variable dependiente, entonces es como si tuviéramos un conjunto de números aleatorios y, en consecuencia, medias aleatorias; las diferencias entre las medias sólo serían fluctuaciones debidas al azar y una prueba F mostraría que no son significativamente diferentes. Si en realidad existe una relación, si existe un vínculo entre las variables independiente y dependiente, la introducción de *diferentes* aspectos de la variable independiente, como serían los distintos métodos de instrucción, debería hacer que las medidas de la variable dependiente variaran en concordancia. El método A_1 podría incrementar las puntuaciones de rendimiento, mientras que el método A_2 podría disminuirlas o hacer que permanezcan casi iguales. Observe que tenemos el mismo fenómeno de variación concomitante que el que tuvimos con el coeficiente de correlación. Considere dos casos extremos: una fuerte relación y una relación de cero. Establecemos una fuerte relación hipotética entre los métodos y el rendimiento en la tabla 13.10. Observe que las puntuaciones de la variable dependiente varían directamente con los métodos de la variable independiente: el método A_1 tiene puntuaciones altas; el método A_2 tiene puntuaciones medianas, y el método A_3 tiene puntuaciones bajas. La relación también se evidencia al comparar los métodos y las medias de la variable dependiente.

Ahora compare el ejemplo de la tabla 13.10 con lo esperado por el azar. Si no existiera relación entre los métodos y el rendimiento, entonces las medias del rendimiento no covarían con los métodos, es decir, las medias serían casi iguales. Para demostrarlo se anotaron en hojas de papel separadas las 12 puntuaciones de rendimiento de la tabla 13.10 y se mezclaron repetidamente dentro de un sombrero; después se arrojaron al suelo todas las hojas y se recogieron cuatro a la vez, asignando las primeras cuatro a A_1, las segundas cuatro a A_2 y las terceras cuatro a A_3. Los resultados se presentan en la tabla 13.11.

Ahora es muy difícil o imposible "percibir" una relación. Las medias difieren, pero no mucho. Ciertamente, la relación entre los métodos y las puntuaciones de rendimiento (y las medias) ya no es tan clara como antes. Aun así, debemos estar seguros. Se realizaron análisis de varianza en ambos conjuntos de datos: la razón F de los datos de la tabla 13.10

FIGURA 13.1



(fuerte relación) fue 57.59, altamente significativa; mientras que la razón F de los datos de la tabla 13.11 (relación baja o de cero) fue 1.29, que no es significativa. Las pruebas estadísticas confirman nuestras impresiones visuales. Sabemos que existe una relación entre los métodos y el rendimiento en la tabla 13.10, pero no en la tabla 13.11.

Sin embargo, el problema consiste en mostrar la relación entre las pruebas de significancia, como la prueba F, y el método de correlación. Esto puede efectuarse de varias maneras; aquí se ilustran dos de ellas, una gráfica y una estadística. En la figura 13.1 se graficaron los datos de las tablas 13.10 y 13.11 como se grafican las medidas continuas X y Y en un problema común de correlación. En cada caso la variable independiente (métodos) se coloca en el eje horizontal; y la variable dependiente (rendimiento), en el eje vertical. Para indicar la relación, se dibujaron líneas lo más cercanas posible a las medias. Una línea diagonal con un ángulo de 45 grados respecto al eje horizontal indicaría una fuerte relación. Una línea horizontal a lo largo de la gráfica indicaría una relación de cero. Observe que la representación gráfica de las puntuaciones de los datos de la tabla 13.10 claramente indican una fuerte relación: la altura de las puntuaciones graficadas (cruces) y las medias (círculos) varían con el método. Aun con el reordenamiento de los métodos con el fin de poder compararlos, el gráfico de los datos de la tabla 13.11 muestra una débil o casi nula relación.

Ahora se enfocará el problema desde un punto de vista estadístico. Es posible calcular coeficientes de correlación con datos de este tipo. Si ya se realizó un análisis de varianza, se puede obtener un coeficiente simple (pero no muy satisfactorio) con la siguiente fórmula:

$$\eta = \sqrt{\frac{sc_e}{sc_t}} \tag{13.4}$$

Por supuesto que sc_e y sc_t son, respectivamente, la suma de cuadrados entre los grupos y la suma de cuadrados total. Sólo se toman estas sumas de cuadrados de la tabla del análisis de varianza para calcular el coeficiente. η, por lo común llamada *razón de correlación*, es un coeficiente general o índice de relación, frecuentemente utilizado con datos no lineales. (En general, *lineal* significa que si dos variables se grafican una respecto a la otra, el gráfico

tiende a formar una línea recta. Ésta es otra forma de explicar lo que se indicó en el capítulo 12 sobre las combinaciones lineales.) Los valores de η varían entre 0 y 1.00. Aquí tan sólo interesa su uso con el análisis de varianza y su poder para demostrar la magnitud de la relación entre las variables independiente y dependiente.

Recuerde que las medias de los datos de la tabla 13.1 eran 3 y 4, y que no resultan significativamente diferentes; por lo tanto no existe una relación entre la variable independiente (métodos) y la variable dependiente (rendimiento). Si se realiza un análisis de varianza con los datos de la tabla 13.1 y se utiliza el método indicado en la tabla 13.7, entonces mediante $sc_e = 2.50$ y $sc_t = 22.50$ se obtiene η:

$$\eta = \sqrt{2.50 / 22.50} = \sqrt{.111} = 0.33$$

que se refiere a la correlación entre los métodos y el rendimiento. Puesto que se sabe que los datos no son significativos ($F = 1$), η no es significativo. En otras palabras, $\eta = 0.33$ es equivalente a una relación de cero. Si no hubiese diferencia alguna entre las medias, entonces, por supuesto que $\eta = 0$. Si la suma de cuadrados entre los grupos fuera igual a la suma de cuadrados total, es decir $sc_e = sc_t$, entonces $\eta = 1.00$. Esto puede ocurrir sólo si todas las puntuaciones de un grupo son iguales entre sí y todas las puntuaciones del otro grupo son iguales entre sí, pero diferentes que las del primer grupo. En la práctica este hecho es bastante improbable. Por ejemplo, si las puntuaciones de A_1 fueran 4, 4, 4, 4, 4 y las puntuaciones de A_2 fueran 3, 3, 3, 3, 3, entonces:

$$SC_e = SC_t = 2.5 \text{ y } \eta = \sqrt{2.5 / 2.5} = 1$$

Parece obvio que no existe varianza dentro de los grupos, lo cual es bastante improbable. Tome los datos de la tabla 13.7, cuyas medias son 6 y 3. Son significativamente diferentes, ya que $F = 9$. Si se calcula η:

$$\eta = \sqrt{\frac{SC_e}{SC_t}} = \sqrt{\frac{22.50}{42.50}} = \sqrt{.529} = 0.73$$

Observe el incremento sustancial en η. Y como F es significativa, $\eta = 0.73$ también es significativa. Existe una relación sustancial entre los métodos y el rendimiento.

El estudio de Hurlock es más interesante:

$$\eta = \sqrt{1\,260.06 / 5\,509.35} = \sqrt{.229} = 0.48$$

que, por supuesto, es significativo. Si lo demás se mantiene constante, el incentivo está altamente relacionado con el rendimiento aritmético, como se definió antes.

Hasta aquí el estudiante ya tiene los antecedentes suficientes para interpretar η^2 en términos de varianza. En el capítulo 6 esto se realizó para r, donde se explicó que r^2 indicaba la varianza compartida por dos variables. Se puede dar una interpretación similar para η^2. Si η se eleva al cuadrado, η^2, indica, en esencia, la varianza compartida por las variables independiente y dependiente. Quizás más claro, η^2 indica la proporción de la varianza de la variable dependiente, por ejemplo rendimiento, determinada por la varianza de la variable independiente, métodos o incentivos. En el ejemplo de Hurlock $\eta^2 = (0.48)^2 = 0.23$, lo que indica que el 23% de la varianza de las puntuaciones de la prueba de sumas se explica por las diferentes formas de incentivos empleadas por Hurlock.

η^2 es un índice de la proporción de la varianza explicada en este ejemplo. Otro índice, ω^2, omega al cuadrado (véase Hays, 1994), constituye un estimado de la fuerza de la aso-

ciación entre la variable independiente y la variable dependiente poblacional. Se recomienda su uso mediante la siguinete fórmula:

$$\omega^2 = \frac{SC_e - (k-1)CM_d}{SC_t + CM_d} \tag{13.5}$$

donde k es igual al número de grupos en el análisis de varianza y los otros términos son las sumas de cuadrados y los cuadrados medios definidos con anterioridad. η^2 es un estimado conservador de la fuerza de la asociación o relación entre la variable independiente X y la variable dependiente Y, o entre la variable que constituye el tratamiento experimental y la medida de la variable dependiente. Si se calcula ω^2 para el ejemplo de Hurlock:

$$\omega^2 = \frac{1\,260.06 - (4-1)(41.66)}{5\,509.35 + 41.66} = 0.205$$

Este valor es bastante cercano al valor η^2 de 0.23. η^2 se compara con ω^2 más que con η. Ambos índices se refieren a la proporción de varianza de una variable dependiente debida a la supuesta influencia de una variable independiente. Existen otros índices disponibles para reportar la cantidad de varianza explicada. En la primera y segunda edición de este libro, se recomendaba el coeficiente de correlación intraclase, *RI*. No obstante, el *RI* es más adecuado para un tipo diferente de modelo de análisis de varianza al que se ha presentado aquí (véase Hays, 1994).

La fórmula de *RI* es:

$$RI = \frac{CM_e - CM_d}{CM_e + (n_j - 1)\,CM_d}$$

La relación entre estas medidas y sus méritos relativos no conforman problemas sencillos. Vaughan y Corballis (1969) analizan este problema. Simon (1987) alienta el uso de estas medidas en lugar de las pruebas de significancia. Simon señala que las pruebas de significancia están sujetas a la influencia del tamaño de la muestra; mientras que η^2 y ω^2 no lo están.

El objetivo del análisis anterior consiste en poner de manifiesto la similitud conceptual de éstos y otros índices de asociación o correlación. Un análisis más importante concierne a la similitud entre el principio y la estructura del análisis de varianza y los métodos de correlación. Desde un punto de vista práctico y aplicado debe enfatizarse que η^2, ω^2 y *RI*, u otras medidas de asociación siempre deben calcularse y reportarse. No es suficiente reportar las razones F y su significancia estadística; es necesario saber qué tan fuertes son las relaciones. Después de todo, con N suficientemente grandes, las razones F y t casi siempre pueden ser estadísticamente significativas. Aunque son moderados con respecto a su efecto, especialmente cuando son bajos, los coeficientes de asociación de las variables independiente y dependiente constituyen partes indispensables de los resultados de investigación.

Ampliación de la estructura: pruebas *post hoc* y comparaciones planeadas

El enfoque utilizado en este capítulo y en los dos siguientes es, aunque pedagógicamente útil, demasiado rígido; es decir, se han enfatizado paradigmas ordenados que tienen su culminación en la prueba F y en algunas medidas de relación. Sin embargo, la investigación

real a menudo no se ajusta a formas y razonamientos tan precisos; sin embargo, las nociones básicas del análisis de varianza pueden utilizarse de manera más amplia y libre, con la expansión del diseño y de las posibilidades estadísticas. En este capítulo se examinan dichas posibilidades dentro del marco general de este capítulo.

Pruebas *post hoc*

Suponga un experimento como el realizado por Hurlock, donde se tienen los datos de la tabla 13.9. El investigador sabe que las diferencias globales entre las medias son estadísticamente significativas; pero no sabe cuáles diferencias contribuyen a la significancia. ¿Se pueden simplemente probar las diferencias entre todos los pares de medias para saber cuáles son significativas? Sí y no, pero por lo común no. Tales pruebas no son independientes y, con un número de pruebas suficiente, una podría ser significativa por azar. En pocas palabras, un procedimiento tan "de súbito" como éste se capitaliza por el azar; además de que es ciego y "descabezado" (como se le ha llamado).

Existen varias formas de realizar pruebas *post hoc*; pero sólo se mencionará brevemente una. Zwick (1993), Edwards (1984) y Kirk (1995) ofrecen excelentes descripciones de varias pruebas. La prueba de Scheffé (véase Scheffé, 1959), usada con discreción, constituye un método general que puede aplicarse a todas las comparaciones de medias posteriores a un análisis de varianza. Si y sólo si la prueba F es significativa, se pueden probar todas las diferencias entre medias. Se puede probar la media combinada de dos o más grupos contra la media de otro grupo; o se puede seleccionar cualquier combinación de medias contra cualquier otra combinación. Dicha prueba resulta muy útil porque tiene la habilidad de efectuar muchas cosas, pero la utilidad y la generalidad se pagan: la prueba es bastante conservadora. Para alcanzar la significancia las diferencias deben ser bastante grandes. La prueba de Scheffé es la prueba disponible más conservadora para *pruebas de comparación múltiple*. Linton y Gallo (1975) muestran la relación entre las diferentes pruebas y la probabilidad del error tipo I. La prueba de Scheffé posee la probabilidad más baja de cometer un error tipo I; aunque también tiene la probabilidad más baja de detectar una diferencia existente (poder). La cuestión más importante es que las comparaciones *post hoc* y las pruebas de medias pueden efectuarse principalmente con propósitos exploratorios e interpretativos. Uno examina los datos en detalle y busca indicios que faciliten su comprensión.

La mecánica para realizar la prueba de Scheffé no se explica aquí, ya que nos sacaría del tema (véase la sugerencia de estudio número 6 al final del capítulo, o revise el libro de Comrey y Lee, 1995, capítulos 10 y 11). Es suficiente decir que al aplicar esta prueba a los datos de Hurlock de la tabla 13.9, demuestra que la media del grupo de felicitados es significativamente mayor que las otras tres medias, y que ninguna de las otras diferencias es significativa. Esta información es importante ya que apunta directamente a la fuente principal de significancia de la razón de F global: felicitar *versus* reprender, ignorar y controlar. (Sin embargo, la diferencia entre el promedio de las medias 1 y 2, contra el promedio de las medias 3 y 4, también es estadísticamente significativa.) Aun cuando esto pueda verse a partir de los *tamaños* relativos de las medias, la prueba de Scheffé hace todo con precisión —de forma conservadora—.

Comparaciones planeadas

Aunque las pruebas *post hoc* son importantes en la investigación real, en especial para explorar los datos y para obtener guías respecto a futuras investigaciones, el método de comparaciones

planeadas es, quizá, científicamente más importante. Siempre que se formulan hipótesis, se prueban sistemáticamente y los resultados empíricos las soportan, hay evidencia mucho más poderosa sobre la validez empírica de la hipótesis que cuando se encuentran resultados "interesantes" (algunas veces entendidos como "apoyan mis predicciones") *después* de que se obtuvieron los datos. Esto se señaló en el capítulo 2 cuando se explicó el poder de las hipótesis.

En el análisis de varianza, si una prueba *F* resulta significativa, ello simplemente indica que existen diferencias significativas en alguna parte de los datos. Una inspección de las medias puede revelar, aunque de forma imprecisa, qué diferencias son importantes. Sin embargo, para probar hipótesis se requieren pruebas estadísticas más o menos precisas y controladas. Existe una gran variedad de comparaciones posibles en cualquier conjunto de datos que se prueben, ¿pero cuáles deben aplicarse? Como de costumbre, el problema de investigación y la teoría que lo subyace deberían determinar las pruebas estadísticas adecuadas. Uno diseña la investigación, en parte, para probar hipótesis sustantivas.

Suponga que la teoría del reforzamiento en que se basa el estudio de Hurlock señala, en efecto, que cualquier forma de atención, ya sea positiva o negativa, mejorará el desempeño; y que el reforzamiento positivo lo mejorará más que el castigo. Esto significaría que las medias de los grupos E_1 y E_2 de la tabla 13.9, juntos o separados, serían significativamente mayores que las medias de los grupos E_3 y C, juntos o separados; es decir, que tanto el grupo felicitado (reforzamiento positivo) como el grupo reprendido (castigo) resultarían significativamente mayores que el grupo ignorado (sin reforzamiento) y el grupo control (sin reforzamiento). Además, la teoría afirma que el efecto del reforzamiento positivo es mayor que el efecto del castigo, de tal manera que el grupo felicitado será significativamente mayor que el reprendido. Estas pruebas implícitas pueden escribirse de manera simbólica:

$$H_1\colon C_1 = \frac{M_1 + M_2}{2} > \frac{M_3 + M_4}{2}$$
$$H_2\colon C_2 = M_1 > M_2$$

donde C_1 indica la primera comparación y C_2 la segunda. Aquí se tienen los elementos de un análisis de varianza de un factor; pero la prueba global simple y su democracia de medias han sido radicalmente cambiadas; es decir, que el plan y el diseño de la investigación han cambiado bajo el impacto de la teoría y del problema de investigación.

Cuando se utiliza la prueba de Scheffé, la razón *F* global debe ser significativa porque ninguna de las pruebas de Scheffé puede ser significativa si la *F* general no lo es. Sin embargo, cuando se utilizan comparaciones planeadas, no es necesario hacer una prueba *F* global, ya que el punto nodal son las comparaciones planeadas y las hipótesis. El número de pruebas y comparaciones realizadas están limitados por los grados de libertad. En el ejemplo de Hurlock existen tres grados de libertad para el cálculo entre grupos ($k - 1$), por lo tanto, se pueden realizar tres pruebas. Éstas deben ser *ortogonales* entre sí —es decir, que deben ser independientes—. Las comparaciones se mantienen ortogonales mediante el uso de los llamados *coeficientes* o *contrastes ortogonales*, que son pesos que se añaden a las medias en la comparación. En otras palabras, los coeficientes especifican las comparaciones. Los coeficientes o pesos para las anteriores hipótesis son:

H_1:	1/2	1/2	−1/2	−1/2
H_2:	1	−1	0	0

Para que las comparaciones sean ortogonales deben cumplirse dos condiciones: la suma de cada conjunto de pesos tiene que ser igual a cero, y la suma de los productos de cuales-

quiera dos conjuntos de pesos también debe ser igual a cero. Resulta obvio que los dos conjuntos anteriores suman cero; si se prueba la suma de los productos: (1/2)(1) + (1/2)(–1) + (–1/2)(0) + (–1/2)(0) = 0. Por lo tanto, ambos conjuntos de pesos son ortogonales.

Es importante entender los pesos ortogonales, así como las dos condiciones recién explicadas. El primer conjunto de pesos simplemente se representa: $(M_1 + M_2)/2 - (M_3 + M_4)/2$. El segundo conjunto se representa: $M_1 - M_2$. Ahora suponga que también se desea probar la noción de que la media del grupo ignorado es mayor que la media del grupo control. Esto se prueba por medio de: $M_3 - M_4$, y se codifica: H_3: 0 0 1 –1. De ahora en adelante, a estos pesos se les llamará *vectores*. Los valores de los vectores suman cero. ¿Qué sucede con su suma de productos con los otros dos vectores?

$$H_1 \times H_3\text{: } (1/2)(0) + (1/2)(0) + (-1/2)(1) + (-1/2)(-1) = 0$$
$$H_2 \times H_3\text{: } (1)(0) + (-1)(0) + (0)(1) + (0)(-1) = 0$$

El tercer vector es ortogonal o independiente de los otros dos vectores. Ahora puede realizarse la tercera comparación. Si se efectúan estas tres comparaciones, ya no es posible ninguna otra debido a que los grados de libertad disponibles $k - 1 = 4 - 1 = 3$ ya han sido utilizados.

Suponga ahora que, en lugar del H_3 en la fórmula anterior, se deseara probar la diferencia entre el promedio de las primeras tres medias contra la cuarta media; la codificación sería: 1/3 1/3 1/3 –1, lo cual es equivalente a $(M_1 + M_2 + M_3)/3 + M_4$. ¿El vector es ortogonal respecto a los primeros dos? Para saberlo se calcula:

$$(1/2)(1/3) + (1/2)(1/3) + (-1/2)(1/3) + (-1/2)(-1)$$
$$= 1/6 + 1/6 - 1/6 + 1/2 = 4/6 = 2/3$$

Puesto que la suma de los productos no es igual a cero, entonces no es ortogonal respecto al primer vector y no debe hacerse la comparación, ya que al hacerlo se produciría información redundante; en este caso, la comparación usando el tercer vector ofrece información que en parte ya fue dada por el primer vector.

El método para calcular la significancia de las diferencias de comparaciones planeadas no necesita detallarse. Además, en este momento no se requieren los cálculos reales. Nuestro propósito es mayor: demostrar la flexibilidad y el poder del análisis de varianza cuando se concibe y comprende adecuadamente. Las pruebas F (o las pruebas t) se utilizan con cada comparación o, en este caso, con cada grado de libertad. Los detalles de los cálculos se pueden encontrar en Hays (1994) y en otros textos. La idea básica de las comparaciones planeadas es bastante general y se utilizará posteriormente cuando se estudie el diseño de investigación.

Hasta ahora se ha recorrido un largo, y quizá duro, camino sobre el análisis de varianza. Cabría preguntarse por qué se ha dedicado tanto espacio a este tema; existen varias razones. Primero, el análisis de varianza tiene una amplia aplicabilidad práctica; toma muchas formas que son aplicables en psicología, sociología, economía, ciencias políticas, agricultura, biología, educación y otras disciplinas. Nos libera de trabajar sólo con una variable independiente a la vez y nos ofrece un poderoso apoyo para resolver problemas de medición. Incrementa las posibilidades de realizar experimentos exactos y precisos; también nos permite probar varias hipótesis simultáneamente, así como probar hipótesis que no pueden ser probadas de ninguna otra manera, al menos con precisión. Así que su rango de aplicación es extenso.

Más relacionado con los propósitos de este libro, el análisis de varianza permite el conocimiento de métodos y enfoques modernos de investigación al enfocarse precisa y constantemente en el razonamiento sobre varianza, clarificando la estrecha relación entre

los problemas de investigación y los métodos y la inferencia estadísticos; y clarificando la estructura y arquitectura del diseño de investigación. También constituye un paso importante en el entendimiento de la concepción multivariada contemporánea de la investigación, ya que es una expresión del modelo lineal general.

El modelo de este capítulo es simple y puede anotarse de la siguiente manera:

$$y = a_0 + A + e$$

donde y es la puntuación de la variable dependiente de un individuo, a_0 es un término común a todos los individuos, por ejemplo, la media general de y. A representa el efecto del tratamiento de la variable independiente, y e es el error. El modelo del siguiente capítulo será ligeramente más complejo y, antes de que el libro finalice, los modelos serán mucho más complejos. Como se verá después, el modelo lineal general es flexible y generalmente aplicable a muchos problemas y situaciones de investigación. Quizá de mayor importancia inmediata para nosotros, puede ayudarnos a comprender mejor los detalles comunes de los diferentes enfoques y métodos multivariados.

Anexo computacional

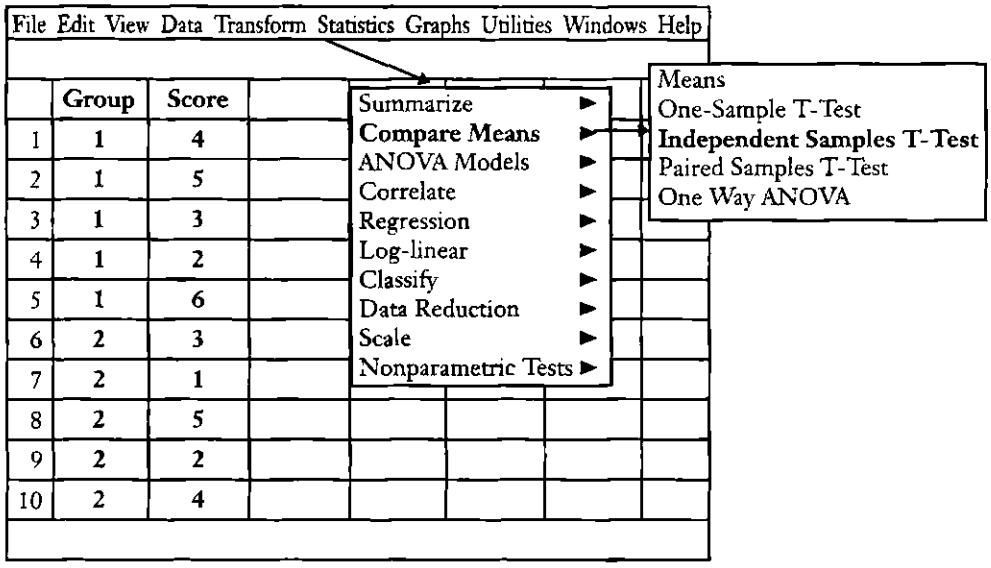
En este capítulo se examinó la razón t que se utilizó para analizar la diferencia entre dos medias y el análisis de varianza de un factor que puede usarse para analizar la diferencia entre dos o más medias grupales. Técnicamente nos referimos a los grupos como niveles de la variable independiente, y a las medidas resultantes como la variable dependiente. Aunque tales cálculos pueden efectuarse con papel y lápiz o con una calculadora, a veces resulta más eficiente utilizar una computadora. En el capítulo 6 se introdujo y en el capítulo

FIGURA 13.2 *Tabla de datos para una prueba t en el SPSS*

File Edit View Data Transform Statistics Graphs Utilities Windows Help

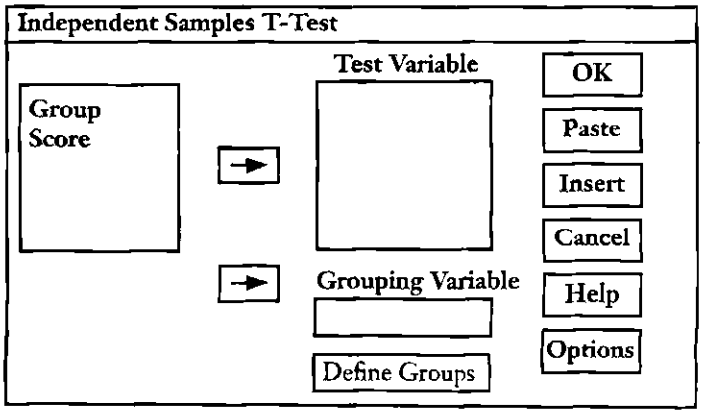
	Group	Score					
1	1	4					
2	1	5					
3	1	3					
4	1	2					
5	1	6					
6	2	3					
7	2	1					
8	2	5					
9	2	2					
10	2	4					

FIGURA 13.3 *Selección del análisis estadístico apropiado en el SPSS*



10 se demostró cómo puede utilizarse la computadora para analizar datos de frecuencias. En este capítulo se demostrará cómo puede usarse para realizar una prueba *t* y un ANOVA de un factor. Se espera que el lector ya haya leído y entendido el material acerca de la computadora de los capítulos 6 y 10, respecto de la creación de la tabla de datos en el SPSS.

FIGURA 13.4 *Pantalla del SPSS para la especificación de las variables independiente y dependiente*



Razón *t* o prueba *t* en el SPSS

Tome los datos de la tabla 13.1 y observe que la variable pertenecía a un grupo, se expresa como una variable categórica. En este caso es la variable independiente "Group" ("grupo" en español) y se maneja como una variable con dos niveles. Para A_1, Grupo = 1; para A_2, Grupo = 2. La segunda variable, "Score" ("puntuación" en español), es la variable dependiente. En el SPSS y otros programas de cómputo de análisis estadísticos, se espera que los datos se introduzcan de esta manera. La figura 13.2 muestra cómo debe aparecer la tabla de datos del SPSS para este problema.

Utilizando el ratón, señale y haga clic en "Statistics". Aparecerá otro menú listando los diferentes análisis que pueden realizarse con los datos. Para la prueba *t* seleccione "Compare Means". Esta selección, a su vez, despliega otro menú en donde puede seleccionar "Independent samples T-Test". Esto se muestra en la figura 13.3.

Al seleccionar "Independent samples T-Test", aparece una nueva pantalla donde se muestran las variables listadas en la tabla de datos y ahí debe especificarse cuál será la variable dependiente, y cuál será la variable independiente. La figura 13.4 muestra esta pantalla aún sin cambios realizados por el usuario. Utilizando la terminología del SPSS "Test variable" se refiere a las variables dependientes; "Grouping variable" se refiere a la variable independiente.

Se puede especificar la variable a prueba o dependiente resaltando la variable "Score" en el cuadro de la extrema izquierda y haciendo clic en el botón de la flecha asociada con el cuadro de "Test variable". Con esto veremos el nombre de la variable "Score" moverse del cuadro de la extrema izquierda al cuadro superior de la extrema derecha. Después se resalta (select) la variable independiente o "Grouping Variable " que en el ejemplo se llama "Group" y se hace clic en el botón de la flecha asociado con la caja de "Grouping Variable"; el nombre de la variable, Group, se moverá desde el cuadro de la extrema izquierda al cuadro de "Grouping Variable". La figura 13.5 muestra cómo aparece esta pantalla después de dichas operaciones.

Note que la variable Group encierra en un paréntesis dos signos de interrogación. Esto indica que necesita especificar los niveles de la variable independiente. Los valores deben corresponder con aquellos utilizados en la tabla de datos original, que en este ejemplo serían 1 y 2. Para indicarle esto al SPSS, haga clic en el botón "Define Groups".

FIGURA 13.5 *Pantalla después de especificar las variables dependiente e independiente*

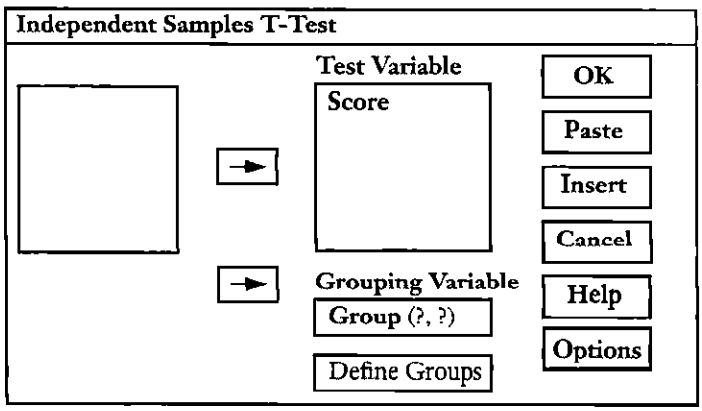
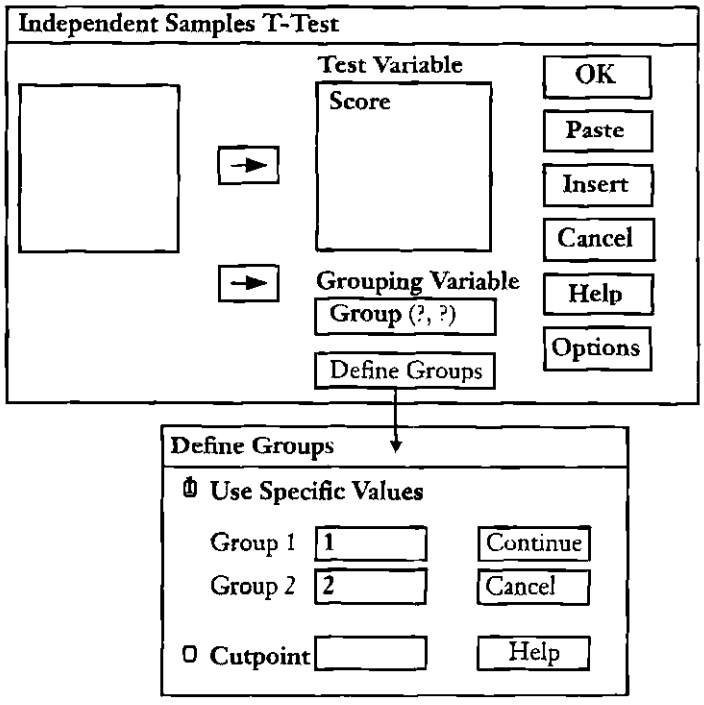


FIGURA 13.6 *Pantalla utilizada para definir niveles de la variable independiente*

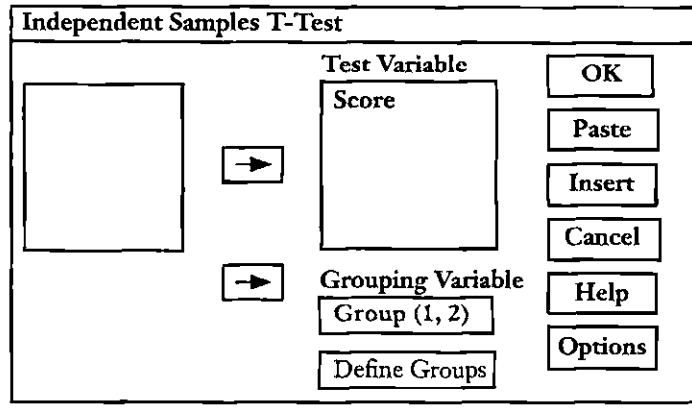


Después de hacerlo aparece otra pantalla que permite definir los niveles de la variable Group. La figura 13.6 muestra dicha pantalla. Observe que se anota "1" para el Grupo 1 y "2" para el grupo 2. Para regresar a la pantalla previa se hace clic en el botón "Continue". Sin embargo, los dos signos de interrogación desaparecen y se reemplazan por la especificación "1, 2".

Ahora es necesario desviarse un poco antes de terminar con el SPSS y la prueba *t*. Suponiendo que se tienen más de dos niveles de la variable independiente (por ejemplo, tres o más grupos), la prueba *t* puede comparar solamente dos niveles (grupos) al mismo tiempo. Si se tuvieran tres grupos, se podría efectuar la prueba *t* entre los grupos 1 y 2, los grupos 1 y 3 o los grupos 2 y 3. En la pantalla desplegada en la figura 13.6 se especificaría el grupo 1 con el índice "1" y el grupo 2 con el índice "3", si estuviera interesado en comparar los grupos 1 y 3. Si el propósito fuera comparar a los grupos 2 y 3, se especificaría "2" para el grupo 1 y "3" para el grupo 2 en la pantalla de la figura 13.6.

La figura 13.7 muestra la pantalla que aparece después de hacer clic en el botón "Continue" de la figura 13.6. Si ahora hace clic en el botón "OK", se realizará el análisis estadístico elegido para los datos. El resultado de este análisis se presenta en el cuadro de la página 301. Observe que el valor *t* calculado es el mismo que el realizado a mano para los datos de la tabla 13.1. El SPSS también calcula la probabilidad de un error tipo I, que en este caso es de .347; como es mayor a .05, la diferencia entre las medias comparadas no es estadísticamente significativa.

FIGURA 13.7 *Pantalla que muestra el resultado de la definición de los grupos para la variable independiente*



ANOVA de un factor en el SPSS

De nueva cuenta se asume que el lector desarrolló una tabla de datos dentro del SPSS y que está listo para seleccionar y realizar un análisis estadístico específico. La figura 13.8 muestra la tabla de datos que se utiliza con el SPSS. Los datos fueron tomados de la tabla 13.7. Aunque hay solamente dos grupos, el procedimiento mostrado aquí sería muy similar para más de dos grupos o más de dos niveles de la variable independiente. Antes, al realizar

FIGURA 13.8 *Tabla de datos para un ejemplo de ANOVA de un factor*

File Edit View Data Transform Statistics Graphs Utilities Windows Help

	Group	Score					
1	1	6					
2	1	7					
3	1	5					
4	1	4					
5	1	8					
6	2	3					
7	2	1					
8	2	5					
9	2	2					
10	2	4					

Prueba *t* para muestras independientes de grupo

Variable	Número de casos	Media	DE	EE de la media
PUNTUACIÓN				
GRUPO 1	5	4.0000	1.581	.707
GRUPO 2	5	3.0000	1.581	.707

Diferencia Media = 1.0000
Prueba de Levene para la igualdad de varianzas: F = .000 p = 1.000

Prueba *t* para la igualdad de medias

Varianzas	Valor *t*	gl	2 colas	Sig. de EE de dif.	95% CI para dif.
Iguales	1.00	8	.347	1.000	(–1.306, 3.306)
No iguales	1.00	8.00	.347	1.000	(–1.306, 3.306)

la prueba *t*, se observó que al hacer clic en "Statistics" aparecía otro menú listando los distintos análisis que pueden realizarse con los datos. La figura 13.3 presenta estos menús.

Para el ANOVA de un factor escoja "Compare Means" "(Comparar Medias"). Esta selección ofrece un nuevo menú, del cual se escoge "One way ANOVA" ("ANOVA de un Factor"). Al hacerlo, se despliega una pantalla que pide especificar cuál variable de la tabla de datos será la variable independiente y cuál la variable dependiente. Esta pantalla se muestra en la figura 13.8. Como se hizo para la prueba *t*, escoja "Score" como variable dependiente y "Group" como variable independiente. Aquí, en la terminología del SPSS, "Dependent list" ("Lista Dependiente") es para la variable dependiente y "Factor" es para la variable independiente (véase figura 13.9a). Como en las pantallas utilizadas para la prueba *t*, resalte el nombre de la variable "Score" en el cuadro de la extrema izquierda y haga clic en la flecha que apunta hacia la caja "Dependent List ". Esto mueve el nombre de la variable "Score" al cuadro asociado con "Dependent List". Haga lo mismo para la eti-

FIGURA 13.9a *Pantalla del SPSS para seleccionar las variables independiente y dependiente*

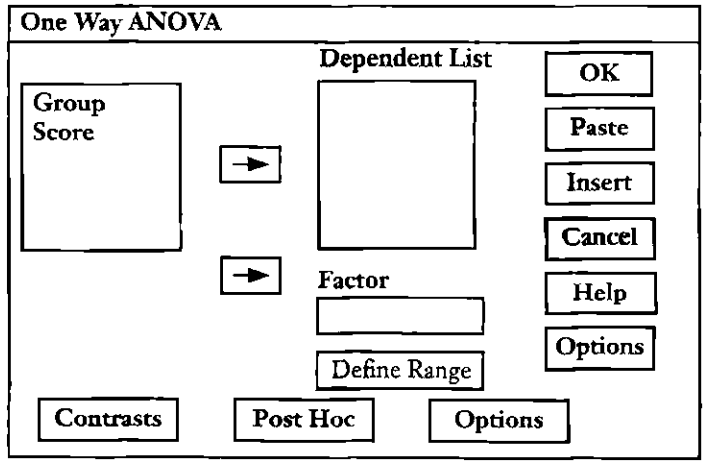


FIGURA 13.9b *Pantalla para especificar las variables dependiente e independiente*

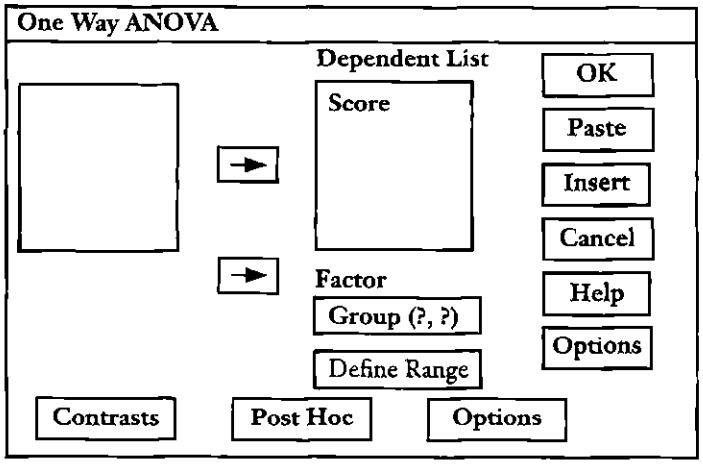


FIGURA 13.10 *Pantalla para definir el rango de valores para la variable independiente*

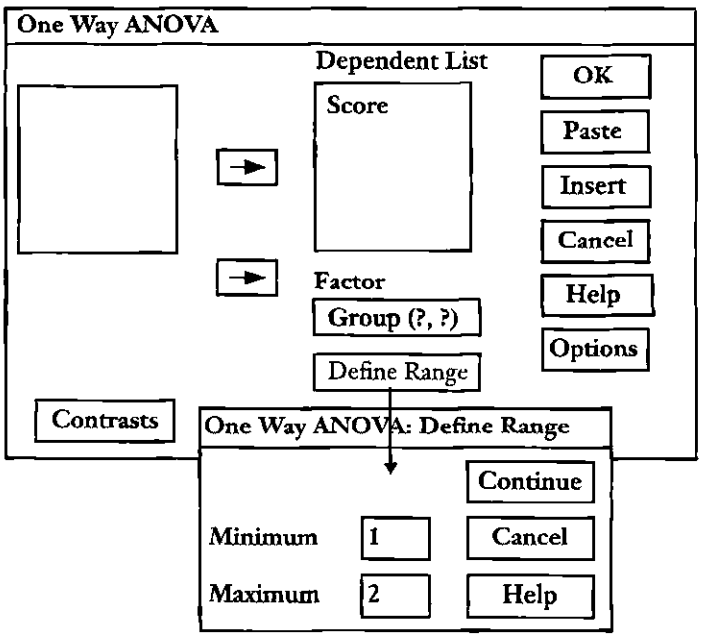
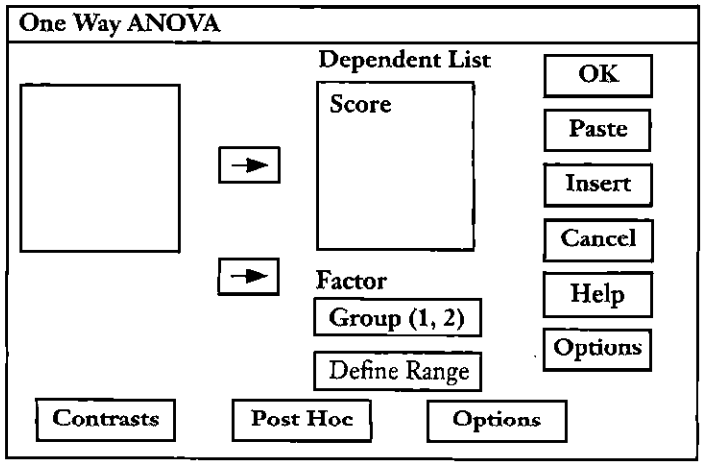


FIGURA 13.11 *Pantalla del ANOVA de un factor después de definir el rango*



queta de la variable "Group"; muévala al cuadro asociado con "Factor". Al hacer esto, el SPSS pide especificar el rango de valores de la variable dependiente. Las figuras 13.9a y 13.9b muestran lo anterior. Para definir los factores (variables independientes) haga clic en el botón etiquetado "Define Range". A partir de esta operación aparece otra pantalla, presentada en la figura 13.10. Anote los números "1" y "2" para los valores mínimo y máximo de la variable independiente. Si se tuvieran tres grupos específicos en la tabla de datos como "1, 2 y 3", se especificaría 1 como mínimo y 3 como máximo. El SPSS espera un ordenamiento sistemático de las categorías de la variable independiente. Una vez terminada la definición de rangos, haga clic en el botón "Continue" para regresar a la pantalla One way ANOVA con el rango definido para la variable independiente (véase figura 13.11). Ahora haga clic en el botón "OK" para iniciar el análisis. Note que si deseara hacer pruebas de comparación múltiple *post hoc*, tendría que hacer clic en "Post Hoc" (véase figura 13.11) antes de indicar al SPSS continuar con el análisis. Cuando se activa "Post Hoc" se presenta una pantalla que contiene una lista de las pruebas *post hoc* más utilizadas; el usuario sólo necesita seleccionar la prueba deseada.

---*ANOVA DE UN FACTOR* ----

Variable PUNTUACIÓN
Por variable GRUPO

		Análisis de varianza			
Fuente	**G.L.**	**Suma de cuadrados**	**Cuadrados medios**	**Razón F**	**Prob. F**
Entre grupos	1	22.5000	22.5000	9.0000	.0171
Dentro de grupos	8	20.0000	2.5000		
Total	9	42.5000			

Los resultados del ANOVA de un factor se presentan en la tabla anterior, los cuales coinciden con los que se realizaron con papel y lápiz. Con el SPSS no es necesario buscar

el valor crítico en la tabla de la razón F para rechazar o no la hipótesis nula, pues la significancia estadística asociada a la F aparece de manera automática en el cuadro de resultados.

Anexo

Cálculos del análisis de varianza con medias, desviación estándar y *n*

En ocasiones resulta útil poder hacer el análisis de varianza a partir de medias, desviaciones estándar y las n de los grupos, en lugar de hacerlo a partir de puntuaciones en bruto. Un método para hacerlo es el siguiente (se utilizan datos de la tabla 13.7 para ilustrar el método):

1. A partir de las n y M calcule la sumatoria de cada uno de los grupos, $\sum X_j$ y súmelas para obtener la $\sum X_t$. Calcule la N total a partir de las n de los grupos.

$$\sum X_t = \sum [M_j n_j] = (5)(6) + (5)(3) = 45; N = 5 + 5 = 10$$

2. Término de corrección: $(\sum X_t)^2/N = 45^2/10 = 202.50$ (C).
3. Calcule la suma de cuadrados dentro de grupos: el promedio de las sumas de cuadrados dentro de grupos:

$$(1.5811^2)(4) + (1.5811^2)(4) = 19.9990 = 20 = SC_d$$

4. Calcule la suma de cuadrados entre grupos:

$$SC_e = \sum [n_j M_j^2] - C$$
$$SC_e = [(6^2)(5) + (3^2)(5)] - C = 225.00 - 202.50 = 22.50$$

5. Elabore la tabla del análisis de varianza (como en la tabla 13.7) y calcule los cuadrados medios y la razón F.

Nota especial: Este método supone que las desviaciones estándar originales fueron calculadas con $n - 1$. Si se hubieran calculado con n, modifique el paso 3: $(1.4142^2)(5) + (1.4142^2)(5) = 20$; es decir, cambie los números 4 por 5, o $n - 1$ por n.

RESUMEN DEL CAPÍTULO

1. La varianza de la variable dependiente puede descomponerse en dos o más componentes.
2. Los componentes se denominan fuentes u origen de la varianza.
3. Las fuentes de la varianza sirven como base para el método estadístico conocido como análisis de varianza o ANOVA.
4. En un ANOVA de un factor, las fuentes de la varianza son aquellas entre los grupos y dentro de los grupos.
5. Una diferencia estadísticamente significativa se presenta cuando la varianza entre los grupos excede en gran cantidad a la varianza dentro de los grupos. Una tabla de la razón F se utiliza para determinar el valor crítico.

6. Se presenta una demostración con datos ficticios y reales respecto a cómo calcular los valores de un análisis de varianza.
7. La fuerza de la relación entre las variables independiente y dependiente está determinada, ya sea por η^2 o por ω^2. Estas medidas no son sensibles al tamaño de la muestra y se interpretan como r^2.
8. Cuando una prueba F es significativa y existen tres o más grupos (o niveles de la variable independiente), se requiere de pruebas de comparación múltiple para determinar qué medias son estadísticamente diferentes entre sí.
9. La prueba de Scheffé es una de las diversas pruebas de comparación múltiple. Cuando no hay un plan predeterminado con respecto a comparaciones, las pruebas se llaman *pruebas post hoc*.
10. A las comparaciones determinadas antes de realizar la prueba se les llama *comparaciones planeadas*.
11. El contenido de este capítulo introduce los conceptos necesarios para los siguientes dos capítulos concernientes al análisis de varianza.

SUGERENCIAS DE ESTUDIO

1. Existen muchas y excelentes referencias sobre el análisis de varianza, con diferentes grados de dificultad y claridad en la explicación. La discusión de Hays (1994) que incluye el modelo lineal general es, como de costumbre, excelente; pero no es fácil. Se recomienda para un cuidadoso estudio. Los siguientes cuatro libros son realmente muy recomendables; son obras primordiales de la estadística. Algunos textos se listan también en la sección de referencias, ya que fueron citados en el texto.

   Edwards, A. L. (1984). *Experimental design in psychological research* (5a. ed.). Reading, Massachusetts: Addison-Wesley.
   Hays, W. L. (1994). *Statistics* (5a. ed.). Fort Worth, Texas: Harcourt Brace.
   Kirk, R. E. (1995). *Experimental designs: Procedures for the behavioral sciences.* Pacific Grove, California: Brooks/Cole.
   Woodward, J. A., Bonett, D. G. y Brecht, M. (1990). *Introduction to linear models and experimental design.* San Diego, California: Harcourt Brace Jovanovich.

   Algunos estudiantes quizá deseen leer una historia interesante sobre el análisis de varianza, especialmente en psicología, seguida de una historia sobre el nivel .05 de significancia estadística. Para ello, se recomiendan los siguientes títulos:

   Cowles, M. (1989). *Statistics in psychology: An historical perspective.* Hillsdale, Nueva Jersey: Lawrence Erlbaum.
   Rucci, A. y Tweny, R. (1980). Analysis of variance and the second discipline of scientific psychology: A historical account. *Psychological Bulletin, 87*, 166-184.

2. Un profesor universitario conduce un experimento para probar la eficacia relativa de tres métodos de enseñanza: A_1, conferencia; A_2, discusión en grupos grandes y A_3, discusión en grupos pequeños. De un universo de estudiantes universitarios de segundo año, seleccionó aleatoriamente a 30 de ellos y los asignó a los tres grupos también de manera aleatoria. Los tres métodos fueron, a su vez, asignados aleatoriamente a los tres grupos. Se evaluó el rendimiento de los estudiantes al final de los cuatro meses que duró el experimento. Las puntuaciones de los tres grupos se presentan a continuación:

	Métodos	
A_1 (conferencia)	A_2 (discusión en grupos grandes)	A_3 (discusión en grupos pequeños)
4	5	3
7	6	5
9	3	1
6	8	4
9	3	4
6	2	5
5	5	7
7	6	3
7	7	5
10	5	3

Pruebe la hipótesis nula utilizando el análisis de varianza de un factor al nivel .01 de significancia. Calcule η^2 y ω^2. Interprete los resultados y estructure una tabla con los datos, similar a las que se presentaron en el texto.
[Respuestas: $F = 7.16(.01)$; $\eta^2 = 0.35$; $\omega^2 = 0.29$.]

3. A partir de una tabla de números aleatorios —puede utilizar aquélla en el apéndice C— obtenga tres muestras de 10 sujetos, de números entre 0 y 9.
   a) Diseñe una investigación con el planteamiento del problema y las hipótesis, e imagine que los tres conjuntos de números son sus resultados.
   b) Realice un análisis de varianza de los tres conjuntos de números. Calcule η, η^2 y ω^2. Estructure una tabla con los resultados, similar a la de la figura 13.1. Interprete los resultados estadística y sustantivamente.
   c) Añada una constante de 2 a cada una de las puntuaciones del grupo con la media más grande. De nuevo siga las instrucciones del inciso *b)*. Interprete. ¿Qué cambios ocurren en los estadísticos? [Examine las sumas de cuadrados y preste atención a las varianzas dentro de los grupos (cuadrados medios) de ambos ejemplos.]
4. Tome las puntuaciones de los grupos más alto y más bajo en la sugerencias de estudio 2 (grupos A_1 y A_2).
   a) Realice un análisis de varianza y calcule la raíz cuadrada de la F, $\sqrt{F}$. Después realice una prueba t como se describió en el capítulo 12. Compare la t obtenida con la raíz cuadrada de la $\sqrt{F}$.
   b) Después de hacer el análisis de varianza de los tres grupos, ¿es legítimo, calcular la razón t como se indicó y después extraer conclusiones acerca de las diferencias entre los dos métodos? (Consulte a su instructor si es necesario; esta pregunta es difícil.)
   [Respuestas: *a)* $F = 14.46$; $\sqrt{F} = 3.80$; $t = 3.80$; *b)* $\eta^2 = .45$; $\omega^2 = .40$.]
5. Aronson y Mills (1959) probaron la interesante y, tal vez, humanamente perversa hipótesis de que los individuos que se someten a una iniciación desagradable para convertirse en miembros de un grupo sienten mayor agrado por el grupo, que aquellos que no se sometieron a dicha iniciación. Tres grupos con 21 mujeres jóvenes cada uno fueron sujetos a tres condiciones experimentales: (i) *condición severa*, donde se les pidió a los sujetos leer palabras obscenas y descripciones vívidas de actividad sexual, para poder convertirse en miembros del grupo; (ii) *condición ligera*, en la cual los sujetos leían palabras relacionadas al sexo, pero no obscenas, y (iii) *condición control*, donde los sujetos no requerían hacer nada para convertirse en miembros del

grupo. Después de un procedimiento bastante elaborado, se les pidió a los sujetos evaluar las discusiones y a los miembros del grupo al que ahora aparentemente, pertenecían. Las medias y las desviaciones estándar de las puntuaciones totales son: *severo*, $M = 195.3$, $DE = 31.9$; *ligero*, $M = 171.1$, $DE = 34.0$; *control*, $M = 166.7$, $DE = 21.6$. Cada n fue de 21.

a) Realice un análisis de varianza con estos datos, utilizando el método explicado en el anexo de este capítulo. Interprete los datos. ¿Se apoya la hipótesis?

b) Calcule ω^2. ¿La relación es fuerte? ¿Esperaría que la relación fuera fuerte en un experimento de este tipo?

[Respuestas: *a*) $F = 5.39$ (.01); *b*) $\omega^2 = .12$.]

6. Utilice la prueba de Scheffé para calcular la significancia de todas las diferencias entre las tres medias de la sugerencia para estudio 2. Una forma de efectuar la prueba de Scheffé consiste en calcular el error estándar de las diferencias entre dos medias con la siguiente fórmula:

$$EE_{M_i - M_j} = \sqrt{CM_d\left(\frac{1}{n_i} + \frac{1}{N_j}\right)} \tag{13.6}$$

donde CM_d es el cuadrado medio dentro de los grupos y n_i y n_j representan el número de casos en los grupos i y j. Para el ejemplo, esto sería:

$$EE_{M_{A1} - M_{A2}} = \sqrt{(3.26)\left(\frac{1}{10} + \frac{1}{10}\right)} = .81$$

Después calcule el estadístico S (por Scheffé):

$$S = \sqrt{(k-1)F_{.05\,(k-1,m)}} \tag{13.7}$$

donde k es el número de grupos en el análisis de varianza, y el término F es la razón F al nivel .05, obtenida de una tabla de la razón F, con $k - 1(3 - 1 = 2)$ y $m = N - k = 30 - 3 = 27$ grados de libertad. Esto es 3.35, por lo tanto:

$$S = \sqrt{(3-1)(3.35)} = \sqrt{6.70} = 2.59$$

El paso final consiste en multiplicar los resultados de las ecuaciones 13.6 y 13.7:

$$S \times EE_{M_i - M_j} = (2.59)(.81) = 2.10$$

Para que cualquier diferencia sea estadísticamente significativa al nivel .05, ésta debe ser tan grande o mayor que 2.10. Ahora utilice el estadístico en el ejemplo.

7. Los estudios que utilizan el análisis de varianza de un factor son menos frecuentes que aquellos que utilizan otros métodos. De la siguiente lista de nueve estudios que utilizan el análisis de varianza de un factor, seleccione dos con fines de estudio. Ponga particular atención a las pruebas *post hoc* de la significancia de las diferencias entre medias.

Gibson, R. L. y Hartshorne, T. S. (1996). Childhood sexual abuse and adult loneliness and network orientation. *Child Abuse & Neglect, 20,* 1087-1093.

Goldenberg, D. e Iwasiw, C. (1993). Professional socialization of nursing students as the outcome of a senior clinical preceptorship experience. *Nurse Education Today, 13*, 3-5.

Gupta, S. (1992). Season of birth in relation to personality and blood groups. *Personality and Individual Differences, 13*, 631-633.

Jamal, M. y Baba, V. V. (1992). Shiftwork and department-type related to job stress, work attitudes and behavioral intentions: A study of nurses. *Journal of Organizational Bebavior, 13*, 449-464.

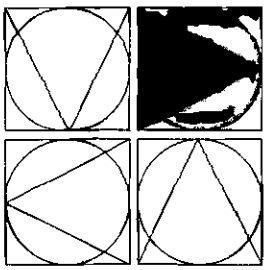
Kirsch, I., Mobayed, C. P., Council, J. R. y Kenny, D. A. (1992). Expert judgments of hypnosis from subjective state reports. *Journal of Abnormal Psychology, 101*, 675-662.

Silverstein, B. (1982). Cigarette smoking, nicotine addiction, and relaxation. *Journal of Personality and Social Psychology, 42*, 946-950.

Sonnenschein, S. (1986). Developing referential communication: Transfer across novel tasks. *Bulletin of Psychonomic Society, 24*, 127-130.

Uddin, M. (1996). College women's sexuality in an era of AIDS. *Journal of American College Health, 44*, 252-261.

Wittrock, M. (1967). Replacement and nonreplacement strategies in children's problem solving. *Journal of Educational Psychology 58*, 69-74.



CAPÍTULO 14

ANÁLISIS FACTORIAL DE VARIANZA

Ahora se estudiará el enfoque estadístico y de diseño que resume el verdadero comienzo de la perspectiva moderna sobre la investigación científica del comportamiento. La idea del diseño factorial y del análisis factorial de varianza es una de las ideas de investigación creativas propuestas en los pasados 60 años o más. Su influencia en la investigación del comportamiento contemporáneo, especialmente en psicología y educación, ha sido formidable. No es una exageración señalar que los diseños factoriales son los diseños experimentales más utilizados, y que el análisis factorial de varianza se emplea en

investigación psicológica experimental más que cualquier otro tipo de análisis. Éstas son afirmaciones importantes que requieren de una explicación; este capítulo se dedica a realizar dicha explicación, junto con la descripción y explicación de la mecánica del análisis factorial de varianza. Su importancia y complejidad hacen necesario extenderse más de lo usual sobre diferentes aspectos del tema; en otras palabras, este capítulo será más complejo que la mayoría de los demás. Por lo tanto, el lector deberá ser persistente, paciente y tolerante, sabiendo que es por una buena causa. Primero se examinarán dos ejemplos de investigación que resultan muy ilustrativos.

Dos ejemplos de investigación

El prejuicio es un fenómeno sutil y profundo. Una vez que surge, penetra grandes partes del pensamiento. Es un hecho obvio que el prejuicio negativo en contra de las minorías es un fenómeno potente y muy extendido. ¿El prejuicio es tan penetrante y sutil que puede funcionar a "la inversa"? ¿La gente que se considera a sí misma libre de prejuicio discrimina positivamente a las minorías? ¿Existe algo como un "prejuicio inverso"? Las compañías y universidades que contratan mujeres y afroamericanos, ¿lo hacen por un prejuicio inverso, o sólo porque resulta un buen negocio? Preguntas como éstas son, por supuesto, fáciles de formular; aunque no son fáciles de responder —al menos no científicamente—.

En un estudio revelador y un tanto desconcertante, Dutton y Lake (1973) hipotetizaron que si las personas se sienten amenazadas por la idea de que tal vez son prejuiciosas, actuarán de manera discriminatoria inversa hacia los miembros de grupos minoritarios; en otras palabras, sí discriminarán, pero favorablemente.

De una población de 500 estudiantes universitarios, 40 hombres y 40 mujeres que se habían autoevaluado como relativamente libres de prejuicios en un cuestionario previo al experimento, fueron asignados a dos condiciones experimentales: "amenaza" y "raza", divididos en alta y baja amenaza; y en pordiosero afroamericano y americano blanco. Por lo tanto, éste representa el diseño factorial más simple posible, llamado de dos por dos (2×2). Éste se presenta en la tabla 14.1, con las medias de la variable dependiente, representada por el dinero (centavos) dado a un pordiosero. Observe que esta tabla de 2×2 se parece a las tablas de contingencia de 2×2, revisadas en el capítulo 10. Sin embargo, en esencia son diferentes y el estudiante debe entender con claridad esa diferencia: las tablas de contingencia incluyen frecuencias o porcentajes en las casillas; mientras que el análisis factorial utiliza medidas de la variable dependiente, generalmente medias, en las casillas. La variable dependiente siempre es una de las variables en los márgenes (fuera) de la *tabla*

TABLA 14.1 *Diseño factorial 2 × 2 del experimento de discriminación inversa de Dutton y Lake*[a]

	Amenaza		
Raza	**Alta amenaza**	**Baja amenaza**	
Pordiosero afroamericano	47.25	16.75	32.00
Pordiosero americano blanco	28.25	27.75	28.00
	37.75	22.25	

[a] Los números en las casillas representan medias de los centavos dados a los pordioseros. El diseño original incluyó sexo, pero tal variable se omitió aquí.

de contingencia; en los diseños factoriales la variable dependiente siempre es la medida dentro de las casillas.

Dutton y Lake supusieron que la discriminación inversa podría ocurrir si a los participantes que se consideraban a sí mismos como no prejuiciosos se les hacía sospechar que en realidad sí lo eran. Esta sospecha representaría una amenaza a sí mismos, y un sujeto que experimentara tal amenaza, bajo las condiciones apropiadas, actuaría por discriminación inversa. A los participantes en el grupo de alta amenaza se les dijo que habían mostrado una alta activación emocional —supuestamente medida por medio de la respuesta galvánica de la piel y por la frecuencia del pulso— al observar diapositivas con escenas interraciales. A los participantes del grupo de baja amenaza no se les dio retroalimentación respecto a las diapositivas. La condición experimental se indica en la parte superior del diseño, en la tabla 14.1.

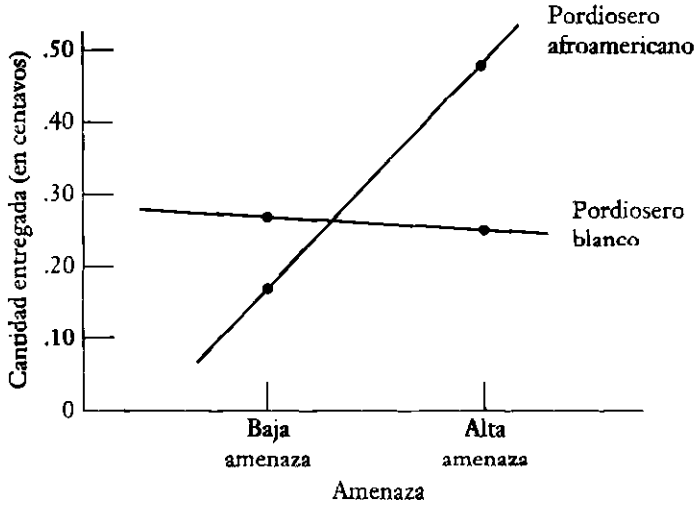
La segunda variable experimental, "raza", se manipuló de la siguiente manera: después de completar la manipulación de la variable *amenaza*, a los sujetos se les pagó con monedas de 25 centavos y se les informó que podían irse. A la salida del laboratorio, un cómplice afroamericano se dirigió a la mitad de los sujetos y un cómplice americano blanco a la otra mitad, para formularles la siguiente pregunta: "¿Podría darme algunas monedas para comprar comida?" Esta segunda variable experimental, "raza", se presenta en el margen lateral de la tabla 14.1, con sus dos niveles: pordiosero afroamericano y pordiosero blanco. Se predijo que los sujetos del grupo de alta amenaza darían más dinero al pordiosero afroamericano que al americano blanco, ya que se supuso que los sujetos del grupo de alta amenaza reaccionarían contra la idea de que eran prejuiciosos, como lo había sugerido el polígrafo de la condición experimental, dando más dinero al pordiosero afroamericano. Los sujetos del grupo de baja amenaza no darían la misma cantidad de dinero puesto que a ellos no se les hizo dudar respecto a su falta de prejuicios. En otras palabras, habría una diferencia entre los grupos de amenaza, respecto al dinero dado al pordiosero afroamericano; pero no existiría una diferencia entre los grupos de amenaza, respecto al dinero ofrecido al pordiosero blanco. Al resultado predicho se le conoce como *interacción*, término que se explicará posteriormente con mayor detalle.

Los datos de la tabla 14.1, tomados de los datos más extensos reportados por Dutton y Lake, parecen apoyar la hipótesis. Las medias del grupo de alta amenaza, en contraposición con la del grupo de baja amenaza en la condición del pordiosero afroamericano fueron 47.25 y 16.75 centavos; mientras que las medias en la condición del pordiosero blanco fueron 28.25 y 27.75. El análisis estadístico indicó que los resultados hipotetizados fueron tal como los autores indicaron que serían. Se trata de enfatizar la naturaleza de los datos obtenidos al graficar las medias en la figura 14.1. Los puntos graficados —indicados por los pequeños círculos negros— son las medias de la tabla 14.1. El eje horizontal representa la "amenaza". Puesto que solamente hay dos "valores", su ubicación en la línea es casi arbitraria. El eje vertical representa la cantidad de dinero entregada al pordiosero.

La relación es notoria: al pordiosero afroamericano le dieron más dinero en la condición de alta amenaza que en la condición de baja amenaza; mientras que virtualmente no existe diferencia entre las dos condiciones de amenaza con el pordiosero blanco. En efecto, se apoya la hipótesis de interacción: se presentó discriminación invertida y, quizá, se puede afirmar que "existe" el prejuicio inverso.

En un interesante estudio experimental sobre los efectos de dos variables, autoexpresión y composición genérica del grupo, Elias (1989) encontró que ambas tenían un efecto sobre la cohesión del grupo, el compromiso con la tarea y la productividad. Elias asignó aleatoriamente a cada uno de los 144 estudiantes universitarios (72 mujeres y 72 hombres) a uno de 36 grupos. Cada grupo se componía de 4 miembros. De los 36 grupos, 12 incluían sólo hombres, 12 sólo mujeres y 12 eran mixtos (hombres y mujeres). Seis grupos

FIGURA 14.1



de cada categoría del género (hombres, mujeres, mixto) fueron asignados aleatoriamente a la condición experimental (autoexpresión) o a la condición control (sin autoexpresión). Todos los grupos completaron la tarea simple de armar un rompecabezas como medida de la productividad y contestaron cuestionarios para evaluar la cohesión y el compromiso con la tarea. A los sujetos se les indicó que no debían comunicarse verbalmente con los demás miembros del grupo y que podían dar piezas del rompecabezas a los otros miembros. En los grupos de autoexpresión, los miembros participaron en una discusión grupal después de completar el rompecabezas; la discusión se centró en los hechos y sentimientos relevantes a la tarea del rompecabezas. Se utilizaron tarjetas con señales de autoexpresión para facilitar la discusión. El grupo control observó una cinta con escenas de la naturaleza y se les indicó que no debían comunicarse entre sí. Después de esto, ambos grupos participaron en una segunda tarea de resolución de un rompecabezas. La cantidad de tiempo empleada para resolver el segundo rompecabezas sirvió como medida de la productividad. Se aplicó un cuestionario para medir la cohesión y el compromiso con la tarea. Los datos mostraron que la intervención de autoexpresión resultó en una mayor cohesión, compromiso con la tarea y productividad. Resulta más fácil comprender esto si se observa la tabla 14.2. Una variable es la composición genérica del grupo, la cual se dividió en "hombres", "mujeres" y "mixto". La otra variable, "expresión", se dividió en "autoexpresión" y "grupo control" (o no autoexpresión). Tales fueron las condiciones experimentales. Las variables dependientes son: cohesión, compromiso con la tarea y productividad. Respecto a cohesión y a compromiso con la tarea, ambas variables independientes tuvieron un efecto estadísticamente significativo, como lo indican sus respectivas medias combinadas. En estas dos variables dependientes, las mujeres reportaron una mayor cohesión (al analizar la tabla 14.2 debe notarse que sólo para la variable cohesión una menor puntuación denota mayor cohesión) y compromiso con la tarea, que los grupos de hombres o los grupos mixtos. En cuanto a la productividad grupal, autoexpresión *versus* control tuvo un efecto estadísticamente significativo, lo que no ocurrió con composición genérica del grupo.

TABLA 14.2 *Diseño factorial de 2 × 3 y resultados (medias) del estudio de Elias*[a]

	Composición genérica del grupo			
Cohesión	**Mujeres**	**Hombres**	**Mixto**	**Media combinada**
Autoexpresión	14.20	15.96	15.61	15.25
Grupo control	15.92	19.08	17.75	17.58
Media combinada	15.06	17.52	16.68	
Compromiso con la tarea	**Mujeres**	**Hombres**	**Mixto**	**Media combinada**
Autoexpresión	77.13	71.80	71.88	73.60
Grupo control	72.08	65.50	68.88	68.68
Media combinada	74.60	68.85	70.17	
Productividad	**Mujeres**	**Hombres**	**Mixto**	**Media combinada**
Autoexpresión	149.17	71.17	143.67	121.00
Grupo control	193.00	217.17	310.17	240.11
Media combinada	171.09	144.17	226.92	

[a] Puntuaciones bajas indican mayor cohesión.

La naturaleza del análisis factorial de varianza

En el análisis factorial de varianza dos o más variables independientes varían de manera independiente o interactúan entre sí para generar una variación en una variable dependiente. *El análisis factorial de varianza es el método estadístico que analiza los efectos independientes e interactivos de dos o más variables independientes sobre una variable dependiente.*

Si se trata de dos variables independientes, como en el ejemplo que se acaba de discutir, el modelo lineal en cuestión es una extensión del modelo lineal del capítulo anterior:

$$y = a_o + A + B + AB + e \qquad \textbf{(14.1)}$$

donde y, como siempre, es una puntuación de un individuo en la variable dependiente; a_o es el término común para todos los individuos, por ejemplo, la media general; A es el efecto de una variable independiente; B es el efecto de otra variable independiente; AB es el efecto de ambas variables trabajando juntas o interactuando, y e es el error. Además del efecto particular de una variable (A) y del error (e) en el análisis de varianza de un factor, ahora se tiene un segundo efecto (B) y un tercer "efecto" que es el trabajo o influencia que en conjunto ejercen A y B o AB sobre y. No existe límite teórico respecto al número de variables independientes en los diseños factoriales. Aquí se presenta el modelo para tres variables independientes:

$$y = a_o + A + B + C + AB + AC + BC + ABC + e \qquad \textbf{(14.2)}$$

Aquí hay tres variables independientes, A, B y C, sus interacciones AB, AC y BC, y la interacción simultánea de las tres, ABC. Tan complejo como este modelo pueda parecer, en la literatura existen muchas aplicaciones de él (se darán ejemplos posteriormente); y

también se pueden añadir más variables independientes. Las únicas limitaciones son de tipo práctico: cómo manejar tantas variables a la vez y cómo interpretar las interacciones, en especial las triples y las cuádruples. Sin embargo, lo que se busca aquí son las ideas básicas que subyacen a los modelos y diseños factoriales.

Uno de los acontecimientos más significativos y revolucionarios en el diseño de investigación moderno y en la estadística consiste en la planeación y el análisis de la operación e interacción simultáneas de dos o más variables. Desde hace mucho tiempo los científicos saben que las variables no actúan de forma independiente, sino que lo hacen de forma conjunta. La virtud de un método de enseñanza, en contraste con otro método, depende de los maestros que los utilicen. El efecto educativo de cierto tipo de maestro depende, en gran medida, del tipo de alumno a quien enseña. Un maestro ansioso puede ser muy efectivo con alumnos ansiosos; pero menos efectivo con alumnos no ansiosos. Los diferentes métodos de enseñanza en las universidades dependen de la inteligencia y personalidad tanto de los maestros como de los estudiantes. En el estudio de Dutton y Lake (1973), el efecto de la amenaza dependió de la raza del pordiosero (véase tabla 14.1 y figura 14.1). En el estudio de Elias (1989) la interacción fue diferente; no hubo interacciones en ninguno de los análisis de las tres variables dependientes. El efecto conjunto de las variables independientes —expresión y composición genérica del grupo— fue acumulativo; el efecto fue más fuerte cuando ambas estuvieron presentes (tabla 14.2).

Antes de la invención del análisis de varianza y de los diseños sugeridos por el método, la postura tradicional en la investigación experimental consistía en estudiar el efecto de una variable independiente sobre una variable dependiente. Aquí no se está afirmando, por cierto, que dicho método sea erróneo, sino tan sólo que es limitado; no obstante, muchas preguntas de investigación pueden contestarse adecuadamente utilizando este esquema de "uno a uno". Muchas otras preguntas de investigación pueden responderse adecuadamente sólo si se consideran las influencias múltiples e interactivas. Los científicos educativos sabían que el estudio de los efectos de distintos métodos y técnicas pedagógicas sobre los resultados educativos era, en parte, función de otras variables como la inteligencia de los estudiantes, la personalidad de los maestros, los antecedentes sociales de los maestros y de los estudiantes, y el ambiente general de la clase y de la escuela. Sin embargo, muchos investigadores consideraban que el método de investigación más efectivo consistía en hacer variar una variable independiente, mientras se controlaban lo mejor posible otras variables independientes que podían estar contribuyendo a la varianza de la variable dependiente. Simon (1976; 1987) discrepa con dicho esquema tradicional y recomienda el uso de diseños multifactoriales económicos. Estos diseños, no obstante, requieren de una cuidadosa planeación y ejecución del experimento; pero pueden brindar información útil sobre un gran número de variables.

En los estudios que se resumieron antes, las conclusiones van más allá de las simples diferencias entre efectos o grupos. Fue posible calificar las conclusiones de maneras importantes a causa de que los autores estudiaron los efectos simultáneos de las dos variables independientes y, en consecuencia, fueron capaces de hablar del *efecto diferencial* de sus variables. Ellos podían afirmar, por ejemplo, que el tratamiento A_1 es efectivo cuando se combina con el nivel B_1, pero no es efectivo cuando se presenta solo o cuando se combina con el nivel B_2 y que, quizás, A_2 resulta efectivo sólo cuando se combina con B_1.

La lógica implícita detrás de este tipo de pensamiento científico puede comprenderse mejor al regresar a las proposiciones y pensamientos condicionales de un capítulo previo. Recuerde que un enunciado condicional toma la forma "si p entonces q" o "si p entonces q, bajo las condiciones r y s". En notación lógica: $p \rightarrow q$ y $p \rightarrow q \mid r,s$. Esquemáticamente, la proposición condicional detrás de los problemas de análisis de varianza de un factor del capítulo 13, es la proposición simple: si p entonces q. En el estudio de Hurlock: si ciertos

incentivos, entonces cierto rendimiento. En el estudio de Aronson y Mills (véase la sugerencia para estudio número 5, capítulo 13), si severidad en la iniciación, entonces agrado por el grupo.

Los enunciados condicionales asociados con los problemas de investigación de este capítulo son, sin embargo, más complejos y sutiles: si p entonces q, bajo las condiciones r y s, o, $p \rightarrow q \mid r,s$, donde "$\mid$" significa "bajo la(s) condición(es)". En el estudio de Dutton y Lake (1973) el enunciado condicional sería $p \rightarrow q > \mid r$; o, si amenaza entonces discriminación inversa, bajo las condiciones de que el objetivo (el pordiosero) sea afroamericano. Aunque estructuralmente similar, la lógica "acumulativa" de Elias (1989) resulta diferente: si p y r, entonces q; o, si autoexpresión y composición genérica del grupo, entonces mayor será la cohesión y el compromiso con la tarea; o, en símbolos lógicos: $(p \cap r) \rightarrow q$ (léase: si p y r, entonces q). Aquí no puede decirse "bajo la condición", porque p y q (autoexpresión y composición genérica del grupo) son copartícipes y se combinan para afectar la cohesión y el compromiso con la tarea. En otro estudio que se analizará más adelante en este capítulo, Martin y Seneviratne (1997) plantean la proposición: si hambriento, entonces sobrevienen dolores de cabeza.

El significado de la interacción

Interacción es la acción conjunta de dos o más variables independientes en su influencia sobre una variable dependiente. Siendo más precisos, interacción significa que la operación o influencia de una variable independiente sobre una variable dependiente depende del nivel de otra variable independiente. Ésta es una manera un poco torpe de decir lo que se expresó antes al hablar de los enunciados condicionales; por ejemplo, si p entonces q, bajo la condición r. En otras palabras, la interacción ocurre cuando una variable independiente tiene diferentes efectos sobre una variable dependiente, con diferentes niveles de otra variable independiente.

Esa definición de interacción comprende dos variables independientes y se le llama una interacción de primer orden. Es posible que tres variables independientes interactúen en su influencia sobre una variable dependiente; ésta es una interacción de segundo orden. Son posibles interacciones de orden mayor; pero interpretarlas se vuelve muy difícil; a diferencia de las interacciones de primer orden que se han presentado aquí en una figura bidimensional. Las interacciones de orden mayor son difíciles de visualizar y graficar. Cuando se tiene un efecto de interacción significativo, se sabe que hay una diferencia en los tratamientos. Sin embargo, para determinar exactamente cómo difieren los tratamientos, se necesitaría examinar los niveles de las otras variables independientes. Para predecir el resultado del tratamiento para un solo individuo, la predicción únicamente puede realizarse si se conoce la situación de ese individuo en todas las variables independientes. Algunos autores de libros de texto incluso han llegado a decir que los efectos de interacciones de orden superior son carentes de importancia. Esto puede ser verdadero si el estudio se diseña apropiadamente; pero puede no serlo para todos los estudios. En una breve inspección de bastantes libros de estadística intermedia y avanzada, utilizados a nivel de posgrado, el análisis sobre la interpretación de los efectos de interacciones de orden superior es muy escasa (véase Hays, 1994; Kirk, 1995; Howell, 1997). Sin embargo, los trabajos de Daniel (1976) y Simon (1976) sugieren cómo manejar los efectos de interacción de orden superior. Antes de pasar a los aspectos computacionales, el lector debe estar consciente de que la interacción puede ocurrir en la ausencia de los efectos separados de las variables independientes. (La interacción también puede estar ausente cuando una o más variables independientes tienen efectos significativos separados.) A los efectos separados de las variables

independientes se les denomina *efectos principales*. Ahora se mostrará esta posibilidad utilizando un ejemplo ficticio y después utilizando un ejemplo proveniente de investigación publicada.

Un ejemplo ficticio simple

Como siempre, se utiliza un ejemplo simple aunque no realista que resalta los problemas y características básicas del análisis factorial de varianza. Suponga que un investigador educativo está interesado en la eficacia relativa de dos métodos de enseñanza: A_1 y A_2. A esta variable se le llamará *métodos*. El investigador considera que los métodos de enseñanza no difieren mucho entre sí, sino que solamente difieren cuando se utilizan con cierto tipo de estudiantes, por cierto tipo de maestros, en ciertas situaciones educativas y por cierta clase de motivos. Estudiar todas estas variables de forma simultánea implica un orden alto, pero no necesariamente imposible. Así, se toma la decisión de estudiar los métodos y las motivaciones, lo que representa dos variables independientes y una dependiente. La variable dependiente se llama desempeño y se utilizará algún tipo de medida de rendimiento, quizás las puntuaciones en una prueba estandarizada.

El investigador conduce un experimento con ocho niños de sexto grado (un experimento real se realizaría con mucho más de ocho niños) y asigna aleatoriamente a los ocho niños a cuatro grupos, dos por grupo. También asigna aleatoriamente los métodos A_1 y A_2 y las motivaciones B_1 y B_2 para los cuatro grupos. Recuerde el análisis previo sobre la partición de conjuntos: es posible dividir y subdividir conjuntos de objetos. Los objetos pueden asignarse a una división o subdivisión con base en la posesión de ciertas características; pero también pueden asignarse aleatoriamente —y después el experimentador les "asigna", supuestamente, ciertas características—. En cualquiera de los dos casos la lógica de esta división es la misma. El experimentador terminará con cuatro subparticiones: A_1B_1, A_1B_2, A_2B_1 y A_2B_2. El paradigma experimental se ilustra en la figura 14.2.

Cada casilla en el diseño representa la intersección de dos subconjuntos. Por ejemplo, el método A_1 combinado con la motivación B_2 conceptualmente es $A_1 \cap B_2$. El método A_2 combinado con la motivación B_2 es la intersección $A_2 \cap B_2$. En tal diseño por simplicidad se anota solamente A_1B_2 y A_2B_2. Ahora, se han asignado aleatoriamente dos niños a cada una de las cuatro casillas; lo cual quiere decir que cada niño recibirá una combinación de dos manipulaciones experimentales, y que cada par de niños recibirá una combinación diferente.

Llame *recitación* a A_1 y *no recitación* a A_2; *elogio* a B_1 y *crítica* a B_2. Después a los niños de las casillas A_1B_1 se les enseñará a recitar y serán elogiados por su trabajo. A los niños de la casilla A_1B_2 se les enseñará a recitar pero serán criticados por su trabajo; se realizaría algo similar para las otras dos casillas. Si los procedimientos experimentales han sido manejados adecuadamente es posible considerar a las variables como independientes, es decir, que dos experimentos separados en realidad se efectúan con los mismos participantes. Un experimento manipula los *métodos*; el otro, los tipos de *motivaciones*. En otras palabras, el

FIGURA 14.2

		Métodos	
		A_1	A_2
Motivaciones	B_1	A_1B_1	A_2B_1
	B_2	A_1B_2	A_2B_2

TABLA 14.3 *Datos del experimento factorial hipotético con los cálculos del análisis de varianza*

	Métodos	
Tipos de motivación	A_1	A_2
B_1	8, 6	4, 2
B_2	8, 6	4, 2

		Métodos		
Tipos de motivación		A_1	A_2	
B_1	ΣX	14	6	$\Sigma X_{B_1} = 20$
	$(\Sigma X)^2$	196	36	$(\Sigma X_{B_1})^2 = 400$
	M	7	3	$M_{B_1} = 5$
B_2	ΣX	14	6	$\Sigma X_{B_2} = 20$
	$(\Sigma X)^2$	196	36	$(\Sigma X_{B_2})^2 = 400$
	M	7	3	$M_{B_2} = 5$
		$\Sigma X_{A_1} = 28$	$\Sigma X_{A_2} = 12$	$\Sigma X_t = 40$
		$(\Sigma X_{A_1})^2 = 784$	$(\Sigma X_{A_2})^2 = 144$	$(\Sigma X_t)^2 = 1\,600$
		$M_{A_1} = 7$	$M_{A_2} = 3$	$M_t = 5$
				$\Sigma X_t^2 = 240$

diseño del experimento permite al investigador probar *de forma independiente* los efectos de 1) *método* y 2) *tipo de motivación* sobre una variable dependiente, en este caso el desempeño. Para mostrar ésta y otras importantes facetas de los diseños factoriales, ahora se analizarán los datos ficticios del experimento. Tales "datos" se reportan en la tabla 14.3 junto con los cálculos necesarios para el análisis factorial de varianza. Primero se calculan las sumas de cuadrados, como se haría en un análisis de varianza de un factor. Existe, por supuesto, una suma de cuadrados *total* calculada a partir de todas las puntuaciones, utilizando C, el término de corrección:

$$C = \frac{(40)^2}{8} = \frac{1\,600}{8} = 200$$

o

$$C = M^2(N) = 5^2(8) = 200$$

$$\text{Total} = 240 - 200 = 40$$

o

$$\text{Total} = DE^2(N) = \left[\frac{240 - \frac{40^2}{8}}{8}\right](8) = 40$$

Puesto que hay cuatro grupos, existe una suma de cuadrados asociada con las medias de los cuatro grupos. Tan sólo se consideran los cuatro grupos ubicados lado a lado como en el

análisis de varianza de un factor, y se calcula la suma de cuadrados como en el capítulo anterior. Sin embargo, ahora se le llama suma de cuadrados *entre grupos* para distinguirla de las sumas de cuadrados que se calcularán más adelante.

$$SC \text{ entre grupos} = \sum \frac{(\Sigma X)^2}{n_i} - C$$

$$SC \text{ entre grupos} = \left(\frac{196}{2} + \frac{36}{2} + \frac{196}{2} + \frac{36}{2}\right) - 200 = 32$$

Esta suma de cuadrados es una medida de la variabilidad de las cuatro medias grupales; por lo tanto, si se resta tal cantidad de la suma de cuadrados total se debe obtener la suma de cuadrados debida al error, las fluctuaciones aleatorias de las puntuaciones dentro de las casillas (grupos). Lo anterior resulta familiar: es la suma de cuadrados *dentro de grupos*:

$$SC \text{ dentro de grupos} = 40 - 32 = 8$$

Para calcular la suma de cuadrados para métodos, se procede exactamente igual que en el análisis de varianza de un factor: se trata a las puntuaciones (X) y a las sumas de las puntuaciones (ΣX) de las columnas (métodos), como si no hubiera tipos de motivación B_1 y B_2:

Métodos

	A_1	A_2
	8	4
	6	2
	8	4
	6	2
ΣX	28	12

El cálculo es el siguiente:

$$\text{Entre métodos } (A_1, A_2) = \left(\frac{(28)^2}{4} + \frac{(12)^2}{4}\right) - 200$$

$$= \left(\frac{784}{4} + \frac{144}{4}\right) - 200 = 32$$

De manera similar, se tratan los tipos de motivación (B_1 y B_2) como si no hubiera métodos:

Motivación		**ΣX**
B_1	8 6 4 2	20
B_2	8 6 4 2	20

El cálculo de la suma de cuadrados entre tipos no es realmente necesario. Puesto que las sumas (y las medias) son las mismas, la suma de cuadrados entre tipos es cero:

$$\text{Entre tipos } (B_1, B_2) = \left[\frac{(20)^2}{4} + \frac{(20)^2}{4}\right] - 200 = 0$$

Existe otra posible fuente de varianza, la varianza debida a la *interacción* de las dos variables independientes. La suma de cuadrados entre todos los grupos incluye la variabilidad debida a las medias de los cuatro grupos: 7, 3, 7 y 3. La suma de cuadrados es 32. Si éste no fuese un ejemplo inventado, parte de dicha suma de cuadrados se debería a los métodos, parte al tipo de motivación y una parte restante *debida a la acción conjunta o interacción* de los métodos y los tipos. En muchos casos sería relativamente pequeña, no mayor que lo esperado por el azar. En otros casos sería lo bastante grande para ser estadísticamente significativa; excedería la expectativa por el azar. En el problema presente claramente es cero, ya que la suma de cuadrados entre métodos fue 32, lo que es igual a la suma de cuadrados entre todos los grupos. Para completar los cálculos:

Interacción: métodos × tipos = entre todos los grupos
$$- (\text{entre métodos} + \text{entre tipos}) = 32 - (32 + 0) = 0$$

Note que en los análisis factoriales de varianza más complejos, las interacciones no resultan tan fáciles de calcular. El lector debe consultar a Hays (1994) o a Kirk (1995) para mayor información. Ahora ya es posible elaborar la tabla final del análisis de varianza; aunque esto debe posponerse hasta realizar una operación menor sobre estas puntuaciones.

Se utilizan exactamente las mismas puntuaciones, aunque se reordenan un poco: se invierten las puntuaciones A_1B_2 y A_2B_2. Puesto que todas las puntuaciones individuales (X) son exactamente las mismas, la suma de cuadrados total debe también ser exactamente la misma. Además, las sumas y las sumas de cuadrados de B_1 y B_2 (tipos) deben también ser exactamente las mismas. La tabla 14.4 muestra lo que se realizó y su efecto sobre las medias de los cuatro grupos.

TABLA 14.4 *Datos de un experimento factorial hipotético de la tabla 14.3 con B_2. Números reordenados*

	Métodos		
Tipo de motivación	A_1	A_2	
B_1	8	4	
	6	2	
ΣX	14	6	$\Sigma X_{B_1} = 20$
M	7	3	$M_{B_1} = 5$
B_2	4	8	
	2	6	
ΣX	6	14	$\Sigma X_{B_2} = 20$
M	3	7	$M_{B_2} = 5$
ΣX_A	20	20	$\Sigma X_t = 40$
M_A	5	5	$M_t = 5$
			$\Sigma X_t^2 = 240$

TABLA 14.5 *Medias de los datos de las tablas 14.3 y 14.4*

	Medias de la tabla 14.3				Medias de la tabla 14.4		
	A_1	A_2			A_1	A_2	
B_1	7	3	5	B_1	7	3	5
B_2	7	3	5	B_2	3	7	5
	7	3			5	5	

Si se estudian los números de las tablas 14.3 y 14.4 se notarán las diferencias. Para enfatizar las diferencias, las medias aparecen en **negritas** en ambas tablas. Para volver aún más claras las diferencias, se presentan las medias de ambas tablas en la tabla 14.5. La tabla de la izquierda presenta dos fuentes de variación: aquellas entre las cuatro medias, y entre las medias de A_1 y de A_2. En la tabla de la derecha solamente hay una fuente de variación, aquella entre las cuatro medias. En ambas tablas la variabilidad de las cuatro medias es la misma, ya que las dos poseen las mismas cuatro medias: 7, 3, 7 y 3. De hecho, no hay variabilidad de las medias de B en ambas tablas. Entonces existen dos diferencias entre las tablas: las medias de A y el arreglo de las cuatro medias dentro de los recuadros. Si se analiza la suma de cuadrados de las cuatro medias (las sumas de cuadrados entre todos los grupos), se encuentra que B_1 y B_2 no contribuyen en absoluto en ambas tablas, ya que no hay variabilidad entre 5 y 5, las medias de B_1 y B_2. En la tabla de la derecha, las medias de A_1 y A_2, 5 y 5, no contribuyen a la variabilidad. Sin embargo, en la tabla de la izquierda las medias A_1 y A_2 difieren considerablemente, 7 y 3; por lo tanto, sí contribuyen a la varianza.

Si se asume por el momento que las medias de 7 y 3 difieren significativamente, se puede afirmar que los métodos de la tabla 14.3 tienen un efecto, sin tomar en cuenta el tipo de motivación. Esto es, $\mu_{A1} \neq \mu_{A2}$ o $\mu_{A1} > \mu_{A2}$. En lo que concierne a este experimento, los métodos difieren significativamente *sin importar el tipo de motivación*. De hecho, el tipo de motivación no tuvo efecto alguno, ya que $\mu_{B1} = \mu_{B2}$. Por otro lado, en la tabla 14.4 la situación es bastante diferente: ni los métodos ni el tipo de motivación tuvieron un efecto *por sí mismos*; pero aún así hay varianza. El problema es: ¿cuál es la fuente de la varianza? Es la ***interacción de las dos variables***, la interacción de los métodos y los tipos de motivación.

Si se hubiera realizado un experimento y se hubieran obtenido datos como los de la tabla 14.4, entonces se llegaría a la posible conclusión de que hubo una interacción del efecto de las dos variables sobre la variable dependiente. En ese caso, los resultados se interpretarían de la siguiente manera: los métodos A_1 y A_2, al operar por sí mismos, no difieren en su efecto. Los tipos de motivación B_1 y B_2, por sí mismos, no difieren en su efecto. Cuando a los métodos y al tipo de motivación se les permite "actuar juntos", si se les permite interactuar, existen diferencias significativas en su efecto. Específicamente, el método A_1 resulta superior al método A_2, cuando se combina con el tipo de motivación B_1. Al combinarse con el tipo de motivación B_2, resulta inferior a A_2. Este efecto de interacción está indicado en el lado derecho de la tabla 14.5 con las flechas cruzadas. Al interpretar cualitativamente los métodos originales se encuentra que la recitación parece ser superior a la no recitación bajo las condiciones de elogio; pero que es inferior a la no recitación bajo la condición de crítica.

Antes de continuar, es ilustrativo notar que la interacción puede estudiarse y calcularse mediante un procedimiento de sustracción. En un diseño de 2×2 este procedimiento es simple. Se resta una media de la otra en cada renglón y, después, se calcula la varianza de estas diferencias. Considere las medias ficticias de la tabla 14.5: si se restan las medias de la tabla 14.3, se obtiene $7 - 3 = 4$; $7 - 3 = 4$. Con claridad se ve que el cuadrado medio es cero y, por lo tanto, la interacción es cero. Si se sigue el mismo procedimiento con las medias

de la tabla 14.4 (parte derecha de la tabla): 7 – 3 = 4; 3 – 7 = – 4. Si ahora se trata a estas dos diferencias como se trató a las medias en el capítulo anterior, y se calculan la suma de cuadrados y el cuadrado medio, se llega a la suma de cuadrados y al cuadrado medio de la interacción, 32 en cada caso. La lógica detrás de este procedimiento es simple: si no hubiesen interacciones, se esperaría que las diferencias entre las medias de los renglones fueran aproximadamente iguales entre sí y respecto a la diferencia entre las medias en la parte inferior de la tabla, las medias de los métodos en este caso. Note que eso sucede con las medias de la tabla 14.3: la diferencia del último renglón es 4, al igual que las diferencias de cada uno de los renglones. Sin embargo, las diferencias entre los renglones de la tabla 14.4 se desvían de la diferencia entre las medias del último renglón (métodos). Éstas son 4 y –4; mientras que la diferencia del último renglón es 5 – 5 = 0. A partir de esta discusión y un poco de reflexión, puede verse que una interacción significativa puede ser causada por un renglón desviado. Por ejemplo, las medias del ejemplo anterior podrían ser:

7	3	5
5	5	5
6	4	

Se restan los renglones; 7 – 3 = 4; 5 – 5 = 0, y 6 – 4 = 2; de hecho, existe algo de varianza en tales residuos.

Es útil anotar las tablas finales de los análisis de varianza, donde se calcularon las diferentes varianzas y las razones *F*. La tabla 14.6 incluye las tablas finales de los análisis de varianza de los dos ejemplos. Las sumas de cuadrados entre los grupos no fueron incluidas en la tabla, tan sólo son útiles para calcular las sumas de cuadrados dentro de grupos. Los grados de libertad para los efectos principales (métodos y tipos), y para aquellos entre grupos y dentro de grupos, se calculan de la misma forma que en el análisis de varianza de un factor. Lo anterior resulta obvio al estudiar la tabla. Los grados de libertad de la interacción son el producto de los grados de libertad de los efectos principales, es decir, 1 × 1 = 1. Si la variable métodos tuviera cuatro grupos y tipos tuviera tres grupos, entonces los grados de libertad de la interacción hubieran sido 3 × 2 = 6.

La suma de cuadrados, el cuadrado medio y la razón *F* resultante de 16 en la parte izquierda de la tabla, indican lo que ya se sabía del análisis previo: los métodos son significativamente diferentes (al nivel .05) y los tipos de motivación y la interacción no son significativos. Los números semejantes en la parte derecha de la tabla indican que solamente la interacción es significativa.

TABLA 14.6 *Tablas finales de los análisis de varianza: datos de las tablas 14.3 y 14.1*

		Datos de la tabla 14.3			*Datos de la tabla 14.4*		
Fuente	***gl***	***sc***	***cm***	***F***	***sc***	***cm***	***F***
Entre métodos (A_1, A_2)	1	32	32	16(.05)	0	0	
Entre tipos (B_1, B_2)	1	0	0		0	0	
Interacción $A \times B$	1	0	0		32	32	16(.05)
Dentro de grupos	4	8	2		8	2	
Totales	7	40			40		

Interacción: un ejemplo

En el capítulo anterior se indicó que si el muestreo era aleatorio, las medias de los k grupos serían aproximadamente iguales. Si, por ejemplo, hubiera cuatro grupos y la media general M_t fuera 4.5, entonces se esperaría que cada una de las medias fuese aproximadamente 4.5. De la misma forma, si en el análisis factorial de varianza se extraen muestras aleatorias de números para cada casilla, entonces las medias de las casillas deben ser aproximadamente iguales. Si la media general M_t fuera 10, entonces la mejor expectativa para cualquier media de casilla en el diseño factorial sería 10. Por supuesto que estas medias rara vez serían exactamente de 10; de hecho, algunas podrían ser muy diferentes de 10. La pregunta estadística fundamental es: ¿difieren significativamente de 10? Las medias de combinaciones de medias también deben mantenerse alrededor de 10. Por ejemplo, en un diseño como el del ejemplo previo, las medias A_1 y A_2 deberían ser aproximadamente 10, y las medias B_1 y B_2 deberían ser aproximadamente 10. Además, las medias de cada una de las casillas A_1B_1, A_1B_2, A_2B_1 y A_2B_2 deberían mantenerse alrededor de 10.

Utilizando una tabla de números aleatorios, se extrajeron 60 dígitos, del 0 al 24, para llenar las seis casillas de un diseño factorial. El diseño resultante tiene dos niveles o variables independientes, A y B. A se subdivide en A_1, A_2 y A_3; B se subdivide en B_1 y B_2. Tal diseño se denomina diseño factorial de 3×2. (Los ejemplos de las tablas 14.3 y 14.4 son diseños de 2×2.)

Para el siguiente ejemplo los datos son ficticios. El ejemplo se basa en un estudio real de Pury y Mineka (1997), en el cual se examina el efecto de dos variables independientes sobre la reacción emocional. Una variable independiente, grado de temor, no se manipuló (atributo); la segunda variable independiente es el tipo de estímulo visual. Se podría hipotetizar que personas con diferentes niveles de temor a heridas sangrantes tendrían una respuesta emocional distinta a diferentes tipos de estímulos. Para la variable temor se examinan los niveles alto y bajo; para los estímulos visuales se utilizan fotografías de 1) heridas menores (como cortadas, mordidas y hematomas), 2) flores y 3) conejos. La variable dependiente serían las calificaciones combinadas en las tres dimensiones emocionales. El diseño del estudio es un diseño factorial de 3×2. Suponga que se realizó el experimento y que se obtuvieron los resultados de la tabla 14.7, que ofrece el paradigma del diseño y las medias de cada casilla, así como las medias de las dos variables, A y B, y la media general, M_t. Estas medias fueron calculadas a partir de los 60 números aleatorios extraídos en grupos de 10 cada uno e insertados en las casillas.

Difícilmente se requiere de una prueba de significancia estadística para saber que estas medias no difieren significativamente. Su rango total es de 10.4 a 13.6. La media esperada, por supuesto, es la media de los números 0 al 24, es decir 12.0. La cercanía de las medias a la $M_t = 12.00$ es notable, aun para el muestreo aleatorio. De cualquier forma, si

TABLA 14.7 *Diseño factorial de dos factores: medias de los nueve grupos de números aleatorios*

	Tipo de estímulo visual			
Temor	A_1 Heridas menores	A_2 Flores	A_3 Conejos	Medias de temor
B_1 alto	12.9	13.3	10.4	12.2
B_2 bajo	10.5	11.5	13.6	11.9
Medias visuales	11.7	12.4	12.0	$M_t = 12.03$

TABLA 14.8 *Medias de la tabla 14.7 alteradas sistemáticamente al sumarles y restarles constantes*

	Tipo de estímulos visuales			
Temor	A_1	A_2	A_3	**Medias de temor**
B_1	12.9 + 2 = 14.9	13.3	10.4 − 2 = 8.4	12.2
B_2	10.5 − 2 = 8.5	11.5	13.6 + 2 = 15.6	11.9
Medias visuales	11.7	12.4	12.0	12.03

éstos fueran los resultados de un experimento real, el investigador quizás estaría muy disgustado; el tipo de estímulo visual, el grado de temor y la interacción, entre ellos, no son significativos.

Considere cuántos resultados posibles, distintos al azar, habría si una o ambas variables hubiesen sido efectivas. Las tres medias de estímulo visual (M_{A1}, M_{A2} y M_{A3}) podrían haber resultado significativamente diferentes, mientras que las medias de miedo (M_{B1} y M_{B2}) no hubieran sido significativamente diferentes. O las medias de miedo podrían haber sido significativamente diferentes, mientras que las medias de estímulo visual no hubieran sido significativamente diferentes, o ambos conjuntos de medias podrían ser diferentes; o ambos podrían resultar no diferentes, mientras sus interacciones hubieran sido significativas. Las posibilidades de los *tipos* de diferencias e interacciones son considerables también; aunque tomaría demasiadas palabras y números ilustrar incluso a un pequeño número de ellas. Si el estudiante juega un poco con los números, puede lograr bastante conocimiento sobre la estadística y las posibilidades de los diseños. Puesto que la preocupación más importante aquí es la interacción, se alterarán las medias para crear una interacción significativa. Se incrementa en 2 la media de A_1B_1; se decrementa en 2 la media de A_1B_2; se incrementa en 2 la media de A_3B_2, y se decrementa en 2 la media de A_3B_1. Se deja como está la media de A_2, y se alteran los efectos principales de acuerdo a ello. Los cambios se presentan en la tabla 14.8.

La tabla 14.8 debe estudiarse cuidadosamente y compararse con la tabla 14.7. Con las alteraciones arbitrarias se produjo una interacción. Las medias de las casillas de desbalancearon, por decirlo así; mientras que las medias marginales (A_1, A_2, A_3, B_1, B_2) casi no se alteraron. La media total permanece sin cambio en 12.03. Las tres medias de A son iguales, ¿por qué? Las dos medias de B cambiaron muy poco. Un análisis factorial de varianza de los números aleatorios apropiadamente alterados —los cuales, por supuesto, ya no son números aleatorios— produce la tabla final del análisis de varianza incluida en la tabla 14.9.

TABLA 14.9 *Análisis de varianza final: tabla de datos[a] alterados de números aleatorios*

Fuente	***gl***	***sc***	***cm***	***F***
Entre todos los grupos	5	485.13		
Dentro de grupos	54	2 984.80	55.27	
Entre estímulos (A_1, A_2, A_3)	2	4.93	2.47	< 1.0 (n.s)
Entre temores (B_1, B_2)	1	1.67	1.67	< 1.0 (n.s)
Interacción: $A \times B$	2	478.53	239.27	4.33 (0.05)
Totales	59	3 469.93		

[a] n.s. = no significativo.

Ninguno de los efectos principales (temor y estímulos visuales) es significativo; es decir, las medias de A_1, A_2 y A_3 no difieren significativamente del azar. Lo mismo sucede con las medias de B_1 y B_2. La única razón F significativa es la de la interacción, que es significativa al nivel de .05. Obviamente la alteración de las puntuaciones tuvo un efecto. Si se estuvieran interpretando los resultados, como en las tablas 14.8 y 14.9, se diría que ningún tipo de estímulo visual, dentro de sí mismo y entre ellos mostró diferencias, y lo mismo sucedió con el temor. El análisis no reveló diferencias entre alto y bajo nivel de temor, ni entre los tres estímulos visuales. Sin embargo, las personas con alto nivel de temor perciben el conejo con respuestas emocionales menos negativas que el grupo de bajo nivel de temor. Por el otro lado, las personas con alto nivel de temor perciben a las heridas menores de forma más negativa que las personas con bajo nivel de temor.

Tipos de interacción

Hasta ahora no se ha dicho nada acerca de los tipos de interacción de las variables independientes en su influencia conjunta sobre una variable dependiente. Para llegar al meollo de la cuestión de las interacciones, se presentan varios conjuntos de medias para explicar las principales posibilidades. Por supuesto, existen muchas posibilidades, especialmente cuando se incluyen interacciones de orden superior. Los seis ejemplos en la tabla 14.10 indican las principales posibilidades con dos variables independientes. Las primeras tres agrupaciones indican las tres posibilidades de efectos principales significativos; son tan obvios que no requieren analizarse. (De hecho, existe otra posibilidad: ni A ni B son significativas.)

Por otro lado, cuando existe una interacción significativa la situación no es tan obvia. Las agrupaciones *d*), *e*) y *f*) muestran tres posibilidades comunes. En *d*) las medias se cruzan, como indican las flechas en la tabla. Se puede afirmar que A es efectiva en una dirección en B_1, pero que no es efectiva en la otra dirección en B_2; o que $A_1 > A_2$ en B_1, pero que $A_1 < A_2$ en B_2. A este tipo de interacción, con este patrón de cruce, se le llama interacción

TABLA 14.10 *Varios conjuntos de medias que muestran diferentes tipos de efectos principales e interacción*

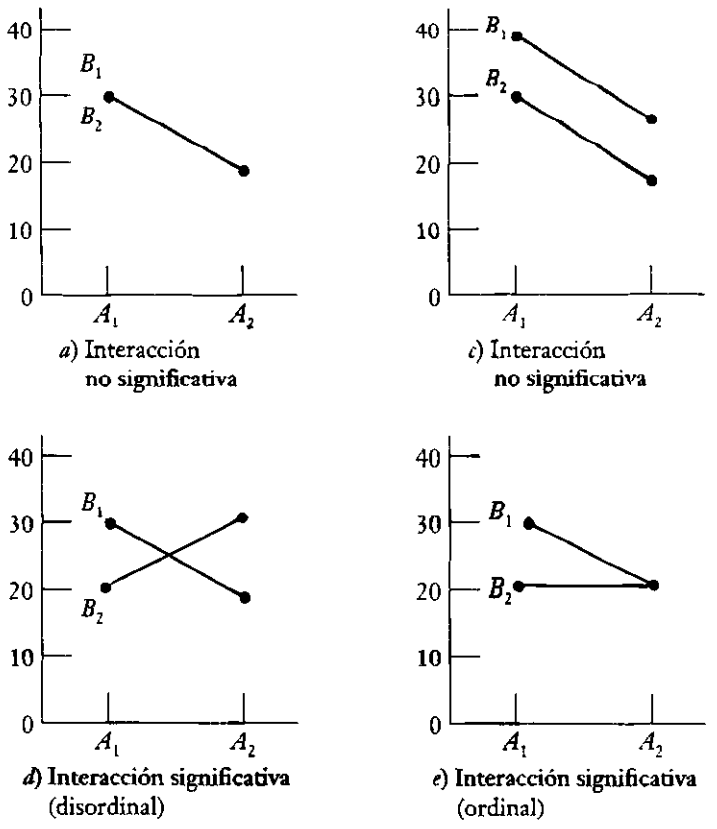
	A_1	A_2		A_1	A_2		A_1	A_2	
B_1	30	20	25	30	30	30	30	20	25
B_2	30	20	25	20	20	20	40	30	35
	30	20		25	25		35	25	
	a) A significativa; B no significativa; interacción no significativa			*b*) A no significativa; B significatitiva; interacción no significativa			*c*) A significativa; B significativa; interacción no significativa		
	A_1	A_2		A_1	A_2		A_1	A_2	
B_1	30	20	25	30	20	25	20	20	20
B_2	20	30	25	20	20	20	30	20	25
	25	25		25	20		25	20	
	d) Interacción significativa (disordinal)			*e*) Interacción significativa (ordinal)			*f*) Interacción significativa (ordinal)		

disordinal (véase abajo de la tabla 14.10). En este capítulo, el ejemplo ficticio de la tabla 14.4 era una interacción disordinal (véase también la tabla 14.5). El ejemplo ficticio en la tabla 14.8, donde la interacción fue deliberadamente inducida al sumar y restar constantes, es otro caso de interacción disordinal.

Sin embargo, difieren las presentaciones en *e)* y en *f)*. Aquí una variable independiente es efectiva sólo en un nivel de la otra variable independiente. En *e)*, $A_1 > A_2$ en B_1, pero $A_1 = A_2$ en B_2. En *f)*, $A_1 = A_2$ en B_1, pero $A_1 > A_2$ en B_2. La interpretación cambia en consecuencia. En el caso de *e)* se diría que A_1 es efectiva al nivel de B_1; pero no provoca una diferencia al nivel de B_2. El caso de *f)* tendría una interpretación similar. Dichas interacciones se denominan interacciones *ordinales*.

Una forma simple para estudiar la interacción con un arreglo de 2×2 (es más complejo con modelos más complejos) consiste en restar un registro de otro en cada renglón, como se hizo antes. Si esto se realiza para *a)*, se obtiene, para los renglones B_1 y B_2, 10 y 10. Para *b)* se obtiene 0 y 0; y para *c)* 10 y 10 nuevamente. Cuando estas dos diferencias son iguales, como en este caso, no existe interacción. Pero si ahora se intenta con *d)*, *e)* y *f)*, se obtiene 10 y −10 para *d)*, 10 y 0 para *e)*, y 0 y 10 para *f)*. Cuando estas diferencias son

FIGURA 14.3



a) Interacción no significativa

c) Interacción no significativa

d) Interacción significativa (disordinal)

e) Interacción significativa (ordinal)

significativamente desiguales, está presente una interacción. El lector puede interpretar estas diferencias como ejercicio.

También es posible —y muchas veces muy útil— graficar las interacciones, como se hizo antes en la figura 14.1. Se establece una variable independiente al colocar los grupos experimentales (A_1, A_2, etcétera) en intervalos similares sobre el eje horizontal y valores apropiados de la variable dependiente en el eje vertical. Después se grafican, contra las posiciones grupales del eje horizontal (A_1, A_2, etcétera), los valores de las medias en la tabla a los niveles de la otra variable independiente (B_1, B_2, etcétera). Dicho método puede ser fácilmente utilizado con diseños de 2×3, 3×2 y otros parecidos. Las gráficas de *a*), *c*), *d*) y *e*) se presentan en la figura 14.3.

Estas gráficas se analizarán sólo brevemente, ya que tanto las gráficas como las relaciones gráficas se han discutido. En efecto, primero se pregunta si existe una relación entre los efectos principales (variables independientes) y las medidas de las variables dependientes. Se grafican cada una de estas relaciones como en el capítulo anterior, excepto que la relación entre una variable independiente y la variable dependiente se grafica en ambos niveles de las otras variables independientes; por ejemplo, A se grafica contra la variable dependiente (eje vertical) en B_1 y B_2. La pendiente de las líneas indica aproximadamente la magnitud de la relación. En cada caso se eligió graficar las relaciones utilizando A_1 y A_2 sobre el eje horizontal. Si la línea graficada es horizontal, obviamente no existe una relación. No existe relación entre A y la variable dependiente al nivel B_2 en *e*) de la figura 14.3; pero sí existe una relación al nivel B_1. En *a*) existe una relación entre A y la variable dependiente en ambos niveles, B_1 y B_2. Lo mismo sucede con *c*). Cuanto más diagonal sea la línea, mayor será la relación. Si las dos líneas tienen aproximadamente el mismo ángulo, en la misma dirección (es decir, que sean paralelas), como sucede en *a*) y en *c*), la relación tiene aproximadamente la misma magnitud en cada nivel. Dependiendo de que las líneas formen diferentes ángulos con el eje horizontal (no paralelas), una interacción estará presente.

Si las gráficas de la figura 14.3 se hubieran realizado a partir de datos reales de investigación, se podrían interpretar de la siguiente forma: llámense a las medidas de la variable dependiente (en el eje vertical) Y; en *a*), A se relaciona con Y a pesar de B; no hace ninguna diferencia lo que B sea; A_1 y A_2 difieren significativamente. La interpretación de *c*) resulta similar: A se relaciona con Y a ambos niveles de B. No hay interacción ni en *a*) ni en *c*). Sin embargo, en *d*) y en *e*) el caso es distinto; la gráfica de *d*) presenta interacción. A se relaciona con Y, pero el tipo de relación depende de B. Bajo la condición B_1, A_1 es mayor que A_2, pero bajo la condición B_2, A_2 es mayor que A_1. La gráfica de *e*) indica que A se relaciona con Y al nivel de B_1 pero no al nivel de B_2; o A_1 es mayor que A_2 en B_1, pero son iguales al nivel de B_2. (Observe que es posible graficar B sobre el eje horizontal, aunque las interpretaciones diferirían en concordancia.)

Notas de precaución

La interacción no siempre es resultado de la interacción "verdadera" de los tratamientos experimentales. Más bien, existen tres posibles causas de una interacción significativa. Una es la "verdadera" interacción, la varianza aportada por la interacción que "realmente" existe entre dos variables en su efecto mutuo sobre una tercera variable. Otra es el error; una interacción significativa puede suceder por el azar, tal como las medias de los grupos experimentales pueden diferir significativamente debido al azar. Una tercera posible causa de interacción es un efecto extraño, indeseable y no controlado, que opera a un nivel de un experimento pero no a otro. Tal causa de interacción debe vigilarse en usos no experimen-

tales del análisis de varianza, esto es, en el análisis de varianza de datos reunidos después de que las variables independientes ya han operado. Suponga, por ejemplo, que los niveles de un experimento sobre métodos son las escuelas. Factores extraños, en este caso, pueden generar una interacción significativa. Suponga también que el director de una escuela, aunque hubiese permitido que se realizara el experimento en su escuela, tuviera una actitud negativa hacia la investigación. Tal actitud podría transmitirse fácilmente a los maestros y a los alumnos, contaminando el tratamiento y los métodos experimentales. En pocas palabras, las interacciones significativas deben manejarse con el mismo cuidado que cualesquiera otros resultados de investigación. Son interesantes y aun dramáticas, como se ha visto, y quizá provoquen la pérdida momentánea de la acostumbrada precaución. Un precepto que los investigadores deben tomar seriamente es: siempre que sea posible, *replique* los estudios de investigación. La réplica debe planearse de forma rutinaria; especialmente cuando se encuentran relaciones complejas. Si se encuentra una interacción en un estudio original y en su réplica, entonces tal vez no se deba al azar, aunque puede aun deberse a otras causas. La palabra *réplica* se utiliza en lugar de *repetición* ya que aunque en una réplica se estudia nuevamente la relación original, se puede estudiar con diferentes tipos de participantes, bajo condiciones relativamente diferentes, e incluso con menos, más o diferentes variables. La tendencia en la literatura sobre investigación psicológica es, felizmente, llevar a cabo dos o más estudios relacionados sobre el mismo problema básico. Dicha tendencia está muy relacionada a la comprobación de hipótesis alternativas, cuya virtud y necesidad se discutieron en capítulos previos.

Dos dificultades relacionadas del análisis factorial son: las n desiguales en las casillas de un diseño, y el uso experimental y no experimental del método. Si las n en las casillas de un diseño factorial no son iguales (y están desproporcionadas, es decir, no están en proporción de un renglón a otro o de una columna a otra), se deteriora la ortogonalidad o independencia de las variables independientes. En ocasiones, incluso se obtendrán sumas de cuadrados negativas y aunque es factible realizar un ajuste, son un poco extrañas y no muy satisfactorias.[1] Al realizar experimentos, el problema no es tan severo porque los participantes pueden ser asignados aleatoriamente a las casillas —excepto, por supuesto, en el caso de variables atributivas— y las n se mantienen iguales o casi iguales. Pero en el uso no experimental del análisis factorial, las n en las casillas se salen del control del investigador. De hecho, aun en estudios experimentales, donde se incluye más de una variable categórica (como raza y sexo), las n casi necesariamente se tornan desiguales.

Para entender esto, tome un ejemplo simple. Suponga que se dividió un grupo en dos, de acuerdo al sexo: 50 hombres y 50 mujeres. Una segunda variable son las preferencias políticas, y se desea tener dos grupos iguales de republicanos y demócratas. Pero suponga también que el sexo está correlacionado con la preferencia política; entonces, habría, por ejemplo, más hombres republicanos comparados con mujeres republicanas, ocasionando una desproporción. Esto se ilustra en la tabla 14.11. Si se añade otra variable independiente, las dificultades se incrementan exponencialmente.

Entonces, ¿qué se puede hacer en la investigación no experimental? ¿No puede usarse el análisis factorial de varianza? La respuesta es compleja y evidentemente no se entiende con claridad. Los paradigmas del análisis factorial de varianza pueden y deben utilizarse, porque guían y clarifican la investigación. Existen estrategias para superar la dificultad de las n desiguales; pueden hacerse ajustes a los datos, o igualarse los grupos eliminando participantes aleatoriamente; pero éstas son estrategias complicadas. Una solución analítica que tiene potencial es el análisis de regresión múltiple; aunque no desaparecen todos

[1] Los programas de cómputo como el SPSS realizan tales ajustes; pero puede ser confuso, ya que la suma de cuadrados ajustada no corresponde con la suma de cuadrados real para las variables independientes.

TABLA 14.11 *Ejemplo de desproporción y n desiguales en las casillas que surgen de variables no experimentales*[a]

	Republicano	Demócrata	
Hombre	30	20	50
Mujer	20	30	50
	50	50	

[a] Las cifras en las casillas son frecuencias.

los problemas, muchos son minimizados con el esquema de la regresión múltiple. En general, el análisis factorial de varianza resulta más conveniente para la investigación experimental, donde los participantes pueden ser asignados aleatoriamente a las casillas, lo cual mantiene iguales las *n* y satisface más o menos los supuestos que subyacen al método. La investigación experimental o no experimental que utiliza muchas variables no experimentales (atributos) podría servirse mejor del análisis de regresión múltiple (véase Keith, 1988). Con *n* iguales y variables experimentales, el análisis de regresión múltiple genera exactamente las mismas sumas de cuadrados, cuadrados medios y razones *F*, incluyendo las razones *F* de interacción, que el análisis factorial estándar. Las variables no experimentales, que son un problema para el análisis factorial, representan menos problema en el análisis de regresión múltiple. No obstante, Simon (1975) y Lee (1995) han señalado que la regresión múltiple no constituye una panacea para la investigación pobremente diseñada. Todo esto se retoma en un capítulo posterior.

Interacción e interpretación

Esta sección sobre interacción se termina con un complejo y difícil problema: la interpretación de los resultados del análisis factorial de varianza, cuando las interacciones son significativas. Suponga que se tienen dos variables, *A* y *B*. Ambas razones *F* son estadísticamente significativas y la razón *F* de la interacción no es significativa. Esto es sencillo y no hay problemas de interpretación. Si, por el otro lado, *A* o *B*, o ambas, son significativas y la interacción de *A* y *B* también es significativa, hay razones para preocuparse. Algunos autores afirman que no es posible la interpretación de efectos principales significativos en la presencia de una interacción, y que si se hace puede llevar a conclusiones incorrectas. La razón es que cuando se dice que un efecto principal es significativo, se implica que es significativo bajo todas las condiciones, que M_{A1} es mayor que M_{A2} con todo tipo de individuos y en todo tipo de lugares, por ejemplo. Sin embargo, si la interacción entre *A* y *B* resulta significativa, la conclusión no es válida empíricamente; por lo menos debe ser aclarada: existe por lo menos una condición, dígase *B*, que tiene que tomarse en cuenta. En lugar de declarar el enunciado simple "si *p* entonces *q*", se dice "si *p* entonces *q*, bajo la condición *r*" o, por ejemplo, M_{A1} es mayor que M_{A2} bajo la condición B_1 pero no bajo la condición B_2. Un método de reforzamiento (elogiar, por ejemplo) resulta efectivo con niños de clase media, pero no con niños de clase trabajadora.

Se pueden encontrar extensos análisis sobre las interacciones en Edwards (1984). Aunque antiguo, el libro de Lubin (1961) presenta una discusión valiosa y clara de las interacciones ordinales y disordinales, además de mostrar las virtudes de la graficación de interacciones significativas. Pedhazur (1996) también analiza la interpretación de los efectos principales cuando las interacciones son significativas. La discusión de Pedhazur es

especialmente convincente cuando ataca la dificultad de la interpretación de las interacciones en la investigación no experimental.

Una regla general es que cuando una interacción es significativa, no se recomienda interpretar los efectos principales, ya que éstos no son constantes sino que varían de acuerdo con las variables que interactúan con ellos; esto es especialmente verdadero si la interacción es disordinal [véase la figura 14.3 *d*)] o si el efecto principal bajo estudio es débil. Si el efecto principal es fuerte —las diferencias entre las medias son grandes— y la interacción es ordinal [véase la figura 14.3 *e*)], entonces quizás se pueda interpretar un efecto principal. Obviamente, la interpretación de los datos de investigación, cuando se estudia más de una variable independiente, resulta a menudo compleja y difícil. Sin embargo, ésta no debería ser una razón para desanimarse. Dicha complejidad tan sólo refleja la naturaleza multivariada y compleja de la realidad psicológica, sociológica y educativa. La tarea de la ciencia consiste en entender tal complejidad; dicho entendimiento nunca podrá ser completo, por supuesto, pero puede lograrse un progreso sustancial con la ayuda de los modernos métodos de diseño y análisis. Los diseños factoriales y el análisis de varianza son grandes logros que incrementan de manera importante nuestra habilidad para entender la compleja realidad psicológica, sociológica y educativa.

Análisis factorial de varianza con tres o más variables

El análisis factorial de varianza funciona con más de dos variables independientes. Es posible utilizar tres, cuatro y más variables y, de hecho, aparecen en la literatura. Sin embargo, diseños con más de cuatro variables son poco comunes. Ello no se debe tanto a que las estadísticas se vuelvan complejas y difíciles de manejar, sino más bien es una cuestión de sentido práctico y de tradición. Con el uso de los paradigmas de investigación actuales se vuelve muy difícil conseguir suficientes participantes para llenar las casillas de los diseños complejos; y es todavía más difícil manipular cuatro, cinco o seis variables independientes al mismo tiempo. Por ejemplo, considere un experimento con cuatro variables independientes. El arreglo más pequeño posible es de $2 \times 2 \times 2 \times 2$, que produce 16 casillas, dentro de las cuales debe ubicarse un número mínimo de participantes. Si se incluyeran 10 *sujetos* en cada casilla, sería necesario manejar un total de 160 sujetos de cuatro formas diferentes. Aun así no se debe ser dogmático respecto al número de variables; quizás dentro de los próximos años los diseños factoriales con más de cuatro variables se volverán comunes. Simon (1987) se ha manifestado durante años para que los experimentos utilicen más variables independientes. De hecho, Simon y Roscoe (1981) han demostrado

especialmente convincente cuando ataca la dificultad de la interpretación de las interacciones en la investigación no experimental.

Una regla general es que cuando una interacción es significativa, no se recomienda interpretar los efectos principales, ya que éstos no son constantes sino que varían de acuerdo con las variables que interactúan con ellos; esto es especialmente verdadero si la interacción es disordinal [véase la figura 14.3 *d*)] o si el efecto principal bajo estudio es débil. Si el efecto principal es fuerte —las diferencias entre las medias son grandes— y la interacción es ordinal [véase la figura 14.3 *e*)], entonces quizás se pueda interpretar un efecto principal. Obviamente, la interpretación de los datos de investigación, cuando se estudia más de una variable independiente, resulta a menudo compleja y difícil. Sin embargo, ésta no debería ser una razón para desanimarse. Dicha complejidad tan sólo refleja la naturaleza multivariada y compleja de la realidad psicológica, sociológica y educativa. La tarea de la ciencia consiste en entender tal complejidad; dicho entendimiento nunca podrá ser completo, por supuesto, pero puede lograrse un progreso sustancial con la ayuda de los modernos métodos de diseño y análisis. Los diseños factoriales y el análisis de varianza son grandes logros que incrementan de manera importante nuestra habilidad para entender la compleja realidad psicológica, sociológica y educativa.

Análisis factorial de varianza con tres o más variables

El análisis factorial de varianza funciona con más de dos variables independientes. Es posible utilizar tres, cuatro y más variables y, de hecho, aparecen en la literatura. Sin embargo, diseños con más de cuatro variables son poco comunes. Ello no se debe tanto a que las estadísticas se vuelvan complejas y difíciles de manejar, sino más bien es una cuestión de sentido práctico y de tradición. Con el uso de los paradigmas de investigación actuales se vuelve muy difícil conseguir suficientes participantes para llenar las casillas de los diseños complejos; y es todavía más difícil manipular cuatro, cinco o seis variables independientes al mismo tiempo. Por ejemplo, considere un experimento con cuatro variables independientes. El arreglo más pequeño posible es de $2 \times 2 \times 2 \times 2$, que produce 16 casillas, dentro de las cuales debe ubicarse un número mínimo de participantes. Si se incluyeran 10 *sujetos* en cada casilla, sería necesario manejar un total de 160 sujetos de cuatro formas diferentes. Aun así no se debe ser dogmático respecto al número de variables; quizás dentro de los próximos años los diseños factoriales con más de cuatro variables se volverán comunes. Simon (1987) se ha manifestado durante años para que los experimentos utilicen más variables independientes. De hecho, Simon y Roscoe (1984) han demostrado el uso de un nuevo paradigma de investigación que puede ser fructífero en términos de producción de buena información. Sin embargo, semejante a la protesta manifestada por Cohen (1994), la psicología académica parece resistirse a tales cambios. Efectivamente, cuando se estudie el análisis de regresión múltiple más adelante, se verá que el análisis factorial de varianza puede realizarse con análisis de regresión múltiple, y que cuatro o cinco factores pueden acomodarse *analíticamente* con facilidad; es decir, las complejidades de los cálculos del análisis de varianza con cuatro o cinco variables independientes se simplifican de manera considerable. Sin embargo, esta facilitación analítica de los cálculos de ninguna forma cambia las dificultades *experimentales* de manejar diversas variables independientes, manipuladas a través de métodos más tradicionales.

La forma más simple de un análisis factorial de varianza de tres variables es un diseño de $2 \times 2 \times 2$. El estudio de Little, Sterling y Tingstrom (1996) utiliza este diseño. La tabla 14.12 presenta, de forma tabular, el diseño de tal estudio. Little, Sterling y Tingstrom

TABLA 14.12 *Diseño de análisis factorial de varianza con tres variables*[a]

		Ubicación del participante			
		A_1 *(noreste de EUA)*		A_2 *(sureste de EUA)*	
		Raza del actor			
		C_1 **Afro-americano**	C_2 **Americano blanco**	C_1 **Afro-americano**	C_2 **Americano blanco**
Ubicación del actor	B_1 (Norte de EUA)	$A_1B_1C_1$	$A_1B_1C_2$	$A_2B_1C_1$	$A_2B_1C_2$
	B_2 (Sur de EUA)	$A_1B_2C_1$	$A_1B_2C_2$	$A_2B_2C_1$	$A_2B_2C_2$

[a] Estudio de Little, Sterling y Tingstrom (1996).

estudiaron los efectos del sesgo dentro del grupo sobre la atribución. Ellos deseaban determinar si el apareamiento del lugar de origen del actor y el lugar de origen del participante resultaría en una evaluación más alta. La ubicación del actor y su raza se variaron para integrar cuatro viñetas escritas. Las viñetas eran descripciones breves de un comportamiento que reflejaba homogeneidad dentro del grupo, homogeneidad fuera del grupo, o uno de dos tipos de individuo con membresía grupal heterogénea, involucrado en una conducta negativa (pelear). Los participantes fueron reclutados de dos ubicaciones en Estados Unidos: noreste y sureste. A cada participante se le pidió leer una breve descripción de un comportamiento y evaluar a la persona descrita. Las evaluaciones se hicieron por medio de un cuestionario de atributos, donde las calificaciones altas indicaban una alta responsabilidad personal, y las calificaciones bajas revelaban baja responsabilidad personal.

Ahora el investigador puede probar siete hipótesis: las diferencias entre A_1 y A_2 (ubicación del participante), entre B_1 y B_2 (ubicación del actor), y entre C_1 y C_2 (raza del actor). Éstos son los efectos principales. También se pueden probar cuatro interacciones: $A \times B$, $A \times C$, $B \times C$ y $A \times B \times C$. La tabla final del análisis de varianza se vería como la tabla 14.13. Es evidente que es posible obtener una gran cantidad de información de este experimento. Si se contrasta con el experimento de una variable, donde sólo se puede probar una hipótesis, la diferencia no sólo es grande, sino que indica una forma fundamentalmente diferente de conceptualizar los problemas de investigación.

TABLA 14.13 *Tabla final del análisis de varianza para el diseño de 2 × 2 × 2 de la figura 14.4*

Fuente	***gl***	***sc***	***cm***	***F***
Entre ubicación del participante (A_1, A_2)	1			
Entre ubicación del actor (B_1, B_2)	1			
Entre raza del actor (C_1, C_2)	1			
Interacción: $A \times B$		1		
Interacción: $A \times C$		1		
Interacción: $B \times C$		1		
Interacción: $A \times B \times C$		1		
Dentro de grupos		$N - 7$		
Total	$N - 1$			

Las interacciones significativas de primer orden se reportan cada vez más en los estudios de investigación publicados. Hace algunos años se les consideraba un fenómeno raro; aunque esto ya no es así (véase Gresham y Witt, 1997). La mayoría de las preocupaciones metodológicas y sustantivas respecto a la interacción en la literatura ocurren en el terreno de la educación. Incluso tiene un nombre: investigación ATI, Aptitude-Treatment Interaction (Interacción Aptitud-Tratamiento, en español). Evidentemente ha florecido debido a que mucha o la mayoría de la investigación educativa se preocupa por mejorar la instrucción; se cree que la interacción de las aptitudes de los alumnos y los métodos de instrucción constituyen una clave importante para lograrlo. Sin embargo, Gresham y Witt (1997) señalaron que la investigación ATI no ha sido fructífera.

En efecto, ahora resulta evidente que las interacciones de las variables se hipotetizan con base en la teoría (véase Tingstrom, 1989; Martin y Seneviratne, 1997). Parte de la esencia de la teoría científica es, por supuesto, especificar las condiciones bajo las cuales un fenómeno puede ocurrir. Por ejemplo, Christenfeld (1997) estaba interesado en el efecto de las distracciones en el manejo del dolor. Christenfeld creía que la memoria jugaba un papel en el reporte de las personas sobre la efectividad de la distracción sobre el dolor. Este estudio probó la noción de que el verdadero efecto de la distracción puede no ser detectable hasta después de una demora. Se produjo dolor a todos los participantes al pedirles que introdujeran una mano dentro de una tina de hielo durante 90 segundos. En el estudio de Christenfeld se asignó a los participantes a una de dos condiciones: de baja distracción o a otra de alta distracción. A su vez, la mitad de los participantes de cada grupo calificó su dolor en uno de dos momentos: inmediatamente después de que terminaron los 90 segundos (grupo de calificación inmediata); la otra mitad contestó un formato idéntico después de realizar una tarea cognitiva irrelevante (grupo de calificación demorada). Christenfeld encontró un efecto de interacción entre la distracción y el momento de la evaluación del dolor. El grupo de alta distracción, que calificó su dolor inmediatamente después de sacar la mano de la tina de hielo, asignó calificaciones más altas que el grupo de baja distracción. Con el grupo que experimentó un periodo de demora antes de calificar su dolor, el patrón fue inverso. Aunque no son comunes, las interacciones de orden superior significativas ocurren; el problema es que frecuentemente resultan difíciles de interpretar. Las interacciones de primer y segundo orden pueden manejarse; pero las de tercer orden y de orden superior vuelven la investigación incómoda a causa de que uno se siente desorientado con respecto a su significado. La literatura reporta algunos estudios con efectos de interacción de tercer orden (véase Bente, Feist y Elder, 1996; Bjorck, Lee y Cohen, 1997).

Hasta el momento el lector sin duda se da cuenta de que en principio la división de las variables independientes no se restringe solamente a dos o tres subparticiones. Es muy posible tener divisiones de 2×4, 2×5, 4×6, $2 \times 3 \times 3$, $2 \times 5 \times 4$, $4 \times 4 \times 3 \times 5$. Blanton y Gerrard (1997) utilizan un diseño de $2 \times 2 \times 3 \times 3$ para estudiar la motivación sexual y la percepción de riesgo de los hombres. Como siempre, el problema que está siendo investigado y el juicio del (los) investigador(es) conforma(n) los criterios para determinar qué diseño y análisis concomitante usarán.

Ventajas y virtudes del diseño factorial y del análisis de varianza

El análisis factorial de varianza, como se ha estudiado, logra muchas cosas, todas las cuales representan ventajas importantes de este enfoque y método. Primero, permite al

investigador manipular y controlar dos o más variables simultáneamente. En la investigación educativa no sólo es posible estudiar los efectos de los métodos de enseñanza sobre el rendimiento, también se pueden estudiar los efectos de dos métodos y, por ejemplo, tipos de reforzamiento. En la investigación psicológica se pueden estudiar los efectos separados y combinados de muchos tipos de variables independientes, tales como ansiedad, culpa, reforzamiento, prototipos, clases de persuasión, raza y atmósfera grupal, sobre muchos tipos de variables dependientes, tales como obediencia, conformidad, aprendizaje, transferencia, discriminación, percepción y cambio de actitud. Además, es factible controlar variables tales como el sexo, la clase social y el ambiente del hogar.

Una segunda ventaja consiste en que el análisis factorial es más preciso que el análisis de un factor. Aquí se aprecia una de las virtudes de combinar el diseño de investigación con las consideraciones estadísticas. Puede decirse que, en situaciones similares, los diseños factoriales son mejores que los diseños de un factor. Este juicio de valor ha estado implícito en la mayor parte de la discusión anterior. El argumento de la precisión le añade peso y será elaborado brevemente.

Una tercera ventaja —y, desde un punto de vista científico amplio, quizás la más importante— es el estudio de los efectos interactivos de las variables independientes sobre las variables dependientes. Esto ya ha sido discutido; pero se debe agregar un punto sumamente importante: el análisis factorial posibilita al investigador *hipotetizar sobre las interacciones*, ya que los efectos interactivos pueden probarse directamente. Si se regresa a los enunciados condicionales, se percibe el fundamento de la importancia de esta afirmación. En un análisis de un factor tan sólo se dice: si p, entonces q; si tales y cuales métodos, entonces tales y cuales resultados. Sin embargo, en el análisis factorial se establecen enunciados condicionales más ricos; como sería si p, entonces q y si r, entonces q, que es equivalente a hablar sobre los efectos principales en un análisis factorial. En el problema de la tabla 14.4, por ejemplo, p son los métodos (A) y r es el tipo de motivación (B). Sin embargo, también podría decirse: si p y r, entonces q, que es equivalente a la interacción de los métodos y los tipos de motivación. La interacción también se expresa como: si p, entonces q bajo la condición r.

Con base en la teoría, en la investigación previa o en corazonadas, los investigadores hipotetizan acerca de las interacciones. Uno hipotetiza que una variable independiente tendrá un cierto efecto sólo en la presencia de otra variable independiente. Christenfeld (1997), en el estudio de la distracción y el dolor percibido, se preguntaba si las personas que reportaban su dolor inmediatamente después de la suspensión del estímulo doloroso tendían a reportar niveles más altos de dolor que la gente que respondía después de una demora. Christenfeld encontró un efecto de interacción entre la condición inmediata y la de demora, y entre la de alta y baja distracción. Parte de estos resultados se presentan en la tabla 14.14; las medias en la tabla reflejan la cantidad de dolor percibido. Ninguno de los efectos principales —tiempo en que se calificó o cantidad de distracción— fue estadís-

TABLA 14.14 *Calificaciones medias del dolor realizadas inmediatamente después del baño de hielo o después de una demora, de los participantes en condiciones de baja y alta distracción (estudio de Christenfeld)*[a]

	Inmediata	Demorada
Alta distracción	5.61	4.67
Baja distracción	5.44	5.67

[a] A mayor calificación mayor intensidad del dolor.

ticamente significativo; pero la interacción entre ellos sí fue significativa. Cuando la distracción era alta, la condición de respuesta inmediata generó calificaciones de dolor más altas. Sin embargo, cuando la distracción era baja, la condición de respuesta demorada produjo calificaciones más altas de dolor. La hipótesis de la interacción fue apoyada —un hallazgo de significancia tanto teórica como práctica—.

Se ha vuelto práctica común dividir una variable continua en dicotomías u otras policotomías. En el estudio de Christenfeld, por ejemplo, una medida continua —cantidad de distracción— se dicotomizó. Observe que antes se señaló que crear una variable categórica a partir de una variable continua elimina la varianza y, por lo tanto, debe evitarse esta práctica. Los investigadores deben considerar el poder que brinda la regresión múltiple, en lugar del análisis de varianza. Se aprenderá en un capítulo próximo que el análisis factorial de varianza puede realizarse con análisis de regresión múltiple, y que con este análisis no es necesario sacrificar porciones de la varianza por la conversión de variables. No obstante, hay argumentos compensatorios: 1) si una diferencia es estadísticamente significativa y la relación es sustancial, no afecta la conversión de variables; el peligro reside en ocultar una relación que, de hecho, existe. 2) Hay ocasiones en que es recomendable realizar la conversión de una variable —por ejemplo, para la exploración de un nuevo campo o problema, y cuando la medición de una variable es, en el mejor de los casos, burda e imperfecta—. En otras palabras, aunque la regla es benéfica, es mejor no ser inflexible respecto a su uso. Se ha realizado buena investigación —incluso excelente— utilizando variables continuas que por una u otra razón se han dividido.

Análisis factorial de varianza: control

En un análisis de varianza de un factor existen dos fuentes de varianza *identificables*: aquella que se presume ocurre por los efectos experimentales y aquella que presumiblemente se debe al error o a la varianza por el azar. Ahora se estudiará más de cerca esta última.

Cuando se han asignado aleatoriamente los sujetos a los grupos experimentales, el único estimado posible de la variación por el azar es la varianza dentro de los grupos. Pero (y esto es importante) queda claro que la varianza dentro de los grupos no contiene solamente la varianza debida al error, sino que también contiene la varianza debida a las diferencias individuales entre los participantes. Dos ejemplos simples son la inteligencia y el género; existen, por supuesto, muchas otras. Si en un experimento se utilizan tanto niños como niñas, la aleatorización puede servir para balancear las diferencias individuales que son concomitantes al género. Entonces, el número de niños y niñas en cada grupo experimental sería casi igual. También se puede asignar arbitrariamente el mismo número de niños y de niñas a los grupos; sin embargo, este método no logra el propósito general de la aleatorización, que es igualar los grupos *en todas* las variables posibles. *Sí* iguala a los grupos en lo que respecta a la variable género; pero no podemos tener la seguridad de que las otras variables queden distribuidas de la misma forma en los grupos. Lo mismo sucede con la inteligencia. Si la aleatorización es exitosa, igualará a los grupos de tal forma que las medias y las desviaciones estándar de la prueba de inteligencia de los grupos serán aproximadamente iguales. Aquí de nuevo es posible asignar a los jóvenes arbitrariamente a los grupos, de tal forma que queden casi iguales; pero entonces no se puede estar seguro de que otras variables posibles estén controladas de la misma forma, debido a que se ha interferido con la aleatorización.

Suponga que la aleatorización ha sido "exitosa"; entonces en teoría no habría diferencias entre los grupos respecto a la inteligencia ni a todas las otras variables. Pero *aún habrá diferencias individuales en inteligencia —y otras variables— dentro de cada grupo.* Con dos gru-

pos, por ejemplo, el *grupo* 1 puede tener calificaciones de inteligencia que vayan de, digamos, 88 a 145, y el *grupo* 2 tendría calificaciones en inteligencia de 90 a 142. Este rango de calificaciones muestra en sí mismo, tal como lo hace la presencia de niños y niñas dentro de los grupos, que hay diferencias individuales en inteligencia *dentro de* los grupos. Si ello es verdad, ¿cómo puede decirse que la varianza dentro de los grupos puede ser un estimado del error, de la variación por el azar? La respuesta es que esto es lo mejor que puede hacerse bajo las circunstancias del diseño. Si el diseño es del tipo de un factor simple, no existe otra medida de error que se pueda obtener; por lo tanto, se calcula la varianza dentro de los grupos y se trata como si fuera una medida "verdadera" de la varianza del error. Debe quedar claro que la varianza dentro de los grupos será mayor que la varianza del error "verdadera", puesto que contiene varianza debida a las diferencias individuales, así como varianza del error. Por ende, una razón F puede no ser significativa cuando, de hecho, sí existe una diferencia entre los grupos. Obviamente si la razón F resulta significativa, no hay mucho de qué preocuparse, porque la varianza entre los grupos es suficientemente grande para superar la varianza del error sobrestimada.

Para resumir lo que se ha expuesto, de nuevo se presenta una ecuación teórica previa:

$$V_t = V_e + V_d \tag{14.3}$$

Puesto que la varianza dentro de los grupos contiene más varianza que la varianza del error, la varianza debida a las diferencias individuales, se escribe como sigue:

$$V_d = V_i + V_{error} \tag{14.4}$$

donde V_i es igual a la varianza debida a las diferencias individuales y V_{error} es igual a la varianza "verdadera" del error. Si esto es verdad, entonces puede sustituirse la parte derecha de la ecuación 14.4 por la V_d en la ecuación 14.3 de la siguiente manera:

$$V_t = V_e + V_i + V_{error} \tag{14.5}$$

En otras palabras, la ecuación 14.5 es una forma abreviada de decir lo que antes se explicó.

La significancia práctica de investigación de la ecuación 14.5 es considerable. Si se puede encontrar la forma de controlar o medir V_i para separarla de V_d, entonces se hace posible obtener una medida más precisa de la varianza "verdadera" del error. Dicho de otra forma, la ignorancia del investigador respecto a la situación de la variable disminuye porque se identifica y aísla más varianza sistemática. Se identifica una porción de la varianza que fue atribuida al error; en consecuencia se reduce la varianza dentro de los grupos.

Muchos de los principios y de la práctica del diseño de investigación se ocupan de este problema, que es esencialmente un problema de control —el control de la varianza—. Cuando se afirmó antes que el análisis factorial de varianza era más preciso que el análisis de varianza de un factor simple, se quiso decir que al establecer niveles de una variable independiente, por ejemplo sexo o clase social, se disminuye el estimado del error, la varianza dentro de los grupos y así nos acercamos a la varianza del error "verdadera". En lugar de escribir la ecuación 14.5, ahora se anotará una ecuación más específica, sustituyendo para V_i la varianza de las diferencias individuales, V_{cs}, la varianza de la clase social y reintroduciendo V_d:

$$V_t = V_e + V_{cs} + V_d \tag{14.6}$$

Compare esta ecuación con la ecuación 14.3. Se ha identificado y denominado más de la varianza total, aparte de la varianza entre grupos. Esta varianza, V_{ew}, en efecto, ha sido sacada de la V_d de la ecuación 14.3.

Ejemplos de investigación

En años recientes, se han reportado un gran número de usos interesantes del análisis factorial de varianza en la literatura sobre investigación del comportamiento; en realidad uno se confronta con una desconcertante abundancia. Se han seleccionado varios ejemplos de diferentes tipos para ilustrar la utilidad y la fuerza del método. Se incluyen más ejemplos de los usuales a causa de la complejidad del análisis factorial, la frecuencia de su uso y su importancia manifiesta.

Raza, sexo y admisión universitaria

En un estudio clásico, ingenioso y elegantemente concebido, Walster, Cleary y Clifford (1970) se preguntaron si en las universidades de Estados Unidos se discrimina en contra de los aspirantes femeninos y afroamericanos. Ellos utilizaron un diseño factorial de $2 \times 2 \times 3$, donde raza (americanos blancos, afroamericanos), género (hombres, mujeres) y habilidad (alta, media, baja) eran las variables independientes; y admisión (elevada en una escala de cinco puntos, donde 1 es igual a rechazo, hasta el 5 que equivale a aceptación con entusiasmo) era la variable dependiente. Seleccionaron aleatoriamente 240 universidades de una lista, y mandaron cartas de solicitud preparadas de forma especial a las universidades, de parte de individuos ficticios que poseían, entre otras cosas, la raza, el sexo y los niveles de habilidad mencionados antes. Por ejemplo, el aspirante podría ser un hombre afroamericano con un nivel medio de habilidad. Observe la inteligente manipulación de las variables que por lo general no son sujetas a la manipulación experimental. También es importante notar que la unidad del análisis fueron instituciones.

El análisis factorial de varianza mostró que ninguno de los tres efectos principales fue estadísticamente significativo. Si ésta fuera toda la información que tuvieran los investigadores, podrían haber concluido que no se había practicado discriminación; sin embargo, una de las interacciones —género por habilidad— fue estadísticamente significativa. Las medias de género y habilidad se presentan en la tabla 14.15. (Se omitió la variable raza porque el efecto principal de raza y sus interacciones con otras variables no fueron signi-

TABLA 14.15 *Resultados del estudio de Walster, Cleary y Clifford sobre sexo, habilidad y admisión (medias)*[a]

Género	Habilidad			
	Alta	Media	Baja	
Hombre	3.75	3.48	3.00	3.41
Mujer	4.05	3.48	1.93	3.15
	3.90	3.48	2.47	

[a] Las medias marginales se calcularon a partir de las medias de las casillas. A mayor valor de la media, mayor aceptación.

ficativas.) ¡Un hallazgo intrigante! Parece que se discrimina a las mujeres con bajo nivel de habilidad, pero no con los niveles medio y alto.

El efecto del género, tipo de violación e información sobre la percepción

La percepción de la gente hacia una víctima de violación ha recibido mucha atención por parte de los medios de comunicación. Se ha realizado investigación para determinar el proceso de toma de decisión del jurado en juicios de violación. Johnson (1994) llevó a cabo un estudio de este tipo utilizando tres variables independientes y dos dependientes. Johnson quería determinar el efecto del género (hombre contra mujer), el tipo de violación (de un conocido contra un extraño) y admisibilidad de la información (sí contra no) sobre el disfrute percibido de la víctima y la atribución de la responsabilidad. Se utilizó un diseño factorial de $2 \times 2 \times 2$.

Para reducir sesgos por posibles demandas en el estudio, Johnson dio a los participantes tres pasajes para leer, y luego les pidió contestar varias preguntas acerca del contenido de la lectura. A los participantes se les hizo creer que el estudio era respecto a la formación de impresiones. Dos de las lecturas eran irrelevantes al estudio; en la lectura experimental, se daba una descripción de una estudiante universitaria que había sido violada. La lectura variaba en el tipo de violación: cometida por un conocido o por un extraño. La lectura también hablaba de las reacciones de los compañeros de clase de la víctima; se implicaba que la víctima de violación tenía un historial de promiscuidad sexual. La mitad de los participantes fueron explícitamente instruidos para ignorar los comentarios de los compañeros de clase, al formarse una percepción (opinión) sobre la víctima (inadmisible); la otra mitad no recibió tales instrucciones (admisible). A cada sujeto se le pidió responder preguntas sobre si la víctima disfrutó la violación y sobre la cantidad de responsabilidad atribuida a la víctima por el hecho de la violación.

Parte del resumen de los datos del estudio se presentan en la tabla 14.16. Los valores incluidos son medias. Los valores mayores indican probabilidades más altas de disfrute y mayor atribución de la responsabilidad. Los participantes hombres percibieron una mayor probabilidad de que la víctima disfrutara de la violación, que las mujeres. Los participantes que no fueron instruidos para ignorar los comentarios de los compañeros de clase de la víctima percibieron una mayor probabilidad de disfrute de la víctima, y de atribución de la responsabilidad que aquellos a quienes se les indicó no tomar en cuenta los comentarios. De la misma manera, los participantes de la condición de violación por un conocido reportaron una mayor probabilidad de disfrute de la víctima y atribución de la responsabilidad, que los de la condición de violación por un extraño. Considere la conveniencia de

TABLA 14.16 *Percepciones medias por tipo de violación y admisibilidad de la información (estudio de Johnson)*[a]

	Admisibilidad de la información	
Tipo de violación	**Admisible**	**Inadmisible**
Por un conocido	*4.0* **4.8**	*3.7* **3.9**
Por un extraño	*3.8* **3.5**	*1.6* **1.6**

[a] Los números en itálicas registran la percepción del disfrute; los valores en negritas, la atribución de la responsabilidad.

un análisis factorial de varianza para el problema analítico y la aplicabilidad de la idea de interacción en esta situación.

Ensayos del estudiante y evaluación del profesor

Los ejemplos anteriores estaban limitados a dos o tres variables independientes. Ahora se analizará brevemente un ejemplo más complejo con más de tres variables independientes. El tema de la investigación siempre ha representado gran interés para los educadores: la lectura, la puntuación y la evaluación de ensayos de los estudiantes. En el que probablemente sea un importante estudio sobre el problema, Freedman (1979) manipuló el contenido, organización, mecánica y estructura de las oraciones de los ensayos. Ella reescribió ocho ensayos de estudiantes "de moderada calidad", para que resultaran como fuertes o débiles en las cuatro características mencionadas. (Ésta fue una tarea difícil, que Freeman realizó admirablemente.) Los ensayos a evaluar incluyeron tanto los ensayos originales como los reescritos. Después fueron evaluados por 12 lectores (otra variable del diseño). La variable dependiente era la calidad, evaluada en una escala de cuatro puntos. Se tiene, entonces, un diseño de $2 \times 2 \times 2 \times 2 \times 12$ (el 12 representa a los 12 lectores). El análisis factorial de varianza se resume en la tabla 14.17.

Estos resultados son interesantes y potencialmente importantes. Primero, los lectores (L) no difirieron, tal como debía de ser. Segundo, el contenido y la organización fueron altamente significativos. (El autor habla de "el mayor efecto principal" que podía haber sido juzgado por ω^2.) La mecánica (M) también resultó significativa; la estructura de la oración (EO) no fue significativa. Pero las interacciones significativas $O \times EO$ y $O \times M$ mostraron que la fuerza o debilidad de la mecánica y la estructura de la oración eran importantes cuando los ensayos tenían una organización fuerte. Dicho estudio y la evaluación de sus ensayos están ciertamente a otro nivel de discurso que los métodos sencillos y más o menos intuitivos que la mayoría usa al juzgar la escritura de los estudiantes.

Anexo computacional

Se estudió cómo utilizar el SPSS para analizar los datos con la prueba t y con el ANOVA de un factor en el capítulo anterior. El uso del SPSS para el análisis factorial de varianza

TABLA 14.17 *Resultados del análisis factorial de varianza de los efectos de la reescritura (estudio de Freedman sobre la evaluación de ensayos)**

Fuente	*gl*	*cm*	*F*
Lector (L)	11	.448	
Contenido (C)	1	9.860	37.78**
Organización (O)	1	5.195	29.69**
Estructura de la oración (EO)	1	1.500	2.54
Mecánica (M)	1	5.042	9.77**
$C \times EO$	1	1.960	6.30
$C \times M$	1	.990	3.18
$O \times EO$	1	3.767	12.11*
$O \times M$	1	6.155	19.79**
$EO \times M$	1	.001	

* Significativo al nivel .01; ** significativo al nivel .001.

TABLA 14.18 *Diseño factorial con datos ficticios*

		Dificultad		
		B_1 (baja)	B_2 (media)	B_3 (alta)
Método de enseñanza	A_1 (tradicional)	18,17,17	17,16,15	11,12,10
	A_2 (mejorado)	18,18,16	14,15,16	12,10,10

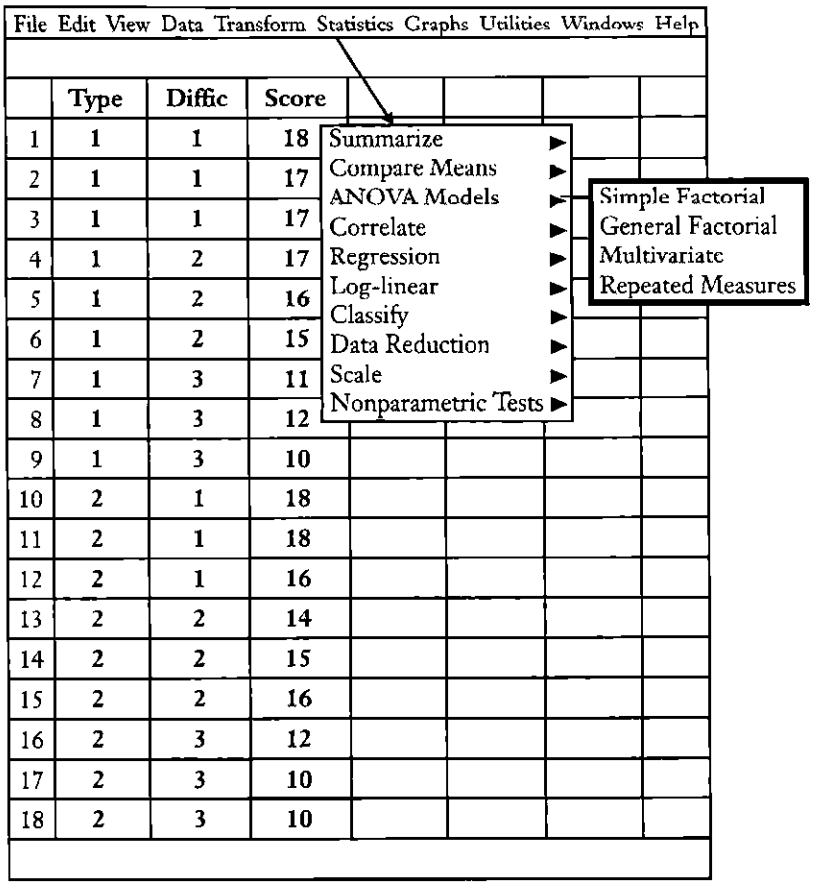
resulta muy similar. La tabla 14.18 presenta datos ficticios en la tradicional forma de tabla. La figura 14.4 muestra cómo se reestructuraron tales datos en el formato de la hoja de cálculo del SPSS. Es muy importante que el lector sepa cómo moverse de la presentación de los datos en la tabla 14.18 a la tabla de datos utilizada por el SPSS. Ese estudio ficticio

FIGURA 14.4

File Edit View Data Transform Statistics Graphs Utilities Windows Help

	Type	Diffic	Score				
1	1	1	18				
2	1	1	17				
3	1	1	17				
4	1	2	17				
5	1	2	16				
6	1	2	15				
7	1	3	11				
8	1	3	12				
9	1	3	10				
10	2	1	18				
11	2	1	18				
12	2	1	16				
13	2	2	14				
14	2	2	15				
15	2	2	16				
16	2	3	12				
17	2	3	10				
18	2	3	10				

FIGURA 14.5



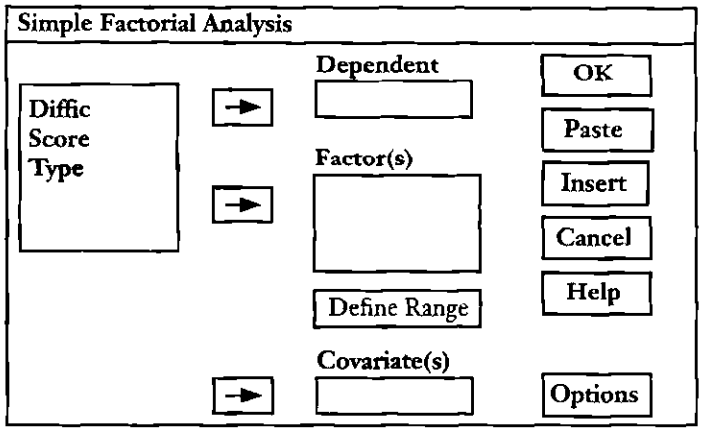
	Type	Diffic	Score
1	1	1	18
2	1	1	17
3	1	1	17
4	1	2	17
5	1	2	16
6	1	2	15
7	1	3	11
8	1	3	12
9	1	3	10
10	2	1	18
11	2	1	18
12	2	1	16
13	2	2	14
14	2	2	15
15	2	2	16
16	2	3	12
17	2	3	10
18	2	3	10

incluyó los efectos de dos variables independientes sobre el rendimiento. Una variable independiente (*A*) era el tipo de método de enseñanza (tradicional, mejorado). La segunda variable independiente era la dificultad de la prueba (baja, media, alta). La variable dependiente fue la calificación en la prueba.

Para realizar el análisis de varianza de dos factores deseado, haga clic en "Statistics". Esto despliega un menú de análisis estadísticos. Elija "ANOVA Models" (figura 14.5) y aparece otro menú. De éste escoja "Simple Factorial". Esto se presenta en la figura 14.5.

Al escoger esa opción aparece una nueva pantalla (figura 14.6) donde se especifica cuáles de las variables son las dependientes, y cuáles las independientes. En el cuadro de la extrema izquierda hay una lista de las tres variables: "Diffic", "Score" y "Type". Primero realce "Score" y haga clic en la flecha que apunta hacia la derecha del cuadro etiquetado "Dependent". Después realce la variable llamada "Diffic"; para introducir "Diffic" en el

FIGURA 14.6



cuadro etiquetado "Factor(s)" haga clic en la flecha que apunta hacia la derecha asociada con el cuadro "Factor(s)" (la figura 14.7 muestra esto).

Después de escoger la variable "Diffic" (figura 14.7), necesita indicarle al SPSS cuántos niveles tiene la variable "Diffic". Haga esto con un clic en el botón "Define Range", con lo cual aparece otra pantalla (mostrada en la figura 14.8). Especifique los valores mínimo y máximo para la variable "Diffic". Existen tres niveles de dificultad, así que se puede anotar "1" para el valor mínimo y "3" para el valor máximo. Cuando se está satisfecho con las anotaciones se hace clic en "Continue". El SPSS ahora regresará a la pantalla previa, y se apreciará un cambio grande; los signos de interrogación ya no siguen al nombre de la variable Diffic en el cuadro Factor(s), en lugar de ello aparece "(1, 3)".

FIGURA 14.7

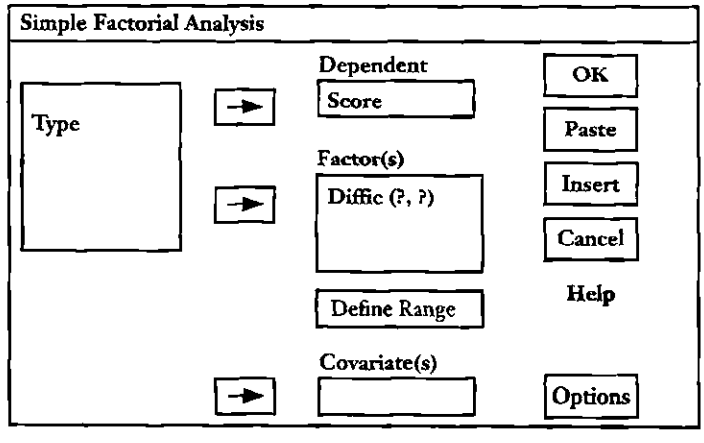
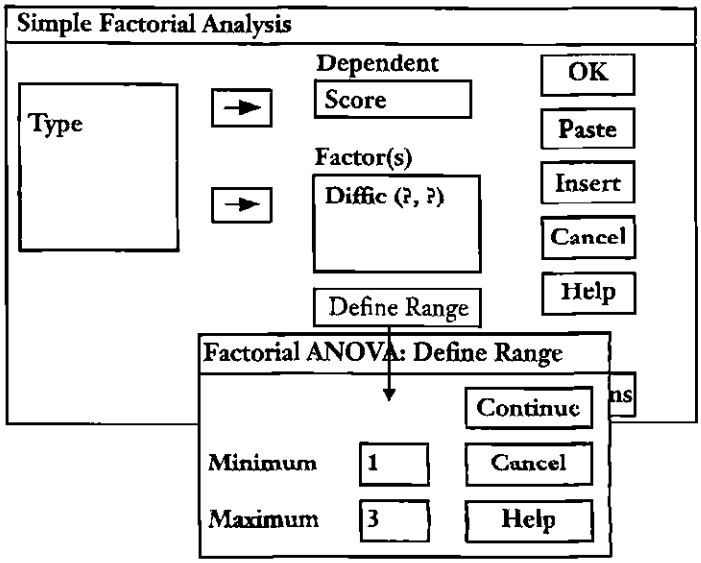


FIGURA 14.8



La siguiente tarea consiste en seleccionar la variable "Type". Resáltela y haga clic en la flecha que apunta hacia la derecha, asociada con el cuadro "Factor(s)". Al hacer esto, el nombre de la variable "Type" aparece en dicho cuadro, seguida de signos de interrogación dentro de paréntesis (figura 14.9). Repita los pasos previos haciendo clic nuevamente en el botón "Define Range" para obtener una pantalla donde puede especificar los niveles de la variable "Type". Puesto que "Type" tiene sólo dos niveles, anote un "1" para el valor mínimo y un "2" para el valor máximo (figura 14.10).

FIGURA 14.9

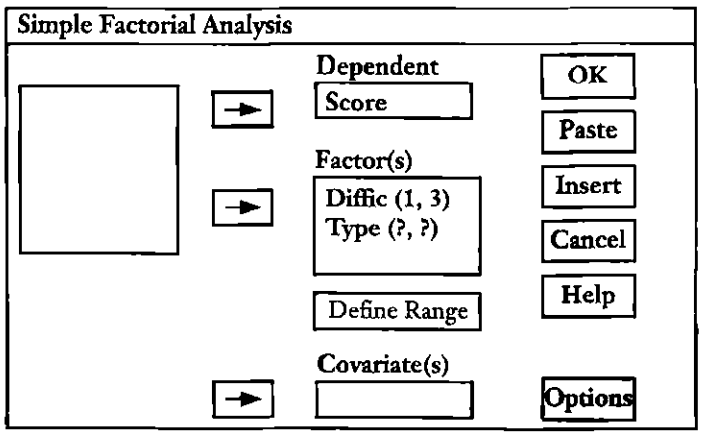
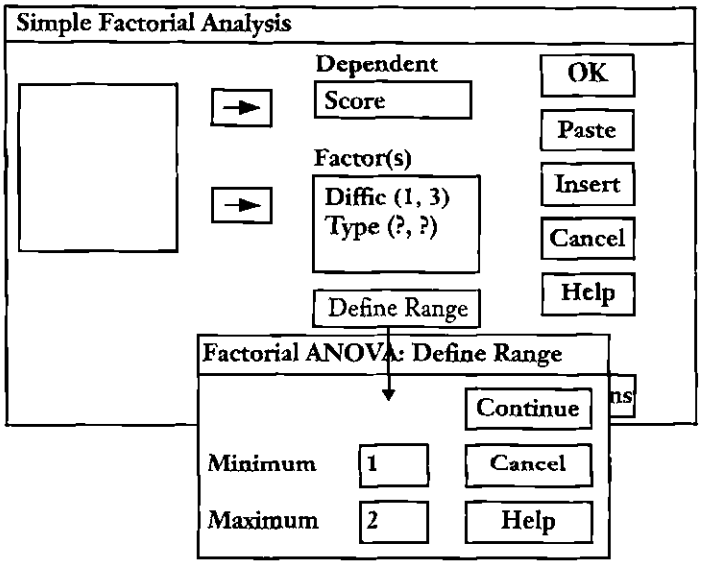
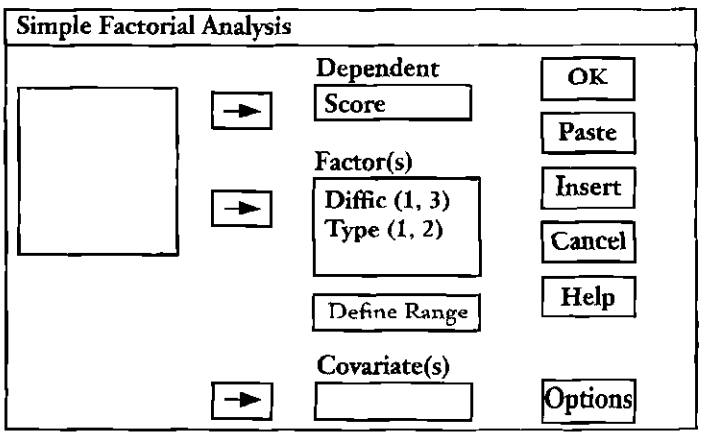


FIGURA 14.10



La figura 14.11 ilustra una pantalla donde ya están definidas todas las variables. Al hacer clic en "OK", SPSS realizará el análisis. Los resultados del análisis se presentan en el cuadro sombreado en la página 343. La pantalla de resultados de arriba muestra la tabla del análisis de varianza y las medias apropiadas de las casillas. La especificación de las medias de las casillas se logró al seleccionar "Options" en la pantalla que se muestra en la figura 14.11. Cuando se selecciona el botón "Options" aparece la pantalla mostrada en

FIGURA 14.11



*** CELL MEANS ***

SCORE
by DIFFIC
TYPE

Total Population	14.56 (18)		
DIFFIC			
	1	2	3
	17.33	15.50	10.83
	(6)	(6)	(6)
TYPE			
	1	2	
	14.78	14.33	
	(9)	(9)	

TYPE DIFFIC	1	2
1	17.33 (3)	17.33 (3)
2	16.00 (3)	15.00 (3)
3	11.00 (3)	10.67 (3)

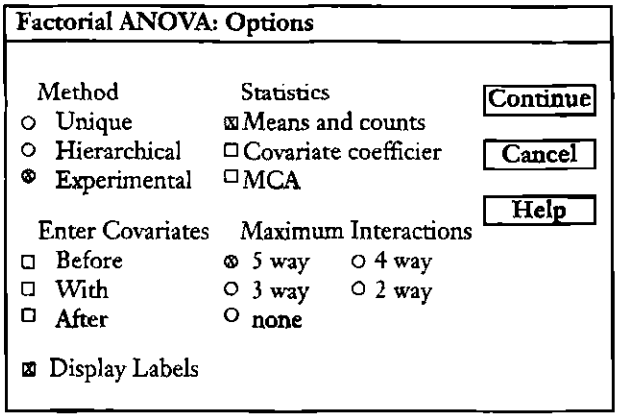
ANALYSIS OF VARIANCE

SCORE
by DIFFIC
TYPE

EXPERIMENTAL sums of squares
Covariates entered FIRST

Source of Variation	**Sum of** *Squares*	*DF*	**Mean** *Square*	*F*	**Sig** *of F*
Main Effects	*135.667*	*3*	*45.222*	*45.222*	*.000*
DIFFIC	*134.778*	*2*	*67.389*	*67.389*	*.000*
TYPE	*.889*	*1*	*.889*	*.889*	*.364*
2-Way Interactions	*.778*	*2*	*.389*	*.389*	*.686*
DIFFIC TYPE	.778	2	.389	.389	.686
Explained	136.444	5	27.289	27.289	.000
Residual	12.000	12	1.000		
Total	**148.444**	**17**	**8.732**		

FIGURA 14.12



la figura 14.12. Para conseguir que las medias aparezcan en el análisis de varianza, seleccione "Experimental" como método a emplear y después escoja "Means and counts".

RESUMEN DEL CAPÍTULO

1. Los diseños factoriales se utilizan con frecuencia en la investigación de las ciencias del comportamiento para analizar dos o más variables independientes simultáneamente. Es posible medir el efecto conjunto de las variables independientes (interacción) sobre la variable dependiente.
2. Todos los niveles de cada variable independiente se cruzan con todos los niveles de las otras variables independientes.
3. Los diseños factoriales son capaces de manejar diseños complejos.
4. Los diseños factoriales están limitados sólo por cuestiones prácticas.
5. Estos diseños pueden manejar los efectos diferenciales de las variables y utilizar enunciados condicionales.
6. La *interacción* se define como la influencia combinada de dos o más variables independientes sobre una variable dependiente.
7. La interacción puede ocurrir en ausencia de cualquier efecto separado de las variables independientes.
8. Los efectos independientes separados se denominan *efectos principales*.
9. En el ANOVA para diseños factoriales, la suma de cuadrados total se separa en: efectos principales, efecto(s) de interacción y efecto del error (dentro de grupos). La tabla de resumen del ANOVA muestra una forma conveniente de presentar el análisis de los datos.
10. Existen dos tipos básicos de los efectos de interacción: (i) ordinal, donde una de las variables independientes es significativa junto con un efecto de interacción significativo; y (ii) disordinal, donde hay un patrón de cruce cuando se grafican las medias de las casillas.
11. Los diseños factoriales y el ANOVA para dos variables independientes se anotan como "*i* por *j*", donde *i* es el número de niveles de la primera variable independiente y *j* es el número de niveles de la segunda variable independiente.

SUGERENCIAS DE ESTUDIO

1. Aquí se presentan algunos estudios psicológicos o educativos variados e interesantes que de una u otra forma han utilizado el análisis factorial de varianza. Lea y estudie dos de ellos y pregúntese: ¿el análisis factorial fue el apropiado?; es decir, ¿los investigadores podrían haber utilizado, por ejemplo, una forma más simple de análisis?

   Behling, D. (1995). Influence of dress on perception of intelligence and scholastic achievement in urban schools with minority populations. *Clothing and Textiles Research Journal, 13*, 11-16. Este estudio examina el efecto de "halo", utilizando un diseño de $6 \times 2 \times 2 \times 3 \times 3$ (estilo de vestuario × sexo del modelo × estatus × escuela × raza). Los resultados mostraron que los maestros y los estudiantes fueron influenciados de forma diferente por el estilo de vestuario.

   Cairns, E. (1990). Impact of television news exposure on children's perceptions of violence in Northern Ireland. *Journal of Social Psychology, 130*, 447-452. Evaluó el impacto de la exposición a las noticias televisivas sobre la percepción que tienen niños irlandeses sobre el nivel de violencia en sus barrios. Se utilizó un ANOVA de cuatro factores (área × sexo × edad × exposición a las noticias). Los resultados mostraron un efecto para área y sexo con respecto al área de alta violencia y los niños varones. Dos interacciones de segundo orden también alcanzaron significancia estadística.

   Langer, E. e Imber, L. (1980). When practice makes imperfect: Debilitating effects of overlearning. *Journal of Personality and Social Psychology, 37*, 2014-2024. Utiliza un diseño factorial de 3×3 y de 3×2, con resultados poco comunes.

   Many, J. E. (1991). The effects of stance and age level on children's literary responses. *Journal of Reading Behavior, 23*, 61-85. Este estudio exploró los efectos del uso de posturas estéticas y eferentes en respuesta a la literatura. Todos los participantes leyeron las mismas tres historias cortas y dieron respuestas libres a cada una. El ANOVA de dos factores reveló efectos significativos para la postura y el nivel de calificación del en tendimiento. El grado de entendimiento se incrementaba con el nivel de calificación. No se encontraron efectos de interacción.

   Wayne, S. J., Kaemar, K. M. y Ferris, G. R. (1995). Coworker responses to others' ingratiation attempts. *Journal of Management Issues*, 7, 277-289. Este estudio utiliza un diseño factorial de $2 \times 2 \times 2 \times 2$ (congraciarse × desempeño objetivo × recompensa × tiempo) para estudiar la satisfacción y la percepción de justicia de los compañeros de trabajo.

2. Estamos interesados en probar la eficacia relativa de diferentes métodos de enseñanza para idiomas extranjeros (o cualquier otra materia). Se cree que la aptitud para los idiomas es posiblemente una variable de influencia. ¿Cómo podría diseñarse un experimento para probar la eficacia de los métodos? Ahora añada una tercera variable, género, y establezca el paradigma para ambas investigaciones. Discuta la lógica de cada diseño desde el punto de vista estadístico. ¿Qué prueba de significancia estadística utilizaría? ¿Qué papel juegan en la interpretación de los resultados?
3. Escriba dos problemas y las hipótesis respectivas, utilizando cualesquiera tres (o cuatro) variables que usted desee. Explore los problemas e hipótesis de las sugerencias de estudio 2 y 3, del capítulo 2, y las variables dadas en el capítulo 3. También puede utilizar cualquiera de las variables de este capítulo. Escriba por lo menos una hipótesis que sea de interacción.

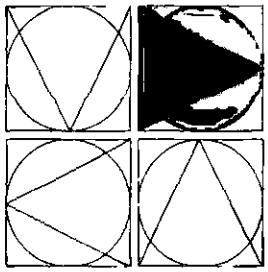
4. A partir de los números aleatorios del Apéndice a, obtenga 40 números, del 0 al 9, en grupos de 10. Considere a los cuatro grupos como A_1B_1, A_1B_2, A_2B_1 y A_2B_2.
   a) Realice un análisis factorial de varianza como se explicó en el capítulo. ¿Cómo deben ser las razones F de A, B y $A \times B$ (interacción)?
   b) Sume 3 a cada una de las puntuaciones en el grupo con la media más alta. ¿Cuál o cuáles razones F deben ser afectadas? ¿Por qué? Realice el análisis factorial de varianza. ¿Se cumplieron sus expectativas?
5. Quizás algunos estudiantes deseen ampliar su lectura y estudio del diseño de investigación y del análisis factorial de varianza. Se ha escrito mucho, por lo que resulta difícil recomendar obras y artículos. Sin embargo, existen cuatro libros que incluyen grandes recursos y capítulos interesantes sobre diseño, problemas estadísticos, suposiciones y su prueba, e historia del análisis de varianza y métodos relacionados.

Collier, R. y Hummel, T. (1977). ***Experimental Design and Interpretation***. Berkeley, California: McCutchan. Este libro fue patrocinado por la American Educational Research Association.

Harlow, L. L., Mulaik, S. A. y Steiger, J. H. (1997). *What if there were no significance test?* Hillsdale, Nueva Jersey: Lawrence Erlbaum.

Keren, G. y Lewis, C. (1993). ***A handbook for data analysis in the behavioral sciences: Statistical issues***. Hillsdale, Nueva Jersey: Lawrence Erlbaum.

Kirk, R. E. (1972). *Statistical issues: A reader for the behavioral sciences*. Monterey, California: Brooks/Cole.



CAPÍTULO 15

ANÁLISIS DE VARIANZA: GRUPOS CORRELACIONADOS

En los capítulos anteriores los grupos en el ANOVA eran independientes. Los participantes que conformaban un grupo no estaban, de forma lógica o significativa, relacionados con los participantes de los otros grupos. Por ejemplo, en un factorial de 2 × 3 existen seis grupos separados. Cada grupo recibe una combinación de tratamientos (variables independientes) diferente a la de los otros grupos. Para grupos independientes por lo común se utilizan participantes diferentes en cada combinación de tratamiento. En este capítulo se considerará la situación en que los participantes no son independientes. Se utiliza el término "grupos correlacionados" porque expresa mejor la naturaleza básica y distintiva del tipo de análisis de varianza que se estudia en este capítulo. Otros términos que se utilizan con más frecuencia son "bloques aleatorizados", "dentro de sujetos" y "medidas repetidas"; aunque estos términos no son completamente generales.

Suponga que un equipo de investigación desea probar los efectos de la marihuana y del alcohol sobre la conducción de un automóvil. Por supuesto que el equipo puede establecer un diseño de un factor o un diseño factorial; pero en su lugar, los investigadores

TABLA 15.1 *Diseño de un experimento sobre los efectos de la marihuana y el alcohol en la conducción: medidas repetidas (puntuaciones ficticias)*[a]

	Marihuana	Alcohol	Control	
Sujetos	(A_1)	(A_2)	(A_3)	*Sumas*
1	18	27	16	61
2	24	29	21	74
.	.	.	.	.
36	21	25	20	66
Sumas	710	820	680	$\Sigma X_t = 2\,210$

[a] Aunque se utilizaron datos ficticios, el diseño fue tomado de un estudio real de investigación efectuado por Crancer, Dille, Delay, Wallace y Haykin (1969).

deciden utilizar a los participantes como sus propios controles; es decir, a cada sujeto le serán aplicados los tres tratamientos o condiciones experimentales: marihuana (A_1), alcohol (A_2) y control (A_3). Después de la aplicación de cada uno de los tratamientos, los participantes operarán un simulador de conducción de un automóvil. La medida de la variable dependiente es el *número de errores al conducir*. En la tabla 15.1 se muestra un paradigma del diseño del experimento, con algunas puntuaciones ficticias. Observe que las sumas tanto de las columnas como de los renglones se incluyen en la tabla. También debe notarse que el diseño se parece al del análisis de varianza de un factor, con una excepción: las sumas de los renglones; éstas son las sumas de las puntuaciones de cada sujeto durante los tres tratamientos.

Se trata de una situación bastante distinta de la de los modelos anteriores, donde los participantes eran asignados aleatoriamente a los grupos experimentales. Aquí a todos los participantes se les aplicaron todos los tratamientos, haciendo de cada sujeto su propio control. De manera general, en lugar de independencia, ahora se tiene dependencia o correlación entre grupos. ¿Qué quiere decir correlación entre grupos? No es sencillo responder tal pregunta con una simple afirmación.

Definición del problema

En el análisis de varianza de uno o más factores, la independencia de los grupos, de los participantes y de las observaciones constituye un factor necesario en los diseños. En ambos métodos se asigna aleatoriamente a los participantes a los grupos experimentales. No interviene la noción de correlación entre los grupos —por definición—. Excepto para las variables incluidas específicamente en el diseño (como añadir género a los tratamientos), la varianza debida a las diferencias individuales se distribuye aleatoriamente entre los grupos experimentales y, por lo tanto, los grupos se "igualan". Se sabe que la varianza debida a las diferencias individuales resulta sustancial sí puede aislarse y extraerse de la varianza total. Entonces debería haber un incremento sustancial en la precisión, ya que dicha fuente de variación en las puntuaciones puede restarse de la varianza total. Así, se crea un error de varianza más pequeño para utilizarse para evaluar los efectos de los tratamientos.

En el capítulo 14 uno de los ejemplos del análisis factorial de varianza identificó y sustrajo la varianza debida a la clase social, a partir de la varianza total (véase ecuaciones 14.3 y 14.6, así como el análisis subyacente) reduciendo así la varianza dentro de los grupos, es decir el término del error. La lógica de este capítulo es similar: aislar y extraer la

varianza de la variable dependiente debida a las diferencias individuales. Para hacer concreta tal discusión abstracta, se utiliza un ejemplo fácil donde se introduce la idea de "apareamiento": utilizar a los mismos participantes en los diferentes grupos experimentales y aparearlos en una, dos o más variables. Esto involucra la misma idea básica de la correlación entre grupos. En el siguiente ejemplo, el apareamiento se utiliza para mostrar la aplicabilidad del análisis de grupos correlacionados a situaciones comunes de investigación, pues ciertos aspectos acerca de la correlación y sus efectos pueden realizarse convenientemente. Sin embargo, por lo común no se recomienda al apareamiento como herramienta de investigación, por razones que se expondrán en un capítulo posterior.

Un ejemplo ficticio

El director de una escuela y los miembros del personal decidieron introducir un programa de educación en relaciones intergrupales, como agregado al currículum escolar. Uno de los problemas que encontraron estaba relacionado con el empleo de películas. Se mostraron videos en las fases iniciales del programa, pero los resultados no fueron muy alentadores. El personal hipotetizó que la falla de los videos para provocar un impacto pudo deberse a no haber realizado un esfuerzo particular para resaltar las posibles aplicaciones del video en las relaciones intergrupales. Ellos decidieron probar la hipótesis de que observar los videos y después discutirlos mejoraría la actitud de los espectadores hacia los miembros de grupos minoritarios, más que simplemente ver los videos.

Para un estudio preliminar, el personal seleccionó aleatoriamente un grupo de estudiantes del cuerpo total de estudiantes y los apareó respecto a su inteligencia hasta obtener 10 pares, de manera que cada par tuviera aproximadamente el mismo nivel de inteligencia. La lógica detrás de este experimento fue que la inteligencia se relaciona con las actitudes hacia los grupos minoritarios, y que necesitaba ser controlada. Se asignó aleatoriamente a cada miembro de cada par al grupo experimental o al grupo control y, después se mostró a ambos grupos un video sobre relaciones intergrupales. El grupo A_1 (experimental) tuvo una sesión de análisis después de que se le mostró el video; el grupo A_2 (control) no tuvo tal análisis después de ver el video. Ambos grupos fueron evaluados con una escala

TABLA 15.2 *Puntuaciones de actitud y cálculos del análisis de varianza (ejemplo ficticio)*

	Grupos		
Pares	A_1 *(Experimental)*	A_2 *(Control)*	Σ
1	8	6	14
2	9	8	17
3	5	3	8
4	4	2	6
5	2	1	3
6	10	7	17
7	3	1	4
8	12	7	19
9	6	6	12
10	11	9	20
ΣX	70	50	$\Sigma X_t = 120$
M	7	5	$\Sigma X_t^2 = 930$

diseñada para medir actitudes hacia los grupos minoritarios. Las puntuaciones de actitud y los cálculos para un análisis de varianza se presentan en la tabla 15.2.

Primero se realiza un análisis de varianza de un factor, como si los investigadores no hubiesen apareado a los participantes. Se hace caso omiso del procedimiento de apareamiento y se analizan las puntuaciones como si todos los participantes hubiesen sido asignados aleatoriamente a los dos grupos, sin importar su inteligencia. Los cálculos son:

$$C = \frac{14\,400}{20} = 720$$

$$Total = 930 - 720 = 210$$

$$Entre\ columnas\ (A_1, A_2) = \left(\frac{70^2}{10} + \frac{50^2}{10}\right) - 720 = 20$$

La tabla final de este análisis de varianza se presenta en la tabla 15.3. Puesto que la razón F de 1.89 no es significativa, las dos medias grupales de 7 y 5 no difieren significativamente. La interpretación de estos datos llevaría a los investigadores a creer que el video con la discusión no tuvo efecto alguno; la conclusión sería errónea. La diferencia en este caso en realidad es significativa al nivel 0.01. Suponga que esta afirmación es verdadera; si lo es, entonces algo debe estar mal en el análisis.

Una digresión explicativa

Cuando se aparea a los sujetos en variables *relacionadas significativamente* con la variable dependiente, entonces se introduce la correlación en el panorama estadístico. En el capítulo 14 se demostró que con frecuencia era posible identificar y controlar una porción mayor de la varianza total de una situación experimental, al considerar varios niveles de una o más variables supuestamente relacionadas con la variable dependiente. Por ejemplo, la inclusión de dos o tres niveles de clase social hace posible identificar la varianza en las puntuaciones de la variable dependiente debida a la clase social. El apareamiento del presente experimento ha determinado en realidad 10 niveles, uno por cada par. Los miembros del primer par tenían puntuaciones de inteligencia de, por ejemplo, 130 y 132; los miembros del segundo par, 124 y 125, y así sucesivamente hasta el décimo par, cuyos miembros presentaban puntuaciones de 89 y 92. Cada par (nivel) tiene una media diferente. Si la inteligencia se correlaciona de manera sustancial y positiva con la variable dependiente, entonces los pares de puntuaciones de la variable dependiente deberían reflejar el apareamiento realizado en inteligencia; es decir, que las puntuaciones de la variable dependiente dentro de cada par deben parecerse más entre sí de lo que se parecen a otras puntuaciones de la variable dependiente. Entonces, el apareamiento en inteligencia ha "introducido" varianza entre los pares en la variable dependiente o varianza *entre renglones*.

TABLA 15.3 *Tabla final del análisis de varianza, análisis de un factor sobre datos ficticios de la tabla 15.2*

Fuente de la variación	*gl*	*ss*	*CM*	*F*
Entre grupos (A_1, A_2)	1	20.00	20.0	1.89 (n.s.)
Dentro de grupos	18	190.00	10.56	
Total	19	210.00		

Considere otro ejemplo hipotético para ilustrar lo que sucede cuando existe correlación entre conjuntos de puntuaciones. Suponga que un investigador apareó tres grupos de sujetos respecto a su inteligencia, y que la inteligencia está perfectamente correlacionada con la variable dependiente, un cierto tipo de rendimiento. Esto es altamente improbable; sin embargo, continuemos con el ejemplo para obtener la idea. El primer trío de sujetos tuvo puntuaciones de inteligencia de 141, 142 y 140; el segundo trío, de 130, 126 y 128, y así sucesivamente, hasta el quinto trío, cuyas puntuaciones fueron de 82, 85 y 82. Al verificar en las columnas el orden de los rangos de los tres conjuntos de puntuaciones, se verá que son exactamente iguales: 141, 130,..., 82; 142, 126,..., 85; 140, 128,..., 82. Puesto que se asume que $r = 1.00$ entre inteligencia y rendimiento, entonces el orden de los rangos de las puntuaciones de rendimiento sería el mismo en los tres grupos. Las puntuaciones asumidas de la prueba de rendimiento se presentan en el lado izquierdo de la tabla 15.4. El orden de los rangos de estos datos ficticios, de mayor a menor, aparece en paréntesis junto a cada puntuación de rendimiento. Note que el orden de los rangos es el mismo en los tres grupos.

Ahora suponga que la correlación entre inteligencia y logro fuera aproximadamente cero. En tal caso no se podría hacer predicción alguna del orden de los rangos de las puntuaciones de logro o, dicho de otra forma, las puntuaciones de logro no estarían apareadas. Para simular tal condición de cero correlación, se rompió el orden de los rangos de las puntuaciones del lado izquierdo de la tabla 15.4, con la ayuda de una tabla de números aleatorios. La operación para realizar este "desordenamiento" fue la siguiente: se extrajeron tres conjuntos de números del 1 al 5 y las puntuaciones de cada columna se ordenaron de acuerdo al nuevo orden señalado por los números aleatorios para sus rangos. (Antes de hacer esto, todos los rangos de las columnas fueron 1, 2, 3, 4, 5.) El primer conjunto de números aleatorios fue 2, 5, 4, 3 y 1, por lo tanto el número que antes ocupaba el segundo lugar en la columna A_1 ahora ocupa el primer lugar en la misma columna. Después se tomó el quinto número de A_1 y ahora se anotó en segundo lugar. Este proceso se continuó con las demás puntuaciones de la columna hasta terminar con el primer número que ahora se convirtió en el quinto número. Se realizó el mismo procedimiento con los otros dos grupos de números con, por supuesto, diferentes conjuntos de números aleatorios. Los resultados del nuevo ordenamiento de los rangos se muestran en el lado derecho de la tabla 15.4. También se incluyen las medias de los renglones, así como los rangos de las puntuaciones de las columnas (entre paréntesis).

Primero es necesario estudiar los rangos de los dos conjuntos de puntuaciones. Las puntuaciones correlacionadas se encuentran en la porción izquierda de la tabla, identificada como I. Puesto que los rangos son los mismos para cada columna, la correlación promedio entre las columnas es 1.00. Los números del conjunto identificado como II, que

TABLA 15.4 *Puntuaciones correlacionadas y no correlacionadas (ejemplo ficticio)*

I. Grupos correlacionados				II. Grupos no correlacionados			
A_1	A_2	A_3	M	A_1	A_2	A_3	M
73 (1)	74 (1)	72 (1)	73	63 (2)	74 (1)	46 (5)	61.00
63 (2)	65 (2)	61 (2)	63	45 (5)	55 (3)	61 (2)	53.67
57 (3)	55 (3)	59 (3)	57	50 (4)	50 (4)	59 (3)	53.00
50 (4)	50 (4)	53 (4)	51	57 (3)	65 (2)	53(4)	58.33
45 (5)	44 (5)	46 (5)	45	73 (1)	44 (5)	72 (1)	63.00
	$M_c = 57.80$					$M_c = 57.80$	

son esencialmente aleatorios, presentan una situación bastante diferente; los 15 números de ambos conjuntos son exactamente los mismos; al igual que los números en cada columna (y sus medias). Únicamente los números en renglones y, por supuesto, las medias de renglones, son diferentes. Observe los órdenes de rango de II; no puede hallarse una relación sistemática entre ellos. La correlación promedio debe ser aproximadamente cero, a causa de que los números fueron seleccionados aleatoriamente, de hecho, ésta es de 0.11.

A continuación es necesario estudiar la variabilidad de las medias por renglón. Note que la variabilidad de las medias de I es considerablemente mayor que la de II. Si los números son aleatorios, la media esperada de cualquier renglón es la media general. La media de los renglones de II ronda bastante cerca de la media general de 57.80. El rango es 63 – 53 = 10. No obstante las medias de los renglones de I no se encuentran cercanas a 57.80; su variabilidad es mucho mayor, como lo indica el rango de 73 – 45 = 28. Al calcular las varianzas de los dos conjuntos de medias (llamadas *varianza entre renglones*) se obtiene 351.60 para I y 58.27 para II; la varianza de I es seis veces mayor que la varianza de II. Esta diferencia tan grande constituye un efecto directo de la correlación presente en las puntuaciones de I, pero no en las de II, lo cual indica que la varianza entre renglones es un índice directo de las diferencias individuales. El lector debe realizar una pausa aquí para revisar este ejemplo, especialmente las cifras de la tabla 15.4, hasta que sea claro el efecto de correlación sobre la varianza.

¿Cuál es el efecto de la estimación de la varianza del error de las puntuaciones correlacionadas? Claramente, la varianza debida a la correlación es la varianza *sistemática*, la que debe sustraerse de la varianza total si se desea obtener un estimado más preciso de la varianza del error. De otra manera, la estimación de la varianza del error incluirá a la varianza debida a las diferencias individuales y, por lo tanto, el resultado será demasiado grande. En el ejemplo de la tabla 15.4 se sabe que el procedimiento de mezcla de los datos ocultó la varianza sistemática debida a la correlación. Al reordenar las puntuaciones se elimina la posibilidad de identificar dicha varianza; la varianza está todavía en las puntuaciones de II, pero no puede ser extraída. Para demostrarlo, se calculan las varianzas de los términos del error de I y de II; la de I es 3.10 y la de II es 149.77. Al remover la varianza debida a la correlación de la varianza total, es posible reducir sustancialmente el término de error, con el resultado de que la varianza del error de I resulta 48 veces menor que la varianza del error de II. Si existe una varianza sistemática sustancial en los conjuntos de medidas y es factible aislar e identificar esta varianza, claramente vale la pena hacerlo.

En datos de investigación reales, la situación no es tan dramática como en el ejemplo anterior; las correlaciones casi nunca son de 1, pero con frecuencia son mayores que .50 o .60. *Mientras mayor sea la correlación, mayor será la varianza sistemática que puede extraerse de la varianza total, y mayor será la reducción que se puede lograr en el término del error.* Tal principio es muy importante no sólo en el diseño de investigación, sino también en la teoría y en la práctica de la medición. En ocasiones es posible construir correlaciones entre los datos y después extraer la varianza debida a las puntuaciones correlacionadas resultantes. Por ejemplo, es posible obtener una medida "pura" de las diferencias individuales utilizando a los mismos participantes en diferentes ensayos; obviamente las puntuaciones de un participante serán más semejantes entre sí que con las puntuaciones de otros.

Re-examen de los datos de la tabla 15.2

Ahora regresamos a los datos de investigación ficticios de la tabla 15.2: los efectos de los videos sobre las actitudes hacia los grupos minoritarios. Antes se calculó la suma de cuadrados entre columnas (entre grupos) y la varianza, exactamente de la misma forma

como se realizó en el análisis de varianza de un factor. Se encontró que la diferencia entre las medias no era significativa al utilizar dicho método. A partir del análisis anterior se puede suponer que si hay correlación entre los dos conjuntos de puntuaciones, entonces la varianza debida a la correlación debe sustraerse de la varianza total y, por supuesto, de la estimación de la varianza del error. Si la correlación es alta, este procedimiento debe marcar una diferencia: el término del error debe hacerse considerablemente menor. La correlación entre los conjuntos de las puntuaciones A_1 y A_2 de la tabla 15.2 es .93; puesto que éste es un alto nivel de correlación, el término del error (cuando se calcula de forma apropiada) es mucho menor que antes.

La operación adicional requerida es simple: tan sólo se suman las puntuaciones de cada renglón de la tabla 15.2 y se calcula la suma de cuadrados entre renglones y la varianza. La suma de cada renglón se eleva al cuadrado y el resultado se divide entre el número de puntuaciones en dicho renglón; por ejemplo, en el primer renglón: $8 + 6 = 14$; $(14)^2 \div 2 = 196 \div 2 = 98$. Se repite este procedimiento para cada renglón, se suman los cocientes y después se resta el término de corrección C. Esto produce la suma de cuadrados entre renglones. (Ya que el número de puntuaciones en cada renglón es siempre 2, resulta más fácil, en especial con una calculadora de mano, sumar todas las sumas de cuadrados y después dividirlas entre 2.)

$$\text{Entre renglones } (1, 2, 3, \ldots, 10) = \left[\frac{(14)^2 + (17)^2 + \cdots + (20)^2}{2}\right] - 720$$

$$= 920 - 720 = 182$$

Esta suma de cuadrados entre renglones es una medida de la variabilidad debida a diferencias individuales, como se indicó antes.

Ya se extrajo la suma de cuadrados entre columnas y entre renglones de la suma de cuadrados total, ahora se establece la ecuación ya familiar utilizada en el análisis de varianza de un factor:

$$sc_t = sc_c + sc_d \qquad \textbf{(15.1)}$$

El análisis de la tabla 15.3 es un ejemplo. Dicha ecuación debe alterarse para que se adecue a las presentes circunstancias. La anterior suma de cuadrados entre grupos, sc_g, se designa de nuevo como sc_c, que es la suma de cuadrados de las columnas. Luego se debe sumar la suma de cuadrados de los renglones, sc_r, y la que antes se llamaba sc_d ahora debe designarse de otra manera, pues ya no se tiene varianza dentro de grupos. (¿Por qué?) Ahora se denomina como sc_{res}, que se refiere a la suma de cuadrados de los *residuos*. Como su nombre lo indica, la suma de cuadrados *residual* se refiere a la suma de cuadrados que queda después de que las sumas de cuadrados de columnas y renglones han sido extraídas de la suma de cuadrados total. Entonces se tiene la siguiente ecuación:

$$sc_t = sc_c + sc_r + sc_{res} \qquad \textbf{(15.2)}$$

En resumen, la varianza total se ha separado en dos varianzas sistemáticas e identificables y una varianza del error, la cual constituye un estimado más preciso del error o variación de las puntuaciones por el azar, que el de la tabla 15.3.

En lugar de sustituir en la ecuación, se incluyó la tabla final del análisis de varianza (tabla 15.5). La razón F de las columnas es ahora $20.00 \div .89 = 22.47$, que es significativo al nivel .001. En la tabla 15.3 la razón F no fue significativa. —

TABLA 15.5 *Tabla completa de análisis de varianza: datos de la tabla 15.2*

Fuente de la variación	*gl*	*sc*	*CM*	*F*
Entre columnas (A_1, A_2)	1	20	20.0	22.47 (0.001)
Entre renglones (1, 2, 3, ..., 10)	9	182	20.22	22.72 (0.001)
Residual	9	8	0.89	
Totales	19	210		

Esto implica una gran diferencia. Ya que la varianza entre columnas es la misma, la diferencia se debe en gran medida al término del error, disminuido en gran cantidad, puesto que ahora es igual a .89 y antes era igual a 10.56. Al calcular la suma de cuadrados de los renglones y la varianza, ha sido posible reducir el término del error a cerca de 1/12 de su magnitud anterior. En esta situación, obviamente, la varianza del error anterior igual a 10.56 estaba enormemente sobrestimada. Algunos textos de estadística (por ejemplo, Kirk, 1990; Mendenhall y Beaver, 1997) se refieren a las columnas como "tratamientos" y a los renglones como "bloques". Regresando al problema original, ahora se puede afirmar que añadir la discusión después del video parece haber logrado un efecto significativo sobre las actitudes hacia los grupos minoritarios.

Consideraciones adicionales

Antes de abandonar el ejemplo anterior, es necesario resaltar algunos puntos adicionales. El primero incluye al término del error y las varianzas dentro de grupos y residuales. Cuando se calculan las varianzas de las columnas y de los renglones, no es posible calcular la varianza dentro de grupos, ya que tan sólo hay una puntuación por casilla. También es necesario tener en mente que tales cálculos de ambas varianzas del error, *son sólo estimados de la varianza del error*. En el caso del análisis de un factor, el único estimado posible es la varianza dentro de grupos. En el caso presente se puede obtener un mejor estimado; "mejor" en el sentido de que hay más varianza sistemática. Cuando es posible extraer varianza sistemática, se hace. Con los datos de la tabla 15.2 fue posible hacerlo.

Un segundo punto es: ¿Por qué no utilizar la prueba *t*? La respuesta resulta simple: se puede hacer si así se desea. Si únicamente hay un grado de libertad, es decir, dos grupos; entonces la *t* es igual a la raíz cuadrada de *F*, o $F = t^2$. La razón *t* de los datos de la tabla se obtiene fácilmente mediante la raíz cuadrada de 22.47 = 4.74. Pero si existe más de un grado de libertad, entonces debe abandonarse la prueba *t* y recurrir a la prueba *F*. Por otra parte, el análisis de varianza provee mayor información; el análisis de la tabla 15.5 indica que la diferencia entre el promedio de las puntuaciones de actitud de los grupos experimental y de control es significativamente diferente. La prueba *t* habría ofrecido la misma información; pero la tabla 15.5 también indica clara y simplemente que el apareamiento resultó efectivo o que la correlación entre las puntuaciones de la variable dependiente de los dos grupos es significativa. Si la razón *F* entre renglones no hubiese resultado significativa, se sabría que el apareamiento no había sido exitoso, lo cual representa una información muy valiosa. Por último, una vez comprendidos los cálculos del análisis de varianza, éstos son fáciles de recordar; mientras que las ecuaciones utilizadas para estimar el error estándar de las diferencias entre medias parecen confundir al estudiante novato. (La fórmula simple que se dio anteriormente tiene que alterarse debido a la correlación.)

Punto tres: las pruebas *post hoc* de la significancia de la diferencia entre medias individuales pueden realizarse con más de dos grupos. Las pruebas de Sheffé, Tukey y otras

utilizadas para comparaciones múltiples pueden aplicarse. La prueba de Scheffé se estudió en el capítulo 13.

Finalmente, y de gran importancia, los principios analizados anteriormente son aplicables a una variedad de situaciones de investigación y su aplicación al apareamiento es quizás la menos importante; aunque tal vez sea la más fácil de entender. Siempre que se utilicen los mismos sujetos y medidas repetidas, se aplican tales principios. Cuando se usan diferentes grupos de clase o diferentes escuelas en la investigación educativa, se aplican estos principios: la varianza debida a las diferencias de grupos escolares o escuelas puede extraerse de los datos. De hecho, los principios pueden utilizarse en cualquier investigación donde se empleen diferentes tratamientos experimentales en diferentes unidades de una mayor organización, institución o área geográfica —siendo que estas unidades difieran en variables de significancia para la investigación—.

Para entender lo que esto significa, imagine que los renglones del lado izquierdo de la tabla 15.4 son diferentes escuelas o grupos de clase en un sistema escolar, que las escuelas o clases difieren significativamente respecto al rendimiento, como lo indican las medias por renglón, y que A_1, A_2 y A_3 son tratamientos experimentales de un estudio realizado en cada una de las escuelas o de los grupos de clase (véase la sugerencia de estudio 2).

El análisis de varianza de dos factores (dos variables independientes) es útil para la solución de ciertos problemas de medición, en especial en psicología y educación, como se verá en capítulos posteriores. Las diferencias individuales son una fuente de varianza constante que necesita ser identificada y analizada. Un buen ejemplo lo constituye el estudio de calificadores y calificaciones. Se puede separar la varianza de los calificadores (jueces) de la varianza de los objetos que se están calificando. Se puede estudiar la confiabilidad de los intrumentos de medición, ya que la varianza de los reactivos puede separarse de la varianza de las personas que responden los reactivos. Se regresará continuamente a estos importantes puntos y a los principios subyacentes.

Para ilustrar el uso de jueces o calificadores como "bloques", considere el siguiente ejemplo. Ocho diferentes jueces evalúan cuatro videos y cada video cubre el mismo material. Cada juez asigna una calificación entre 0 y 20 a cada video en términos de la efectividad de presentación. Cada juez vio los videos en orden aleatorio. La tabla que se presenta a continuación contiene los datos, el análisis y el resumen; el análisis revela que los jueces difieren en las calificaciones asignadas a los videos; por lo tanto, separar las varianzas incrementa el efecto entre los videos.

	Videos				
Jueces/Bloques	**A**	**B**	**C**	**D**	**Totales por renglón**
1	6	4	14	8	32
2	8	2	10	7	27
3	7	8	10	7	32
4	12	6	11	12	41
5	5	0	9	8	22
6	7	3	10	7	27
7	10	9	16	11	46
8	9	4	12	9	34
Totales por columna	64	36	92	69	**261**

$$SC_{Total} = DE^2(N) = 3.3737^2\,(32) = 364.22$$

$$SC_{Video} = \left[\frac{64^2 + 36^2 + 92^2 + 69^2}{8}\right] - M^2(N) = 2\,327.13 - 8.15625(32) = 198.34$$

$$SC_{Jueces} = \left[\frac{32^2 + 27^2 + \ldots + 34^2}{4}\right] - 2\,128.78125 = 106.97$$

$$SC_{Residual} = SC_{Total} - SC_{Video} - SC_{Jueces} = 58/91$$

Fuente de la variación	*gl*	*sc*	*CM*	*F*
Videos	3	198.34	66.11	23.53 (0.01)
Jueces (bloques)	7	106.97	15.28	5.44 (0.01)
Residual	21	58.91	2.81	
Total	31	364.22		

Extracción de varianzas por sustracción

Para asegurarse de que el lector entiende los puntos explicados, aquí se repiten ejemplos previos. En la tabla 15.6 se presentan dos conjuntos de números designados como I y II. Los números en dichos conjuntos son exactamente los mismos, lo único que difiere es el orden que tienen. En I no existe correlación entre las dos columnas de números; el coeficiente de correlación es exactamente cero, lo cual resulta análogo a la asignación aleatoria de los participantes a los dos grupos. El análisis de varianza de un factor puede aplicarse. Por otro lado, en II los números A_2 han sido reordenados de tal manera que haya correlación entre los números de A_1 y A_2. (Verifique el orden de los rangos.) De hecho, r

TABLA 15.6 *Análisis de varianza de datos ficticios aleatorizados (I) y correlacionados (II)*

	I $r = 0.00$			*II* $r = 0.90$		
	A_1	A_2	Σ	A_1	A_2	Σ
	1	5	6	1	2	3
	2	2	4	2	4	6
	3	4	7	3	3	6
	4	6	10	4	5	9
	5	3	8	5	6	11
ΣX	15	20	$\Sigma X_t = 35$	15	20	$\Sigma X_t = 35$
M	3	4	$\Sigma X_t^2 = 145$	3	4	$\Sigma X_t^2 = 145$
			$M_t = 3.5$			$M_t = 3.5$

$$C = \frac{(35)^2}{10} = 122.50$$

$$Total = 145 - 122.50 = 22.50$$

$$Entre\ columnas\ C = \left[\frac{15^2 + 20^2}{5}\right] - 122.50 = 2.50$$

$$Entre\ renglones\ R = \left[\frac{6^2 + 4^2 + .. + 8^2}{2}\right] - 122.50$$

$$= 132.50 - 122.50 = 10$$

$$C = \frac{(35)^2}{10} = 122.50$$

$$Total = 145 - 122.50 = 22.50$$

$$Entre\ columnas\ C = \left[\frac{15^2 + 20^2}{5}\right] - 122.50 = 2.50$$

$$Entre\ renglones\ R = \left[\frac{3^2 + 6^2 + \ldots + 9^2}{2}\right] - 122.50$$

$$= 141.50 - 122.50 = 19$$

TABLA 15.7 *Tablas finales de los análisis de varianza*

		I ($r = 0.00$)			*II* ($r = 0.90$)		
Fuente de la variación	***gl***	***sc***	***CM***	***F***	***sc***	***CM***	***F***
Entre *columnas*	1	2.50	2.50	1.0	2.50	2.50	10.0 (0.05)
Entre *renglones*	4	10.00	2.50	(n.s.)	19.00	4.75	
Residual $C \times R$	4	10.00	2.50		1.00	0.25	
Totales	9	22.50			22.50		

= 0.90; el análisis de varianza de un factor no se puede aplicar aquí. Si se utiliza con los números de II, los resultados serán exactamente los mismos que resultarían con los números de I, pero entonces se estaría pasando por alto la varianza debida a la correlación.

Los cálculos en la tabla 15.6 producen todas las sumas de cuadrados excepto las residuales, que se obtienen mediante la sustracción. Puesto que los cálculos son tan sencillos, se procedió directamente a las tablas finales del análisis de varianza, presentadas en la tabla 15.7. Las sumas de cuadrados totales, de las columnas y de los renglones se incluyen como se indica, con los grados de libertad apropiados. Los grados de libertad entre renglones son el número de renglones menos uno (5 – 1 = 4). Los grados de libertad residuales, como los grados de libertad de la interacción en el análisis factorial de varianza, se obtienen al multiplicar los grados de libertad entre columnas por los grados de libertad entre renglones: $1 \times 4 = 4$; o sólo restando los grados de libertad entre columnas y entre renglones de los grados de libertad totales: 9 – 1 – 4 = 4. De la misma forma, las sumas de cuadrados residuales se obtienen restando las sumas de cuadrados entre columnas y entre renglones de las sumas de cuadrados totales. Para I, 22.5 – 2.5 = 10.0 = 10; para II, 22.5 – 2.5 – 19.0 = 1.

Estos análisis requieren de poca elaboración. Observe en especial que donde existe correlación, la razón *F* entre columnas es significativa; pero cuando la correlación es cero, no lo es. Resulta importante notar también el término del error: para I ($r = .00$), es de 2.5; para II ($r = .90$), es de .25, lo cual es 10 veces más pequeño.

Eliminación de fuentes sistemáticas de varianza

Ahora se utiliza el proceso de sustracción del capítulo 6 para eliminar las dos fuentes sistemáticas de varianza en los dos conjuntos de puntuaciones. Primero se elimina la varianza entre columnas corrigiendo cada media para que sea igual a la media general de 3.5. Después se corrige cada puntuación en cada columna de la misma forma (como se efectuó para I y II en la tabla 15.8).

Si ahora se calculan las sumas de cuadrados totales de I y II, en ambos casos se obtiene 20. Compare este resultado con la cifra anterior de 22.5. El procedimiento de corrección ha reducido las sumas de cuadrados totales en 2.5; por supuesto que éstas son las sumas de cuadrados entre columnas. Note de nuevo que el procedimiento de corrección no ha tenido ningún efecto en la varianza dentro de cada uno de los cuatro grupos de puntuaciones; tampoco tuvo efecto alguno sobre las medias de los renglones.

Después se elimina la varianza de los renglones al dejar la media de cada renglón igual a 3.5, que es la media general, y corrigiendo las puntuaciones por renglón en concordancia. Esto se realizó en la tabla 15.9, la cual debe estudiarse con cautela. Note que la variabilidad de ambos conjuntos de puntuaciones se ha reducido, pero la variabilidad del conjunto correlacionado (II) se redujo drásticamente. De hecho, las puntuaciones de II tienen un

TABLA 15.8 *Eliminación de la varianza entre columnas, mediante la igualación de las medias y las puntuaciones de las columnas*

	I $r = 0.00$			II $r = 0.90$		
Corrección	.5	−.5		.5	−.5	
	A_1	A_2	M	A_1	A_2	M
	1.5	4.5	3.0	1.5	1.5	1.5
	2.5	1.5	2.0	2.5	3.5	3.0
	3.5	3.5	3.7	3.5	2.5	3.0
	4.5	5.5	5.0	4.5	4.5	4.5
	5.5	2.5	4.0	5.5	5.5	5.5
M	3.5	3.5	$M_t = 3.5$	3.5	3.5	$M_t = 3.5$

rango de sólo 4 − 3 = 1; mientras que el rango de las puntuaciones de I es 5 − 2 = 3. El apareamiento de las puntuaciones en II y su correlación concomitante permite, por medio del procedimiento correctivo, reducir de manera importante el término del error al "corregir" la varianza debida a la correlación. La única varianza ahora en las puntuaciones corregidas dos veces es la varianza residual.

"Varianza residual" es un término apropiado para la varianza que permanece después de que se han eliminado las dos varianzas sistemáticas. Si se calculan las sumas de cuadrados *totales* de I y de II, resultan ser 10 y 1, respectivamente. Si se calculan las sumas de cuadrados *dentro* de grupos como se hace con el análisis de varianza de un factor, también resultan ser 10 y 1. En efecto, ya no queda más varianza sistemática en las puntuaciones - únicamente queda la varianza del error. El punto más importante es que la suma de cuadrados residual de las puntuaciones no correlacionadas es 10 veces mayor que la suma de cuadrados residual de las puntuaciones correlacionadas. Se realizó exactamente la misma operación para los dos conjuntos de puntuaciones; sin embargo, con las puntuaciones no correlacionadas no es factible extraer tanta varianza como con las puntuaciones correlacionadas.

Otros diseños correlacionales del análisis de varianza

Hasta ahora se han estudiado, en el análisis concerniente al análisis de varianza, tres de los cinco diseños básicos. El capítulo 13 cubrió el diseño completamente aleatorizado; éste

TABLA 15.9 *Eliminación de la varianza entre columnas al igualar las medias y las puntuaciones de los renglones*

	I $r = 0.00$				II $r = 0.90$		
Corrección	A_1	A_2	M	Corrección	A_1	A_2	M
+0.5	2.0	5.0	3.5	+2.0	3.5	3.5	3.5
+1.5	4.0	3.0	3.5	+0.5	3.0	4.0	3.5
0	3.5	3.5	3.5	+0.5	4.0	3.0	3.5
−1.5	3.0	4.0	3.5	−1.0	3.5	3.5	3.5
−0.5	5.0	2.0	3.5	−2.0	3.5	3.5	3.5
M	3.5	3.5	$M_t = 3.5$		3.5	3.5	$M_t = 3.5$

era el ANOVA de un factor con grupos independientes, los cuales se conforman por lo general a través de la selección aleatoria de sus participantes y a través de la asignación aleatoria de los participantes a las condiciones de tratamiento. El capítulo 14 presentó el diseño factorial aleatorizado. Aquí se estudiaron dos o más variables independientes o experimentales al mismo tiempo. Tal como el diseño completamente aleatorizado, los grupos incluidos en los análisis eran independientes; cuando se incluyen dos variables independientes el análisis se llama ANOVA de dos factores. En este capítulo se ha estudiado el diseño de bloques aleatorizados, que es un ANOVA de un factor donde los sujetos no son independientes. Tales diseños incluyen el uso de participantes apareados o el uso del mismo sujeto en diferentes condiciones de tratamiento; se le llama *bloque aleatorizado* porque los tratamientos se asignan a cada sujeto en orden aleatorio.

Los dos diseños básicos restantes del ANOVA son variaciones del diseño de bloque aleatorizado. Uno de estos diseños es el llamado ANOVA de diagrama dividido o factorial mezclado. El otro es el diseño dentro de participantes de *n* factores. De forma conceptual el más simple de los dos es este último, el cual es parecido al diseño factorial aleatorizado; sin embargo, los participantes no son asignados aleatoriamente a los tratamientos. En este diseño se expone a un grupo de participantes a todas las combinaciones de tratamiento. Recuerde que en el diseño factorial aleatorizado se utilizaron diferentes grupos de participantes en cada combinación de distintos tratamientos. Con dos variables independientes, este análisis se llamaría ANOVA de dos factores dentro de sujetos o ANOVA de sujetos-por-tratamiento-por-tratamiento.

Con el ANOVA factorial mezclado con dos variables independientes, cada sujeto es expuesto a todos los niveles de una variable independiente; pero sólo a un nivel de la segunda variable independiente. A este diseño se le llama "mezclado" porque tiene características tanto del ANOVA correlacionado como del no correlacionado. Se describe alternativamente como un diseño que tiene, por lo menos, un factor entre sujetos y, por lo menos, un factor dentro de sujetos. La tabla 15.10 (a), (b), (c), (d) y (e) indica la diferencia entre los cinco diseños del ANOVA.

En este capítulo se revisó el procedimiento para separar la suma de cuadrados en el ANOVA completamente aleatorizado y en el ANOVA de bloque aleatorizado. Una lógica similar hacia la partición se aplica también en los diseños mezclados o en los diseños dentro de sujetos. Purdy, Avery y Cross (1978) ofrecen una buena explicación de ello, así como la presentación de los datos y la tabla del resumen del ANOVA. Otras referencias excelentes son Hays (1994), Kirk (1995), Linton y Gallo (1975), McGuigan (1997) y Howell (1997).

Ejemplos de investigación

Efectos irónicos del intento de relajarse bajo estrés

Cuando estamos en una situación estresante, ¿ayuda el decirse a uno mismo: *relájate* y *cálmate*? Uno pensaría que éste es el mejor procedimiento que se puede utilizar para estar más sanos. Recientemente, un jugador profesional de baloncesto tuvo una acalorada discusión con su entrenador; después de que fueron separados, el jugador se dirigió a los vestidores, pero regresó 20 minutos después a atacar al entrenador nuevamente. Numerosos investigadores han documentado el hecho de que decirse a sí mismo *relájate* y *cálmate* no es sencillo. En un estudio sobre el fenómeno, Wegner, Broome y Blumberg (1997) demostraron que los esfuerzos conscientes por relajarse por lo común llevan a un estado de mayor exaltación. Dicho estudio encontró que cuando se les indicaba a los participantes

TABLA 15.10 *Presentación de los cinco diseños del ANOVA*

(a) El diseño completamente aleatorizado (ANOVA de un factor)

Variable independiente		
A_1	A_2	A_3
S_1	S_4	S_7
S_2	S_5	S_8
S_3	S_6	S_9
Grupo 1	Grupo 2	Grupo 3

(b) Diseño de bloque aleatorizado (ANOVA de un factor)

Variable independiente		
A_1	A_2	A_3
S_1	S_1	S_1
S_2	S_2	S_2
S_3	S_3	S_3
S_4	S_4	S_4
S_5	S_5	S_5
Grupo 1	Grupo 1	Grupo 1

(c) Diseño factorial aleatorizado (ANOVA de dos factores)

Variable independiente 2	**Variable independiente 1** A_1	A_2	A_3
	S_1	S_7	S_{13}
	S_2	S_8	S_{14}
B_1	S_3	S_9	S_{15}
	Grupo 1	Grupo 3	Grupo 5
	S_4	S_{10}	S_{16}
	S_5	S_{11}	S_{17}
B_2	S_6	S_{12}	S_{18}
	Grupo 2	Grupo 4	Grupo 6

(d) Diseño dentro de sujetos de dos factores (ANOVA de dos factores)

Variable independiente 2	**Variable independiente 1** A_1	A_2	A_3
	S_1	S_1	S_1
	S_2	S_2	S_2
B_1	S_3	S_3	S_3
	Grupo 1	Grupo 1	Grupo 1
	S_1	S_1	S_1
	S_2	S_2	S_2
B_2	S_3	S_3	S_3
	Grupo 2	Grupo 2	Grupo 2

(*continúa*)

TABLA 15.10 *(continuación)*

(e) Diseño factorial mezclado o de diagrama dividido (ANOVA de dos factores)

Variable independiente 2	Variable independiente 1 A_1	A_2	A_3
	S_1	S_1	S_1
	S_2	S_2	S_2
B_1	S_3	S_3	S_3
	Grupo 1	Grupo 1	Grupo 1
	S_4	S_4	S_4
	S_5	S_5	S_5
B_2	S_6	S_6	S_6
	Grupo 2	Grupo 2	Grupo 2

que se relajaran bajo una elevada carga mental, ellos exhibían un nivel más alto de exaltación. Por otro lado, los participantes que estaban sometidos a una menor carga mental o que no se les indicó relajarse tendieron a estar menos exaltados. Wegner y sus colaboradores utilizaron el nivel de conductancia de la piel (NCP) como medida de la exaltación; valores altos del NCP indicaban niveles altos de exaltación. Mientras que el NCP sirvió como variable dependiente en el estudio, las variables independientes fueron la carga (alta contra baja), la instrucción (indicación de relajarse contra no indicación de relajarse) y periodo (pre: primeros 5 minutos; prueba: siguientes 3 minutos; y post: últimos 5 minutos). Esta última variable independiente (periodo), son las medidas repetidas y se expuso a cada sujeto a los tres niveles de la variable. Las otras dos variables independientes eran variables entre sujetos; se expuso a cada sujeto únicamente a una condición de carga y a una condición de instrucciones. El diseño de este estudio puede clasificarse como un ANOVA mezclado o de diagrama dividido. La presentación del diseño con las medias se incluye en la tabla 15.11. Los análisis mostraron un efecto significativo para periodo, $F_{(2,166)} = 137.7$, $p < 0.001$.

Conjuntos de aprendizaje de isópodos

En una demostración interesante y efectiva del uso de participantes como sus propios controles, donde se utilizaron el análisis de varianza de dos factores y la prueba de una teoría de aprendizaje con organismos inferiores, Morrow y Smithson (1969) mostraron que los isópodos (pequeños crustáceos) pueden aprender a aprender. Muchos estudiantes,

TABLA 15.11 *Medias de carga, instrucciones y periodo (estudio de Wagner, Broome y Blumberg)*[a]

	Periodo		
Instrucción/Carga	**Pre**	**Prueba**	**Post**
Ninguna/baja	–0.51	4.27	1.58
Ninguna/alta	–0.46	3.80	1.68
Relajarse/baja	0.38	3.50	1.80
Relajarse/alta	0.10	5.70	2.58

[a] Estos valores se estimaron a partir de las cifras de Wegner, Broome y Blumberg.

humanistas, sociólogos, educadores y aun psicólogos han criticado a los teóricos del aprendizaje y a otros investigadores en psicología por el hecho de utilizar animales en sus investigaciones. Aunque puede existir crítica legítima sobre la investigación en psicología y en otras áreas del comportamiento, criticarla por el uso de animales es parte de la irracionalidad frustrante, pero aparentemente inevitable, que plaga todo el esfuerzo humano. Aun así tiene cierto encanto y puede, en sí misma, ser objeto de investigación científica. Bugelski (1956) escribió una defensa excelente sobre el uso de ratas en la investigación sobre aprendizaje, la cual debe ser revisada por estudiantes de investigación del comportamiento. Otro excelente ensayo sobre una base más amplia es el realizado por Hebb y Thompson (1968). En cualquier caso, una de las razones para probar hipótesis similares con diferentes especies es la misma razón por la que se replican investigaciones en diferentes partes de Estados Unidos y en otros países: la generalidad. Una teoría es mucho más poderosa si se sustenta con personas del norte, del sur, del este y del oeste; con alemanes, japoneses, israelíes y estadounidenses; y con ratas, pichones, caballos y perros. El estudio de Morrow y Smithson (1969) intentó extender la teoría del aprendizaje a criaturas pequeñas, cuyo aprendizaje puede considerarse gobernado por leyes diferentes de las del aprendizaje del hombre y de las ratas. Ellos tuvieron éxito, al menos hasta cierto punto.

Los investigadores entrenaron ocho isópodos, a través de privación de agua y reforzamiento subsecuente por desempeño exitoso con papel húmedo, para revertir sus "preferencias" por uno u otro camino de un laberinto en *T*. Cuando los *sujetos* cumplían un criterio, previamente especificado, de giros correctos en el laberinto, el entrenamiento era revertido, es decir, el cambio de dirección hacia el otro camino del laberinto en *T* se reforzaba hasta que se cumplía el criterio. Esto se hizo con cada isópodo en nueve reversiones. La pregunta es: ¿aprendieron los animales a revertir más rápido conforme progresaban los ensayos? Un aprendizaje de este tipo debe mostrarse con menos errores cada vez.

Morrow y Smithson analizaron los datos mediante un análisis de varianza de dos factores. El número promedio de errores del ensayo inicial y de los nueve ensayos de reversión disminuyeron de forma consistente: 27.5, 23.6, 18.6, 14.3, 16.8, 13.9, 11.1, 8.5, 8.6, 8.6. El resultado del análisis de varianza de dos factores se presenta en la tabla 15.12. El análisis de varianza fue calculado por el primer autor (FNK), a partir de los datos de Morrow y Smithson en su tabla 1.

Las diez medias difieren significativamente, ya que la razón *F* para las columnas (ensayos de reversión), 4.78, es significativa al nivel 0.01. La *F* indica que existe correlación entre las columnas y así las diferencias individuales entre los isópodos fueron significativas al nivel 0.01. ¡Es una nota cautivadora el que aun los pequeños crustáceos sean individuos!

Negocios: conducta de licitación

El siguiente ejemplo se tomó de la literatura sobre investigación en mercadotecnia. El estudio del comportamiento es predominante en la investigación sobre negocios. Numerosos corporativos reconocidos, como Procter & Gamble, contratan científicos del

TABLA 15.12 *Análisis de varianza de los datos de Morrow y Smithson*

Fuente de la variación	*gl*	*sc*	*CM*	*F*
Ensayos de reversión	9	3 095.95	343.994	4.78 (0.01)
Isópodos (bloques)	7	1 587.40	226.771	3.15 (0.01)
Residual	63	4 532.85	71.950	
Totales	79	9 216.20		

TABLA 15.13 *Análisis de varianza de los datos de Reinmuth y Barnes*

Propuestas	Compañía A	Compañía B	Compañía C	Σ
1	$45.00	$42.50	$39.75	127.25
2	45.00	40.25	42.70	127.95
3	46.00	45.50	40.00	131.50
4	43.75	43.50	40.20	127.45
5	46.00	44.50	40.65	131.15
6	43.50	43.25	40.00	126.75
7	44.50	40.90	41.45	126.85
8	45.50	45.00	45.75	136.25
9	50.00	45.50	45.60	141.10
10	46.50	44.50	44.15	135.15
Σ	455.75	435.40	420.25	

comportamiento para ayudar a realizar investigación conductual sobre productos al consumidor.

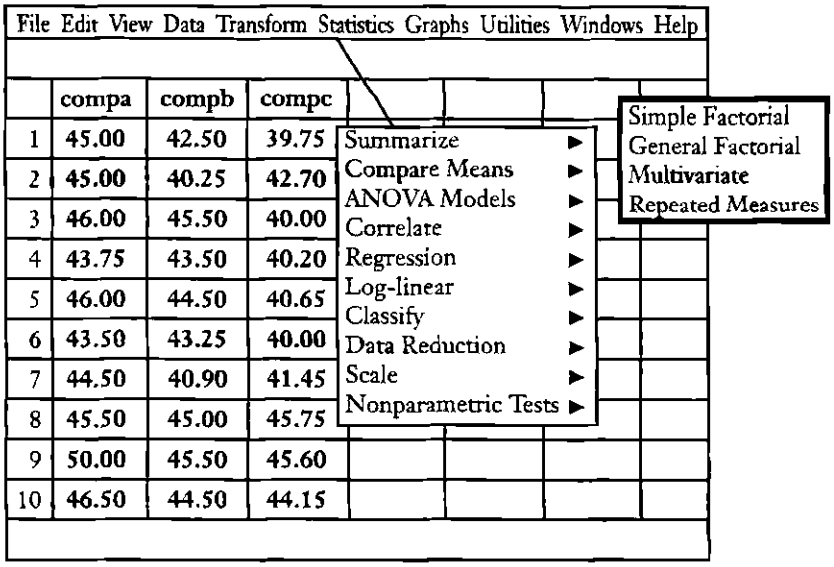
En un estudio de Reinmuth y Barnes (1975) no se utilizó el ANOVA de bloque aleatorizado para analizar sus datos. Sin embargo, los datos que recolectaron en su estudio sobre las propuestas de tres compañías extractoras de petróleo se ajusta al diseño de bloques aleatorizados. El estudio constituía en realidad un problema de mercadotecnia que incluía el desarrollo de un modelo matemático de propuestas competitivas en un proceso de licitación. Los datos presentados por Reinmuth y Barnes eran 10 propuestas aleatoriamente seleccionadas a partir de 35 propuestas posibles. Los datos representan los costos estimados más la ganancia por el uso de una plataforma petrolera y un equipo de cuatro hombres, por hora. La tabla 15.13 presenta los datos recolectados. El objetivo del uso de un análisis de bloques aleatorizados es ver si las tres compañías difieren en sus propuestas al considerar las diferencias de sus ensayos individuales. Los ensayos son las 10 mediciones o propuestas tomadas de cada compañía.

La tabla sumaria del análisis de varianza se presenta en la tabla 15.14. El ANOVA muestra que la diferencia entre las propuestas de las compañías contratantes son altamente significativas ($p < 0.001$). Los bloques o ensayos de propuestas también fueron estadísticamente significativos. La compañía A fue consistentemente el contratista más alto; mientras que la compañía C fue consistentemente el contratista más bajo. Puesto que la fuente de varianza del bloque fue estadísticamente significativa, ello indica que la correlación entre las compañías contratistas y las propuestas contribuyó significativamente a la varianza sistemática. La η^2 para los datos fue .363. Las correlaciones entre las tres compañías contratistas fueron $r_{AB} = .55$, $r_{AC} = .64$ y $r_{BC} = .29$.

TABLA 15.14 *Análisis de varianza de los datos de Reinmuth y Barnes*

Fuente de la variación	*gl*	*sc*	*CM*	*F*
Contratistas	2	63.463	31.7215	14.75 (0.001)
Ensayos de propuestas (bloques)	9	72.708	8.0787	3.76 (0.01)
Residual	18	38.7237	2.1513	
Totales	29	174.8947		

FIGURA 15.1



Anexo computacional

Para demostrar el uso del SPSS en la realización de un análisis de varianza para diseños correlacionados, se eligieron los datos del estudio de Reinmuth y Barnes. Los datos de la tabla 15.13 se vacían en una tabla desglosada del SPSS, como se observa en la figura 15.1. En esta figura también se ilustran los menús y las pantallas que aparecen cuando se seleccionan (resaltan) "statistics" y "ANOVA Models".

Seleccione "Repeated Measures" del menú de ANOVA Models. Después de seleccionar la opción "Repeated Measures", se presenta una nueva pantalla (mostrada en la figura 15.2). En el primer cuadro de "Within-Subject Factor Name" se escribe la etiqueta "Type"

FIGURA 15.2

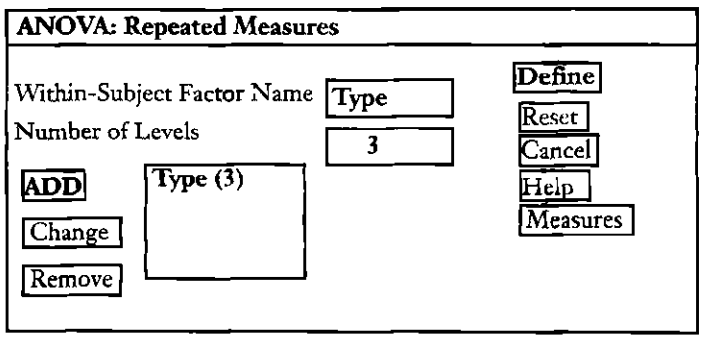
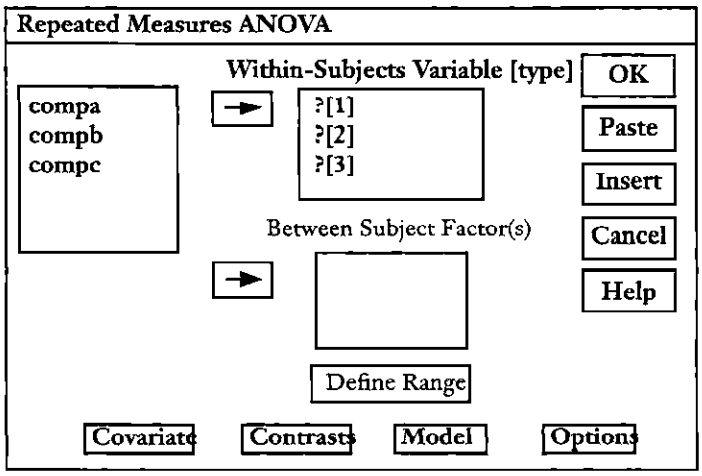


FIGURA 15.3



(para representar "tipo de compañía"). En el cuadro que está debajo de ella se anota el número "3"; esto le indica al SPSS que hay tres grupos de bloques. Después haga clic en el botón "ADD"; entonces verá "Type(3)" aparecer en el cuadro junto al botón "ADD".

A continuación haga clic en el botón "Define", lo que producirá una nueva pantalla (mostrada en la figura 15.3). En el cuadro de la extrema izquierda aparecen los nombres de los tres grupos: "compa", "compb" y "compc". Resalte cada uno de ellos, uno a la vez, y haga clic en la flecha que apunta hacia la derecha, asociada con el cuadro "Within-Subjects Variable". El resultado de este proceso se presenta en la figura 15.4.

FIGURA 15.4

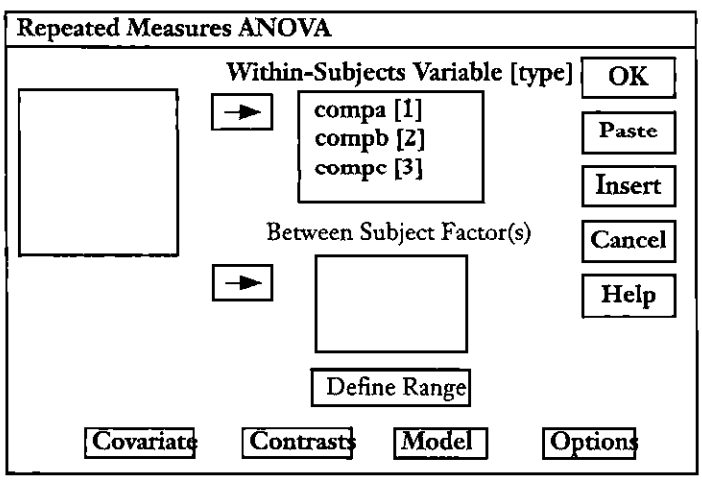


FIGURA 15.5

Source of Variation	SS	DF	MS	F	Sig of F
WITHIN+RESIDUAL	72.71	9	8.08		
CONSTANT	57325.67	1	57325.67	7095.93	000
Source of Variation	SS	DF	MS	F	Sig of F
WITHIN+RESIDUAL	38.72	18	2.15		
TYPE	63.46	2	31.73	14.75	000

Después de haber completado esto para cada grupo que le interese, haga clic en el botón "OK" y el SPSS ejecutará y producirá el resultado deseado.

El resultado abreviado del SPSS se incluye en la figura 15.5. La variable de bloque se representa como Within + Residual en la mitad superior de la tabla. La variable de bloque Within + Residual de la mitad inferior es el componente de error; se verá que corresponde al resumen del cálculo realizado a mano en la tabla 15.14.

RESUMEN DEL CAPÍTULO

1. Se examina el análisis de varianza para sujetos que no fueron asignados aleatoriamente, es decir, grupos no independientes (correlacionados).
2. Los sujetos o los grupos son apareados o utilizados en una situación de medidas repetidas.
3. Se demuestra el ANOVA de un factor con sujetos apareados a lo largo de condiciones de tratamiento. Cuando esto ocurre, una diferencia significativa puede ser enmascarada por la correlación entre las condiciones de tratamiento y los sujetos.
4. Otra fuente sistemática de varianza es la separación de la contribución de los sujetos o bloques (correlación).
5. Con una alta correlación entre sujetos y condiciones, la cantidad de varianza sistemática extraída de la varianza no explicada, o del error, puede ser sustancial.
6. Se presenta un resumen de los tipos de ANOVA cubiertos en los capítulos 13 y 14: diseño completamente aleatorizado, diseño de bloque aleatorizado, diseño factorial aleatorizado, diseño factorial mezclado y el diseño dentro de sujetos.
7. El diseño factorial mezclado contiene tanto grupos independientes como correlacionados.

SUGERENCIAS DE ESTUDIO

1. Realice un análisis de varianza de dos factores con los dos conjuntos de datos ficticios de la tabla 15.6. Utilice el texto como ayuda. Interprete los resultados y después realice un análisis de varianza de dos factores para los dos conjuntos de las tablas

15.8 y 15.9. Elabore las tablas finales del análisis de varianza y compare. Piense con cautela cómo es que las correcciones de ajuste han afectado a los datos originales.

2. Se les pidió a tres sociólogos juzgar la efectividad general de las oficinas administrativas de 10 escuelas primarias en un distrito escolar particular. Una de sus medidas era la *flexibilidad administrativa* (a mayor calificación mayor flexibilidad). A continuación se presentan las 10 calificaciones de esta medida de los tres sociólogos:

	S_1	S_2	S_3
1	9	7	5
2	9	9	6
3	7	5	4
4	6	5	3
5	3	4	2
6	5	6	4
7	5	3	1
8	4	2	1
9	5	4	4
10	7	5	5

   a) Realice un análisis de varianza de dos factores en la forma que se describió en el capítulo.
   b) ¿Concuerdan los tres sociólogos en sus calificaciones medias? ¿Alguno de ellos parece más severo en sus calificaciones?
   c) ¿Existen diferencias sustanciales entre las escuelas? ¿Qué escuela parece tener la mayor flexibilidad administrativa? ¿Cuál es la menos flexible?
   [Respuestas: ***a***) F(columnas) = 24.44 (0.001); F(renglones) = 14.89 (0.001), ***b***) no, sí; ***c***) sí, no, 2, no, 8.]

3. Extraiga 30 dígitos del 0 al 9, de una tabla de números aleatorios (utilice el apéndice C si así lo desea, o genere los números en una computadora, microcomputadora o calculadora programable). Divida arbitrariamente los números obtenidos en tres grupos de 10 dígitos cada uno.
   a) Realice un análisis de varianza de dos factores. Suponga que los números en cada renglón son datos de un individuo.
   b) Ahora sume constantes a los tres números de cada renglón como sigue: 20 a los primeros dos renglones, 15 a los siguientes dos, 10 a los siguientes dos, 5 a los siguientes y cero a los últimos dos renglones. Realice un análisis de varianza de dos factores para tales datos.
   c) ¿Qué ha hecho al "sesgar" los números de los renglones de esta manera?
   d) Compare la suma de cuadrados y los cuadrados medios para los datos de los incisos ***a***) y ***b***). ¿Por qué las sumas de cuadrados *totales* y los cuadrados medios son diferentes? ¿Por qué las sumas de cuadrados *entre columnas* y las *residuales*, y los cuadrados medios son iguales? ¿Por qué las sumas de cuadrados *entre renglones* y los cuadrados medios son diferentes?
   e) Elabore un problema de investigación a partir de todo esto e interprete los resultados. ¿El ejemplo es realista?

4. En una extraordinaria serie de estudios, Miller (1969) ha demostrado que, contrario a la creencia tradicional, es posible aprender a controlar respuestas autónomas como el latido cardiaco, la secreción de orina y las contracciones intestinales. En uno de estos estudios Miller y DiCara (1968) publicaron todos sus datos sobre la secreción

TABLA 15.15 *Datos del condicionamiento de secreción de orina (estudio de Miller y DiCare)*[a]

I *Dos muestras antes del condicionamiento*			*II* *Muestra recompensada por incremento de orina*	
Muestra 1	Muestra 2		Antes	Después
.023	.018	1	.023	.030
.014	.015	2	.014	.019
.016	.012	3	.016	.029
.018	.015	4	.018	.030
.007	.030	5	.007	.016
.026	.027	6	.026	.044
.012	.020	7	.012	.026

[a] Las medidas son milímetros por cada 100 gramos del peso corporal. Los datos listados debajo de I son de dos muestras de ratas asignadas aleatoriamente a los dos grupos. Los datos listados bajo II son las medidas de recompensa, antes y después, de la muestra 1 de I. Los datos de I fueron analizados por medio de un análisis de varianza de un factor; los datos de II, con un análisis de varianza de dos factores o de medidas repetidas.

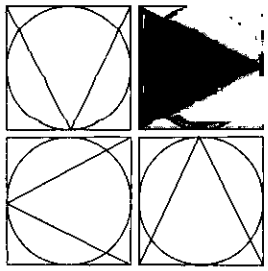
de orina; parte de los datos están reproducidos en la tabla 15.15. Los datos contenidos en II a la derecha son los incrementos en la secreción de orina de siete ratas seleccionadas aleatoriamente de un grupo de 14 ratas, antes y después del "entrenamiento", el cual consistió en un condicionamiento instrumental: siempre que la rata secretaba orina, era reforzada. Entonces, tales datos son medidas repetidas. Si el condicionamiento "funcionaba", las medias posteriores al entrenamiento debían ser significativamente diferentes. Los datos de I (a la izquierda) son las medidas *antes* de dos grupos asignados aleatoriamente (para otro propósito experimental). Debido a que éstas son medidas de secreción de orina *antes* de la manipulación experimental, las medias no deben ser significativamente diferentes. Los análisis antes sugeridos no fueron los que llevaron a cabo Miller y DiCara en su estudio.

a) Realice un análisis de varianza de un factor con las mediciones de I (utilice seis decimales).

b) Elabore un análisis de varianza de dos factores de medidas repetidas, de las mediciones de II (utilice seis decimales). (Nota: puede ser más fácil multiplicar cada una de las puntuaciones por 1 000 antes de realizar el análisis: es decir, mueva el punto decimal tres lugares a la derecha. ¿Afecta esto las razones *F*? Si usted hace esto, entonces tres decimales son suficientes.)

c) Interprete los resultados.

[Respuestas: ***a***) $F = .73$ (n.s); ***b***) $F = 43.88$ ($p < 0.01$).]



CAPÍTULO 16

ANÁLISIS DE VARIANZA NO PARAMÉTRICOS Y ESTADÍSTICOS RELACIONADOS

Es posible, por supuesto, analizar los datos y realizar inferencias acerca de las relaciones entre variables sin utilizar estadísticos. Por ejemplo, algunas veces los datos son tan obvios que en realidad no se necesitan pruebas estadísticas; si todas las puntuaciones de un grupo experimental son mayores (o menores) que las de un grupo control, entonces una prueba estadística resulta superflua. También es posible tener estadísticos de naturaleza bastante diferente a los que se han estudiado; es decir, estadísticos que utilizan otras propiedades de los datos, en lugar de aquellas estrictamente cuantitativas. Se puede inferir un efecto de X sobre Y si las puntuaciones de un grupo experimental son en su mayoría, de cierto tipo (por ejemplo altas y bajas), al contrastarlas con las puntuaciones de un grupo control. Esto se debe a que, con base en la aleatorización y el azar, se espera casi el mismo número de los diferentes tipos de puntuaciones tanto en el grupo control como en el grupo experimental.

De la misma forma, si se ordenan, de la más alta a la más baja, todas las puntuaciones de los grupos control y experimental por el orden de sus rangos, entonces con base únicamente en el azar se puede esperar que la suma o el promedio de los rangos de clase en cada grupo sea aproximadamente el mismo. Si no es así, si los rangos más altos o los más bajos tienden a concentrarse en uno de los grupos, entonces se infiere que ha operado "algo" diferente al azar.

De hecho existen muchas formas de abordar y analizar los datos, además de comparar medias y varianzas; pero el principio básico es siempre el mismo si se trabaja en un mundo probabilístico: comparar los resultados obtenidos con aquellos esperados por el azar o con expectativas teóricas. Por ejemplo, si se administran cuatro tratamientos a los participantes y esperamos que uno de los cuatro sobresalga sobre los demás, se puede comparar la media del grupo favorecido con el promedio de los otros tres grupos por medio de un análisis de varianza o comparaciones planeadas. No obstante, suponga que los datos son muy irregulares en uno o varios aspectos y que se teme por la validez de las pruebas de significancia usuales. ¿Qué se puede hacer? Se pueden ordenar las observaciones de acuerdo al orden de clase de sus rangos. Si ninguno de los cuatro tratamientos tiene mayor influencia que los otros, se espera que los rangos se dispersen aproximadamente igual entre los cuatro grupos. Sin embargo, si el tratamiento A_2 tiene una preponderancia de rangos altos (o bajos), entonces se concluye que se alteró la expectativa común. Este razonamiento constituye buena parte de la base de la llamada estadística no paramétrica y libre de distribución; no existe un nombre único para los estadísticos en cuestión. Los dos nombres más apropiados son "estadística no paramétrica" y "estadística libre de distribución". Esta última, por ejemplo, sugiere que las pruebas estadísticas de significancia no establecen suposiciones sobre la forma precisa de la población muestreada. En este libro se utilizará el término "estadística no paramétrica" para identificar aquellas pruebas estadísticas de significancia que no se basan en la llamada teoría estadística clásica, la cual se fundamenta, en gran parte, en las propiedades de las medias y las varianzas, así como en la naturaleza de las distribuciones.

En este capítulo se examinan ciertas formas interesantes de análisis de varianza no paramétricos. Se mencionarán brevemente otras formas de estadísticos no paramétricos. El capítulo tiene dos propósitos principales: introducir al lector a las ideas que subyacen a la estadística no paramétrica, especialmente al análisis de varianza no paramétrico, y mostrar la semejanza esencial de la mayoría de los métodos que facilitan la inferencia.

El estudiante debe estar consciente de que el estudio cuidadoso de la estadística no paramétrica brinda conocimiento profundo de los estadísticos y de la inferencia estadística. El discernimiento logrado se debe tal vez al considerable relajamiento del pensamiento que parece ocurrir cuando se trabaja tangencialmente a la estructura estadística usual. Se puede observar, por así decirlo, una perspectiva más amplia; inclusive se pueden inventar pruebas estadísticas, una vez que se comprenden bien las ideas básicas. En resumen, las ideas estadísticas e inferenciales se generalizan con base en ideas fundamentales relativamente simples.

Estadística paramétrica y no paramétrica

Una de las preguntas más comunes planteadas a los estadísticos es si se deben utilizar o no métodos estadísticos paramétricos y no paramétricos al analizar datos (véase Allison, Gorman y Primavera, 1993). Una prueba estadística paramétrica, el tipo de pruebas que se han estudiado hasta el momento, depende de un número de supuestos sobre la población de donde se obtienen las muestras utilizadas en la prueba. El supuesto más conocido es

que las puntuaciones de la población están distribuidas normalmente. Una prueba estadística no paramétrica o libre de distribución no depende de supuestos sobre la forma de la población de la muestra o de los valores de los parámetros de la población. Por ejemplo, las pruebas no paramétricas no dependen del supuesto de normalidad de las puntuaciones de la población. El problema de los supuestos es difícil, polémico y controversial. Algunos estadísticos e investigadores consideran que la violación de los supuestos es un asunto serio que lleva a la invalidez de las pruebas estadísticas paramétricas. Otros piensan que, en general, la violación de los supuestos no es tan seria debido a que pruebas como la F y la t son robustas, lo cual, en general, significa que funcionan bien aun bajo la violación de los supuestos, siempre y cuando las violaciones no sean grandes ni múltiples. Sin embargo, otros afirman que este punto de vista equivale a utilizar un zapato como martillo; en efecto, un zapato puede servir como martillo en ciertas situaciones, pero en realidad fue diseñado para usarse para proteger el pie. Hace años, Prokasy (1962) señaló que en ciertas situaciones puede ser correcto utilizar métodos paramétricos para datos dudosos, pero que el poder de esta deducción analítica resulta ilusorio si se usa para hacer inferencias acerca de atributos psicológicos. Brady (1988) afirma que los datos en ciencias sociales generalmente son imprecisos y que con este tipo de datos sólo deben utilizarse los métodos estadísticos más conservadores (no paramétricos). Sin embargo, Toothaker y Newman (1994) apoyan el uso de pruebas paramétricas para datos que no se distribuyen normalmente. La discusión continúa respecto al uso de los estadísticos paramétricos robustos para datos dudosos. Sawilowsky (1993) analiza los mitos que subyacen a la discusión entre el uso de métodos paramétricos y no paramétricos. El trabajo de Zimmerman (véase Zimmerman, 1995a,b; Zimmerman y Zumbo, 1993a, b, 1992) ofrece una solución alternativa para esta discusión. Sin embargo, aquí se examinarán tres supuestos importantes y la evidencia para considerar robustos a los métodos paramétricos. También se analizará un cuarto supuesto —la independencia de las observaciones— debido a su generalidad. Éste se aplica sin importar qué tipo de prueba estadística se utilice. De mayor importancia, es saber que su violación invalida los resultados de la mayoría de las pruebas estadísticas de significancia. Lix, Keselman y Keselman (1996) presentan un análisis de toda la literatura sobre la violación de los supuestos y recomiendan qué método utilizar en ciertas situaciones.

Supuesto de normalidad

El supuesto más conocido que subyace al uso de muchos estadísticos paramétricos es el *supuesto de normalidad*. Al utilizar las pruebas t y F (y por lo tanto, el análisis de varianza), por ejemplo, se asume que las muestras con que se trabaja han sido extraídas de poblaciones normalmente distribuidas; se afirma que si las poblaciones de donde provienen las muestras no son normales, entonces las pruebas estadísticas que dependen del supuesto de normalidad estarán viciadas. Como resultado, los estadísticos y las conclusiones extraídas a partir de las observaciones muestreadas estarán en tela de juicio. Se supone que cuando existe duda respecto a la normalidad de una población, o cuando se sabe que la población no es normal, debe utilizarse una prueba no paramétrica que no se base en el supuesto de normalidad. Algunos maestros exhortan a sus alumnos de pedagogía y psicología a utilizar únicamente pruebas no paramétricas sobre la cuestionable base de que la mayoría de las poblaciones en pedagogía y psicología no son normales. Pero el problema no es tan prosaico.

Homogeneidad de la varianza

El segundo supuesto más importante es aquel que se refiere a la *homogeneidad de la varianza*. En el análisis de varianza se supone que las varianzas dentro de los grupos son estadís-

ticamente las mismas, es decir, se supone que las varianzas son homogéneas de un grupo a otro, dentro de los límites de la variación aleatoria. Si esto no es verdad, la prueba *F* está viciada. Existe una buena razón para afirmar esto: antes se explicó que la varianza dentro de los grupos era un promedio de las varianzas dentro de los dos, tres o más grupos de medidas. Si las varianzas difieren ampliamente, entonces dicho promedio se vuelve cuestionable. El efecto de las varianzas que difieren mucho es inflar la varianza dentro de los grupos; en consecuencia una prueba *F* puede no ser significativa cuando en realidad existan diferencias significativas entre las medias (error tipo II).

Ambos supuestos se han examinado profundamente por medio de métodos empíricos. Se establecieron poblaciones artificiales para extraer muestras de ellas y realizar pruebas *F*. La evidencia que existe hasta la fecha indica que la importancia de la normalidad y de la homogeneidad ha sido sobrestimada; este punto de vista es compartido por el primer autor de este libro, pero no necesariamente por el segundo autor. El artículo de Zimmerman y Zumbo (1993b) muestra situaciones donde los métodos no paramétricos funcionaban mejor que los métodos paramétricos cuando no se cumplían ciertos supuestos, y viceversa. Si las poblaciones no se alejan demasiado de la normalidad, pueden usarse métodos paramétricos en lugar de los no paramétricos sin preocuparse demasiado. La razón de ello es que las pruebas paramétricas casi siempre son más poderosas que las pruebas no paramétricas. (El poder de una prueba estadística se refiere a la probabilidad de que se rechace la hipótesis nula cuando en realidad es falsa.) Existe una situación, o más bien una combinación de situaciones, que pueden ser peligrosas. Boneau (1960) encontró que cuando había heterogeneidad de la varianza y diferencias en los tamaños muestrales de los grupos experimentales, las pruebas de significancia se veían afectadas desfavorablemente. Zimmerman (1995b) también ha señalado que las puntuaciones extremas ejercen una mayor influencia en las pruebas paramétricas como la prueba *t* y la prueba *F*, que en las pruebas no paramétricas.

Continuidad e intervalos iguales de medida

Un tercer supuesto es que las medidas a analizar son medidas continuas con intervalos iguales. Como se verá en un capítulo posterior, este supuesto subyace a las operaciones aritméticas de suma, resta, multiplicación y división. Las pruebas paramétricas como la prueba *F* y la prueba *t* dependen, obviamente, de dicho supuesto, pero muchas pruebas no paramétricas dependen de ello. La importancia de dicho supuesto también ha sido sobrestimada. Anderson (1972) dispuso de él de forma efectiva, y Lord (1972) lo satiriza en un artículo muy conocido sobre las estadísticas en el futbol.

A pesar de estas conclusiones, es aconsejable tener en mente tales supuestos. No resulta sensato utilizar procedimientos estadísticos —o en ese caso, cualquier tipo de procedimiento de investigación— sin el debido respeto por los supuestos que subyacen a estos procedimientos; si éstos son violados seriamente, las conclusiones extraídas a partir de los datos de investigación pueden ser erróneas. Para el lector que ha sido alarmado por algunos textos de estadística, quizá el mejor consejo sea utilizar estadística paramétrica, así como el análisis de varianza de manera rutinaria, pero examinando los datos respecto a alejamientos grandes de la normalidad, homogeneidad de la varianza e igualdad de los intervalos. Es necesario tener cuidado con los problemas de medición y su relación con las pruebas estadísticas, así como estar familiarizados con los estadísticos no paramétricos básicos para utilizarlos cuando sea necesario. También debe tenerse en cuenta que con frecuencia las pruebas no paramétricas son rápidas y fáciles de usar, y que son excelentes para pruebas, si no siempre definitivas, por lo menos preliminares.

Independencia de las observaciones

Otro supuesto importante en medición y en estadística es el de la independencia de las observaciones, también llamada independencia estadística. Ya se estudió la independencia estadística en el capítulo 7, donde se examinó la independencia, la exclusión mutua y la exhaustividad de los eventos y sus probabilidades. (Se invita al lector a revisar esa sección del capítulo 7.) Sin embargo, aquí se reexamina la independencia en el contexto de los estadísticos debido a la importancia especial que tienen los principios involucrados. El supuesto de independencia se aplica tanto para la estadística paramétrica como para la no paramétrica; es decir, no es posible escapar a sus implicaciones utilizando un enfoque estadístico diferente que no involucre este supuesto.

La definición formal de independencia estadística es: si dos eventos, A_1 y A_2, son estadísticamente independientes, la probabilidad de su intersección es: $p(A_1 \cap A_2) = p(A_1) \cup p(A_2)$. Si, por ejemplo, un estudiante contesta una prueba de 10 reactivos, la probabilidad de responder correctamente cualquier reactivo al azar (adivinando) es 1/2. Si los reactivos y sus respuestas son independientes, entonces la probabilidad de responder correctamente dos, tres y siete al azar es: $1/2 \times 1/2 \times 1/2 = .125$; y de forma similar para los 10 reactivos: .001.

En investigación se asume que las observaciones son independientes, es decir, que efectuar una observación no ejerce ninguna influencia sobre la realización de otra observación. Por ejemplo, si se está observando el comportamiento cooperativo de los niños y se nota que Ana parece ser muy cooperadora, entonces existe la posibilidad de violar el supuesto de independencia, pues se *esperará* que su comportamiento futuro sea cooperativo; si, de hecho, la expectativa opera, entonces las observaciones no son independientes.

Las pruebas estadísticas asumen la independencia de las observaciones que producen los números que se incluyen en los cálculos estadísticos. Si las observaciones no son independientes, entonces se vician las operaciones aritméticas y las pruebas estadísticas. Por ejemplo, si el reactivo 3 de la prueba de 10 reactivos en realidad contiene la respuesta correcta del reactivo 9, entonces las respuestas a los dos reactivos no serán independientes. Se altera la probabilidad de tener correctos los 10 reactivos por el azar; en lugar de .001, la probabilidad será alguna cifra más alta, y se contaminarán los cálculos de las medias y otros estadísticos estarán contaminados. La violación de este supuesto parece ser muy común, tal vez porque es muy fácil hacerlo.

En el capítulo 7 se presentó un ejemplo sutil de la violación de este supuesto, cuando se reprodujo una tabla (tabla 7.3), cuyos datos eran actos agresivos en lugar del número de animales que actuaban agresivamente. Suponga que se tiene una tabulación cruzada de frecuencias y que se calcula χ^2 para determinar si los datos de las casillas se apartan significativamente de lo esperado por el azar. La N total debe ser el número total de unidades en la muestra. Las unidades son individuos o algún tipo de agregados (como grupos) que han sido observados de manera independiente. Las N de las fórmulas estadísticas suponen que el tamaño de las muestras son el número de unidades involucradas en el cálculo, donde cada unidad es observada de forma independiente.

Por ejemplo, si se tiene una muestra de 16 participantes, entonces $N = 16$. Suponga que se observaron varios actos de algunos de los participantes y que se registraron las frecuencias de la ocurrencia de dichos actos. Además, suponga que se observaron un total de 54 actos y que este número, 54, se utilizó como N, lo cual implicaría una flagrante violación del supuesto de independencia de las observaciones. En pocas palabras, los datos en las tablas de frecuencias deben ser los números de las observaciones independientes. No es posible contar varias ocurrencias de un tipo de evento de una persona. Si N es el número de personas, entonces no puede convertirse en el número de ocurrencias de even-

tos de las personas. Éste es un punto sutil y, a la vez, peligroso. Los análisis estadísticos de numerosos estudios publicados sufren la violación de dicho principio. Anteriormente se revisó la tabla de un análisis de varianza factorial, cuyas cifras eran el número de ocurrencias de ciertos eventos y no las unidades verdaderas de análisis —los individuos de la muestra—. El problema aquí no es tanto que la violación de independencia sea inmoral; es un delito de investigación, pues puede llevar a conclusiones erróneas acerca de la relación entre variables.

Análisis de varianza no paramétrico

Los métodos no paramétricos del análisis de varianza estudiados aquí dependen del ordenamiento de rangos. Se estudian las formas básicas: análisis de un factor o de dos factores o análisis de medidas repetidas.

Análisis de varianza de un factor: la prueba de Kruskal-Wallis

Un investigador interesado en las diferencias en conservadurismo de tres consejos de educación no puede administrar una medida de conservadurismo a los miembros del consejo; por lo tanto, el investigador pide a un juez experto ordenar por rangos a todos los miembros de los tres consejos, con base en discusiones privadas con ellos. Los tres consejos tienen seis, seis y cinco miembros, respectivamente. Los rangos de todos los miembros se muestran en la tabla 16.1.

Si no hay diferencias respecto al conservadurismo entre los tres consejos, entonces los rangos deben distribuirse aleatoriamente en las tres columnas; por lo tanto, las sumas de los rangos (o sus medias) en las tres columnas deben ser aproximadamente iguales. Por otro lado, si existen diferencias en conservadurismo entre los tres grupos, entonces los rangos en una columna deben ser mayores que los rangos en otra columna, con la consecuente suma o media de rangos clase mayor.

Kruskal y Wallis (1952) ofrecen una fórmula para evaluar la significancia de tales diferencias. Esta fórmula y otras alternativas pueden encontrarse en numerosos libros de texto de estadística (véase Comrey y Lee, 1995; Hays, 1994).

TABLA 16.1 *Rangos de 17 miembros de tres consejos de educación con respecto a su conservadurismo*

	Consejos		
	I	II	III
	12	11	4
	14	16	3
	10	5	8
	17	7	1
	15	6	9
	13	2	
ΣRangos	81	47	25
M	13.50	7.83	5.00

$$H = \frac{12}{N(N+1)} \sum \frac{R_j^2}{n_j} - 3(N+1) \qquad \textbf{(16.1)}$$

donde N es igual al número total de rangos; n_j es igual al número de rangos en el grupo j; y R_j es igual a la suma de los rangos en el grupo j. Al aplicar la ecuación 16.1 a los rangos de la tabla 16.1, primero se calcula $\sum R_j^2/n_j$

$$\sum \frac{R_j^2}{n_j} = \frac{(81)^2}{6} + \frac{(47)^2}{6} + \frac{(25)^2}{5} = 1\,093.5 + 368.17 + 1\,25.0 = 1\,586.67$$

Sustituyendo en la ecuación 16.1 resulta lo siguiente:

$$H = \frac{12}{17(17+1)} \cdot 1\,586.67 - 54 = 62.22 - 54 = 8.22$$

H se distribuye aproximadamente como χ^2. Los grados de libertad son $k - 1$, donde k es el número de columnas o grupos, o $3 - 1 = 2$. Al verificar la tabla de la χ^2 se encuentra que el resultado es significativo al nivel .02; por lo tanto los rangos no son aleatorios.

El método de Kruskal y Wallis es análogo al análisis de varianza de un factor: es simple y efectivo. Algunas veces la medición es tal que vuelve dudosa la legitimidad de la aplicación de los análisis paramétricos; por supuesto que medidas dudosas también pueden transformarse. La esencia de la idea de transformación consiste en alterar medidas que no son respetables (estas medidas pueden carecer de normalidad u otras razones). Se transforman a una forma más respetable por medio de una función lineal del tipo $y = f(x)$, donde y es una puntuación transformada, x es la puntuación original y f es alguna operación ("la raíz cuadrada de") de x (véase Zimmerman, 1995a; Draper y Smith, 1981; Box, Hunter y Hunter, 1978).

Pero en muchos casos es posible ordenar fácilmente las puntuaciones de acuerdo a su rango y realizar el análisis con los rangos. También existen situaciones de investigación en las que la única forma posible de medición es el rango o la medición ordinal; la prueba de Kruskal y Wallis es más útil en tales situaciones. Sin embargo, también es útil cuando los datos son irregulares, pero susceptibles de ser ordenados.

Análisis de varianza de dos factores: la prueba de Friedman

En situaciones donde los participantes están apareados o donde los mismos participantes son observados más de una vez, se utiliza una forma de análisis de varianza de orden de rangos, concebida originalmente por Friedman (1937). También puede emplearse un análisis de varianza ordinario de dos factores de los rangos.

Un investigador educativo, preocupado respecto a la relación entre el desempeño y la percepción de la competencia para la enseñanza, pidió a un grupo de profesores evaluarse entre sí, mediante un instrumento de medición para instructores. También pidió a los administradores y a los estudiantes evaluar a los mismos profesores. Puesto que el número de profesores ("colegas"), administradores y estudiantes era diferente, promedió las calificaciones de los miembros de cada grupo evaluador. La hipótesis establecía que los tres grupos de evaluadores diferirían significativamente en sus evaluaciones. El investigador también deseaba saber si había diferencias significativas entre los profesores. Los datos de una parte del estudio se presentan en la tabla 16.2.

TABLA 16.2 *Medias hipotéticas de las calificaciones de profesores realizadas por sus colegas, administradores y estudiantes, con los rangos de las calificaciones[a] medias de clase de los tres grupos de evaluadores*

Profesores	Colegas		Administradores		Estudiantes	
A	28	(3)	19	(1)	22	(2)
B	22	(1)	23	(2)	36	(3)
C	26	(2)	24	(1)	29	(3)
D	**44**	(2)	34	(1)	**48**	(3)
E	35	(1)	39	(2)	40	(3)
F	40	(2)	38	(1)	45	(3)
ΣRangos		11		8		17

[a] Los números en la tabla son calificaciones compuestas. Los números entre paréntesis son los rangos: a mayor número (o rango) mayor será la competencia percibida. Nota: Las calificaciones de cada *renglón* están ordenadas por rango y reflejan las diferencias entre los tres grupos sobre cada profesor.

Existen diferentes formas de analizar estos datos. Primero, por supuesto, puede usarse un análisis de varianza ordinario de dos factores. Si los números analizados parecen adaptarse razonablemente bien a los supuestos analizados con anterioridad, éste sería el mejor análisis. En el análisis de varianza, la razón F para las columnas (entre evaluadores) es 4.70, que es significativa al nivel .05; y la razón F para los renglones es 12.72, que es significativa al nivel .01. Se apoya la hipótesis del investigador, lo cual está indicado por las diferencias significativas entre las medias de los tres grupos. Los profesores también difieren significativamente.

Ahora considere que el investigador está inquieto por el tipo de datos recopilados y decide utilizar un análisis de varianza no paramétrico; es evidente que no debe emplearse el método de Kruskal-Wallis. El investigador decide utilizar el método de Friedman, ordenando los datos de acuerdo a los rangos por *renglones*: al hacerlo él prueba las diferencias entre las columnas. Si a dos o más evaluadores se les asigna el mismo sistema de ordenamiento de los rangos, digamos 1, 2, 3, 4, 5, es claro que las sumas y las medias de los rangos asignados por los diferentes evaluadores serán siempre las mismas. En este análisis, entonces, la atención se enfoca en las diferencias entre los evaluadores; las diferencias entre los profesores (evaluados) deberían ser ignoradas. De aquí en adelante, el interés se centra en los rangos de los paréntesis ubicados a la derecha de cada calificación compuesta. También resultan de interés las sumas de los rangos en la parte inferior de la tabla.

La fórmula de Friedman es:

$$\chi_r^2 = \frac{12}{kn(n+1)}\Sigma R_j^2 - 3k(n+1) \qquad (16.2)$$

donde $\chi^2 = \chi^2$, rangos; k es igual al número de rangos; n es igual al número de objetos a ordenar mediante los rangos; ΣR_j es igual a la suma de los rangos en la columna (grupo) j; y ΣR_j^2 es igual a la suma de las sumas elevadas al cuadrado. Primero se calcula ΣR_j^2:

$$\Sigma R_j^2 = 11^2 + 8^2 + 17^2 = 474$$

Ahora se determinan k y n. k es el número de ordenamientos, o el número de veces que se utiliza el sistema de ordenamiento de los rangos, cualquiera que éste sea; aquí $k = 6$. El

TABLA 16.3 *Medias compuestas hipotéticas de las evaluaciones de profesores realizadas por sus colegas, administradores y estudiantes, con sus respectivos rangos*[a]

Profesores	Colegas		Administradores		Estudiantes		ΣR
A	28	(3)	19	(1)	22	(1)	5
B	22	(1)	23	(2)	36	(3)	6
C	26	(2)	24	(3)	29	(2)	7
D	44	(6)	34	(4)	48	(6)	16
E	35	(4)	39	(6)	40	(4)	14
F	40	(5)	38	(5)	45	(5)	15

[a] Los números en la tabla representan calificaciones compuestas. Los números entre paréntesis son rangos: a mayor número (o rango de clase), mayor será la competencia percibida. Nota: las calificaciones de cada columna están ordenadas por rango, lo cual refleja las diferencias entre los seis profesores respecto a la calificación que les dio cada grupo.

número de objetos a clasificar, n o el número de rangos es 3. (En realidad no se está clasificando a los evaluadores: 3 es el número de rangos en el sistema utilizado para el ordenamiento de clase.) Ahora se calcula χ_r^2:

$$\chi_r^2 = \frac{12}{(6)(3)(4)} \cdot 474 - (3)(6)(4) = 79 - 72 = 7$$

Este valor se verifica contra la tabla de la χ^2, con $gl = n - 1 = 3 - 1 = 2$. El valor es significativo al nivel .05. Debe prevenirse al lector de que el nivel de significancia es cuestionable, ya que n y k eran relativamente pequeños.

El investigador también estaba interesado en la significancia de las diferencias entre los profesores de acuerdo a cómo fueron evaluados. El asigna rangos a las calificaciones compuestas en columnas (entre paréntesis en la tabla 16.3). Éstos son los rangos que los grupos evaluadores asignaron a los seis profesores. Los profesores con mayor calificación deberían obtener los rangos más altos, lo cual puede determinarse sumando sus rangos a través de los renglones (véase la columna ΣR en el extremo derecho de la tabla). Ahora $k = 3$ y $n = 6$. Se calcula χ^2, utilizando de nuevo la ecuación 16.2:

$$\chi_r^2 = \frac{12}{(3)(6)(7)} \cdot 787 - (3)(3)(7) = 11.95$$

Al verificar este valor en la tabla de χ^2, con $gl = n - 1 = 6 - 1 = 5$, se observa que es significativa al nivel .05. Con base en su evaluación, los profesores parecen ser diferentes.

Compare estos resultados con los resultados del análisis de varianza ordinario. En este último, se encontró que los tres grupos eran significativamente diferentes al nivel .05. En el caso de la significancia de las diferencias entre los profesores, el análisis también arrojó diferencias significativas. En general, los métodos deben concordar bastante bien.

Al utilizar otro método de análisis de varianza con base en rangos más que en varianzas, los resultados de la prueba de Friedman fueron confirmados. Dicho método, llamado prueba *de rangos studentizados* (véase Pearson y Hartley, 1954) es útil. Los rangos son buenas medidas de la variación para muestras pequeñas pero no para muestras grandes. El principio de la prueba de rangos studentizados es similar al de la prueba F en que se utiliza un rango dentro de grupos para evaluar el rango de las medias de los grupos. Otro método

útil, el de Link y Wallace, se describe en detalle en Mosteller y Bush (1954). Ambos métodos tienen la ventaja de que pueden emplearse con análisis de uno o de dos factores. Y aun existe otro método, que tiene la virtud única de probar una hipótesis *ordenada* de los rangos: la prueba L de Page (1963).

El coeficiente de concordancia, W

Quizás el uso de una medida de la asociación de los rangos proporciona una prueba más directa de las hipótesis del investigador. Kendall (1948) diseñó una medida de ese tipo llamada el coeficiente de concordancia, W. Ahora interesa el grado de asociación o de acuerdo de los rangos de las columnas de la tabla 16.2. Cada grupo evaluador ha asignado virtualmente un rango de clase a cada profesor. Si no hubiera asociación alguna entre dos de los grupos evaluadores, y se calculara un coeficiente de correlación del orden de los rangos entre los rangos, éste debería ser cercano a cero. Por otro lado, si existe acuerdo, el coeficiente debería ser significativamente diferente de cero.

El coeficiente de concordancia, W, expresa el acuerdo promedio entre los rangos, en una escala de .00 a 1.00. Existen dos formas de definir W. El método de Kendall se presentará primero. De acuerdo con este método, W puede expresarse como la razón entre la suma de cuadrados *entre grupos* (o rangos) y la suma de cuadrados *total* de un análisis de varianza completo de los rangos. Entonces, esta razón es la razón de la correlación elevada al cuadrado, η^2, de los datos ordenados.

Cuando hay k rangos de n objetos individuales, el coeficiente de concordancia de Kendall se define mediante:

$$W = \frac{12S}{k^2(n^3 - n)} \tag{16.3}$$

S es la suma de las desviaciones elevadas al cuadrado de los totales de los n rangos con respecto a sus medias. S es una suma de cuadrados entre grupos para los rangos; es como sc_r. (De hecho, si se divide S entre k, $S \div k$, se obtiene la misma suma de cuadrados entre grupos que se obtendría con un análisis de varianza completo de los rangos.)

$$S = (5^2 + 6^2 + \ldots + 15^2) - 63^2/6 = 787 - 661.5 = 125.5$$

Puesto que $k = 3$ y $n = 6$:

$$W = \frac{12 \times 125.50}{3^2(6^3 - 6)} = \frac{1\,506}{9(216 - 6)} = \frac{1\,506}{1\,890} = 0.797 = 0.80$$

La relación entre los tres conjuntos de rangos es sustancial. Para evaluar la significancia de W, se puede utilizar la siguiente fórmula, siempre y cuando $k \geq 8$ y $n \geq 7$ (los grados de libertad son $n - 1$):

$$\chi^2 = k(n - 1)W \tag{16.4}$$

Si k y n son pequeños, se pueden utilizar las tablas apropiadas de S (véase Bradley, 1968, pp. 323-325). También se pueden usar razones F; una forma de hacerlo consiste en realizar un análisis de varianza de dos factores utilizando rangos como puntuaciones; entonces $\eta^2 = W$, y la razón F comprueba la significancia estadística tanto de η^2 como de W. La $W =$

0.80 es estadísticamente significativa al nivel .01. La relación es alta: evidentemente existe un alto acuerdo de los tres grupos en sus ordenamientos de los profesores.

Propiedades de los métodos no paramétricos

Un gran número de métodos no paramétricos eficaces están disponibles, la mayoría de ellos pueden encontrarse en el libro de Bradley (1968) o en el de Siegel y Castellan (1988). Por lo común, estos métodos se basan en alguna propiedad de los datos que puede ser probada contra lo esperado por el azar. Por ejemplo, los resultados probables del lanzamiento de una moneda son una propiedad dicotómica que se prueba convenientemente por medio de estadística binomial (véase capítulo 7). Otra propiedad de los datos es el rango; en muestras pequeñas el rango es un buen índice de la variabilidad. Un método rápido para estimar el error estándar de la media, por ejemplo, es:

$$EE_{M(e)} = \frac{\text{observación mayor} - \text{observación menor}}{N}$$

Una prueba t de la diferencia entre dos medias puede realizarse mediante la siguiente fórmula:

$$t_e = \frac{M_1 - M_2}{\frac{1}{2}(R_1 - R_2)}$$

donde t_e es igual a la t estimada; R_1 es igual al rango del grupo 1 y R_2 es igual al rango del grupo 2.

Otra propiedad de los datos es lo que se puede llamar *periodicidad*. Si hay diferentes tipos de eventos (caras y cruces, hombres y mujeres, preferencia religiosa, etcétera), y los datos numéricos de diferentes grupos se combinan y se ordenan de acuerdo a sus rangos, entonces por azar no debería haber series largas de algún evento en particular, como una serie larga de mujeres en un grupo experimental. La prueba de series se basa en esta idea.

Se analizó otra propiedad de los datos en el capítulo 11: la distribución. Las distribuciones de diferentes muestras pueden compararse entre sí o contra un grupo "criterio" (como la distribución normal) con respecto a las desviaciones. La prueba Kolmogorov-Smirnov analiza la bondad de ajuste de las distribuciones. Es una prueba útil, en particular para muestras pequeñas.

La propiedad más omnipresente de los datos es quizá el orden de rango. Siempre que los datos puedan ser ordenados, es posible probarlos contra las expectativas por el azar. Muchas, quizá la mayoría, de las pruebas no paramétricas son pruebas que involucran el ordenamiento de rangos. Las pruebas de Kruskal-Wallis y de Friedman se basan, obviamente, en el orden de los rangos. Los coeficientes de correlación del orden de los rangos son extremadamente útiles; W pertenece a éstos así como también el coeficiente de correlación de orden de los rangos de Spearman y la tau de Kendall.

Los métodos no paramétricos son virtualmente inagotables. Parece no haber fin a lo que se puede hacer, dados los principios relativamente simples implicados, y las diversas propiedades de los datos que pueden explotarse: rango, periodicidad, distribución y orden de los rangos. Aunque las medias y las varianzas poseen propiedades y ventajas estadísticas deseables, no se está de ninguna forma restringido a ellas. Las medianas y los rangos, por ejemplo, con frecuencia son ingredientes apropiados para las pruebas estadísticas. Mu-

chos de los puntos tratados en este capítulo son una repetición del principio enfatizado una y otra vez, quizá de forma un poco tediosa: evaluar los resultados obtenidos contra lo esperado por el azar. No existe magia respecto a los métodos no paramétricos, no se les ha dado un toque divino; se les aplican los mismos principios probabilísticos.

Otro punto que se tocó anteriormente requiere repetirse y enfatizarse: la mayoría de los problemas analíticos de la investigación del comportamiento pueden manejarse adecuadamente por medio de métodos paramétricos. La prueba *F*, la prueba *t* y otros enfoques paramétricos son robustos en el sentido de que tienen un buen desempeño aun cuando los supuestos subyacentes sean violados —a menos, por supuesto, que las violaciones sean grandes y múltiples—. Entonces los métodos no paramétricos constituyen técnicas secundarias o complementarias bastante útiles, que con frecuencia pueden ser valiosas en la investigación del comportamiento. Quizá lo más importante sea que nuevamente muestran el poder, flexibilidad y amplia aplicabilidad de los preceptos básicos de la probabilidad y del fenómeno de aleatoriedad, enunciado en capítulos previos.

Anexo computacional

La prueba de Kruskal-Wallis en el SPSS

Para demostrar cómo se utiliza el SPSS para analizar los datos en la prueba de Kruskal-Wallis, se crearon datos para un estudio ficticio, en el cual se compararon tres planes de dieta con base en el porcentaje de pérdida de peso. La tabla 16.4 muestra la disposición de los datos. Observe que para el plan A había cinco participantes, cuatro para el plan B y tres para el plan C.

La figura 16.1 presenta cómo se anotan los datos en la hoja de cálculo del editor de datos del SPSS para el análisis. A las personas del plan A se les asignó el valor "1" en la variable "plan". El plan B recibió un "2"; y el plan C, un "3". Además de la disposición de los datos, también se presentan los menús y pantallas resultantes al seleccionar la opción "Statistics".

Elija "Nonparametric Test" del primer menú. Esto produce otro menú; de éste seleccione "K Independent Samples"; después de seleccionarlo, el SPSS presenta una pantalla donde debe definir sus variables (figura 16.2); esta pantalla le pide especificar cuál es la variable dependiente ("Test Variable List") y cuál es la variable independiente ("Group Variable"). Seleccione la variable "weight" (véase figura 16.1) y haga clic en el botón asociado con el cuadro "Test Variable List". Después seleccione la variable "plan" y haga clic en el botón asociado con el cuadro "Group Variable". Necesitará definir el rango de los valores para la variable independiente. La figura 16.3 muestra la pantalla que resulta después de la especificación de las variables. Observe que la variable independiente "plan" tiene dos signos de interrogación dentro de un paréntesis; esto indica que debe señalarle al SPSS el rango de valores que ha asignado a los niveles de la variable independiente.

Se tienen tres grupos independientes (es decir, planes de dieta) y se les han asignado los números "1, 2 y 3". Cuando haga clic en el botón "Define Range" surgirá otra pantalla

TABLA 16.4 *Datos de un estudio ficticio de la comparación de planes de dieta*

Plan A	23	41	42	36	30
Plan B	20	24	25	26	
Plan C	40	42	37		

FIGURA 16.1 *Datos relacionados de la tabla 16.4*

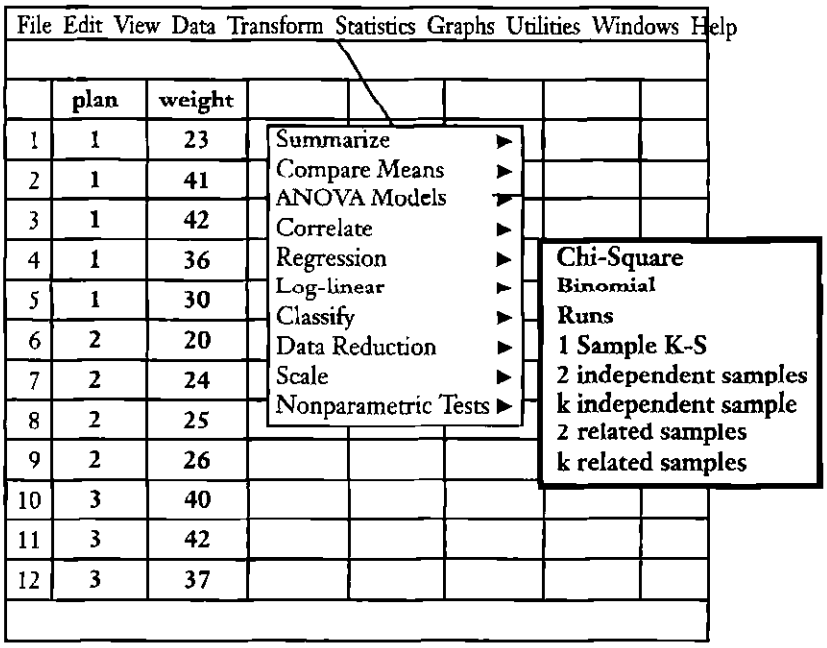


FIGURA 16.2 *Panel del SPSS para la especificación de las variables*

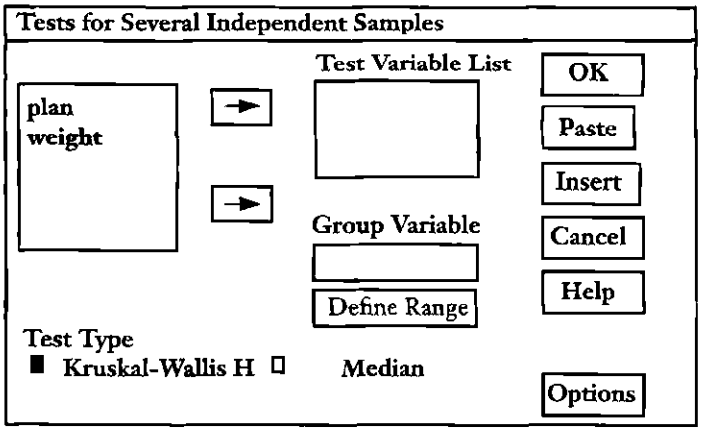
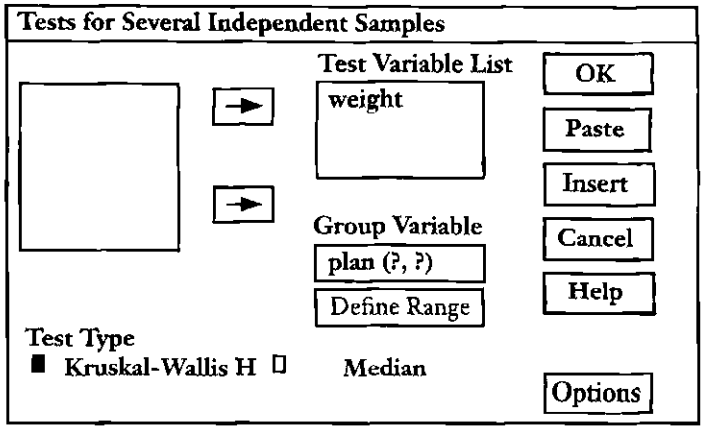


FIGURA 16.3 *Desplazamiento de las variables "peso" y "plan" dentro de los cuadros apropiados*



que le permite definir el rango de valores discretos asignados a los grupos o niveles de la variable independiente (mostrada en la figura 16.4). Teclee un "1" para el valor mínimo y "3" para el valor máximo. Al finalizar, haga clic en el botón "OK" y el SPSS mostrará entonces la pantalla previa con la variable "plan" definida (presentada en la figura 16.5). Una vez que se haya asegurado de seleccionar la opción "Kruskal-Wallis H" (el pequeño cuadro se oscurece), haga clic en el botón "OK" y el SPSS ejecutará el análisis estadístico. Una versión abreviada de los resultados se incluye en la figura 16.6.

FIGURA 16.4 *Pantalla del SPSS para definir el rango de la variable de agrupación*

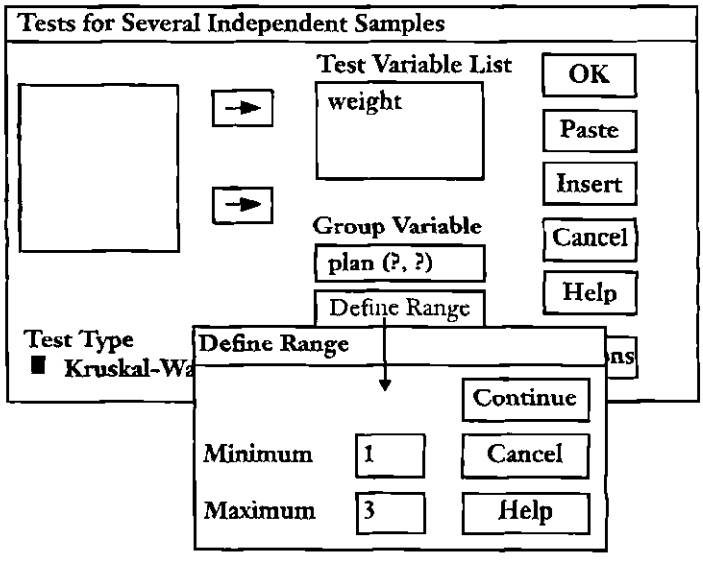
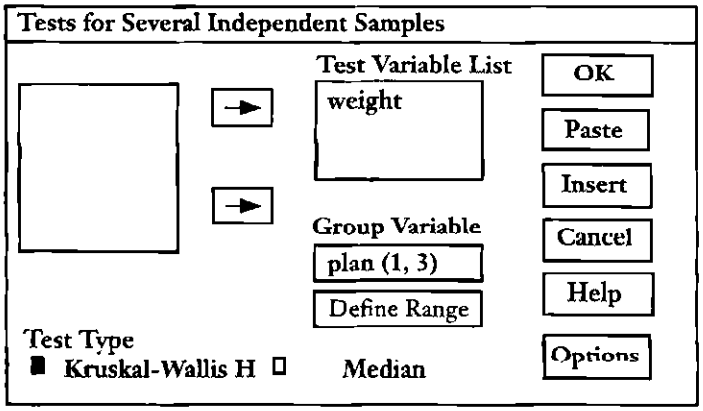


FIGURA 16.5 *Pantalla del SPSS antes de ejecutar el análisis*



La prueba de Friedman en SPSS

Los datos de la tabla 16.2 se utilizaron para demostrar el uso del SPSS para la prueba de Friedman de *k* muestras relacionadas. La figura 16.7 muestra la hoja de cálculo de los datos en el SPSS y presenta también el menú "Statistics". Seleccione de este menú "Nonparametric Test", lo cual lleva a otro menú, donde debe escoger "k related samples". Al hacer esto, el SPSS presenta una nueva pantalla donde se definen las variables. Debajo de "Test Type" elija la prueba "Friedman" haciendo clic en el cuadro pequeño (figura 16.8). Después, seleccione las tres variables: "admin", "peers" y "students", y desplácelas al cuadro "Test Variable" haciendo clic en el botón de la flecha hacia la derecha. El resultado

FIGURA 16.6 *Resultados del SPSS de la prueba de Kruskal-Wallis*

- - - - Kruskal–Wallis One-Way ANOVA

WEIGHT by PLAN

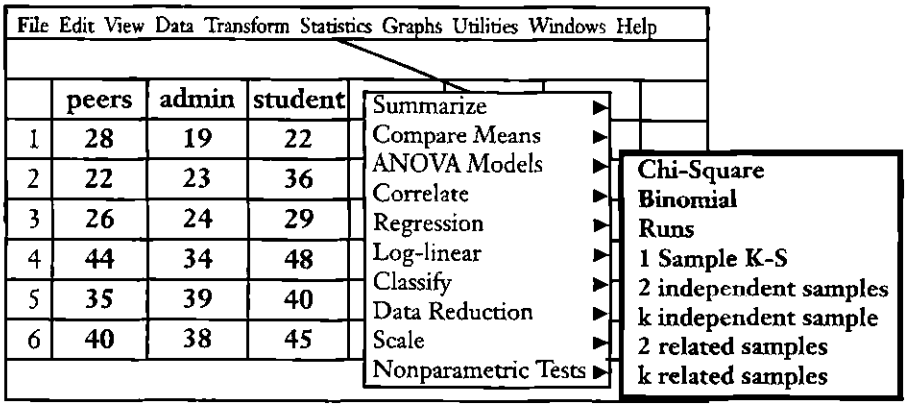
Mean Rank Cases

7.30	5	PLAN = 1
3.25	4	PLAN = 2
9.50	3	PLAN = 3
	12	Total

Corrected for ties

Chi-Square	D.F.	Significance	Chi-Square	D.F.	Significance
5.5731	2	.0616	5.5926	2	.0610

FIGURA 16.7 *Hoja del SPSS para los datos presentados en la tabla 16.2*



de esta operación se presenta en la figura 16.9. Si hace clic en el botón "OK", SPSS realizará la prueba de Friedman con los datos. En la figura 16.10 se muestra una versión editada de la pantalla de resultados del SPSS.

RESUMEN DEL CAPÍTULO

1. En el capítulo se considera el análisis de varianza para datos que provienen de una población desconocida o dudosa.
2. Se analizan las diferencias entre los métodos paramétricos (por ejemplo, prueba *t*, prueba *F*) y los métodos no paramétricos (Wilcoxon, Mann-Whitney, Kruskal-Wallis). (Nota: la T de Wilcoxon y la U de Mann-Whitney no se abarcan en este texto.)
3. Hay cuatro supuestos importantes que algunos autores consideran que deben cumplirse para poder utilizar los métodos paramétricos:

FIGURA 16.8 *Pantalla del SPSS para especificar las variables para el análisis*

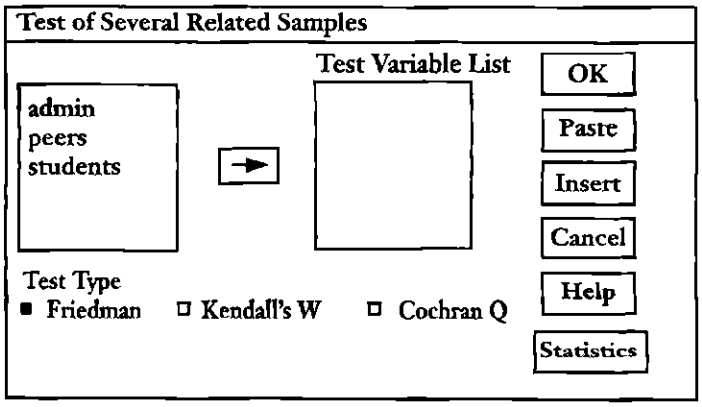
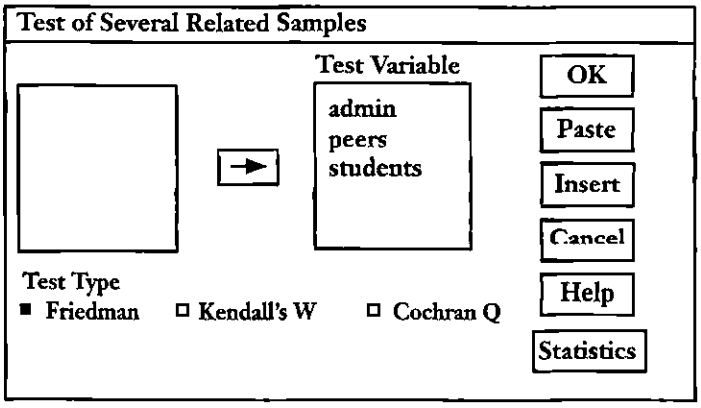


FIGURA 16.9 *Pantalla resultante del SPSS, previa al análisis*



a) Supuesto de normalidad
b) Homogeneidad de la varianza
c) Continuidad e intervalos de medición iguales
d) Independencia de las observaciones

4. Los resultados de la investigación respecto a lo que sucede al utilizar los métodos paramétricos cuando dichos supuestos son violados, han sido contradictorios.
5. Aún existe controversia respecto a cuál de los métodos es superior la mayoría de las veces.
6. La cobertura de los métodos no paramétricos del análisis de varianza incluye:
   a) ANOVA no paramétrico de un factor de Kruskal-Wallis
   b) Prueba de Friedman para un ANOVA de dos factores
   c) Coeficiente de concordancia de Kendall

SUGERENCIAS DE ESTUDIO

1. Una maestra interesada en estudiar el efecto de los libros de trabajo decide conducir un pequeño experimento con su clase. Dividió aleatoriamente la clase en tres gru-

FIGURA 16.10 *Resultados del SPSS para la prueba de Friedman*

```
- - - - Friedman Two-Way ANOVA
-
        Mean Rank  Variable

      1.33  ADMIN
      1.83  PEERS
      2.83  STUDENT

    Cases      Chi-Square      D.F.      Significance
      6          7.0000          2          .00302
```

pos de siete alumnos cada uno y los llamó A_1, A_2 y A_3. Al grupo A_1 le enseñó sin utilizar los libros de trabajo; al grupo A_2, utilizando ocasionalmente los libros de trabajo bajo su dirección, y el aprendizaje del grupo A_3 dependió enormemente del uso de libros de trabajo. Después de cuatro meses, la maestra probó a los alumnos en la materia estudiada. Obtuvo las puntuaciones en forma de porcentajes y pensó que sería cuestionable utilizar el análisis de varianza paramétrico. Ella no sabía que cuando las puntuaciones se encuentran en forma de porcentajes, se transforman fácilmente en puntuaciones que pueden ser sujetas a análisis paramétrico. La transformación apropiada se llama transformación *arco-seno*. Ella utilizó el método de Kruskal-Wallis. Los datos son los siguientes:

A_1	A_2	A_3
55	82	09
32	24	35
74	91	25
09	36	36
48	86	20
61	80	07
12	65	36

Convierta los porcentajes en rangos (del 1 al 21) y calcule H. Interprete. (Para ser significativa al nivel .05, la H debe ser igual o mayor que 5.99, y mayor que 9.21 para el nivel .01; esto es con $k - 1 = 2$ grados de libertad, en la tabla de la χ^2.)

Nota: En estos datos se presentan dos casos de empate de los porcentajes y, en consecuencia, de los rangos. Cuando ocurran empates, tan sólo tome la mediana (o la media) de los rangos que esos porcentajes ocuparían. Por ejemplo, existen tres números 36 en la tabla anterior; la mediana (o la media) del décimo, decimoprimero y decimosegundo rangos es 11; entonces, a los tres números 36 se les asigna un rango de 11. El siguiente rango mayor debe ser 13, puesto que el 10, 11 y 12 ya han sido "utilizados" o asignados. De forma similar, existen dos números 09 en el segundo y tercer rangos; la mediana de 2 y 3 es 2.5. A ambos números 09 se les asigna 2.5, y el siguiente rango mayor, por supuesto, es 4.)

[Respuesta: $H = 7.86(0.05)$.]

2. Un investigador en psicología social estudió la relación entre la conducta de discusión entre miembros del consejo de educación y sus decisiones. En esta investigación se buscaba medir una faceta particularmente compleja de la conducta de discusión, la *conducta antagonista*. El investigador se preguntó si esta conducta podía ser medida de forma confiable. Entrenó a tres observadores y les pidió que ordenaran por rangos las conductas antagonistas de los miembros de un consejo de educa-

	Observadores		
Miembros del consejo	O_1	O_2	O_3
1	3	2	2
2	2	4	1
3	6	6	7
4	1	1	3
5	7	7	6
6	4	3	5
7	5	5	4

ción, durante una sesión de dos horas. Los rangos de los tres observadores se presentan a continuación (rangos altos muestran un alto antagonismo):

a) ¿Cuál es el grado de acuerdo o concordancia entre los tres observadores? (Utilice W.)

b) ¿Es W estadísticamente significativa? (Calcule χ^2 utilizando la ecuación 16.4. Si $\chi^2 = 12.59$, $gl = 6$, entonces es significativa al nivel .05.)

c) ¿Puede el psicólogo social decir que está midiendo de manera confiable el "antagonismo" o la "conducta antagonista"?

[Respuestas: *a*) $W = .86$; $\chi^2 = 15.43$ ($p < .05$); *b*) Sí. *c*) Sí.]

3. Utilizando los datos del ejercicio 2 de las sugerencias de estudio, realice un análisis de varianza de las puntuaciones de *antagonismo* de los miembros del consejo.

a) ¿Cuál es la razón F? ¿Es estadísticamente significativa?

b) Calcule η^2. (Recuerde que $\eta^2 = sc_f/sc_t$.) Compárela con la W calculada en el ejercicio 2.

c) ¿Los miembros del consejo de educación difieren en su conducta antagonista?

[Respuestas: *a*) $F = 14.00$ ($p < .01$), *b*) $\eta^2 = W = .86$; *c*) Sí.]

4. Suponga que obtuvo las siguientes puntuaciones en una medida de complejidad: 27, 21, 14, 12, 6. Obtenga un estimado aproximado y rápido del error estándar de la media (véase el texto).

[Respuesta: $(27 - 6)/5 = 4.20$.]

5. Imagine que usted es un analista especializado y que se le ha pedido inventar y producir un método para evaluar la significancia estadística de series. Una serie es un grupo de valores o identificaciones relacionadas con una población o muestra. Suponga que tiene una muestra de hombres y mujeres, y que está midiendo algún atributo, pero no tiene ningún interés en la variable *género*. Ordene la muestra por rangos de acuerdo con el tamaño de las puntuaciones del atributo. Si el *género* no tiene ninguna relación con el atributo, entonces cuando ordene los casos de acuerdo a los rangos, los hombres y las mujeres deben estar mezclados como si los hubiera colocado en la muestra aleatoriamente. En este caso habría muchas series, por ejemplo, *HH, M, H, MM, H, M, HH, MM, H, M* y, por lo tanto, existe poca o ninguna relación entre el *género* y el atributo. (Recuerde: los casos fueron ordenados de acuerdo con el atributo.) Son 10 series y están en itálicas. Éstas son relativamente muchas series para una muestra de 15 casos. Si, por el otro lado, hubiese relativamente pocas series, por ejemplo: *HHHH, M, HH, M, H, MMMMMM*, o seis series, entonces bien podría existir una relación entre el atributo y el *género*.

a) ¿Qué procedimiento seguiría para diseñar una prueba para evaluar la significancia estadística del número de series en una muestra de n casos? (Sugerencia: Piense en el uso de un generador de números aleatorios de computadora o en una tabla de números aleatorios. No intente encontrar una fórmula. ¡Solamente use la fuerza bruta!)

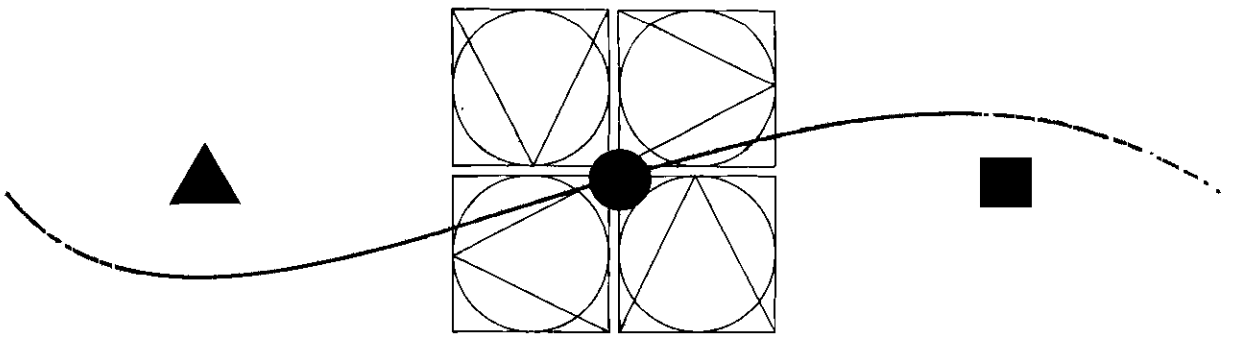
b) Invente dos casos de muestras de 20 cada una, que contenga diferentes números de series y utilice su prueba para evaluar la significancia del número de series en las muestras.

c) Describa los principios básicos de lo que hizo, de tal manera que alguien que no sepa o comprenda la estadística pueda entenderlo. ¿Su prueba es no paramétrica? Explique.

[Nota especial: Éste probablemente es un ejercicio difícil; pero vale la pena trabajar en él y discutirlo con otras personas, especialmente en clase.]

Parte Seis

Diseños de investigación



Capítulo 17

Consideraciones éticas en la realización de investigación en ciencias del comportamiento

Capítulo 18

Diseño de investigación: propósito y principio

Capítulo 19

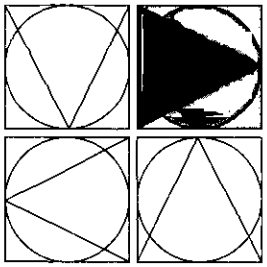
Diseños inadecuados y criterios para el diseño

Capítulo 20

Diseños generales de investigación

Capítulo 21

Aplicaciones del diseño de investigación: grupos aleatorizados y grupos correlacionados



CAPÍTULO 17

CONSIDERACIONES ÉTICAS EN LA REALIZACIÓN DE INVESTIGACIÓN EN CIENCIAS DEL COMPORTAMIENTO

Ficción y realidad

En capítulos anteriores se analizaron la ciencia y las variables involucradas en las ciencias sociales y del comportamiento. También se presentaron algunos de los métodos estadísticos básicos utilizados para analizar los datos reunidos en tales estudios de investigación. En los capítulos posteriores a éste, se analizará la conducción misma del proceso de investigación; antes de hacerlo es necesario presentar un tema importante, el cual incluye la cuestión ética de la investigación. Algunos libros tocan dicho tema en su parte final, después de explicar el plan y los diseños de investigación. Los autores consideramos que este tema debe presentarse antes; el estudiante de investigación necesita esta información para diseñar un estudio adecuado desde el punto de vista ético utilizando los métodos presentados en los siguientes capítulos. Sería ideal que el investigador leyera este capítulo, después los

capítulos sobre los diseños de investigación y posteriormente leyera de nuevo los puntos que se tocan en este capítulo.

¿Qué es la "ética de la investigación"? ¿Qué es "investigación"? Ambos términos son difíciles de definir. Shrader-Frechette (1994) ofrece una definición al distinguir "investigación" de "práctica". Como se estudió en un capítulo anterior, la investigación es una actividad realizada para probar teorías, realizar inferencias y añadir o actualizar información sobre una base de conocimientos. La práctica profesional generalmente no incluye la comprobación de teorías o hipótesis, sino más bien procura incrementar el bienestar de los clientes por medio de acciones e información que han demostrado tener éxito. Algunas de estas acciones fueron establecidas a través de investigación científica previa. Aunque tanto la "investigación" como la "práctica" involucran a la ética, la ética involucrada con el proceso de investigación está dirigida hacia los individuos que realizan investigación y la forma en que conducen el proceso de investigación. Shrader-Frechette señala que la ética de la investigación especifica la conducta que deben mostrar los investigadores del comportamiento durante todo el proceso de su investigación. Keith-Spiegel y Koocher (1985) analizan la ética de la práctica de la psicología; Dawes (1994) ofrece un punto de vista muy crítico sobre la práctica de la psicología y la psicoterapia. Parte de la discusión de Dawes se refiere a la ética de la práctica.

El análisis, el énfasis y la práctica de la ética de la investigación representan eventos relativamente recientes. Antes del siglo xx, se castigaba a los científicos que eran descubiertos experimentando con personas sin el debido consentimiento. Sin embargo, existen ejemplos en la historia donde las violaciones a la ética de la investigación generaron resultados fructíferos. Cuando se piensa acerca de la ética implicada en la investigación con humanos o con animales, no pueden evitarse sentimientos encontrados. Al examinar la historia destacan individuos valientes como Edward Jenner, que inyectó a un niño con una forma debilitada del virus de la viruela, con lo cual desarrolló una vacuna contra la viruela; la historia demuestra que Edward Jenner no solicitó autorización de nadie para hacerlo. O considere al Dr. Barry Marshall quien, para demostrar que las úlceras pépticas eran provocadas por bacterias y no por ácidos, se tragó un cultivo de bacterias y después se trató exitosamente a sí mismo con dosis de antibióticos. Sin embargo, también existen casos documentados de consecuencias trágicas de investigadores que no siguieron los principios éticos de la investigación, y también de quienes cometieron fraude científico. Algunos de estos ejemplos se señalan y analizan por Shrader-Frechette (1994), en un libro excelente que vale la pena leer; también se recomienda el libro de Miller y Hersen (1992) y el de Erwin, Gendin y Kleiman (1994). La evidencia acerca de sospechas de fraude o fraude declarado se remonta a investigaciones realizadas en la antigua Grecia.

Al realizar investigación, el científico sensible a menudo enfrenta dilemas éticos. Antes de 1960 las consideraciones éticas de la investigación se dejaban a la propia conciencia de los investigadores de todos los campos de la ciencia; las publicaciones académicas sobre la conducta apropiada de los científicos brindaban ciertas normas, pero ninguno o pocos de los lineamientos eran obligatorios. La historia ficticia de Martin Arrowsmith, el protagonista de la novela de Sinclair Lewis, *Arrowsmith*, ejemplifica un dilema ético. Aquí el Dr. Martin Arrowsmith, en un estudio de laboratorio, descubre por accidente un principio que es efectivo para destruir bacterias. Arrowsmith lo llama "fago". Cuando la plaga bubónica estalla en un país del tercer mundo, Arrowsmith es enviado a ese país para ayudar a las víctimas y para probar su fago. Arrowsmith sabía que la verdadera efectividad de un fago podía determinarse al aplicarlo solamente a la mitad de la población infectada. A la otra mitad se le daría un placebo o ningún tratamiento. Sin embargo, al ver la alarmante tasa de mortalidad (incluyendo la muerte de su esposa y de un amigo cercano), Arrowsmith decidió administrar el fago a la población completa. Si él hubiese seguido su plan experi-

mental y su fago fuese en realidad efectivo, la gente seleccionada para recibirlo sobreviviría, y quienes hubieran recibido el placebo habrían muerto. La conciencia de Arrowsmith no le permitiría engañar a la mitad de la población y dejarlos morir en nombre de la investigación científica. Él administró el fago a todos; la plaga terminó después de vacunar a los nativos, pero Arrowsmith realmente nunca supo si su fago era o no efectivo. Aunque se trata de ficción, los científicos reales en ocasiones se enfrentan con dilemas similares.

¿Un comienzo?

Estudios realizados en los años sesenta y setenta presentaron evidencia de fraude en investigación y de engaño a los participantes de investigaciones, lo cual condujo a demandar reglas obligatorias específicas para la conducción de la investigación. En 1974 el Congreso de Estados Unidos exigió la creación de consejos de revisión institucionales, cuyo propósito sería revisar la conducta ética de aquellos estudios de investigación que recibieran fondos federales para investigación. Posteriormente, en los años ochenta, se aceptó una legislación que requería que la investigación con fondos federales que incluyera humanos y animales fuera revisada tanto en su conveniencia ética como en su diseño de investigación; en esta década, muchas de las universidades más importantes en Estados Unidos tenían lineamientos para tratar con el mal comportamiento en la investigación. Otros países también empezaron a establecer lineamientos y reglas. Los gobiernos de Suecia y Holanda pidieron que comités de revisión independientes evaluaran todos los estudios biomédicos. Shrader-Frechette (1994) describe dos grandes categorías de cuestiones éticas en la investigación científica: 1) las de los procesos y 2) las de los productos. El proceso de investigación se considera dañino si los participantes no otorgan su consentimiento respecto a los procedimientos que se utilizan en ellos; también se considera perjudicial si se engaña a los participantes o si son reclutados con métodos engañosos. El producto de la investigación es dañino si la conducción de dicha investigación resulta en un ambiente dañino para cualquiera que se ponga en contacto con él. El caso de envenenamiento por radiación a causa de pruebas científicas de armas nucleares constituye un ejemplo de un producto de investigación dañino. Shrader-Frechette describe brevemente esta investigación y sus consecuencias. Saffer y Kelly (1983) ofrecen una explicación más completa en un libro informativo titulado *Countdown Zero*. Ellos describen cómo las consecuencias de las pruebas atmosféricas de la bomba atómica en el desierto de Nevada a finales de los años cuarenta, llegaron a otras partes del desierto. El equipo técnico, el personal y los actores de la película *The Conqueror* estuvieron expuestos a arena radioactiva durante la filmación de la película en el desierto. Todas estas personas desarrollaron cáncer y después murieron de enfermedades relacionadas con ese padecimiento. Algunos de los actores y actrices eran célebres como John Wayne, Susan Hayward y Dick Powell. El libro de Saffer y Kelly también describe cómo la investigación militar estadounidense, sobre cómo llevar a cabo una guerra nuclear en los años cincuenta, condujo a que gran cantidad de personal militar se expusiera a lluvia radioactiva. El propio Saffer fue uno de los soldados que participó en dichos estudios; varios años después de dejar el servicio notó que soldados que habían sido sus compañeros desarrollaron cáncer.

Uno de los casos más infames sobre el uso poco ético del engaño fue el *Estudio Tuskegee* (véase Brandt, 1978). En 1932 el Servicio de Salud Pública de Estados Unidos realizó un experimento con 399 hombres afroamericanos, semianalfabetas y pobres que habían contraído sífilis. Uno de los propósitos de dicho estudio era examinar los efectos de la sífilis en individuos que no recibían tratamiento. Para lograr la participación en el estudio de hombres afroamericanos infectados, se les informó que estaban recibiendo tratamiento cuando, en realidad, no era así. Se midieron y se registraron periódicamente los síntomas de la

sífilis. Se realizaron autopsias en cada individuo después de su muerte. Tomó 40 años para que la población tomara conciencia de esta tragedia en la investigación; cuando se hizo público, el estudio aún continuaba en proceso. La investigación era a todas luces poco ética: una razón es que aún en 1972 el tratamiento les era negado a los supervivientes, mientras pudieron haber sido tratados efectivamente con penicilina, que ya estaba disponible desde los años cuarenta. Una de las principales protestas por conducta poco ética en la investigación se ha enfocado en el empleo del engaño.

El engaño sigue utilizándose en la actualidad en ciertos estudios de investigación; sin embargo, las investigaciones son evaluadas críticamente antes de poder realizarse. Las principales universidades en Estados Unidos tienen un comité de ética de la investigación que monitorea y evalúa los estudios respecto a engaños y efectos perjudiciales potenciales en los participantes; su tarea consiste en asegurar que no se provoque daño en ningún participante.

Uno de los estudios más notorios en psicología que utilizó el engaño fue conducido por el psicólogo social Stanley Milgram, quien reclutó participantes para un experimento de "aprendizaje" (véase Milgram, 1963). A los voluntarios se les dijo que algunos serían maestros y otros serían aprendices; los primeros estaban a cargo de enseñar una lista de palabras a los segundos. A los maestros se les indicó administrar *choques* con un creciente grado de dolor cada vez que el aprendiz cometiera un error. Sin embargo, el propósito real del experimento no era estudiar el aprendizaje, sino la obediencia hacia la autoridad. Milgram estaba muy interesado en saber si había algo de verdad en las afirmaciones de criminales de guerra nazis, quienes declararon haber realizado hechos atroces debido a que sus superiores les habían "ordenado" hacerlo. Sin que los participantes lo supieran, todos ellos funcionaron como "maestros"; es decir, a todos los participantes se les dijo que eran maestros. Ninguno de ellos fungió como "aprendiz"; los aprendices eran cómplices del experimentador, quienes fingieron ser participantes escogidos aleatoriamente para fungir como tales. Además, en realidad no se administraron *choques* en ningún momento; se engañó a los maestros para que creyeran que los gritos de dolor de los aprendices y sus solicitudes de ayuda eran reales. Cuando se les indicó que incrementaran la severidad de los *choques*, algunos de los participantes dudaron; sin embargo, cuando el experimentador les indicó proseguir, ellos continuaron. Incluso siguieron "dando choques" a los aprendices más allá del punto en que ellos "rogaron" que se les liberara del experimento. Los resultados fueron, según Milgram y otros, más allá de lo creíble. Gran cantidad de sujetos (los "maestros") obedecieron sin cuestionar la orden del experimentador: "Por favor continúe" o "No tiene opción, debe continuar", y prosiguieron incrementando el nivel de los *choques* sin importar cuánto rogara el aprendiz al "maestro" que se detuviera. Lo que sorprendió a Milgram en particular fue que ninguno salió del laboratorio disgustado o protestando. Esta notable obediencia se comprobó una y otra vez en diversas universidades donde se repitió el experimento. El enojo público respecto a dicho experimento se centró en el malestar y daño psicológico que pudo haber causado el engaño a los participantes. Aún más, algunas personas sobregeneralizaron y pensaron que se estaban realizando muchos experimentos psicológicos similares.

Años después del ahora célebre estudio, quienes criticaban su estudio, constantemente atacaron a Milgram. Hubo muy poca publicidad alrededor del hecho de que Milgram realizó varios estudios de seguimiento con los participantes, y que no encontró efectos negativos. De hecho, al final de cada sesión experimental se desengañaba a los participantes y se les presentaba con el "aprendiz" para mostrarles que no se habían administrado choques eléctricos peligrosos.

Otra área sensible es aquella dirigida al fraude, que incluye situaciones donde el investigador altera los datos de un estudio de investigación, para demostrar que cierta hipótesis

o teoría es verdadera. Otros casos de fraude incluyen el reporte de hallazgos de investigaciones que nunca se realizaron. La historia muestra que numerosos investigadores prominentes se han involucrado en fraudes (véase Erwin, Gendin y Kleiman, 1994). Uno de los casos más sensacionalistas de acusación de fraude proviene de la psicología. La persona involucrada era Sir Cyril Burt, un prominente psicólogo británico que recibió el título de Sir por su trabajo sobre estadística y la herencia de la inteligencia. Su trabajo se distinguió por el uso de gemelos idénticos, cuya composición genética era la más similar. Burt supuestamente demostró que había un fuerte componente genético en la inteligencia, al examinar la inteligencia de gemelos que se habían criado juntos, contra aquellos que habían sido separados al nacer y que, por lo tanto, fueron criados aparte. El objetivo consistía en determinar qué tanta influencia tenían el ambiente y la herencia sobre la inteligencia. A mediados de los años setenta, después de la muerte de Burt, Leon Kamin (1974) reportó que algunas de las correlaciones reportadas por Burt eran idénticas hasta el tercer decimal; por efecto del azar, ello era altamente improbable. Más adelante se descubrió que algunos de los coautores de Burt en artículos de investigación publicados por la época de la Segunda Guerra Mundial, no pudieron ser localizados. Muchos críticos consideraron que Burt inventó estos coautores para despistar a la comunidad científica; incluso Leslie Hearnshaw, quien fue comisionada por la familia de Burt para escribir su biografía, aseguró haber encontrado evidencia de fraude. Este particular punto de vista sobre el fraude de Burt se detalla en el libro de Gould (1981). Sin embargo, Jensen (1992) presenta un punto de vista sociohistórico diferente sobre Burt, pues afirma que los cargos en contra de Burt nunca se probaron de manera satisfactoria; también ofrece información sobre Burt que nunca se menciona en el libro de Gould ni en otras publicaciones que lo critican.

Casos como el de Tuskegee, Milgram y Burt llevaron a la creación de leyes y reglamentos para restringir o detener el comportamiento de investigación poco ético, en las ciencias médica, del comportamiento y social. Organizaciones profesionales como la American Psychological Association y la American Psychological Society formaron comisiones para investigar y recomendar acciones en casos reportados de comportamiento no ético en la investigación. Sin embargo, la incidencia reportada sobre conducta no ética en científicos de investigación ha sido mínima. Entre los casos que han recibido la publicidad más negativa en investigación de ciencias del comportamiento, se encuentra el de Steven Breuning de la Universidad de Pittsburgh. Breuning fue condenado en 1988 por fabricar datos científicos sobre pruebas de fármacos (Ritalin y Dexedrina) con niños hiperactivos. Los resultados apócrifos de Breuning fueron ampliamente citados y ejercieron influencia para que varios estados de la Unión Americana cambiaran sus reglamentos para el tratamiento de estos niños. El caso de Breuning ilustra cuán peligroso puede resultar el comportamiento fraudulento de un científico.

En las ciencias de la salud y en la medicina, el cardiólogo Maurice Buchbinder fue cuestionado por problemas asociados con sus pruebas del canalizador (Rotablator), un dispositivo que limpia las arterias coronarias. La investigación reveló que el aparato era fabricado por una compañía en la que Buchbinder tenía millones de dólares invertidos en acciones. Algunas de sus violaciones a la ética son: 1) no llevar a cabo exámenes de seguimiento en cerca de 280 pacientes, 2) usar inadecuadamente el aparato en pacientes con enfermedades cardiacas severas y 3) no reportar apropiadamente algunos de los problemas experimentados por los pacientes.

Douglas Richman fue otro médico investigador que adquirió notoriedad por el estudio de un nuevo fármaco para el tratamiento de la hepatitis. Richman fue acusado de no reportar la muerte de pacientes en el estudio, de no informar al productor del fármaco sobre los efectos colaterales perjudiciales y por no explicar adecuadamente los riesgos a los pacientes del estudio. Aunque la incidencia reportada de fraude y comportamiento poco

ético por los científicos es escasa, Shrader-Frechette (1994) ha señalado que muchos comportamientos no éticos pasan inadvertidos o sin reportarse. Incluso las revistas científicas no mencionan nada sobre solicitarle al autor que presente información que avale que el estudio se realizó de manera ética (por ejemplo, el consentimiento de los sujetos por escrito). Es posible que cuando un investigador estudia el comportamiento en humanos, éstos sean puestos en riesgo por medio de coerción, engaño, violación de la privacidad, violación de la confidencialidad, estrés, perjuicio social y falla en la obtención del consentimiento libre informado.

Algunos lineamientos generales

Los siguientes lineamientos consisten en una síntesis del excelente libro de Shrader-Frechette, quien establece los códigos que deben seguir los investigadores en todas las áreas de estudio donde se utilicen participantes humanos y animales. Uno de los temas se centra en las situaciones en las cuales el investigador no debe realizar el estudio. Existen cinco reglas generales a seguir para determinar que el estudio no debe efectuarse.

- Los científicos no deben realizar investigaciones que pongan en riesgo a las personas.
- Los científicos no deben realizar investigaciones que violen las normas del libre consentimiento informado.
- Los científicos no deben realizar investigaciones que conviertan los recursos públicos en ganancias privadas.
- Los científicos no deben realizar investigaciones que puedan dañar seriamente el ambiente.
- Los científicos no deben realizar investigaciones sesgadas.

En el quinto y último punto establecido por Shrader-Frechette, se implican únicamente los sesgos raciales y sexuales. Uno debe tener en cuenta que en todos los estudios de investigación existen sesgos inherentes al diseño de investigación.

No obstante, un criterio importante para decidir acerca de la realización de una investigación son las consecuencias de dicho estudio. Shrader-Frechette afirma que existen estudios que ponen en riesgo al hombre y a los animales, pero que el no realizarlos puede conllevar aun mayores riesgos para los humanos y los animales. En otras palabras, no toda investigación potencialmente peligrosa debe ser condenada. Shrader-Frechette afirma:

> Así como los científicos tienen el deber de realizar investigación pero evitando investigaciones éticamente cuestionables, también tienen la responsabilidad de no tornarse tan escrupulosamente éticos acerca de su trabajo como para amenazar los fines sociales a los que sirve la investigación (p. 37)....

Por lo tanto, el investigador debe ejercitar cierto grado de sentido común al decidir si realiza o no estudios de investigación que involucren la participación de humanos y animales.

Lineamientos de la American Psychological Association

En 1973 la American Psychological Association (APA) publicó lineamientos éticos para los psicólogos. Desde entonces los lineamientos originales se han sometido a una serie de

revisiones. Los últimos lineamientos y principios se publicaron en el ejemplar de marzo de 1990 de la revista *American Psychologist*. Los principios éticos de los psicólogos y el código de conducta pueden encontrarse en la edición de 1994 del manual de estilo de publicaciones de la American Psychological Association. La siguiente sección ofrece una revisión breve de los principios éticos y códigos que son relevantes para la investigación en ciencias del comportamiento. Tales lineamientos están dirigidos hacia la investigación con humanos y con animales. Todas las personas involucradas en un proyecto de investigación están limitadas por los códigos de ética sin importar si son o no psicólogos profesionales o miembros de la American Psychological Association.

Consideraciones generales

La decisión de asumir un proyecto de investigación recae únicamente en el investigador. Algunas preguntas que el investigador debe formularse son: ¿Vale la pena hacerlo? ¿ La información obtenida del estudio será útil y valiosa para la ciencia y el bienestar humano? ¿Ayudará a mejorar la salud de las personas? Si el investigador considera que la investigación es valiosa, entonces debe conducirse con respeto y en consideración al bienestar y la dignidad de los participantes.

El participante con el mínimo riesgo

Una de las consideraciones más importantes sobre si se debe o no realizar el estudio es la decisión concerniente al bienestar del participante: ¿habrá un "sujeto en riesgo" o un "sujeto con el mínimo riesgo"? Si existe la posibilidad de riesgo serio para el participante, el resultado posible de la investigación debe, de hecho, ser de un enorme valor para seguir adelante. Los investigadores que se encuentren en esta circunstancia deben consultar con sus colegas antes de continuar. En la mayoría de las universidades existe un comité especial que revisa los proyectos de investigación para determinar si el valor de dicha investigación amerita el poner en riesgo a los participantes. En toda ocasión, el investigador debe tomar medidas para prevenir el daño a los participantes. Los proyectos de investigación de los estudiantes deben conducirse con la mínima cantidad de riesgo para los participantes.

Justicia, responsabilidad y consentimiento informado

Antes de iniciar el estudio, el investigador y el participante deben realizar un acuerdo que aclare las obligaciones y responsabilidades. En ciertos estudios esto involucra el consentimiento informado, donde el participante expresa su acuerdo en tolerar el engaño, malestar y aburrimiento por el desarrollo de la ciencia. A cambio, el experimentador garantiza la salvaguarda y el bienestar del participante. La investigación en psicología difiere de la investigación médica en este aspecto; la ética de la investigación médica requiere que el investigador informe al participante qué se hará con él y con qué propósito. La mayoría de la investigación en las ciencias sociales y del comportamiento no es tan restrictiva. El investigador en las ciencias del comportamiento necesita hablar sólo de aquellos aspectos del estudio que puedan influir en la voluntad del participante para colaborar. El consentimiento informado no es requerido en investigación de riesgo mínimo. De cualquier manera resulta una buena idea que los investigadores en todos los campos de la investigación establezcan acuerdos claros y justos con los participantes antes de que inicie su participación.

Engaño

Existen requerimientos particulares en muchos estudios de las ciencias del comportamiento. Los participantes colaboran voluntariamente con la creencia de que nada perjudicial les ocurrirá. Sus expectativas y deseos para "hacer lo que el investigador quiere" pueden influir

en el resultado del estudio; por lo que la validez de los resultados puede verse comprometida. El famoso estudio Hawthorne es un caso de tal situación. En este estudio se les dijo con antelación, a los trabajadores de una fábrica, que algunas personas irían a la fábrica a realizar un estudio sobre la productividad de los trabajadores. Éstos, sabiendo que serían evaluados en cuanto a su productividad, se comportaron de manera diferente de como normalmente lo hacían: siendo puntuales, trabajando duro, tomando descansos cortos, etcétera. Como resultado, los investigadores no pudieron obtener una medida verdadera de la productividad de los trabajadores. Aquí entra el engaño; como en un espectáculo de magia, inadvertidamente se desvió la atención de los participantes y esto alteró su comportamiento. Si los investigadores hubiesen asistido a la fábrica como trabajadores "ordinarios", quizá hubieran obtenido una imagen más clara de la productividad de los trabajadores.

Si el investigador puede justificar que el engaño tiene algún valor y no hay procedimientos alternativos disponibles, entonces debe ofrecerse al participante una explicación suficiente tan pronto como sea posible, al finalizar el experimento. Esta explicación se llama *desengaño*. Debe evitarse cualquier procedimiento engañoso que enfrente al participante con una percepción negativa de sí mismo.

Desengaño

Después de recolectar los datos del participante, se le debe explicar cuidadosamente la naturaleza de la investigación. El desengaño es un intento de eliminar cualquier concepto erróneo que el participante pueda tener acerca del estudio. Éste es un elemento extremadamente importante en la conducción de un estudio de investigación. Incluso la explicación sobre el estudio debe realizarse con tiento; necesita explicarse de tal manera que aquellos que acaban de ser engañados no se sientan tontos, estúpidos o avergonzados. En el caso de investigadores estudiantes, sería benéfico tanto para el investigador como para el participante que revisaran juntos los datos. La sesión de desengaño podría utilizarse como una experiencia de aprendizaje, de tal manera que el estudiante participante sienta que adquirió mayor conocimiento sobre la investigación en ciencias del comportamiento. También es aconsejable, si el tiempo lo permite, mostrar al estudiante el laboratorio y explicarle algo acerca de los aparatos.

En el caso de aquellos estudios en los que el desengaño inmediato podría comprometer la validez del estudio, el investigador puede retrasar el desengaño. No obstante, el investigador debe realizar todos los intentos posibles para contactar al participante una vez que se haya completado la recolección de los datos del estudio.

Libertad de coerción

Siempre se debe hacer sentir a los participantes que pueden abandonar el estudio en cualquier momento, sin penalización ni repercusión alguna. Los participantes requieren estar informados de esto antes de comenzar las sesiones experimentales. El investigador de una universidad que utiliza estudiantes de cursos introductorios de psicología como participantes, debe dejarles claro que su participación es voluntaria. En algunas universidades, el curso de introducción a la psicología tiene un componente de investigación en las calificaciones, el cual no puede basarse tan sólo en la participación en estudios de investigación. Para quienes así lo deseen, el componente de investigación puede ser cubierto de otras maneras, tales como con la elaboración de un artículo de investigación. Ofrecer puntos extra por la participación puede ser percibido como coerción.

Protección de los participantes

El investigador debe informar al participante sobre todos los riesgos y peligros inherentes al estudio; debe tener presente que mediante su colaboración, los participantes le están

haciendo un favor. Participar en cualquier investigación quizá provoque algo de estrés. Además, el investigador está obligado a eliminar cualquier consecuencia indeseable de la participación; esto se vuelve relevante en los casos donde se coloca a los participantes en la situación de "hacer nada" o grupo control. En un estudio que examine programas de manejo del dolor sería poco ético colocar personas con dolor crónico en un grupo control, donde no recibirán tratamiento alguno.

Confidencialidad

El principio de la protección del participante contra el daño incluye la confidencialidad. El investigador tiene que garantizarle al participante que los datos que se obtengan de él estarán salvaguardados; es decir, que la información obtenida del participante no será revelada al público de manera que se le pueda identificar. Cuando se trata de información delicada, el investigador debe informar al participante la manera en que ésta será tratada. En un estudio acerca de la conducta sexual y el SIDA, se les pidió a los participantes llenar un cuestionario, ponerlo en un sobre sin marcas y depositarlo dentro de una caja sellada. El investigador aseguró a los participantes que sólo las personas que capturan los datos verían los cuestionarios y ellos "no sabrán y no podrán adivinar quiénes son". Smith y Garner (1976), por ejemplo, tomaron precauciones adicionales para garantizar el anonimato de los participantes en su estudio sobre comportamiento homosexual entre hombres atletas universitarios.

Ética en la investigación con animales

Para algunas personas el empleo de animales en investigación resulta inhumano e innecesario. No obstante, los estudios de investigación que utilizan animales han proporcionado un gran número de avances útiles tanto para los animales como para los humanos. Miller (1985) señala las contribuciones más importantes que la investigación animal ha proporcionado a la sociedad. A diferencia de los participantes humanos, los animales no participan voluntariamente. En oposición a la creencia de los activistas sobre los derechos de los animales, hoy en día muy pocos estudios involucran la situación de infligirles dolor. Los experimentos que utilizan participantes animales en general se permiten siempre y cuando éstos sean tratados humanitariamente. La APA ofrece lineamientos respecto al uso de animales en la investigación del comportamiento y también ofrece recomendaciones logísticas para su alojamiento y cuidados. Existen once puntos importantes que cubren los lineamientos de la APA:

1. General: incluye el código que rige la adquisición, mantenimiento y eliminación de los animales. El énfasis se centra principalmente en la recomendación de familiarizarse con el código.
2. Personal: este punto incluye a las personas que cuidarán de los animales, así como la disponibilidad de un veterinario y un supervisor de las instalaciones.
3. Instalaciones: el alojamiento de los animales debe realizarse de acuerdo con los estándares establecidos por el National Institute of Health (NIH) (*Instituto Nacional de Salud*), que norma su cuidado y uso en el laboratorio.
4. Adquisición de animales: se refiere a la manera en que se adquieren los animales. También se cubren las reglas sobre la crianza y/o la compra de animales.
5. Cuidado y alojamiento de los animales: establece las condiciones de las instalaciones donde se tiene a los animales.
6. Justificación de la investigación: el propósito de la investigación con animales debe quedar claramente establecido.

7. Diseño experimental: el diseño del estudio debe incluir consideraciones de tipo humanitario; esto incluye el tipo y la cantidad de animales a utilizar.
8. Procedimiento experimental: todos los procedimientos experimentales deben tomar en consideración el bienestar del animal. Los procedimientos no deben producir dolor; cualquier cantidad de dolor inducido debe estar justificado por el valor del estudio. Todo estímulo adverso debe presentarse en el nivel más bajo posible.
9. Investigación de campo: los investigadores que realicen investigación de campo tienen que molestar a la población lo menos posible. Debe existir respeto por la propiedad y por la privacidad de los habitantes.
10. Uso educativo de los animales: en primera instancia deben ser considerados estudios alternativos sin animales. Las demostraciones de clase con animales deben realizarse sólo cuando los objetivos educativos no puedan alcanzarse a través del uso de medios de comunicación. Los psicólogos necesitan incluir una presentación sobre la ética en el uso de animales para investigación.
11. Eliminación de los animales: este punto se refiere a lo que se hace con el animal una vez que se finaliza el estudio.

Estos lineamientos (disponibles en la American Psychological Association) deberían darse a conocer a todo el personal implicado en investigación y colocarse en un lugar visible, donde se mantengan y utilicen animales.

Al evaluar una investigación, la posibilidad de incrementar el conocimiento acerca del comportamiento, incluyendo el beneficio para la salud o bienestar de humanos y animales, debe ser suficiente para sobrestimar cualquier perjuicio o sufrimiento hacia los animales. Por lo tanto, siempre deben tenerse en cuenta y prevalecer las consideraciones humanitarias para el bienestar del animal. Si existe la posibilidad de que el animal sea susceptible de daño o aun dolor, deben seguirse con cuidado los procedimientos experimentales especificados en los lineamientos de la American Psychological Association, especialmente en el caso de los procedimientos quirúrgicos. Ningún animal debe desecharse hasta verificar su muerte, lo cual debe realizarse de manera legal y consistente con aspectos de salud, ambiente y estética.

Un libro reciente de Shapiro (1998) presenta la historia y la situación actual del empleo de animales en investigación científica. Este libro contiene artículos que tratan sobre la ética y las situaciones en que la investigación con animales es necesaria y en las que no lo es.

RESUMEN DEL CAPÍTULO

1. Los estudios Tuskegee y Milgrim usaron una forma de engaño y con frecuencia son citados como razones del porqué la investigación científica con humanos y animales necesita ser regulada.
2. El fraude constituye también un asunto de preocupación, puesto que el trabajo de investigadores como Burt y Breuning ejerció gran influencia en la legislación y en cómo la gente se percibía a sí misma y a los demás.
3. Organizaciones como la American Psychological Association establecen lineamientos sobre la ética en la investigación. También han establecido consejos de revisión para evaluar y tomar acciones respecto a quejas de comportamiento no ético.
4. Los investigadores están obligados a no provocar daño físico ni psicológico a los participantes en la investigación.

5. Los investigadores necesitan investigar de manera tal que se produzca información útil.
6. Las normas éticas establecidas por la American Psychological Association incluyen lineamientos para la planeación de la investigación, protección de los participantes, confidencialidad, desengaño, engaño, consentimiento informado y libertad de coerción.
7. También se proporcionan lineamientos para el empleo de animales en investigación, sobre su cuidado, alimentación y alojamiento, y qué hacer con los animales al finalizar el estudio.

SUGERENCIAS DE ESTUDIO

1. Algunas personas piensan que la sociedad ha impuesto demasiadas restricciones a los científicos sobre la manera como conducir sus investigaciones. Liste los puntos fuertes y débiles que subyacen a estas regulaciones.
2. ¿Cuál es el propósito del desengaño? ¿Por qué es necesario?
3. Una estudiante, fanática de los programas de entrevistas diurnos (*talk shows*), desea determinar si la manera en que una mujer viste influye en el comportamiento de los hombres. Ella planea asistir a dos bares en una sola noche. En uno de ellos vestirá de forma provocativa y en el otro usará un traje sastre. La variable dependiente será el número de hombres que se acercan para charlar. ¿Identifica algún problema ético en este diseño de estudio?
4. Visite la biblioteca e intente localizar material respecto a otros casos de fraude y de comportamiento no ético de científicos médicos y del comportamiento. ¿Cuántos pudo encontrar?
5. ¿Podría usted proponer un método alternativo que permitiera a Martin Arrowsmith de la novela *Arrowsmith* probar plenamente su fago?
6. Localice y lea al menos uno de los siguientes artículos:

Braunwald, E. (1987). On analyzing scientific fraud. *Nature, 325*, 215-216.

Broad, W. J. y Wade, N. (1982). *Betrayers of the truth*. Nueva York: Touchstone.

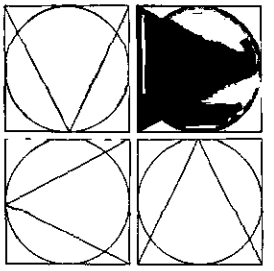
Brody, R. G. y Bowman, L. (1998). Accounting and psychology students' perceptions of whistle blowing. *College Student Journal, 32*, 162-166. (¿Debe el currículum universitario incluir ética?)

Fontes, L. A. (1998). Ethics in family violence research: Cross-cultural issues. *Family Relations: Interdisciplinary Journal of Applied Family Studies, 47*, 53-61.

Herrmann, D. y Yoder, C. (1998). The potential effects of the implanted memory paradigm on child subjects. *Applied Cognitive Psychology, 12*, 198-206. (Analiza el peligro del falso recuerdo.)

Knight, J. A. (1984). Exploring the compromise of ethical principles in science. *Perspectives in Biology and Medicine, 27*, 432-442. (Explora las razones para cometer fraude y la deshonestidad en la ciencia.)

Stark, C. (1998). Ethics in the research context: Misinterpretations and misplaced misgivings. *Canadian Psychology, 39*, 202-211. (Revisión de los códigos éticos de la Canadian Psychological Association.)



CAPÍTULO 18

DISEÑO DE INVESTIGACIÓN: PROPÓSITO Y PRINCIPIO

El *diseño de investigación* constituye el plan y la estructura de la investigación, y se concibe de determinada manera para obtener respuestas a las preguntas de investigación. El plan es el esquema o programa general de la investigación; incluye un bosquejo de lo que el investigador hará, desde formular las hipótesis y sus implicaciones operacionales hasta el análisis final de los datos. La estructura de la investigación resulta más difícil de explicar, ya que el término *estructura* presenta dificultad para ser definido claramente y sin ambigüedades. A causa de que es un concepto que irá tomando gran importancia conforme se continúe el estudio, se realizará una pausa para intentar definirlo y ofrecer una breve explicación. En este momento la disertación será necesariamente un poco abstracta, sin embargo, ejemplos posteriores serán más concretos. Más importante aún, el concepto se encontrará poderoso, útil e incluso indispensable, especialmente en el estudio posterior del análisis multivariado, donde el concepto "estructura" es clave, y cuyo entendimiento se vuelve esencial para comprender la mayoría de la metodología de investigación contemporánea.

Una *estructura* es el marco de referencia, la organización o configuración de los elementos de la estructura, relacionados en formas específicas. La mejor forma de especificar una estructura consiste en escribir una ecuación matemática que relacione las partes de la estructura entre sí. Dicha ecuación matemática, puesto que sus términos están definidos y relacionados específicamente por la ecuación (o conjunto de ecuaciones), no es ambigua.

En resumen, una estructura es un paradigma o modelo de las relaciones entre las variables de un estudio. Los términos *estructura, modelo* y *paradigma* son problemáticos debido a que es difícil definirlos con claridad y sin ambigüedades. Un "paradigma" es un modelo, un ejemplo. Los diagramas, gráficas y bosquejos verbales son paradigmas. Aquí se utiliza "paradigma" en lugar de "modelo" porque "modelo" tiene otro importante significado en la ciencia, significado al que se regresará en el capítulo 37, cuando se analice la comprobación de una teoría utilizando procedimientos multivariados y "modelos" de aspectos de teorías.

Un diseño de investigación expresa tanto la estructura del problema de investigación como el plan de investigación utilizado para obtener evidencia empírica sobre las relaciones del problema. Pronto se presentarán ejemplos del diseño y de la estructura que quizás animen esta discusión abstracta.

Propósitos del diseño de investigación

El diseño de investigación incluye dos propósitos básicos: 1) *proporcionar respuestas a preguntas de investigación* y 2) *controlar la varianza*. El diseño ayuda a los investigadores a obtener respuestas a las preguntas de investigación, y también a controlar las varianzas experimental, extraña y del error del problema de investigación particular en estudio. Ya que puede decirse que toda actividad de investigación tiene el propósito de generar respuestas a preguntas de investigación, es posible omitir este propósito en el análisis y de afirmar que el diseño de investigación tiene un propósito fundamental: controlar la varianza. Sin embargo, tal delimitación del propósito del diseño es peligrosa. Sin un fuerte énfasis en las preguntas de investigación y en el uso del diseño para ayudar a proporcionar respuestas a dichas preguntas, el estudio del diseño puede degenerar en un ejercicio técnico interesante, pero estéril.

Los diseños de investigación se inventaron para permitir a los investigadores responder preguntas de la forma más válida, objetiva, precisa y económica posible. Los planes de investigación se conciben de forma deliberada y específica, y son ejecutados para obtener evidencia empírica que apoye al problema de investigación. Los problemas de investigación pueden ser, y son, expresados en forma de hipótesis; éstas se formulan en un momento de la investigación de manera que puedan ser probadas empíricamente. Los diseños se elaboran con cuidado para que proporcionen respuestas confiables y válidas a las preguntas de investigación contenidas en las hipótesis. Es posible realizar una sola observación e inferir que la relación hipotetizada existe, con base en esta única observación; pero es evidente que no se puede aceptar la inferencia realizada de esa forma. Por otro lado, también es factible realizar cientos de observaciones e inferir que la relación hipotetizada existe, con base en estas múltiples observaciones, en cuyo caso se puede o no aceptar como válida la inferencia. El resultado depende de la manera en que se hicieron las observaciones y la inferencia. Un diseño planeado y ejecutado de forma adecuada ayuda en mucho a permitirse confiar tanto en las observaciones como en las inferencias.

¿Cómo logra esto el diseño? El diseño de investigación establece el marco de referencia para el estudio de las relaciones entre variables. Indica, en cierto sentido, qué observaciones hacer, cómo hacerlas y cómo realizar las representaciones cuantitativas de las observaciones. Estrictamente hablando, el diseño no "dice" precisamente qué hacer, sino qué "sugiere" la dirección de cómo realizar las observaciones y el análisis. Un diseño adecuado "sugiere", por ejemplo, cuántas observaciones deben efectuarse y qué variables son activas y cuáles son atributivas. Entonces se actúa para manipular las variables activas y categorizar y medir las variables atributivas. Un diseño indica qué tipo de análisis estadís-

tico emplear. Por último, un diseño adecuado bosqueja las conclusiones que posiblemente se obtengan del análisis estadístico.

Un ejemplo

Se ha dicho que los colegios y las universidades discriminan a las mujeres respecto a los procesos de contratación y de admisión. Suponga que se desea probar la discriminación en la admisión. La idea para este ejemplo proviene del inusual y extraño experimento de Walster, Cleary y Clifford (1970) citado anteriormente. Se diseña un experimento de la siguiente manera: se envían solicitudes de admisión a una muestra aleatoria de 200 colegios, basando las solicitudes en varios casos modelo seleccionados sobre un rango de habilidades probadas, con todos los detalles iguales excepto el género. La mitad de las solicitudes serán de hombres, y la otra mitad, de mujeres. Manteniendo las otras cuestiones iguales, se espera aproximadamente igual número de aceptaciones y de rechazos; entonces la *aceptación* es la variable dependiente, la cual se mide con una escala de tres puntos: aceptación completa, aceptación con reservas y rechazo. Llámese a los hombres A_1 y a las mujeres A_2. El paradigma del diseño se presenta en la figura 18.1.

El diseño es el más simple posible, dados los requerimientos mínimos de control. Los dos tratamientos se asignan a los colegios aleatoriamente. Entonces, cada colegio recibirá una solicitud, ya sea de un hombre o de una mujer. Se probará la significancia estadística de la diferencia entre las medias M_{A_1} y M_{A_2} con un prueba t o F. La hipótesis sustantiva es: $M_{A_1} > M_{A_2}$, o se admitirán más hombres que mujeres. Si no hay discriminación en la admisión, entonces M_{A_1} sería estadísticamente igual a M_{A_2}. Suponga que una prueba F indica que las medias no son significativamente diferentes. ¿Se puede estar seguro de que no hay práctica de discriminación (en promedio)? Mientras que el diseño de la figura 18.1 es satisfactorio hasta ahora, quizá no llega suficientemente lejos.

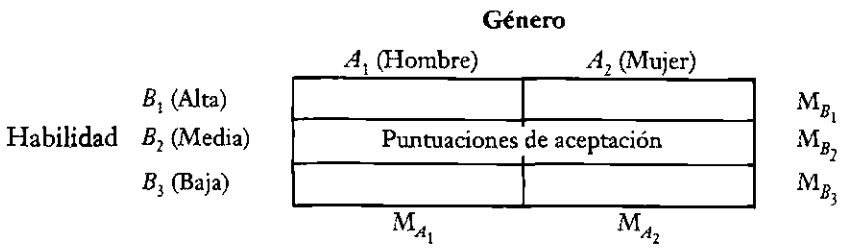
Un diseño más fuerte

Walster y sus colegas utilizaron otras dos variables independientes, la *raza* y la *habilidad*, en un diseño factorial. En el ejemplo se eliminó *raza* —no fue estadísticamente significativa ni interactuó de manera significativa con otras variables— y se enfatizó el género y la habilidad. Si un colegio basa su selección de estudiantes de nuevo ingreso estrictamente en las habilidades, entonces no hay discriminación (a menos, por supuesto, que la selección por habilidades se considere discriminación). Añádase *habilidad* al diseño de la figura 18.1 usando tres niveles; es decir, además de designar a las aplicaciones como *hombre* y *mujer*, también se designan como *habilidad alta*, *habilidad media* y *habilidad baja*. Por ejemplo, tres de los solicitantes pueden ser: hombre con habilidad media, mujer con habilidad alta y

FIGURA 18.1

Tratamientos	
A_1 (Hombre)	A_2 (Mujer)
Puntuaciones de aceptación	
M_{A_1}	M_{A_2}

FIGURA 18.2



mujer con habilidad baja. Ahora, si no existen diferencias significativas entre los géneros ni la interacción de *género* y *habilidad* es significativa, ésta sería una evidencia considerablemente más fuerte de que no hay discriminación que la proporcionada por el diseño y por la prueba estadística de la figura 18.1. Ahora se utiliza el diseño ampliado para explicar esta afirmación y analizar ciertos aspectos del diseño de investigación. El diseño ampliado se presenta en la figura 18.2.

El diseño es un factorial de 2 × 3. Una variable independiente, *A*, es el género, la misma que en la figura 18.1. La segunda variable independiente, *B*, es la habilidad, que se manipuló para indicar, de varias formas, cuáles son los niveles de habilidad de los estudiantes. Es importante no confundirse por el nombre de las variables; *género* y *habilidad* son por lo común variables atributivas, por lo tanto, no experimentales. Sin embargo, en este caso se manipulan. Los registros de los estudiantes enviados a los colegios fueron sistemáticamente ajustados para que se adecuaran a las seis casillas de la figura 18.2. Por ejemplo, un caso en la casilla A_1B_2, sería el registro de un hombre con habilidad media, que es el registro que el colegio evalúa para la admisión.

Suponga que se piensa que la discriminación en contra de las mujeres toma una forma más sutil que la simple exclusión a todos los niveles: se piensa que se discrimina contra las mujeres con habilidad baja (en comparación con los hombres). Ésta es una hipótesis de interacción. De cualquier manera, se utiliza este problema y el paradigma de la figura 18.2 como base para analizar algunos elementos del diseño de investigación.

Los problemas de investigación sugieren diseños de investigación. Puesto que la hipótesis antes discutida es de interacción, evidentemente un diseño factorial es el apropiado. *A* es el *género*; *B* es la *habilidad*; *A* se divide en A_1 y A_2 y *B* en B_1, B_2 y B_3.

El paradigma de la figura 18.2 sugiere varias cosas. La primera, y la más obvia, es que se requiere un gran número de participantes; específicamente se necesitan $6n$ participantes (n es igual al número de sujetos en cada casilla). Si se decide que n debe ser 20, entonces se requiere tener 120 sujetos para el experimento. Observe aquí la "sabiduría" del diseño; si tan sólo se estuvieran probando los tratamientos y se ignorara la habilidad, únicamente se necesitarían $2n$ sujetos. Es preciso observar que algunos autores como Simon (1976, 1987); Simon y Roscoe (1984) y Daniel (1976) discrepan con este enfoque para todo tipo de problemas. Ellos consideran que muchos diseños contienen réplicas ocultas y que serían suficientes mucho menos de 20 participantes por casilla. Tales diseños requieren una planeación mucho más cuidadosa; pero el investigador puede obtener información mucho más útil y estudiar más variables independientes en lugar de sólo dos o tres.

Existen formas para determinar el número de participantes que se requieren en un estudio. Tal determinación forma parte del "poder", que se refiere a la habilidad de una prueba de significancia estadística para detectar diferencias en las medias (u otros estadísticos), cuando en realidad existen tales diferencias. En el capítulo 8 se explica el tamaño de

las muestras y su relación con la investigación. Sin embargo, el capítulo 12 presenta un método para estimar el tamaño de las muestras de manera que se cumplan ciertos criterios. El poder es un valor fraccional entre 0 y 1.00 que se define como $1 - \beta$, donde β es la probabilidad de cometer un error tipo II, el cual sucede cuando no se rechaza una hipótesis nula falsa. Si el poder es alto (cercano a 1.00) indica que si la prueba estadística no fue significativa, la investigación sugiere que la hipótesis nula es verdadera. El poder también indica qué tan sensible es la prueba estadística para detectar diferencias reales. Si la prueba estadística no es lo suficientemente sensible para hacer esto, se dice que la prueba tiene poco poder. Una prueba altamente sensible, que puede detectar diferencias verdaderas, se considera de alto poder. En el capítulo 16 se analizó la diferencia entre las pruebas estadísticas paramétricas y las no paramétricas. Las pruebas no paramétricas son generalmente menos sensibles que las pruebas paramétricas; como resultado, se considera que las primeras tienen menos poder que las segundas. Uno de los libros más completos sobre la cuestión de la estimación del poder es el de Cohen (1988). Jaccard y Becker (1997) ofrecen una introducción fácil de entender al análisis del poder.

En segundo lugar, el diseño indica que los "participantes" (en este caso los colegios) pueden asignarse aleatoriamente tanto a A como a B, ya que ambas son variables experimentales. Sin embargo, si ***habilidad*** fuese una variable no experimental atributiva, entonces los participantes podrían ser asignados de manera aleatoria a A_1 y A_2, pero no a B_1, B_2 ni B_3.

En tercer lugar, de acuerdo al diseño, las observaciones realizadas en los "participantes" deben realizarse de manera independiente. La puntuación de un colegio no debe afectar a la puntuación de otro. Reducir el diseño a un bosquejo como el que se indica en la figura 18.2, en efecto, prescribe las operaciones necesarias para obtener las medidas apropiadas para el análisis estadístico. Una prueba F depende del supuesto de la independencia de las medidas de la variable dependiente. Si aquí ***habilidad*** es una variable atributiva y a los individuos se les mide la inteligencia, por ejemplo, entonces el requisito de independencia está en mayor riesgo debido a la posibilidad de que un sujeto vea los documentos de otro y a que los maestros "ayuden" inconsciente o conscientemente a los estudiantes con las respuestas, entre otras razones. Los investigadores tratan de prevenir este tipo de situaciones, no tanto por razones morales sino para satisfacer los requisitos de un diseño y una estadística sólidos.

Un cuarto punto resulta bastante obvio ahora: la figura 18.2 sugiere un análisis factorial de varianza, pruebas F, medidas de asociación y, quizá, pruebas *post hoc*. Si la investigación está bien diseñada antes de la recolección de los datos —como en realidad lo hicieron Walster *et al.*— la mayoría de los problemas estadísticos pueden resolverse. Además, se evitan ciertos problemas molestos antes de que surjan, o incluso pueden prevenirse del todo. Sin embargo, con un diseño inadecuado, los problemas referentes a las pruebas estadísticas apropiadas se vuelven muy molestos. Una de las razones del gran énfasis de este libro en tratar los problemas de diseño y estadísticos de forma concomitante, es que esto permite señalar maneras de evitar tales problemas. Si el diseño y el análisis estadístico se planean simultáneamente, el trabajo analítico se volverá sencillo y ordenado.

Un dividendo bastante útil del diseño es el siguiente: un diseño claro, como el de la figura 18.2, sugiere qué prueba estadística realizar. Por ejemplo, un diseño aleatorio simple de una variable con dos particiones o tratamientos, A_1 y A_2, permite tan sólo una prueba estadística de la diferencia entre los dos estadísticos producidos por los datos. Dichos estadísticos pueden ser dos medias, dos medianas, dos rangos, dos varianzas, dos porcentajes, etcétera. Sólo una prueba estadística es generalmente posible. Sin embargo, con el diseño de la figura 18.2 existen tres pruebas estadísticas posibles: 1) entre A_1 y A_2; 2) entre B_1, B_2 y B_3, y 3) la interacción entre A y B. En la mayoría de las investigaciones, no

FIGURA 18.3

Condición B_1 Tratamientos		Condición B_2 Tratamientos	
A_1	A_2	A_1	A_2
M_{A_1}	M_{A_2}	M_{A_1}	M_{A_2}

todas las pruebas estadísticas tienen la misma importancia; las importantes, en efecto, son aquellas directamente relacionadas con los problemas e hipótesis de investigación.

En el presente caso, la hipótesis de interacción [la del inciso 3) anterior] es la más importante, ya que se supone que la discriminación depende del nivel de ***habilidad***. Los colegios quizá discriminen a diferentes niveles de habilidad. Como se sugirió antes, las mujeres (A_2) tal vez sean aceptadas más que los hombres (A_1) en el nivel de habilidad más alto (B_1); mientras que quizá sean menos aceptadas en el nivel de habilidad más bajo (B_3). Debería ser evidente que el diseño de investigación no es estático. El tener conocimiento sobre diseño puede ayudar a planear y realizar mejor investigación, y también puede sugerir la comprobación de hipótesis. Y quizá más importante: puede llevar a que uno se dé cuenta de que el diseño de un estudio no es adecuado a las demandas planteadas. ¿Qué significa esta afirmación un tanto peculiar?

Suponga que se formula la hipótesis de interacción como se bosquejó anteriormente, sin saber nada sobre el diseño factorial; en realidad se establece un diseño que consiste de dos experimentos, en uno de los cuales se prueba A_1 contra A_2, bajo la condición B_1. En el segundo experimento se prueba A_1 contra A_2, bajo la condición B_2. El paradigma se vería como el que se muestra en la figura 18.3. (Para simplificar las cosas, únicamente se utilizan dos niveles de B: B_1 y B_2; por lo tanto, el diseño se reduce a uno de 2×2.)

El punto importante a señalar es que no es posible realizar una prueba *adecuada* de la hipótesis con este diseño. A_1 puede probarse contra A_2 bajo las dos condiciones B_1 y B_2 para asegurarse. Pero no es posible saber con claridad y sin ambigüedades, si existe una interacción significativa entre A y B. Aun cuando $M_{A_1} > M_{A_2} \mid B_2$ (M_{A_1} es mayor que M_{A_2}, bajo la condición B_2), como se hipotetizó, el diseño no puede ofrecer una clara posibilidad de confirmación de la interacción hipotetizada, debido a que no se puede obtener información sobre las diferencias entre A_1 y A_2 en los dos niveles de B(B_1 y B_2). Recuerde que una hipótesis de interacción implica, en este caso, que la diferencia entre A_1 y A_2 es distinta en B_1 de lo que es en B_2. En otras palabras, la información tanto de A como de B *juntas en un experimento* es necesaria para probar una hipótesis de interacción. Si los resultados estadísticos de experimentos separados mostraran una diferencia significativa entre A_1 y A_2 en un experimento bajo la condición B_1, y no mostraran diferencias significativas en otro experimento bajo la condición B_2, entonces hay *presunta* evidencia de que la hipótesis de interacción es correcta. Pero no es suficiente contar con presunta evidencia especialmente cuando se sabe que es posible obtener una mejor evidencia.

Suponga que en la figura 18.3, las medias de las casillas fueran, de izquierda a derecha: 30, 30, 40, 30. Tal resultado parecería apoyar la hipótesis de interacción, ya que hay una diferencia significativa entre A_1 y A_2 en el nivel B_2, pero no en el nivel B_1. Pero no puede tenerse la certeza de que esto es así, incluso si la diferencia entre A_1 y A_2 es estadísticamente significativa. La figura 18.4 presenta cómo resultaría esto si se hubiese utilizado un diseño factorial. (Las cifras en las casillas y en los márgenes son medias.) Considerando que los efectos principales, A_1 y A_2; B_1 y B_2, fueran significativos, todavía es posible que la interac-

FIGURA 18.4

	A_1	A_2	
B_1	30	30	30
B_2	40	30	35
	35	30	

ción no sea significativa. A menos que la hipótesis de interacción se pruebe específicamente, la evidencia para determinar la interacción es mera presunción, ya que falta la prueba estadística de la interacción que un diseño factorial proporciona. Debe quedar claro que el conocimiento sobre diseño hubiese mejorado este experimento.

El diseño de investigación como control de la varianza

La principal función técnica del diseño de investigación es *controlar la varianza*. Un diseño de investigación constituye, por así decirlo, un conjunto de instrucciones para que el investigador reúna y analice los datos de cierta forma; por lo tanto, es un mecanismo de control. El principio estadístico que subyace a este mecanismo, como se dijo antes, es: *maximizar la varianza sistemática*, *controlar la varianza sistemática extraña* y *minimizar la varianza del error*. En otras palabras, se debe *controlar* la varianza.

De acuerdo con este principio, al construir un diseño de investigación eficiente, el investigador intenta: 1) maximizar la varianza de la variable o variables de la hipótesis sustantiva de investigación, 2) controlar la varianza de variables extrañas o "indeseables" que puedan tener un efecto en los resultados experimentales y 3) minimizar la varianza del error o aleatoria, incluyendo los llamados errores de medición. Ahora se verá un ejemplo.

Un ejemplo controversial

La controversia abunda en toda la ciencia y parece ser especialmente rica y variada en las ciencias del comportamiento. Dos controversias han surgido a partir de diferentes teorías del comportamiento y aprendizaje humanos. Los teóricos del reforzamiento han demostrado ampliamente que el reforzamiento positivo puede incrementar el aprendizaje. Sin embargo, como siempre, las cuestiones no son tan simples. El supuesto efecto benéfico de las recompensas externas se ha cuestionado; la investigación ha mostrado que la recompensa extrínseca puede tener una influencia perjudicial en la motivación, interés intrínseco y aprendizaje de los niños. En los años setenta, se publicó una serie de artículos y estudios que mostraban los posibles efectos dañinos del uso de la recompensa. En uno de dichos estudios, Amabile (1979) demostró que la evaluación externa tiene un efecto perjudicial sobre la creatividad artística. Otros estudios incluyen el de Deci (1971) y el de Lepper y Greene (1978). Al mismo tiempo, incluso el principio del reforzamiento en apariencia simple, no es tan simple. Sin embargo, en años recientes han aparecido varios artículos que defienden los efectos positivos de la recompensa (véase Eisenberger y Cameron, 1996; Sharpley, 1988; McCullers, Fabes y Moran, 1987; Bates, 1979).

Existen diversas investigaciones y creencias que indican que los estudiantes universitarios aprenden bien bajo el régimen de lo que se ha llamado *aprendizaje de dominio* (*mastery learning*). De manera sintética, diremos que el "aprendizaje de dominio" consti-

tuye un sistema pedagógico basado en instrucciones personalizadas que requiere que los estudiantes aprendan unidades curriculares hasta alcanzar un criterio de dominio (véase Abbott y Falstrom, 1975; Ross y McBean, 1995; Senemoglu y Fogelman, 1995; Bergin, 1995). Aunque parece existir cierta investigación que apoya la eficacia del aprendizaje de dominio, hay por lo menos un estudio —y es un buen estudio— realizado por Thompson (1980), cuyos resultados indican que los estudiantes a quienes se enseñó con el método de aprendizaje de dominio no fueron mejores que los estudiantes a quienes se enseñó con un enfoque convencional de conferencia, discusión y memorización. Éste es un estudio ejemplar, realizado con controles cuidadosos, durante un largo periodo. El ejemplo que se presenta a continuación estuvo inspirado en el estudio de Thompson. Sin embargo, el diseño y los controles del ejemplo son mucho más simples que los de Thompson. Observe también que Thompson tenía una enorme ventaja: realizó su experimento en un establecimiento militar, lo cual, por supuesto, significa que muchos problemas de control, con frecuencia recalcitrantes en la investigación educativa, se resolvieron fácilmente.

La controversia surge porque los partidarios del aprendizaje de dominio parecen estar fuertemente convencidos de sus virtudes; mientras que los escépticos permanecen incrédulos. ¿Decidirá la investigación el asunto? Es difícil. Pero ahora se verá cómo se podría diseñar un estudio relativamente modesto, capaz de proporcionar por lo menos una respuesta *empírica* parcial.

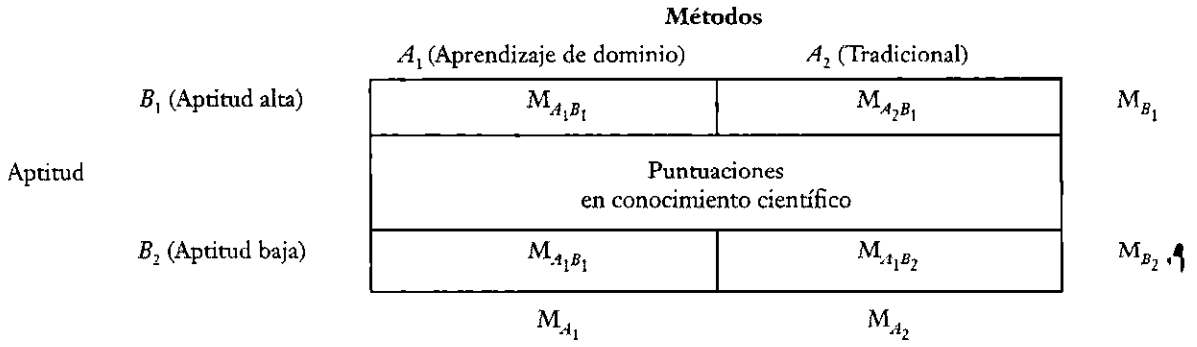
Un investigador educativo decide probar la hipótesis de que el aprovechamiento en ciencia sufre un mayor incremento con un método de aprendizaje de dominio (*AD*), que con un método tradicional (*T*). Se ignoran los detalles de los métodos para concentrarse en el diseño de la investigación. Llámese al método de aprendizaje de dominio A_1 y al método tradicional A_2. Los investigadores saben que otras posibles variables independientes ejercen influencia sobre el aprovechamiento: inteligencia, género, antecedentes de clase social, experiencias previas con la ciencia, motivación, etcétera. Existen razones para creer que los dos métodos funcionan de diferente manera con diferentes tipos de estudiantes. Por ejemplo, quizá funcionen de manera diferente con estudiantes con distintos niveles de aptitud escolar. El enfoque tradicional tal vez resulte efectivo con estudiantes con alta aptitud; mientras que el aprendizaje de dominio sea más efectivo con estudiantes con baja aptitud. Llámese *B* a las aptitudes: aptitud alta es B_1 y aptitud baja es B_2. En este ejemplo la variable aptitud se dicotomizó en los grupos de aptitud alta y baja. Ésta no es la mejor forma de utilizar la variable aptitud; cuando una medida continua se dicotomiza o tricotomiza, se pierde la varianza. En un capítulo posterior se verá que constituye un mejor método respetar el nivel de la medida continua y utilizar una regresión múltiple.

¿Qué tipo de diseño debe establecerse? Para responder es importante etiquetar las variables y saber con claridad cuáles son las preguntas que se formulan. Las variables son:

Variable independiente		**Variable dependiente**
Métodos	*Aptitud*	*Aprovechamiento en ciencias*
Aprendizaje de dominio, A_1	Aptitud alta, B_1	Puntuaciones de la prueba de ciencia
Tradicional, A_2	Aptitud baja, B_2	

Los investigadores pudieron haber incluido otras variables en el diseño, en especial variables potencialmente influyentes sobre el aprovechamiento: inteligencia general, clase social, género, promedio de preparatoria, por ejemplo. También se podría utilizar la asignación aleatoria para ocuparse de la inteligencia y otras posibles variables independientes de influencia. La medida de la variable dependiente se puede obtener mediante una prueba estandarizada de conocimientos de ciencia.

FIGURA 18.5



Parece que el problema requiere de un diseño factorial. Existen dos razones para esta opción: 1) hay dos variables independientes, 2) es claro que se tiene en mente una hipótesis de interacción, aunque no se haya expresado con tantas palabras. Se cree que los métodos funcionarán de manera diferente con distintos tipos de estudiantes. Se establece la estructura de diseño que se representa en la figura 18.5.

Observe que todas las medias marginales y de casilla han sido etiquetadas de forma apropiada. Note también que hay una *variable activa*, métodos; y una *variable atributo*, aptitudes. Quizá recuerde del capítulo 3 que una *variable activa* es una variable experimental o manipulada; una *variable atributo* es una variable medida o una variable que es una característica de personas o grupos; por ejemplo, inteligencia, clase social y ocupación (gente); así como cohesión, productividad y atmósfera restrictiva-permisiva (organizaciones, grupos, etcétera). Todo lo que puede hacerse es categorizar a los participantes como con aptitud alta y aptitud baja, y asignarlos de acuerdo con ello a B_1 y B_2. Sin embargo, es posible asignar a los estudiantes aleatoriamente a A_1 y A_2, los grupos de los métodos. Esto se realiza en dos etapas: 1) los estudiantes de B_1 (aptitud alta) se asignan aleatoriamente a A_1 y A_2, y 2) los estudiantes de B_2 (aptitud baja) se asignan aleatoriamente a A_1 y A_2. Al aleatorizar así a los participantes se puede suponer que antes de que empiece el experimento, los estudiantes en A_1 son aproximadamente iguales a los estudiantes en A_2, en todas las características posibles.

El interés aquí radica en los diferentes papeles de la varianza en el diseño de investigación y en el principio de la varianza. Antes de continuar, al principio de la varianza se le llamará "maxmincon" para su fácil referencia. El origen del nombre es evidente: maximizar la varianza sistemática en estudio; controlar la varianza sistemática extraña y minimizar la varianza del error, con dos sílabas invertidas por eufonía.

Antes de ilustrar la aplicación del principio maxmincon en el presente ejemplo, debe discutirse un punto importante. Siempre que se hable de varianza hay que estar seguro de saber de qué tipo de varianza se habla. Se habla de la varianza de los métodos, de inteligencia, de género, de tipo de hogar, etcétera; parecería que se refiere a la varianza de la variable independiente, lo cual es verdad y, a la vez, no. Siempre se refiere a la *varianza de la variable dependiente, y a la varianza de las medidas de la variable dependiente*, después de que se realizó el experimento. Esto no es cierto en los llamados estudios correlacionales, donde al decir "la varianza de la variable independiente" significa justamente eso. Al correlacionar dos variables, se estudian "directamente" las varianzas de las variables dependiente e independiente. La referencia "varianza de la variable independiente" surge del hecho de que,

mediante la manipulación y el control de las variables independientes, presumiblemente se ejerce influencia sobre la varianza de la variable dependiente. Dicho de manera algo imprecisa, se "hace" que las medidas de la variable dependiente se comporten o varíen como un supuesto resultado de la manipulación y el control de las variables independientes. En un experimento se analizan las medidas de la variable dependiente y, a partir del análisis, se infiere que las varianzas presentes en la varianza total de las medidas de la variable dependiente se deben a la manipulación y control de las variables independientes y no al error. Ahora regresemos al principio en cuestión.

Maximización de la varianza experimental

La preocupación más obvia del investigador, aunque no necesariamente la más importante, consiste en maximizar la llamada *varianza experimental*. Dicho término se introduce para facilitar discusiones subsecuentes y, en general, tan sólo se refiere a la varianza de la variable dependiente, debida a la influencia ejercida por la variable independiente o variables de la hipótesis sustantiva. En este caso en particular, la varianza experimental es la varianza en la variable dependiente, presumiblemente debida a los métodos A_1 y A_2, y a los niveles de aptitud B_1 y B_2. Aunque la varianza experimental puede tomarse para hacer referencia únicamente a la varianza debida a la variable manipulada o *activa*, como los métodos, también se pueden considerar las variables *atributo* como inteligencia, género y, en este caso, aptitud, como variables experimentales. Una de las principales tareas de un experimentador consiste en maximizar esta varianza. Los métodos deben "separarse" lo más posible para hacer a A_1 y A_2 (y A_3, A_4, etcétera, si están en el diseño) tan diferentes como sea posible.

Si la variable independiente no varía de manera sustancial, hay poca posibilidad de separar su efecto de la varianza total de la variable dependiente. Es necesario dar a la varianza de una relación la oportunidad de mostrarse, de separarse, por así decirlo, de la varianza total, la cual es un compuesto de varianzas debidas a numerosas fuentes y al azar. Teniendo presente este subprincipio del principio maxmincon, se puede declarar un precepto de investigación: *diseñar, planear y conducir la investigación de tal forma que las condiciones experimentales sean tan diferentes como sea posible.* Existen, por supuesto, excepciones a este subprincipio, pero probablemente sean poco comunes. Puede ser que un investigador desee estudiar los efectos de pequeñas gradaciones de, digamos, incentivos motivacionales sobre el aprendizaje de algún tema. Aquí no se buscaría que las condiciones experimentales fueran lo más diferentes posibles; aun así, debería asegurarse de que varían un poco o no habría una varianza resultante discernible en la variable dependiente.

En el presente ejemplo de investigación, este subprincipio se refiere a que el investigador debe realizar un esfuerzo para hacer a A_1 y A_2, los métodos de aprendizaje de dominio y el tradicional, tan diferentes como sea posible. Después, B_1 y B_2 también deben ser tan diferentes como sea posible, en la dimensión de aptitud. Este último problema en esencia es uno de medición, como se verá en un capítulo posterior. En un experimento, el investigador es como un titiritero que hace que los títeres de la variable independiente hagan lo que él quiere. Sostiene los hilos de los títeres A_1 y A_2 con la mano derecha; y los hilos de los títeres B_1 y B_2, con la mano izquierda. (Se supone que una mano no ejerce influencia sobre la otra, es decir, las manos deben ser independientes.) Los títeres A_1 y A_2 se ponen a bailar por separado, al igual que los títeres B_1 y B_2; entonces, el investigador presta atención a la audiencia (la variable dependiente) para observar y medir el efecto de las manipulaciones. Si tiene éxito al hacer bailar a A_1 y A_2 por separado, y si existe una relación entre A y la variable dependiente, y, si por ejemplo, separar A_1 de A_2 es gracioso,

la reacción de la audiencia debería ser una carcajada. El investigador incluso puede notar que sólo consigue risas cuando A_1 y A_2 bailan por separado y, al mismo tiempo, B_1 y B_2 bailan por separado (interacción de nuevo).

Control de variables extrañas

El control de variables extrañas se refiere a minimizar, anular o aislar las influencias de aquellas variables independientes extrañas a los propósitos del estudio. Hay tres formas de controlar las variables extrañas. El primero es el más sencillo, si es posible realizarlo: eliminar la variable como tal. Si existe preocupación sobre la inteligencia como un posible factor contribuyente en estudios de aprovechamiento, su efecto sobre la variable dependiente virtualmente puede ser eliminado utilizando participantes con un solo nivel de inteligencia, digamos puntuaciones de inteligencia dentro del rango de 90 a 110. Si se estudia el aprovechamiento, y el origen racial es un posible factor contribuyente a la varianza del aprovechamiento, se elimina utilizando únicamente miembros de una raza. El principio es: *eliminar el efecto de una variable independiente que posiblemente influya sobre la variable dependiente, es decir, elegir a los participantes de manera que sean lo más homogéneos posible en esa variable independiente.*

Este método para controlar la varianza indeseable o extraña es muy efectivo. Si se selecciona solamente un género para un experimento, entonces se tiene la seguridad de que el género no sea una variable independiente contribuyente. Pero entonces se pierde el poder de generalización; por ejemplo, no es factible hablar sobre la relación estudiada respecto a las niñas si únicamente se utilizan niños en el experimento. Si se restringe el rango de inteligencia, entonces solamente se analiza dicho rango restringido. ¿Es posible que la relación, si se descubre una, sea inexistente o muy distinta con niños de alta inteligencia o con niños de baja inteligencia? Simplemente no se sabe; tan sólo se puede conjeturar o suponer.

La segunda forma para controlar la varianza extraña es a través de la aleatorización. Ésta es la mejor manera, en el sentido de que es posible tener el pastel y también comer un poco de él. En teoría, la aleatorización es el único método para controlar todas las variables extrañas posibles. Otra forma de expresarlo es: si se logra una aleatorización adecuada, entonces los grupos experimentales pueden ser considerados estadísticamente iguales en todas las formas posibles. Por supuesto que esto no quiere decir que los grupos sean iguales en todas las variables posibles. Ya se sabe que, por azar, los grupos pueden ser desiguales; pero con la aleatorización adecuada, la probabilidad de que sean iguales es mayor que la probabilidad de que no lo sean. Por tal razón, el control de la varianza extraña por medio de la aleatorización es un poderoso método de control. Todos los otros métodos dejan abiertas muchas posibilidades de desigualdad. Si se aparean los grupos respecto a la inteligencia, quizá se logre exitosamente la igualdad estadística en inteligencia (al menos en aquellos aspectos de inteligencia que se miden), pero se puede sufrir de desigualdad en otras variables independientes significativamente influyentes como aptitud, motivación y clase social. Un precepto que surge a partir de este poder igualador de la aleatorización es: *siempre que sea posible hacerlo, asigne a los sujetos a los grupos y a las condiciones experimentales de manera aleatoria, y asigne las condiciones y otros factores a los grupos experimentales de manera aleatoria.*

El tercer método para controlar una variable extraña es incluirla en el diseño como una variable independiente. Por ejemplo, suponga que en el experimento discutido anteriormente se fuera a controlar el género y que se considerara inoportuno o imprudente eliminarlo. Se podría añadir una tercera variable al diseño: el género. A menos que se estuviera interesado en la diferencia real entre los géneros respecto a la variable depen-

diente, o se deseara estudiar la interacción entre una o dos de las otras variables y el género, es poco probable que se utilice esta forma de control. Se puede desear información del tipo antes mencionado y también desear controlar el género. En tal caso sería deseable añadirlo al diseño como una variable. El punto es que incorporar una variable a un diseño experimental "controla" la variable, ya que, entonces, resulta posible extraer de la varianza total de la variable dependiente, la varianza debida a la variable. (En el caso anterior se trataría de la varianza "entre géneros".)

Tales consideraciones llevan a otro principio: *una variable extraña puede ser controlada al incorporarla al diseño de investigación como una variable atributo, logrando así control y proporcionando información adicional de investigación sobre el efecto de la variable sobre la variable dependiente y sobre su posible interacción con otras variables independientes.*

La cuarta forma para controlar la varianza extraña consiste en aparear a los participantes. El principio de control detrás del apareamiento es el mismo que aquel para cualquier otra forma de control: el control de la varianza. El apareamiento es similar —de hecho podría llamarse un corolario— al principio del control de la varianza de una variable extraña mediante su incorporación al diseño. El principio básico consiste en dividir una variable en dos o más partes en un diseño factorial, digamos en inteligencia alta y baja, y después aleatorizarla dentro de cada nivel, como se describió antes. El apareamiento es un caso especial de este principio. Sin embargo, en lugar de dividir a los participantes en dos, tres o cuatro partes, se dividen en $N/2$ partes; donde N es el número de participantes involucrados; de esta manera se incorpora al diseño el control de la varianza.

Con el uso del método de apareamiento pueden surgir varios problemas. En primer lugar, la variable respecto a la cual se aparean los participantes debe estar sustancialmente relacionada con la variable dependiente o el apareamiento es una pérdida de tiempo; aun peor, puede causar confusión. Además, el apareamiento tiene limitaciones severas. Si se intenta aparear, digamos, más de dos variables, o incluso más de una, se pierden participantes. Es difícil encontrar participantes apareados en más de dos variables. Por ejemplo, si se decide aparear a los sujetos en cuanto a inteligencia, género y clase social, se puede tener éxito al aparear las dos primeras variables, pero se puede fracasar al intentar encontrar pares que sean bastante iguales en las tres variables. Añádase una cuarta variable y el problema se torna difícil, incluso con frecuencia imposible de resolver.

Sin embargo, no se tire al bebé con el agua del baño. Cuando existe una correlación sustancial entre la variable o variables apareadas y la variable dependiente (mayor que .50 o .60), entonces el apareamiento reduce el término del error y aumenta la precisión de un experimento, un resultado deseable. Si se utilizan los mismos participantes con diferentes tratamientos experimentales —llamados medidas repetidas o diseño de bloque aleatorizado—, entonces se tienen poderosos controles de la varianza. ¿Cómo se puede realizar un mejor apareamiento en todas las variables posibles que apareando un sujeto consigo mismo? Por desgracia, otras consideraciones negativas generalmente impiden dicha posibilidad. Debe enfatizarse con vigor que el apareamiento de cualquier tipo no sustituye la aleatorización. Si se aparea a los participantes, *entonces ellos deben ser asignados aleatoriamente a los grupos experimentales.* A través de un procedimiento aleatorio, como lanzar una moneda o utilizar números pares o nones aleatorios, los miembros de los pares apareados son asignados a los grupos experimental y control. Si los mismos participantes son sometidos a todos los tratamientos, entonces el orden de los tratamientos debe asignarse aleatoriamente. Esto añade control de aleatorización al apareamiento, o control de medidas repetidas.

Un principio sugerido por el presente análisis es: *cuando una variable apareada se correlaciona sustancialmente con la variable dependiente, el apareamiento, como forma de control de la varianza, resulta útil y deseable.* Sin embargo, antes de realizar un apareamiento se

deben sopesar cuidadosamente sus ventajas y desventajas en una situación de investigación particular. La aleatorización completa o el análisis de covarianza pueden ser mejores métodos de control de la varianza.

Otra forma de control, el control estadístico, se analizó en capítulos previos; pero uno o dos comentarios son pertinentes en este momento. Los métodos estadísticos constituyen, por así decirlo, formas de control en el sentido de que aíslan y cuantifican las varianzas. Pero el control estadístico es inseparable de otras formas de control de diseño. Por ejemplo, si se utiliza el apareamiento, debe usarse una prueba estadística apropiada, de otro modo se perderá el efecto del apareamiento y, por lo tanto, el control.

Minimización de la varianza del error

La *varianza del error* es la variabilidad de las medidas debidas a fluctuaciones aleatorias, cuya característica básica es que son autocompensatorias, es decir, en un momento varían de una forma, luego de otra, a veces positiva, a veces negativa, a veces hacia arriba, a veces hacia abajo. Los errores aleatorios tienden a equilibrarse entre sí, de tal manera que su media es cero.

Existen varios determinantes de la varianza del error, por ejemplo, factores asociados con las diferencias individuales entre los participantes. Por lo común, a esta varianza debida a diferencias individuales se le llama "varianza sistemática". Pero cuando dicha varianza no puede identificarse ni controlarse, debe agruparse con la varianza del error. Puesto que muchos determinantes interactúan y tienden a cancelarse entre sí (o al menos se supone que lo hacen), la varianza del error tiene estas características aleatorias.

Otra fuente de la varianza del error es aquella asociada con los llamados errores de medición: variación de las respuestas de un ensayo a otro, adivinación, inatención momentánea, fatiga ligera temporal, fallas de la memoria, estados emocionales transitorios de los participantes, etcétera.

Minimizar la varianza del error tiene dos aspectos principales: 1) la reducción de los errores de medición a través de condiciones controladas y 2) un aumento en la confiabilidad de las medidas. Mientras menor sea el control de las condiciones de un experimento, mayor será la influencia de los muchos determinantes de la varianza del error. Ésta es una de las razones para establecer con cuidado condiciones experimentales controladas. En estudios bajo condiciones de campo, por supuesto, dicho control se vuelve difícil; aun así, deben realizarse esfuerzos constantes para disminuir los efectos de los múltiples determinantes de la varianza del error. Esto puede hacerse, en parte, dando instrucciones claras y específicas a los participantes, y excluyendo de la situación experimental los factores que sean extraños al propósito de la investigación.

Incrementar la confiabilidad de las medidas implica reducir la varianza del error. Aunque en posteriores capítulos se incluyen análisis más completos sobre el tema, de momento diremos que la confiabilidad se considera como la precisión de un conjunto de puntuaciones. Hasta el punto en que las puntuaciones no fluctúen aleatoriamente, son confiables. Imagine un instrumento de medición no confiable por completo; dicho instrumento no permite predecir el desempeño futuro de los individuos; en un momento brinda un ordenamiento de los rangos para una muestra de participantes y un ordenamiento diferente de los rangos en otro momento. Con un instrumento de este tipo no sería posible identificar ni extraer las varianzas sistemáticas, debido a que las puntuaciones generadas por el instrumento serían como números en una tabla de números aleatorios, que es el caso extremo. Ahora imaginen cantidades diferentes de confiabilidad y de no confiabilidad en las medidas de la variable dependiente. Cuanto más confiables sean las medidas, mejor se

podrán identificar y extraer las varianzas sistemáticas y menor será la varianza del error en relación con la varianza total.

Otra razón para reducir la varianza del error tanto como sea posible, es darle la oportunidad a la varianza sistemática para que se muestre a sí misma, lo cual no puede hacerse si la varianza del error y, por lo tanto, el término del error, son demasiado grandes. Si existe una relación, se busca descubrirla. Una forma de descubrir la relación consiste en encontrar diferencias significativas entre las medias. Pero si la varianza del error es relativamente grande debido a errores de medición no controlados, la varianza sistemática —llamada antes varianza "entre"— no tendrá la oportunidad de aparecer. Por lo tanto, aunque existente, la relación probablemente no será detectada.

El problema de la varianza del error puede expresarse con claridad en forma matemática: recuerde la siguiente ecuación:

$$V_t = V_e + V_e$$

donde V_t es la varianza total en un conjunto de medidas; V_e es la varianza entre grupos, la varianza presumiblemente debida a la influencia de las variables experimentales; y V_e es la varianza del error (en el análisis de varianza, la varianza dentro de grupos y la varianza residual). En efecto, a mayor valor de V_e, menor deberá ser V_e, con una cantidad dada de V_t.

Considere la siguiente ecuación: $F = V_e/V_e$. Para que el numerador de la fracción a la derecha sea evaluado con precisión respecto a una desviación significativa de las expectativas por el azar, el denominador debe ser una medida exacta del error aleatorio.

Un ejemplo familiar sirve para aclarar esto. Recuerde que en la discusión sobre el análisis de varianza factorial y sobre el análisis de varianza de grupos correlacionados, se habló sobre la varianza debida a las diferencias individuales presentes en las medidas experimentales. Se afirmó que, mientras que la aleatorización adecuada puede igualar efectivamente a los grupos experimentales, habrá varianza en las puntuaciones debida a las diferencias individuales; por ejemplo, diferencias debidas a la inteligencia, aptitud, etcétera. Ahora, en algunas situaciones, estas diferencias individuales pueden ser bastante grandes. Si lo son, entonces la varianza del error y, en consecuencia, el denominador de la ecuación F anterior, serán "demasiado grandes" en relación con el numerador; es decir, las diferencias individuales habrán sido aleatoriamente dispersadas entre, digamos, dos, tres o cuatro grupos experimentales. Aun así, son fuentes de varianza y, como tales, inflarán la varianza dentro de los grupos o la residual, es decir, el denominador de la ecuación anterior.

RESUMEN DEL CAPÍTULO

1. Los diseños de investigación son planes y estructuras utilizados para responder preguntas de investigación.
2. Los diseños de investigación tienen dos propósitos básicos: (i) proporcionar respuestas a preguntas de investigación, y (ii) controlar la varianza.
3. Los diseños de investigación funcionan en conjunto con las hipótesis de investigación para generar una respuesta confiable y válida.
4. Los diseños de investigación también pueden indicar qué prueba estadística emplear para analizar los datos recolectados a partir de ese diseño.
5. Hablar de controlar la varianza, puede referirse a una o más de tres cuestiones:
   - maximizar la varianza sistemática
   - controlar la varianza extraña
   - minimizar varianza del error

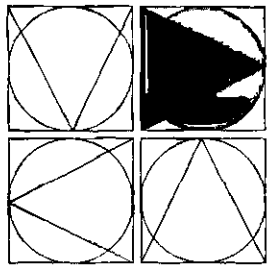
6. Para maximizar la varianza sistemática debe tenerse una variable independiente, cuyos niveles sean muy distintos entre sí.
7. Para controlar la varianza extraña, el investigador necesita eliminar los efectos de una variable independiente potencial sobre la variable dependiente, lo cual puede hacerse por medio de:

   - mantener constante la variable independiente; por ejemplo, si se sabe que el género tiene un efecto potencial, entonces puede mantenerse constante al realizar el estudio incluyendo sólo un género (por ejemplo mujeres).
   - la aleatorización, que se refiere a elegir participantes de manera aleatoria y después asignar aleatoriamente a cada grupo de participantes a las condiciones de tratamiento (niveles de la variable independiente).
   - incorporar la variable extraña al diseño, convirtiéndola en una variable independiente.
   - apareando a los participantes; este método de control puede resultar difícil en ciertas situaciones; un investigador nunca podrá estar muy seguro de que se realizó un apareamiento exitoso en todas las variables importantes.

8. Minimizar la varianza del error incluye la medición de la variable dependiente. Al reducir el error de medición se reduce la varianza del error. El incremento en la confiabilidad de la medición también conlleva una reducción de la varianza del error.

SUGERENCIAS DE ESTUDIO

1. Hemos notado que el diseño de investigación tiene el propósito de obtener respuestas a las preguntas de investigación y controlar la varianza. Explique en detalle qué significa esta afirmación. ¿Cómo controla la varianza un diseño de investigación? ¿Por qué un diseño factorial debería controlar más varianza que un diseño de un factor? ¿Cómo es que un diseño que utiliza participantes apareados o medidas repetidas de los mismos participantes controla la varianza? ¿Cuál es la relación entre las preguntas de investigación, las hipótesis de investigación y un diseño de investigación? Invente un problema de investigación para ilustrar sus respuestas a estas preguntas (o utilice un ejemplo del texto).
2. Sir Ronald Fisher (1951), el inventor del análisis de varianza, en uno de sus libros dijo que debe aclararse que la hipótesis nula nunca se confirma o establece; pero que es posible refutarla en el curso de la experimentación. Se puede decir que todo experimento existe para dar a los hechos la oportunidad de refutar la hipótesis nula. Ya sea que usted esté de acuerdo o no con la afirmación de Fisher, ¿qué piensa que quiso decir con ello? Para estructurar su respuesta, recuerde el principio maxmincon y las pruebas F y t.



CAPÍTULO 19

DISEÑOS INADECUADOS Y CRITERIOS PARA EL DISEÑO

- ENFOQUES EXPERIMENTAL Y NO EXPERIMENTAL
- SIMBOLOGÍA Y DEFINICIONES
- DISEÑOS DEFECTUOSOS
  - **Medición, historia, maduración**
  - **El efecto de regresión**
- CRITERIOS DEL DISEÑO DE INVESTIGACIÓN
  - **¿Responder preguntas de investigación?**
  - **Control de variables independientes extrañas**
  - **Posibilidad de generalización**
  - **Validez interna y externa**

Todas las creaciones de las disciplinas de los seres humanos tienen forma. La arquitectura, la poesía, la música, la pintura, las matemáticas, la investigación científica, todas tienen forma. La gente pone gran énfasis en el contenido de sus creaciones, frecuentemente sin darse cuenta de que sin una estructura fuerte, no importa cuán rico y significativo sea el contenido, las creaciones pueden resultar débiles y estériles.

Lo mismo sucede con la investigación científica. El científico requiere de una forma viable y flexible con la cual expresar las metas científicas. Sin contenido —sin una buena teoría, buenas hipótesis, buenos problemas— el diseño de investigación está vacío. Pero sin forma, sin una estructura concebida y creada adecuadamente para el propósito de la investigación, pueden lograrse pocas cosas valiosas. De hecho, no es exagerado afirmar que muchos de los fracasos en la investigación del comportamiento han sido fallas en las formas de disciplina e imaginación.

El enfoque principal de este capítulo son los diseños de investigación inadecuados. Tales diseños han sido tan comunes que requieren analizarse. Más importante aún, es tener presente que el estudiante debe ser capaz de reconocerlos y de entender por qué son inadecuados. Este enfoque negativo tiene una virtud: el estudio de las deficiencias obliga a preguntarse por qué algo es deficiente, lo que a su vez centra la atención en los criterios

6. Para maximizar la varianza sistemática debe tenerse una variable independiente, cuyos niveles sean muy distintos entre sí.
7. Para controlar la varianza extraña, el investigador necesita eliminar los efectos de una variable independiente potencial sobre la variable dependiente, lo cual puede hacerse por medio de:

   - mantener constante la variable independiente; por ejemplo, si se sabe que el género tiene un efecto potencial, entonces puede mantenerse constante al realizar el estudio incluyendo sólo un género (por ejemplo mujeres).
   - la aleatorización, que se refiere a elegir participantes de manera aleatoria y después asignar aleatoriamente a cada grupo de participantes a las condiciones de tratamiento (niveles de la variable independiente).
   - incorporar la variable extraña al diseño, convirtiéndola en una variable independiente.
   - apareando a los participantes; este método de control puede resultar difícil en ciertas situaciones; un investigador nunca podrá estar muy seguro de que se realizó un apareamiento exitoso en todas las variables importantes.

8. Minimizar la varianza del error incluye la medición de la variable dependiente. Al reducir el error de medición se reduce la varianza del error. El incremento en la confiabilidad de la medición también conlleva una reducción de la varianza del error.

SUGERENCIAS DE ESTUDIO

1. Hemos notado que el diseño de investigación tiene el propósito de obtener respuestas a las preguntas de investigación y controlar la varianza. Explique en detalle qué significa esta afirmación. ¿Cómo controla la varianza un diseño de investigación? ¿Por qué un diseño factorial debería controlar más varianza que un diseño de un factor? ¿Cómo es que un diseño que utiliza participantes apareados o medidas repetidas de los mismos participantes controla la varianza? ¿Cuál es la relación entre las preguntas de investigación, las hipótesis de investigación y un diseño de investigación? Invente un problema de investigación para ilustrar sus respuestas a estas preguntas (o utilice un ejemplo del texto).
2. Sir Ronald Fisher (1951), el inventor del análisis de varianza, en uno de sus libros dijo que debe aclararse que la hipótesis nula nunca se confirma o establece; pero que es posible refutarla en el curso de la experimentación. Se puede decir que todo experimento existe para dar a los hechos la oportunidad de refutar la hipótesis nula. Ya sea que usted esté de acuerdo o no con la afirmación de Fisher, ¿qué piensa que quiso decir con ello? Para estructurar su respuesta, recuerde el principio maxmincon y las pruebas F y t.

utilizados para juzgar tanto las adecuaciones como las inadecuaciones. Así, el estudio de diseños inadecuados conduce al estudio de los criterios del diseño de investigación. También se aprovecha la ocasión para describir el sistema simbológico a utilizar, así como para identificar una distinción importante entre investigación experimental y no experimental.

Enfoques experimental y no experimental

La discusión sobre el diseño se inicia por medio de una distinción importante: aquella entre los enfoques experimental y no experimental de la investigación. De hecho, tal distinción es tan importante que un capítulo separado (capítulo 23) se dedicará a dicho tema. Un *experimento* es una investigación científica donde un investigador manipula y controla una o más variables independientes y observa la(s) variable(s) dependiente(s) para determinar si hay variación concomitante a la manipulación de las variables independientes. Un *diseño experimental*, entonces, es aquel en el que el investigador *manipula* por lo menos una variable independiente. En un capítulo anterior se analizó brevemente el estudio clásico de Hurlock (1925), quien manipuló incentivos para producir diferentes cantidades de retención. En el estudio de Walster, Cleary y Clifford (1970) (capítulo 18), se manipularon género, raza y niveles de habilidad para estudiar sus efectos en la aceptación universitaria: las solicitudes enviadas a las universidades difirieron en las descripciones de los solicitantes como hombre-mujer; blanco-negro; y niveles altos, medios y bajos de habilidad.

En la investigación no experimental no es posible manipular las variables o asignar aleatoriamente a los participantes o tratamientos debido a que la naturaleza de las variables es tal que imposibilita su manipulación. Los participantes llegan al investigador con sus características distintivas intactas, por así decirlo. Vienen con su "ya presente" sexo, inteligencia, nivel ocupacional, creatividad o aptitud. Wilson (1996) utilizó un diseño no experimental para estudiar la legibilidad, contenido étnico y sensibilidad cultural del material educativo sobre pacientes, utilizado por los enfermeros del departamento local de salud y centros comunitarios de salud. En dicho caso el material ya existía; no hubo asignación o selección aleatorias. Edmondson (1996) también usó un diseño no experimental para comparar el número de errores de medicación cometidos por enfermeros, médicos y boticarios en ocho unidades hospitalarias de dos hospitales urbanos de enseñanza. Edmondson no eligió de manera aleatoria estas unidades u hospitales, ni a los profesionales médicos. De la misma forma, en muchas áreas de investigación, por desgracia no es posible realizar asignaciones aleatorias, como se verá más adelante. Aunque la investigación experimental y la no experimental difieren en estos aspectos cruciales, comparten características estructurales y de diseño que se indicarán en éste y subsecuentes capítulos. Además, su propósito básico es el mismo: estudiar relaciones entre fenómenos. Su lógica científica también es la misma: obtener evidencia empírica para realizar proposiciones condicionales de la forma si p, entonces q. En algunos campos de las ciencias sociales y del comportamiento, las estructuras no experimentales son inevitables. Keith (1988) afirma que muchos estudios conducidos por psicólogos escolares son de naturaleza no experimental. Los investigadores de psicología escolar, así como muchos en psicología educativa deben trabajar dentro de una estructura práctica. Muchas veces, escuelas, salones de clase e incluso estudiantes son dados al investigador "como son". Stone-Romero, Weaver y Glenar (1995) sintetizaron casi 20 años de artículos del *Journal of Applied Psychology*, respecto al uso de diseños de investigación experimentales y no experimentales.

El ideal de la ciencia es el experimento controlado. Excepto, quizás, en investigación taxonómica —aquella que tiene el propósito de descubrir, clasificar y medir fenómenos naturales y los factores que subyacen a dichos fenómenos— donde el modelo de ciencia

deseado es el experimento controlado. Puede ser difícil para muchos estudiantes aceptar esta afirmación más bien categórica, puesto que su lógica aún no es aparente. Anteriormente se indicó que la meta principal de la ciencia era descubrir relaciones entre fenómenos; entonces, ¿por qué dar prioridad al experimento controlado? ¿No existen otros métodos para descubrir relaciones? Sí, por supuesto que existen. Sin embargo, la principal razón para la preeminencia del experimento controlado es que los investigadores pueden tener más confianza en que las relaciones que ellos estudian son las relaciones que creen que son. La razón no es difícil de ver: ellos estudian las relaciones bajo las condiciones más cuidadosamente controladas de indagación que se conocen. Así, la virtud única y abrumadoramente importante del estudio experimental es el control. En un estudio experimental perfectamente controlado, el investigador puede confiar en que la manipulación de la variable independiente es lo que afectó a la variable dependiente, y nada más. En resumen, un estudio experimental perfectamente conducido es más confiable que un estudio no experimental perfectamente conducido. La razón de ello debe volverse más obvia conforme se avance en el estudio del diseño de investigación.

Simbología y definiciones

Antes de discutir los diseños inadecuados, resulta necesario explicar la simbología utilizada en estos capítulos. X se utiliza para definir una variable (o variables) independiente que es *experimentalmente manipulada*. X_1, X_2, X_3, etcétera, representan las variables independientes 1, 2, 3, etcétera, aunque por lo común se utiliza la X sola, aun cuando pueda significar más de una variable independiente. (También se utiliza $X_1, X_2, \ldots$, para representar las particiones de una variable independiente; pero la diferencia siempre se hará clara.) El símbolo (X) indica que la variable independiente no está *manipulada* —no está bajo el control directo del investigador, sino que es *medida o imaginada*—. La variable dependiente es Y: Y_a es la variable independiente *antes* de la manipulación de X, y Y_d es la variable dependiente *después* de la manipulación de X. Con $\sim X$ se toma prestado el signo de negación de la teoría de conjuntos: $\sim X$ ("no-X") para indicar que la variable experimental (la variable independiente X) *no* está manipulada. [Nota: (X) es una variable no manipulable y $\sim X$ es una variable manipulable que no está manipulada.] El símbolo (A) se utilizará para la asignación aleatoria de los participantes a los grupos experimentales y para la asignación aleatoria de los tratamientos experimentales a los grupos experimentales.

La explicación dada respecto a $\sim X$ no es muy precisa puesto que en algunos casos $\sim X$ puede representar un aspecto diferente del tratamiento X, más que la simple ausencia de tratamiento. En el lenguaje científico usado antes, el grupo experimental era el grupo al que se le daba el llamado tratamiento experimental, X; mientras que el grupo control no lo recibía, $\sim X$. Para los propósitos del texto, sin embargo, $\sim X$ será suficiente, especialmente si se entiende el significado generalizado de *control* explicado antes. Entonces, un *grupo experimental* es un grupo de participantes que reciben algún aspecto o tratamiento de X. En la comprobación de la hipótesis de frustración-agresión, el grupo experimental es aquel a cuyos participantes se les induce frustración sistemáticamente. En contraste, el grupo control es aquel al que "no" se le da tratamiento.

En la investigación multivariada moderna es necesario expandir estos conceptos. En esencia no cambian, sino que se expanden. Como se ha visto, es muy posible tener más de un grupo experimental. No sólo son posibles diferentes grados de manipulación de la variable independiente, sino que con frecuencia son deseables, e incluso imperativos. Además, es posible incluir más de un grupo control, afirmación que de entrada parece absurda. ¿Cómo es posible tener diferentes grados de "no" tratamiento experimental? Esto

sucede porque el concepto de *control* se generaliza. Cuando hay más de dos grupos, y cuando cualquier par de éstos es tratado de manera diferente, uno o más grupos sirven como "controles" para los otros. Recuerde que el control se refiere siempre al control de varianza. Con dos o más grupos tratados de forma diferente, la varianza es generada por la manipulación experimental. Así, el concepto tradicional de X y $\sim X$ (tratamiento y no tratamiento) se generaliza a X_1, X_2, X_3,..., X_k, formas o grados diferentes de tratamiento.

Si X está entre paréntesis (X), significa que el investigador "se imagina" la manipulación de X, o supone que X ocurrió y que se trata de la X de la hipótesis. También puede significar que X está siendo medida y no manipulada. En realidad aquí se está señalando lo mismo de diferente forma; el contexto del análisis debería dejar en claro la distinción. Suponga que un sociólogo estudia la delincuencia y la hipótesis de frustración-agresión. El investigador observa la delincuencia, Y, e imagina que los participantes delincuentes sufrieron frustración en sus primeros años, o (X). Todos los diseños no experimentales tendrán (X); entonces, (X) por lo común representa una variable independiente *que no está bajo el control experimental del investigador.*

Un punto más: en general cada diseño en este capítulo tendrá una forma *a* y una *b*. La forma *a* será la forma experimental, o aquella en la cual se manipula X. La forma *b* será la forma no experimental, en la cual X no está bajo el control del investigador, o (X). Obviamente también es posible ($\sim X$).

Diseños defectuosos

Existen cuatro (o más) diseños de investigación inadecuados que con frecuencia se han usado —y aún se utilizan ocasionalmente— en la investigación del comportamiento. Los defectos de los diseños conducen a un control pobre de las variables independientes. A continuación se enumera cada uno de estos diseños, se le da un nombre, se esquematiza su estructura y después se analiza.

Diseño 19.1: De un grupo		
(*a*) X	*Y*	(Experimental)
(*b*) (X)	*Y*	(No experimental)

El diseño 19.1 ha sido llamado "estudio de caso de un disparo", un nombre pertinente asignado por Campbell y Stanley (1963). La forma (*a*) del diseño es experimental y la forma (*b*) es no experimental. Un ejemplo de investigación con un diseño de la forma (*a*): el cuerpo docente de una facultad instituye un nuevo currículum y busca evaluar sus efectos. Después de un año, se mide Y, el aprovechamiento de los estudiantes. Se concluye, digamos, que el aprovechamiento se ha incrementado con el nuevo programa. Con un diseño de este tipo, la conclusión es débil. El diseño 19.1(*b*) es la forma no experimental del diseño de un grupo. Se estudia Y, el resultado; y X se supone o imagina. Un ejemplo sería el estudio de la delincuencia al analizar el pasado de un grupo de delincuentes juveniles para identificar los factores que probablemente los hayan conducido a su comportamiento antisocial. El método es problemático debido a que pueden confundirse los factores (variables). Cuando el efecto de dos o más factores (variables) no puede separarse, los resultados se vuelven difíciles de interpretar; cualquier número de posibles explicaciones serían plausibles.

Desde el punto de vista científico, el diseño 19.1 carece de valor. Virtualmente no hay control de otras posibles influencias sobre el resultado. Como Campbell señaló (1957), el

mínimo de información científica útil requiere de por lo menos una comparación formal. El ejemplo del currículum requiere, *por lo menos*, una comparación entre el grupo que experimentó el nuevo currículum con otro que no lo haya experimentado. El supuesto efecto del nuevo curriculum, digamos tal y cual aprovechamiento, muy bien pudo haber sido el mismo como resultado de cualquier tipo de currículum. El punto no es si el curriculum tuvo o no un efecto, sino que sin una comparación formal y controlada del desempeño de los miembros del grupo "experimental", contra el desempeño de los miembros de algún otro grupo que no experimentó el nuevo currículum, poco es lo que puede decirse acerca de su efecto.

Una distinción importante requiere tomarse en cuenta. No es que el método carezca por completo de valor, sino que *científicamente* carece de valor. En la vida diaria, por supuesto, dependemos de este tipo de evidencia científicamente cuestionable; tenemos que hacerlo. Actuamos, digamos, con base en nuestra experiencia; tenemos la esperanza de que utilizamos nuestra experiencia de forma racional. No se critica el paradigma del pensamiento diario implicado en el diseño 19.1; únicamente que cuando un paradigma de ese tipo se utiliza o se considera científico, entonces comienzan las dificultades. Aun en tareas intelectuales elevadas se utiliza el pensamiento implícito en este diseño. Las observaciones cuidadosas y los análisis brillantes y creativos de Freud sobre el comportamiento neurótico parecen caer dentro de esta categoría. La queja no es en contra de Freud, sino en contra de las suposiciones de que estas conclusiones están "científicamente establecidas".

Diseño 19.2: De un grupo, antes-después (pretest, postest)			
(*a*) Y_a	X	Y_d	(Experimental)
(*b*) Y_a	(X)	Y_d	(No experimental)

El diseño 19.2 representa sólo una pequeña mejoría del diseño 19.1. La característica esencial de esta forma de investigación es que un grupo se compara consigo mismo. Teóricamente, no existe una mejor opción puesto que se controlan todas las variables independientes posibles asociadas con las características de los participantes. El procedimiento sugerido por un diseño de este tipo es el siguiente: se mide un grupo en su variable dependiente, Y, antes de la manipulación experimental, lo cual se llama generalmente *pretest*. Suponga que se miden las actitudes de un grupo de participantes hacia las mujeres; se utiliza una manipulación experimental diseñada para cambiar dichas actitudes. Un experimentador podría exponer al grupo a una opinión experta sobre los derechos de la mujer, por ejemplo. Después de la interposición de esta X, las actitudes de los participantes se miden nuevamente. Se examinan las diferencias de las puntuaciones sobre el cambio de actitud, o $Y_d - Y_a$.

Aparentemente, ésta parecería una buena manera de lograr el propósito experimental. Después de todo, si las diferencias entre las puntuaciones son estadísticamente significativas, ¿esto no indica un cambio en las actitudes? La situación no resulta tan sencilla. Existen otros factores que quizás hayan contribuido al cambio en las puntuaciones; por lo tanto, se confunden los factores. Campbell (1957) ofrece una excelente y detallada discusión de estos factores; aquí únicamente se presenta un resumen sobre ello.

Medición, historia, maduración

Primero está el posible efecto del procedimiento de medición: el hecho de medir a los participantes los cambia. ¿Es posible que las medidas post-X se vieran influenciadas no

mínimo de información científica útil requiere de por lo menos una comparación formal. El ejemplo del currículum requiere, *por lo menos*, una comparación entre el grupo que experimentó el nuevo currículum con otro que no lo haya experimentado. El supuesto efecto del nuevo curriculum, digamos tal y cual aprovechamiento, muy bien pudo haber sido el mismo como resultado de cualquier tipo de currículum. El punto no es si el curriculum tuvo o no un efecto, sino que sin una comparación formal y controlada del desempeño de los miembros del grupo "experimental", contra el desempeño de los miembros de algún otro grupo que no experimentó el nuevo currículum, poco es lo que puede decirse acerca de su efecto.

Una distinción importante requiere tomarse en cuenta. No es que el método carezca por completo de valor, sino que *científicamente* carece de valor. En la vida diaria, por supuesto, dependemos de este tipo de evidencia científicamente cuestionable; tenemos que hacerlo. Actuamos, digamos, con base en nuestra experiencia; tenemos la esperanza de que utilizamos nuestra experiencia de forma racional. No se critica el paradigma del pensamiento diario implicado en el diseño 19.1; únicamente que cuando un paradigma de ese tipo se utiliza o se considera científico, entonces comienzan las dificultades. Aun en tareas intelectuales elevadas se utiliza el pensamiento implícito en este diseño. Las observaciones cuidadosas y los análisis brillantes y creativos de Freud sobre el comportamiento neurótico parecen caer dentro de esta categoría. La queja no es en contra de Freud, sino en contra de las suposiciones de que estas conclusiones están "científicamente establecidas".

Diseño 19.2: De un grupo, antes-después (pretest, postest)

(*a*) Y_a	X	Y_d	(Experimental)
(*b*) Y_a	(X)	Y_d	(No experimental)

El diseño 19.2 representa sólo una pequeña mejoría del diseño 19.1. La característica esencial de esta forma de investigación es que un grupo se compara consigo mismo. Teóricamente, no existe una mejor opción puesto que se controlan todas las variables independientes posibles asociadas con las características de los participantes. El procedimiento sugerido por un diseño de este tipo es el siguiente: se mide un grupo en su variable dependiente, Y, antes de la manipulación experimental, lo cual se llama generalmente *pretest*. Suponga que se miden las actitudes de un grupo de participantes hacia las mujeres; se utiliza una manipulación experimental diseñada para cambiar dichas actitudes. Un experimentador podría exponer al grupo a una opinión experta sobre los derechos de la mujer, por ejemplo. Después de la interposición de esta X, las actitudes de los participantes se miden nuevamente. Se examinan las diferencias de las puntuaciones sobre el cambio de actitud, o $Y_d - Y_a$.

Aparentemente, ésta parecería una buena manera de lograr el propósito experimental. Después de todo, si las diferencias entre las puntuaciones son estadísticamente significativas, ¿esto no indica un cambio en las actitudes? La situación no resulta tan sencilla. Existen otros factores que quizás hayan contribuido al cambio en las puntuaciones; por lo tanto, se confunden los factores. Campbell (1957) ofrece una excelente y detallada discusión de estos factores; aquí únicamente se presenta un resumen sobre ello.

Medición, historia, maduración

Primero está el posible efecto del procedimiento de medición: el hecho de medir a los participantes los cambia. ¿Es posible que las medidas post-X se vieran influenciadas no

sólo por la manipulación de X sino por un incremento en la sensibilización por el pretest? Campbell (1957) llama a dichas medidas *reactivas*, ya que por sí mismas provocan que el sujeto reaccione. Por ejemplo, las actitudes controvertidas parecen ser especialmente susceptibles a dicha sensibilización. Las medidas de aprovechamiento, aunque quizá menos reactivas, también se afectan. Las medidas que involucran a la memoria son susceptibles; si se responde un examen ahora, es más probable que se recuerden las últimas cosas incluidas en el examen. En resumen, los cambios observados pueden deberse a efectos reactivos.

Otras dos fuentes importantes de varianza extraña son la *historia* y la *maduración*. Entre las pruebas Y_a y Y_d pueden ocurrir muchas cosas diferentes a X. A mayor periodo de tiempo, mayor será la posibilidad de que variables extrañas afecten a los participantes y, por lo tanto, a las medidas de Y_d. Esto es lo que Campbell (1957) llama *historia*. Dichas variables o eventos son *específicos* para la situación experimental particular. La *maduración*, por otro lado, cubre eventos que son *generales* —no son específicos de cualquier situación particular, sino que reflejan cambio o crecimiento en el organismo estudiado—. La edad mental se incrementa con el tiempo, un incremento que fácilmente afecta el aprovechamiento, la memoria y las actitudes. La gente puede aprender en cualquier intervalo de tiempo dado, y el aprendizaje puede afectar las medidas de la variable dependiente. Ésta es una de las dificultades exasperantes de la investigación, que perdura por periodos considerables. Mientras más prolongado sea el intervalo de tiempo, mayor será la posibilidad de que fuentes extrañas e indeseables de varianza sistemática influyan en las medidas de la variable dependiente.

El efecto de regresión

Un fenómeno estadístico que ha confundido a los investigadores es el llamado *efecto de regresión*. Las puntuaciones de las pruebas cambian como un hecho de la vida estadística: en el retest, en general, los sujetos tienden a regresar a la media. El efecto de regresión opera a causa de la correlación imperfecta entre las puntuaciones del pretest y del postest. Si $r_{da} = 1.00$, entonces no hay efecto de regresión; si $r_{da} = .00$, entonces el efecto es máximo en el sentido de que la mejor predicción de cualquier puntuación del postest, a partir de la puntuación del pretest, es la media. Con la correlación encontrada en la práctica, el efecto neto es que las puntuaciones más bajas en el pretest tienden a ser altas, y las puntuaciones más altas tienden a ser más bajas en el postest —cuando, de hecho, no ha ocurrido un cambio real en la variable dependiente—. De este modo, si en un estudio se utilizan participantes con bajas puntuaciones, sus puntuaciones en el postest probablemente serán más altas que en el pretest, debido al efecto de regresión. Lo anterior puede engañar al investigador al hacerlo creer que la intervención experimental resultó efectiva, cuando en realidad no fue así. De la misma forma, se puede concluir erróneamente que una variable experimental ha tenido un efecto depresor en los sujetos con altas puntuaciones en el pretest, lo cual no es así necesariamente. Las puntuaciones más altas y más bajas de los dos grupos quizá se deban al efecto de regresión. ¿Cómo funciona esto? Existen muchos factores del azar que influyen en cualquier conjunto de puntuaciones. Dos excelentes referencias sobre la discusión del efecto de regresión son la de Anastasi (1958) y la de Thorndike (1963). Para una presentación más compleja desde el punto de vista estadístico, véase Nesselroade, Stigler y Baltes (1980). En el pretest, algunas puntuaciones altas son mayores de lo que "deberían ser" a causa del azar, y lo mismo sucede con algunas puntuaciones bajas. En el postest es poco probable que se mantengan las puntuaciones altas, ya que los factores que las hicieron altas eran factores del azar —los cuales no están correlacionados en el pretest y postest—. De este modo, el sujeto con una puntuación alta tenderá a bajar en el postest. Un argumento similar se aplica al sujeto con baja puntuación, pero de manera inversa.

Los diseños de investigación deben construirse con el efecto de regresión en mente. No hay manera de controlarlo en el diseño 19.2. Si hubiera un grupo control, entonces se podría "controlar" el efecto de regresión, ya que ambos grupos, el control y el experimental, cuentan con un pretest y un postest. Si la manipulación experimental hubiese tenido un efecto "real", entonces ello debería notarse por encima del efecto de regresión. Es decir, las puntuaciones de ambos grupos, manteniendo igual lo demás, se afectan de la misma manera por la regresión y por otras influencias. Así, si los grupos difieren en el postest, debe ser por la manipulación experimental.

El diseño 19.2 resulta inadecuado, no tanto porque puedan operar variables extrañas y el efecto de regresión (las variables extrañas operan siempre que hay un intervalo de tiempo entre el pretest y el postest), sino *porque no se sabe si éstos han operado, si han afectado las medidas de la variable dependiente*. El diseño no brinda oportunidad alguna para controlar o probar tales posibles influencias.

Diseño 19.3: Simulación de antes-después		
	X	Y_a
Y_a		

El título peculiar del diseño 19.3 surge en parte de su propia naturaleza. Como el diseño 19.2, es un diseño antes-después. En lugar de utilizar las mediciones previas y posteriores (o pretest-postest) de un grupo, se emplean como medidas del pretest las medidas de otro grupo, el cual se elige para ser tan similar como sea posible al grupo experimental y, por lo tanto, constituye algo parecido a un grupo control. (La línea entre los dos niveles en el esquema indica grupos separados.) Este diseño satisface la condición de tener un grupo control y, por lo tanto, es un paso más hacia la comparación necesaria en la investigación científica. Por desgracia, los controles son débiles como resultado de la imposibilidad que enfrenta el investigador para saber si los dos grupos eran equivalentes antes de X, la manipulación experimental.

Diseño 19.4: De dos grupos, sin control			
(*a*)	X	Y	(Experimental)
	$\sim X$	$\sim Y$	
(*b*)	(X)	Y	(No experimental)
	($\sim X$)	$\sim Y$	

El diseño 19.4 es común. En (*a*) al grupo experimental se le administra el tratamiento X. El grupo "control", al que se toma o asume como similar al grupo experimental, no recibe X. Las medidas Y se comparan para comprobar el efecto de X. Los grupos o participantes se toman "como son" o pueden ser apareados. La versión no experimental del mismo diseño se clasifica como (*b*). Se observa si un efecto, Y, ocurre en un grupo (línea superior), pero no en otro grupo; o si ocurre en menor grado en el otro grupo (indicado por $\sim Y$ en la línea inferior). Se descubre que el primer grupo experimentó X y el segundo grupo no.

Este diseño tiene una debilidad básica: se *asume* que los dos grupos son iguales respecto a las variables independientes, excepto por X. Algunas veces es posible verificar la igualdad de los grupos de manera general, al compararlos respecto a diferentes variables pertinentes, por ejemplo, edad, sexo, ingresos, inteligencia, habilidad, etcétera. Esto debe hacerse si es posible, pero como Stouffer afirma (1950, p. 522), "con demasiada frecuencia

existe una puerta muy abierta, a través de la cual otras variables no controladas pueden entrar". Puesto que no se utiliza la aleatorización —es decir, los participantes no son asignados aleatoriamente a los grupos—, no es posible suponer que los grupos sean iguales. Ambas versiones del diseño padecen seriamente de falta de control de las variables independientes por la falta de aleatorización.

Criterios del diseño de investigación

Después de examinar algunas de las principales debilidades de los diseños de investigación inadecuados, ahora es un buen momento para discutir lo que puede llamarse *criterios* del diseño de investigación. Junto con los criterios se enunciarán ciertos principios para guiar a los investigadores. Por último, los criterios y principios se relacionarán con las nociones de validez interna y externa de Campbell (1957), las cuales en cierto sentido expresan los criterios de otra forma.

¿Responder preguntas de investigación?

El criterio principal de un diseño de investigación puede expresarse en una pregunta: *¿el diseño responde a la pregunta de investigación?* O *¿el diseño prueba adecuadamente las hipótesis?* Quizá la debilidad más seria de los diseños, con frecuencia propuesta por los neófitos, es que no son capaces de responder adecuadamente las preguntas de investigación. Un ejemplo común de esta falta de congruencia entre las preguntas de investigación y las hipótesis, por un lado, y el diseño de investigación, por el otro, es el apareamiento de los participantes por razones que son irrelevantes a la investigación, y luego el uso de un grupo experimental del tipo de diseño con grupo control. Por ejemplo, los estudiantes a menudo suponen que, debido a que aparean a los sujetos con respecto a inteligencia y género, sus grupos experimentales son iguales. Ellos han escuchado que se requiere aparear a los participantes como "control" y que se necesita un grupo experimental y un grupo control. Sin embargo, frecuentemente las variables apareadas resultan irrelevantes para los propósitos de la investigación. Es decir, si no existe relación entre, digamos, el género y la variable dependiente, el apareamiento por género es irrelevante.

Otro ejemplo de esta debilidad es el caso donde se necesitan tres o cuatro grupos experimentales. Por ejemplo, con tres grupos experimentales y un grupo control, o cuatro grupos con diferentes cantidades o aspectos de X, se requiere el tratamiento experimental. Sin embargo, el investigador usa sólo dos porque ha escuchado que un grupo experimental y un grupo control son necesarios y deseables.

El ejemplo que se presentó en el capítulo 18, referente a la comprobación de una hipótesis de interacción realizando dos experimentos separados, es otro ejemplo. La hipótesis a prueba era que la discriminación en las admisiones a la universidad es una función de género y del nivel de habilidad; que se excluye a las mujeres con baja habilidad (en contraste con los hombres de baja habilidad). Ésta es una hipótesis de interacción y probablemente requiera de un diseño de tipo factorial. Establecer dos experimentos, uno para los aplicantes con alta habilidad y otro para los aplicantes con baja habilidad, constituye un procedimiento pobre porque dicho diseño, como se mostró anteriormente, no prueba en definitiva la hipótesis planteada. De la misma manera, aparear a los participantes respecto a su habilidad y después establecer un diseño de dos grupos, perdería por completo la pregunta de investigación. Tales consideraciones conducen a un precepto general y aparentemente obvio:

Diseñar la investigación para responder preguntas de investigación.

Control de variables independientes extrañas

El segundo criterio es el *control*, que se refiere al control de variables independientes: las variables independientes del estudio de investigación y las variables independientes extrañas. Las variables independientes extrañas son, por supuesto, variables que pueden influir en la variable dependiente; pero que no son parte del estudio. Dichas variables se confunden con la variable independiente bajo estudio. En el estudio sobre admisiones del capítulo 18, por ejemplo, la ubicación geográfica (de las universidades) quizá sea una variable extraña potencialmente influyente, que opaque los resultados del estudio. Es decir, si las universidades del este rechazan más mujeres que las universidades del oeste, entonces la localización geográfica es una fuente de varianza extraña en las medidas de admisión, que debe ser controlada de alguna manera. El criterio se refiere también al control de las variables del estudio. Ya que este problema se ha discutido y continuará en análisis, no es necesario decir más aquí. Pero la pregunta debe plantearse: *¿este diseño controla adecuadamente las variables independientes?*

La mejor forma de responder satisfactoriamente esta pregunta se expresa en el siguiente principio:

> *Aleatorizar siempre que sea posible: seleccionar aleatoriamente a los participantes; asignar aleatoriamente a los participantes a los grupos; asignar aleatoriamente los tratamientos experimentales a los grupos.*

Mientras que quizá no sea posible seleccionar aleatoriamente a los participantes, puede ser posible asignarlos aleatoriamente a los grupos, "igualando" así los grupos en el sentido estadístico analizado en capítulos previos. Si tal asignación aleatoria de los participantes a los grupos no es factible, entonces debe realizarse un gran esfuerzo para asignar aleatoriamente los tratamientos experimentales a los grupos experimentales. Y, si los tratamientos experimentales se administran en diferentes momentos con diferentes experimentadores, entonces los momentos y los experimentadores deben asignarse de forma aleatoria.

El principio que vuelve pertinente la aleatorización es complejo y difícil de aplicar:

> *Controlar las variables independientes para que las fuentes extrañas e indeseables de varianza sistemática tengan la mínima oportunidad de operar.*

Como se aprendió antes (capítulo 8), en teoría la aleatorización satisface este principio. Cuando se prueba la validez empírica de una proposición: si p entonces q, se manipula p y se observa que q covaría con la manipulación de p. ¿Pero qué tanta confianza se puede tener en que la proposición si p entonces q sea realmente "verdadera"? La confianza está directamente relacionada con qué tan completos y adecuados son los controles. Si se utiliza un diseño similar a los diseños 19.1 a 19.4, no se puede tener demasiada confianza en la validez empírica de la proposición si p entonces q, debido a que el control de variables independientes extrañas es débil o inexistente. Puesto que dicho control no es siempre posible en gran parte de la investigación psicológica, sociológica y educativa, entonces, ¿hay que abandonar la investigación por completo? En lo absoluto. Sin embargo, es necesario estar consciente de las debilidades del diseño intrínsecamente pobre.

Posibilidad de generalización

El tercer criterio, la *generalización*, es independiente de otros criterios pues es diferente tipo. Éste es un punto importante que pronto quedará claro. Tan sólo significa: *¿es posible generalizar los resultados de un estudio a otros participantes, otros grupos y otras*

condiciones? Quizá la pregunta se plantee mejor así: *¿qué tanto* pueden generalizarse los resultados del estudio? Quizás ésta sea la pregunta más compleja y difícil que pueda hacerse respectoa los datos de investigación, ya que no sólo toca cuestiones técnicas (como el muestreo y el diseño de investigación), sino también problemas más amplios de la investigación básica y aplicada. En la investigación básica, por ejemplo, la posibilidad de generalización no es la primera consideración, pues el interés central son las relaciones entre variables y por qué tales variables se relacionan como lo hacen. Lo anterior enfatiza los aspectos internos en vez de los aspectos externos del estudio. Estos estudios frecuentemente se diseñan para examinar cuestiones teóricas tales como la motivación o el aprendizaje. La meta de la investigación básica consiste en aportar información y conocimiento a un campo de estudio pero, en general, sin un propósito práctico específico. Sus resultados son generalizables; aunque no en el mismo terreno que los resultados encontrados en estudios de investigación aplicada, en la cual, por otro lado, el interés central obliga a preocuparse más por la generalización, puesto que en efecto se desea aplicar los resultados a otras personas y a otras situaciones. Los estudios de investigación aplicada por lo común se fundamentan en estudios de investigación básica. Con el uso de información encontrada en un estudio de investi-gación básica, los estudios de investigación aplicada utilizan dichos hallazgos para determinar si pueden resolver un problema práctico. Por ejemplo, considere el trabajo de B. F. Skinner; sus primeras investigaciones son generalmente consideradas como investigación básica. Fue a partir de su investigación que los programas de reforzamiento fueron establecidos. Sin embargo, más adelante Skinner y otros (Skinner, 1968; Garfinkle, Kline y Stancer, 1973) aplicaron los programas de reforzamiento a problemas militares, educativos y de comportamiento. Quienes realizan investigación sobre la modificación del comportamiento están aplicando muchas de las teorías e ideas probadas y establecidas por B. F. Skinner. Si el lector pondera los siguientes dos ejemplos de investigación básica y aplicada, entonces podrá acercarse a esta distinción.

En el capítulo 14 se examinó un estudio de Johnson (1994) respecto al tipo de violación, admisibilidad de información y percepción de las víctimas de violación. Esta es claramente investigación básica: el interés central fueron las relaciones entre el tipo de violación, admisibilidad de información y percepción. A pesar de que nadie sería tan insensato para afirmar que a Johnson no le preocupaba el tipo de violación, la admisibilidad de información y la percepción, en general, el énfasis recayó en las relaciones entre las variables del estudio. Contraste este estudio con el esfuerzo de Walster *et al.* (1970) para determinar si las universidades discriminan en contra de las mujeres. Naturalmente, Walster y sus colegas fueron exigentes respecto a los aspectos internos de su estudio; pero ellos por fuerza debían tener otro interés: ¿se practica la discriminación entre las universidades en general? Su estudio es claramente investigación aplicada, aunque no puede decirse que había ausencia de interés de realizar investigación básica. Las consideraciones de la siguiente sección ayudan a explicar la posibilidad de generalización.

Validez interna y externa

Dos criterios generales del diseño de investigación se han discutido con profundidad por Campbell (1957) y por Campbell y Stanley (1963). Estos conceptos constituyen una de las contribuciones más significativas, importantes e informativas a la metodología de la investigación durante las pasadas tres o cuatro décadas.

La *validez interna* plantea la pregunta: ¿La manipulación experimental, *X*, realmente causó una diferencia significativa? Los tres criterios del capítulo 18 en realidad son aspectos de la validez interna. De hecho, cualquier cuestión que afecte los *controles* de un diseño

se convierte en un problema de validez interna. Si un diseño es tal que sólo es posible tener poca o ninguna confianza en las relaciones, como se muestra por las diferencias significativas entre grupos experimentales, entonces se trata de un problema de validez interna.

Con anterioridad en este capítulo se presentaron cuatro amenazas posibles a la validez interna. Algunos autores de libros de texto se han referido a ello como "explicaciones alternativas" (véase Dane, 1990) o "hipótesis rivales" (véase Graziano y Raulin, 1993). Éstas fueron enlistadas como medición, historia, maduración y regresión estadística. Campbell y Stanley (1963) enlistan también otras cuatro amenazas: instrumentación, selección, abandono y la interacción entre algunas de las amenazas mencionadas (un total de ocho).

La instrumentación es un problema del dispositivo utilizado para medir los cambios de la variable dependiente a través del tiempo. Esto es especialmente verdadero en estudios que usan observadores humanos. Los observadores humanos o jueces quizá se vean afectados por eventos previos o por fatiga. Los observadores pueden volverse más eficientes a través del tiempo, de tal manera que las últimas mediciones sean más precisas que las primeras. Por otro lado, los observadores humanos con fatiga se volverían menos precisos en los últimos ensayos que en los primeros; cuando así sucede, los valores de la variable dependiente cambiarán, y dicho cambio no se deberá sólo a la manipulación de la variable independiente.

Con el término selección, Campbell y Stanley (1963) se refieren al tipo de participantes que el experimentador selecciona para el estudio, lo cual ocurre así cuando el investigador no es precavido en estudios que no utilizan selección o asignación aleatorias. El investigador pudo haber seleccionado participantes en cada grupo que fueran muy diferentes en algunas características y, de esta manera, encontrar una diferencia en la variable dependiente. Es importante que el investigador tenga igualados los grupos previamente a la administración del tratamiento. Si los grupos son iguales antes del tratamiento, la lógica indica que si son diferentes después del tratamiento, entonces fue el tratamiento (variable independiente) lo que causó las diferencias y no otra cosa. Sin embargo, si los grupos son diferentes al inicio y diferentes después del tratamiento, es muy difícil afirmar que la diferencia se debió al tratamiento. Más adelante, cuando se discutan los diseños cuasiexperimentales, se verá cómo se puede fortalecer la situación.

El abandono o la mortandad experimental se refiere al retiro de los participantes. Si demasiados participantes en una condición de tratamiento abandonan el estudio, el desequilibrio se vuelve una posible razón del cambio en la variable dependiente. El abandono también incluye la salida de los participantes con ciertas características.

Cualquiera de estas siete amenazas a la validez interna también pueden interactuar entre sí. La selección puede interactuar con la maduración. La amenaza es especialmente posible cuando se utilizan participantes voluntarios. Si el investigador está comparando dos grupos —un grupo formado por voluntarios (autoseleccionados) y el otro grupo, por no voluntarios— la diferencia en el desempeño de ambos en la variable dependiente quizá se deba al hecho de que los voluntarios estén más motivados. Los investigadores estudiantes algunas veces utilizan sujetos voluntarios o miembros de su propia familia o círculo social como participantes. Puede haber un problema de validez interna si se ubica a los voluntarios en un grupo de tratamiento y si sus amigos son colocados en otro.

Un criterio difícil de satisfacer —*la validez externa*— es la *representatividad* o *posibilidad de generalización*. Cuando se completa un estudio y se encuentra una relación, ¿a qué población podría generalizarse? ¿Puede decirse que *A* se relaciona con *B* en todos los alumnos? ¿En todos los alumnos de octavo grado? ¿En todos los alumnos de octavo grado en este sistema escolar? O, ¿en todos los alumnos de octavo grado únicamente de esta

escuela? ¿O los hallazgos deben limitarse a los alumnos de octavo grado con quienes se trabajó? Siempre *deben preguntarse y responderse* estas importantes preguntas científicas.

No sólo debe cuestionarse la generalización de la muestra, sino que también es necesario plantear preguntas acerca de la representatividad ecológica y de las variables de los estudios. Si cambia el escenario social donde se condujo el experimento, ¿se mantendrá aún la relación entre *A* y *B*? ¿Estarán *A* y *B* relacionadas si se replica el estudio en una escuela de menor clase social? ¿En una escuela occidental? ¿En una escuela del sur? Las anteriores son preguntas sobre la *representatividad ecológica*.

La *representatividad de la variable* es un término más sutil. Una pregunta que no se plantea frecuentemente, pero que debe hacerse, es: ¿las variables de este estudio son representativas? Cuando un investigador trabaja con variables psicológicas y sociológicas, se asume que las variables son "constantes". Si el investigador encuentra una diferencia en aprovechamiento entre niños y niñas, se asume que el género, como variable, es "constante".

En el caso de variables como aprovechamiento, agresión, aptitud y ansiedad, ¿el investigador puede suponer que la "agresión" de los participantes suburbanos es la misma "agresión" que se encontraría en los barrios bajos citadinos? ¿La variable es igual en los suburbios europeos? La representatividad de la "ansiedad" es más difícil de determinar. Cuando se habla de "ansiedad", ¿a qué tipo de ansiedad se refiere? ¿Todos los tipos de ansiedad son iguales? Si la ansiedad se manipula en una situación por medio de instrucciones verbales y en otra situación por medio de un choque eléctrico, ¿las dos ansiedades inducidas son iguales? Si la ansiedad es manipulada por, digamos, la instrucción experimental, ¿ésta es la misma ansiedad que la medida por medio de una escala de ansiedad? La representatividad de las variables es, entonces, otro aspecto del gran problema de la validez externa y, por lo tanto, de la generalización.

A menos que se tomen precauciones especiales y que se realicen esfuerzos considerables, los resultados de investigación con frecuencia no son representativos y, por lo tanto, no son generalizables. Campbell y Stanley (1963) afirman que la validez interna es una condición indispensable del diseño de investigación, pero que el diseño ideal debe ser fuerte tanto en la validez interna como en la externa, aun cuando éstas sean frecuentemente contradictorias. En estos capítulos el principal énfasis recae en la validez interna, con un ojo vigilante sobre la validez externa.

Campbell y Stanley (1963) presentan cuatro amenazas a la validez externa. Son los efectos reactivos o de interacción de la prueba, los efectos de interacción de los sesgos de selección y la variable independiente, los efectos reactivos de los arreglos experimentales y la interferencia de tratamiento múltiple.

El efecto reactivo o de interacción de la prueba se refiere al uso de un pretest antes de la administración del tratamiento. Aplicar un pretest quizá disminuya o incremente la sensibilidad del participante a la variable independiente; esto haría que los resultados de la población evaluada en el pretest no sean representativos del efecto del tratamiento para la población que no fue evaluada con anterioridad. La probabilidad de una interacción entre el tratamiento y el pretest parece haber sido señalada en primera instancia por Solomon (1949).

El efecto de interacción del sesgo en la selección y la variable independiente indica que la selección de los participantes puede muy bien afectar la generalización de los resultados. Un investigador que utiliza sólo participantes de la población de sujetos de una universidad en particular, que generalmente consiste de estudiantes de primero y de segundo año, encontrará difícil generalizar los resultados del estudio a otros alumnos de la universidad o de otras universidades.

La mera participación en un estudio de investigación puede constituir un problema en términos de la validez externa. La presencia de observadores, la instrumentación o el

ambiente de laboratorio pueden tener un efecto sobre el participante, lo que no ocurriría si el participante estuviera en un escenario natural. El hecho de participar en un estudio experimental puede alterar la conducta normal del sujeto. Si el experimentador es hombre o mujer, afroamericano o blanco también puede tener un efecto.

Si los participantes son expuestos a más de una condición de tratamiento, el desempeño en ensayos posteriores se ve afectado por el desempeño en los primeros ensayos. Por lo tanto, los resultados sólo pueden generalizarse a personas que han tenido múltiples exposiciones, presentadas en el mismo orden.

El enfoque negativo de este capítulo se hizo con la creencia de que una exposición sobre procedimientos pobres, pero comúnmente utilizados y *aceptados*, junto con una discusión sobre sus mayores debilidades, proporcionarían un buen punto de inicio para el estudio del diseño de investigación. Otros diseños inadecuados son posibles; aunque todos ellos son inadecuados únicamente en sus principios estructurales. Dicho punto debe enfatizarse, ya que en el capítulo 20 se observará que una estructura de diseño perfecta puede ser utilizada pobremente. Por lo tanto, es necesario aprender y entender las dos fuentes de debilidad en la investigación: los diseños intrínsecamente pobres y los diseños intrínsecamente buenos pero pobremente utilizados.

RESUMEN DEL CAPÍTULO

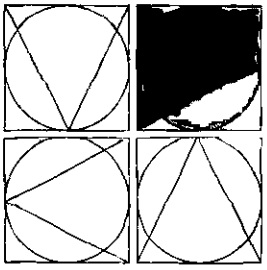
1. El estudio de diseños inadecuados ayuda al investigador a diseñar mejores estudios al saber qué dificultades evitar.
2. Los diseños no experimentales son aquellos con variables independientes no manipuladas y con ausencia de asignación o selección aleatorias.
3. Los diseños inadecuados incluyen el diseño de "estudio de caso de un disparo", el diseño pretest-postest, el diseño pretest-postest simulado y el diseño de dos grupos sin control.
4. Los diseños inadecuados se analizan en términos de su validez interna.
5. La validez interna implica qué tanto el experimentador puede establecer el efecto de la variable independiente sobre la variable dependiente. A mayor confianza del investigador respecto a la variable independiente manipulada, más fuerte será la validez interna.
6. Los estudios no experimentales son más débiles en cuanto a la validez interna que los estudios experimentales.
7. Existen ocho clases básicas de variables extrañas, las cuales, si no son controladas, pueden confundirse con la variable independiente. Las ocho clases básicas se denominan amenazas a la validez interna.
8. Las amenazas a la validez interna, según Campbell, se enumeran como sigue:
   - Historia
   - Maduración
   - Prueba o medición
   - Instrumentación
   - Regresión estadística
   - Selección
   - Mortalidad experimental o abandono
   - Interacción selección-maduración
9. La validez externa implica qué tan fuerte es la afirmación que el experimentador puede hacer respecto a la generalización de los resultados del estudio.

10. Campbell y Stanley señalan cuatro fuentes posibles de amenaza a la validez externa:
    - Efecto reactivo o de interacción de la prueba
    - Efectos de interacción de los sesgos de selección y la variable independiente
    - Efectos reactivos de los arreglos experimentales
    - Interferencia de tratamiento múltiple

SUGERENCIAS DE ESTUDIO

1. Suponga que una universidad de arte decide iniciar un nuevo currículum para todos los estudiantes de pregrado. Se pide al profesorado formar un grupo de investigación para estudiar la efectividad del programa durante dos años. Con el objetivo de tener un grupo con el cual comparar al grupo del nuevo currículum, el grupo de investigación solicita que el programa actual se continúe por dos años y que se permita a los estudiantes elegir el programa actual o el nuevo. El grupo de investigación considera que así tendrán un grupo experimental y un grupo control.

   Analice críticamente la propuesta del grupo de investigación. ¿Qué tanta confianza tendría usted en los hallazgos al final de los dos años? Mencione las razones de su reacción positiva o negativa hacia la propuesta.
2. Imagine que usted es profesor de una escuela de posgrado y se le pide juzgar el valor de una tesis doctoral propuesta. La estudiante de doctorado es una jefa escolar que está instituyendo un nuevo tipo de administración dentro de su sistema escolar. Ella planea estudiar los efectos de la nueva administración durante un periodo de tres años y, después, escribir la tesis. Ella dice que no estudiará ninguna otra situación escolar durante el periodo para no sesgar los resultados. Discuta la propuesta y, al hacerlo, plantéese la pregunta: ¿la propuesta es adecuada para un trabajo doctoral?
3. En su opinión, ¿debe basarse estrictamente toda investigación en el criterio de generalización? Explique por qué sí o por qué no. ¿Qué campo puede ser que tenga más investigación básica: psicología o educación? ¿Por qué? ¿Qué implicaciones tienen sus conclusiones para la generalización?
4. ¿Qué tiene que ver la replicación de investigación con la generalización? Explique. Si fuese posible, ¿debería replicarse toda investigación? Explique por qué sí o por qué no. ¿Qué tiene que ver la replicación con la validez interna y externa?



CAPÍTULO 20

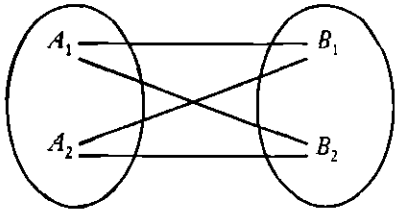
DISEÑOS GENERALES DE INVESTIGACIÓN

El diseño constituye una disciplina de datos. El propósito implícito de todo diseño de investigación consiste en imponer restricciones controladas sobre observaciones de fenómenos naturales. El diseño de investigación, en efecto, le dice al investigador: Haga esto y aquello; no haga esto ni aquello; tenga cuidado con esto; ignore aquello; etcétera. Es el proyecto del arquitecto e ingeniero de investigación. Si el diseño está concebido de forma estructuralmente pobre, el producto final será defectuoso. Si al menos está bien concebido desde el punto de vista estructural, el producto final tiene una mayor probabilidad de alcanzar atención científica seria. En este capítulo, la principal preocupación son distintos diseños básicos de investigación "buenos". También se analizan ciertos fundamentos conceptuales de investigación y algunos problemas relacionados con el diseño; por ejemplo, la lógica de los grupos control y los pros y los contras del apareamiento.

Fundamentos conceptuales del diseño de investigación

Los fundamentos conceptuales para entender el diseño de investigación se establecieron en los capítulos 4 y 5, donde se definieron y analizaron los conjuntos y las relaciones. Recuerde que una *relación* es un conjunto de pares ordenados y también que un *producto cartesiano* son todos los pares ordenados posibles de dos conjuntos. Una *partición* divide un conjunto universal U en subcojuntos que están *separados* y son *exhaustivos*. Una *partición cruzada* es una partición nueva que surge de partir sucesivamente U formando todos los subconjuntos de la forma $A \cap B$. Tales definiciones se explicaron en los capítulos 5 y 6. Ahora se aplicarán al diseño y a las ideas de análisis.

Tome dos conjuntos, A y B, divididos en A_1 y A_2, B_1 y B_2. El producto cartesiano de los dos conjuntos es:



Los pares ordenados, entonces, son: A_1B_1, A_1B_2, A_2B_1, A_2B_2. Puesto que hay un conjunto de pares ordenados, esto es, una relación; también es una partición cruzada. El lector debe revisar las figuras 4.7 y 4.8 del capítulo 4 para ayudarse a aclarar estas ideas y para conocer la aplicación del producto cartesiano y las ideas de relación en el diseño de investigación. Por ejemplo, A_1 y A_2 pueden ser dos aspectos de cualquier variable independiente: experimental-control, dos métodos, hombre y mujer, etcétera.

Un *diseño* es algún subconjunto del producto cartesiano de las variables independientes y la variable dependiente. Es posible aparear cada medida de la variable dependiente, a la cual se le llama Y en este análisis, con algún aspecto o partición de una variable independiente. Los casos más simples posibles ocurren con una variable independiente y una variable dependiente. En el capítulo 10, una variable independiente, A, y una variable dependiente, B, se dividieron en $[A_1, A_2]$ y $[B_1, B_2]$, y después se realizó una partición cruzada para formar la ahora familiar tabulación cruzada de 2×2, con frecuencias o porcentajes en las casillas. Sin embargo, el interés está en particiones cruzadas similares de A y B; pero con medidas continuas en las casillas.

Tome solamente a A, utilizando un diseño de análisis de varianza de un factor. Suponga que se tienen tres tratamientos experimentales, A_1, A_2 y A_3 y, para simplificar, dos puntuaciones Y en cada casilla, lo cual se presenta a la izquierda de la figura 20.1, denominada (*a*). Digamos que seis participantes han sido asignados aleatoriamente a tres tratamientos y que las puntuaciones de los seis individuos después de los tratamientos *experimentales* son las que aparecen en la figura.

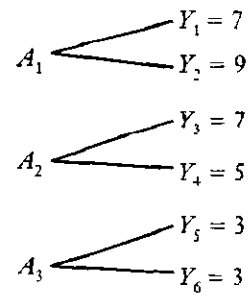
La parte derecha de la figura 20.1, designada como (*b*), muestra la misma idea pero como pares ordenados o en forma de relación. Los pares ordenados son A_1Y_1, A_1Y_2, A_2Y_3, ..., A_3Y_6. Esto no es, por supuesto, un producto cartesiano, el cual aparearía a A_1 con todas las puntuaciones Y, a A_2 con todas las puntuaciones Y, y A_3 con todas las puntuaciones Y, dando un total de $3 \times 6 = 18$ pares. De manera más precisa, la figura 20.1(*b*) es un subconjunto del producto cartesiano $A \times B$. Los diseños de investigación son subconjuntos de $A \times B$, y el diseño y el problema de investigación definen o especifican cómo se establecen los subconjuntos. Los subconjuntos del diseño de la figura 20.1 están presumiblemente dictados por el problema de investigación.

FIGURA 20.1

(*a*)

A_1	A_2	A_3
7	7	3
9	5	3

(*b*)



Cuando hay más de una variable independiente, el problema se vuelve más complejo. Tome dos variables independientes, *A* y *B*, divididas en $[A_1, A_2]$ y $[B_1, B_2]$. El lector no debe confundir esto con el paradigma de frecuencias *AB* previo, en el cual *A* era la variable independiente y *B* la variable dependiente.

Ahora deben tenerse tríos ordenados (o dos conjuntos de pares ordenados): *ABY*. Analice la figura 20.2; en el costado izquierdo de la figura, denominada (*a*) se presenta el diseño de 2 × 2 del análisis de varianza factorial y el ejemplo utilizado en el capítulo 14 (véase la figura 14.2, y las tablas 14.3 y 14.4), con las medidas de la variable dependiente *Y*, insertadas en las casillas; es decir, ocho participantes fueron asignados aleatoriamente a las cuatro casillas. Sus puntuaciones después del experimento, son Y_1, Y_2,..., Y_8. El costado derecho de la figura, denominado (*b*), muestra los tríos ordenados, *ABY*, como un árbol. Obviamente éstos son subconjuntos de $A \times B \times Y$, y son relaciones. El mismo razonamiento es extensible a diseños más grandes y más complejos, como un factorial de 2 × 2 × 3 (*ABCY*) o uno de 4 × 3 × 2 × 2 (*ABCDY*). (En dichas designaciones generalmente se omite *Y* debido a que está implícita.) Otros tipos de diseños se conceptualizan de manera similar, aunque su descripción por medio de árboles puede resultar laboriosa.

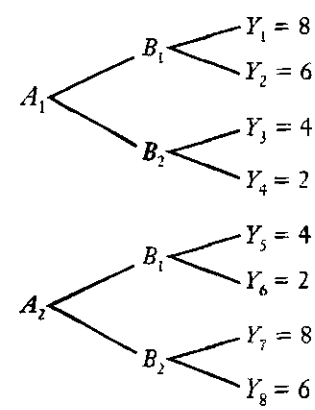
En resumen, un diseño de investigación es algún subconjunto del producto cartesiano de las variables independiente y dependiente. Con sólo una variable independiente, se

FIGURA 20.2

(*a*)

	B_1	B_2
A_1	$Y_1 = 8$ $Y_2 = 6$	$Y_3 = 4$ $Y_4 = 2$
A_2	$Y_5 = 4$ $Y_6 = 2$	$Y_7 = 8$ $Y_8 = 6$

(*b*)



divide la única variable; con más de una variable independiente, las variables independientes se convierten en particiones cruzadas. Con tres o más variables independientes la conceptualización es la misma; sólo difieren las dimensiones, por ejemplo, $A \times B \times C$ y $A \times B \times C \times D$, y sus particiones cruzadas. Siempre que sea posible, es deseable tener diseños "completos" —un diseño completo es una partición cruzada de las variables independientes— y observar las dos condiciones básicas de separación y exhaustividad; es decir, los diseños no deben tener un caso (la puntuación de un participante) en más de una casilla de una partición o de una partición cruzada; y deben utilizarse todos los casos. Además, el mínimo básico de cualquier diseño es, por lo menos, una partición de la variable independiente en dos subconjuntos, por ejemplo, partir A en A_1 y A_2. Existen también diseños "incompletos", aunque en este libro se enfatizan los diseños "completos". Véase Kirk (1995) para estudiar de manera más profunda los diseños incompletos.

El término "diseños generales" establece que los diseños incluidos en el capítulo se simbolizan o expresan en su forma más general y abstracta. Cuando se da una X sencilla (que representa una variable independiente), debe tomarse como un indicador de más de una X —es decir, la X se divide en dos o más grupos experimentales—. Por ejemplo, el diseño 20.1, que se estudiará en breve, tiene X y $\sim X$, que quiere decir que existen grupos control y experimental y, por lo tanto, es una partición de X. Pero X puede dividirse en varias categorías de X, quizás cambiando el diseño, de uno simple con una variable a, digamos, un diseño factorial. No obstante, la simbología básica asociada con el diseño 20.1 permanece igual. Tales complejidades se aclararán en éste y posteriores capítulos.

Una nota preliminar: diseños experimentales y análisis de varianza

Antes de examinar los diseños de este capítulo, es necesario aclarar uno o dos puntos confusos y potencialmente controversiales que por lo común no se consideran en la literatura. La mayoría de los diseños que aquí se estudian son experimentales. Como generalmente se piensa, la lógica de los diseños de investigación está basada en condiciones e ideas experimentales; también están íntimamente ligados a los paradigmas del análisis de varianza. Esto, por supuesto, no es accidental. Las concepciones modernas del diseño, en especial los diseños factoriales, nacieron cuando se inventó el análisis de varianza. Aunque no existe una ley dura que diga que el análisis de varianza sea aplicable únicamente a situaciones experimentales —de hecho, ha sido utilizado muchas veces en investigación no experimental—, por lo común es verdad que resulta más apropiado para los datos de experimentos. Esto es especialmente cierto para los diseños factoriales, donde hay igual número de casos en las casillas del paradigma del diseño, y donde los participantes son asignados aleatoriamente a las condiciones experimentales (o casillas).

Cuando no es posible asignar aleatoriamente a los participantes, y cuando, por una razón u otra, existe un número desigual de casos en las casillas de un diseño factorial, el uso del análisis de varianza se vuelve cuestionable e incluso inapropiado. También puede ser torpe y poco elegante. Lo anterior es así porque el uso del análisis de varianza supone que las correlaciones entre las variables independientes de un diseño factorial son iguales a cero. La asignación aleatoria hace que se mantenga este supuesto, ya que dicha asignación presumiblemente divide las fuentes de varianza de forma igualitaria entre las casillas. Sin embargo, la asignación aleatoria sólo puede lograrse en experimentos. En la investigación no experimental, las variables independientes son características más o menos estables de los participantes (por ejemplo, inteligencia, sexo, clase social y otras similares), que por lo común están correlacionadas sistemáticamente. Considere dos variables indepen-

dientes manipuladas, digamos, reforzamiento y ansiedad. Puesto que los participantes con cantidades variables de características correlacionadas con tales variables están distribuidos aleatoriamente en las casillas, se supone que las correlaciones entre aspectos de reforzamiento y de ansiedad son iguales a cero. Si, por el otro lado, las dos variables independientes son inteligencia y clase social, ambas generalmente no manipuladas y correlacionadas, no se cumple el supuesto de correlación cero entre ellas, necesario para el análisis de varianza. Debe utilizarse algún método de análisis que justifique la correlación entre ellas. Se verá más adelante en el libro que está disponible un método de dicho tipo: la regresión múltiple.

Hasta ahora no se ha alcanzado un estado de madurez de investigación para apreciar la profunda diferencia entre las dos situaciones. Por ahora, sin embargo, acepte la diferencia y la afirmación de que el análisis de varianza es básicamente una concepción y una forma de análisis experimentales. Estrictamente hablando, si las variables independientes son no experimentales, entonces el análisis de varianza no es el tipo de análisis apropiado. No obstante, existen excepciones a esta afirmación; por ejemplo, si una variable independiente es experimental y la otra es no experimental, el análisis de varianza es apropiado. Además, en el análisis de varianza de un factor, ya que sólo hay una variable independiente, el análisis de varianza puede utilizarse con una variable independiente no experimental, aunque quizás el análisis de regresión sería más apropiado. En el número 3 de las sugerencias para estudio se cita un uso interesante del análisis de varianza con datos no experimentales.

De manera similar, si por alguna razón el número de casos en las casillas no es igual (y es desproporcionado), entonces habrá correlación entre las variables independientes y no es sostenible el supuesto de la correlación cero. Esta abstracta y abstrusa digresión del tema principal del diseño puede parecer un poco confusa en este punto del estudio; el problema involucrado deberá quedar claro después de estudiar la investigación experimental y la no experimental y, posteriormente en el libro, el fascinante y poderoso enfoque conocido como regresión múltiple.

Los diseños

En lo que resta de este capítulo se analizan varios diseños básicos de investigación. Recuerde que un diseño es un plan, un proyecto para conceptualizar la estructura de las relaciones entre las variables de un estudio de investigación. Un diseño no sólo dispone las relaciones del estudio, también implica cómo se controla la situación de investigación y cómo se analizarán los datos. Un diseño, en el sentido utilizado en este capítulo, constituye el armazón de la investigación, el cual se recubre con las variables y relaciones de la misma. Los esquemas presentados en los diseños 20.1 a 20.8 representan la estructura simple y abstracta de la investigación. Algunas veces las tablas analíticas, como la figura 20.2 (a la izquierda) y las figuras del capítulo 18 (por ejemplo, las figuras 18.2, 18.3 y 18.5) se llaman *diseños*. Mientras que el llamarles diseños no causa mucho daño, estrictamente hablando son paradigmas analíticos. Sin embargo, para no ser demasiado exigentes, a ambos tipos de representaciones se les llamará "diseños".

Diseño 20.1: Grupo experimental-grupo control: participantes aleatorizados

[*A*]	*X*	*Y*	(Experimental)
	~*X*	*Y*	(Control)

El diseño 20.1, con dos grupos como en la tabla anterior, y sus variantes con más de dos grupos, probablemente son los "mejores" diseños para muchos propósitos experimentales en la investigación del comportamiento. Campbell y Stanley (1963) llaman a este diseño el *diseño de grupo control sólo con postest*, mientras que Isaac y Michael (1987) se refieren a él como *diseño de grupo control aleatorizado sólo con postest*. La [*A*] que aparece antes del paradigma indica que los participantes son asignados aleatoriamente al grupo experimental (línea superior) y al grupo control (línea inferior). Dicha aleatorización elimina las objeciones al diseño 19.4 mencionado en el capítulo 19. En toería todas las variables independientes posibles están controladas; por supuesto que en la práctica puede no ser así. Si se incluyen suficientes participantes en el experimento para darle a la aleatorización una oportunidad de "operar", entonces se tiene un control fuerte y se satisfacen bastante bien las demandas de validez interna. Este diseño controla los efectos de la historia, la maduración y el pretest, pero no mide tales efectos.

Si se extiende a más de dos grupos y si es capaz de responder las preguntas de investigación planteadas, el diseño 20.1 tiene diversas ventajas: 1) tiene el mejor sistema de control teórico integrado que cualquier otro diseño, con una o dos excepciones posibles en casos especiales; 2) resulta flexible, es teóricamente capaz de extenderse a cualquier número de grupos con cualquier número de variables; 3) si se extiende a más de una variable, puede probar varias hipótesis al mismo tiempo; y 4) es elegante estadística y estructuralmente.

Antes de estudiar otros diseños, es necesario examinar el concepto de grupo control, una de las invenciones creativas de los pasados cien años, y ciertas extensiones del diseño 20.1.

La noción del grupo control y las extensiones del diseño 20.1

Evidentemente el término *control* y la expresión "grupo control" no aparecían en la literatura científica a finales del siglo XIX, lo cual está documentado por Boring (1954). Sin embargo, la noción de experimentación controlada es más antigua. Boring afirma que Pascal lo utilizó tan temprano como 1648. Solomon (1949) buscó en la literatura psicológica y no pudo encontrar un solo caso del uso de grupo control antes de 1901. Quizá la noción de grupo control fue utilizada en otros campos, aunque es dudoso que la idea estuviese bien desarrollada. Solomon (p. 175) también señala que el estudio sobre actitudes de Peterson y Thurstone de 1933 fue el primer intento serio por emplear grupos control en la evaluación de los efectos de procedimientos educativos. No es posible encontrar la expresión "grupo control" en la famosa decimoprimera edición (1911) de la *Encyclopedia Britannica*, aunque sí se analiza el método experimental. Solomon también afirma que el diseño de grupo control aparentemente tuvo que esperar desarrollos estadísticos y el avance de la sofisticación estadística entre los psicólogos.

Quizás el primer uso de grupos control en psicología y educación ocurrió en 1901 con la publicación de Thorndike y Woodworth (1901). Uno de los hombres que realizó esta investigación, Thorndike, extendió las ideas básicas y revolucionarias de esta primera investigación a la educación (Thorndike, 1924). En el gigantesco estudio de Thorndike de 8 564 alumnos de muchas escuelas en varias ciudades, los controles fueron grupos educativos independientes. Entre otras comparaciones, él contrastó las ganancias en las puntuaciones de pruebas de inteligencia, presumiblemente generadas por el estudio de inglés, historia, geometría y latín con las presumiblemente generadas por el estudio de inglés, historia, geometría y taller. En efecto, él intentó comparar la influencia del latín y del

taller. También realizó otras comparaciones de naturaleza similar. A pesar de la debilidad del diseño y del control, sus experimentos y otros realizados por quienes él mismo estimuló, eran excepcionales por su discernimiento. Thorndike incluso criticó a sus colegas por no admitir estudiantes de estenografía y mecanografía que no habían estudiado latín, debido a que él reclamaba haber demostrado que la influencia de otros factores sobre la inteligencia era similar. Es interesante el hecho de que él pensara que se necesitaba un número enorme de participantes —exigió 18 000 casos más—. En 1924 estaba bastante consciente de la necesidad de emplear muestras aleatorias.

La noción del grupo control requiere generalización. Suponga que en un experimento educativo se tienen cuatro grupos experimentales como siguen. En A_1 se da un reforzamiento por cada respuesta, en A_2 se da en intervalos regulares de tiempo, en A_3 se da en intervalos aleatorios, y en A_4 no se da el reforzamiento. Técnicamente hay tres grupos experimentales y un grupo control, en el sentido tradicional del grupo control. Sin embargo, A_4 podría ser otro "tratamiento experimental"; podría ser algún tipo de reforzamiento mínimo. Entonces, en el sentido tradicional no habría grupo control. El sentido tradicional del término "grupo control" carece de generalidad. Si se generaliza el concepto de grupo control, la dificultad desaparece. Siempre que haya más de un grupo experimental y a cualesquiera dos grupos se les apliquen diferentes tratamientos, el control está presente en el sentido comparativo antes mencionado. Mientras exista un intento por hacer a dos grupos sistemáticamente diferentes en una variable dependiente, será posible una comparación. Por lo tanto, el concepto tradicional de que un grupo experimental debe recibir el tratamiento, y que éste no se da a un grupo control, es un caso especial de la regla más general de que se necesitan grupos de comparación para la validez interna de la investigación científica.

Si el razonamiento es correcto, es posible establecer diseños como el siguiente:

	X_1	Y
$[A]$	X_2	Y
	X_3	Y

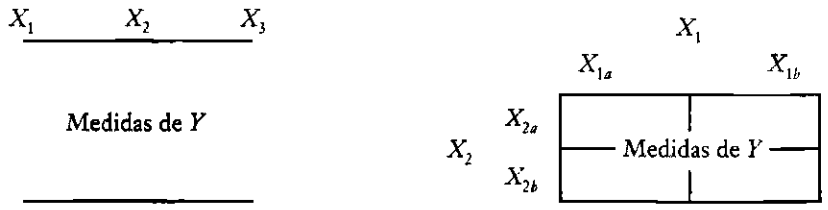
o

	X_{1a}	Y
$[A]$	X_{1b}	Y
	X_{2a}	Y
	X_{2b}	Y

Estos diseños se reconocerán más fácilmente si se construyen a la manera del análisis de varianza, como en la figura 20.3. El diseño de la izquierda es un diseño simple de análisis de varianza de un factor; y el de la derecha, un diseño factorial de $2 \times 3 \times 2$. En el diseño del lado derecho, X_{1a} puede ser el grupo experimental y X_{1b} el control, con X_{2a} y X_{2b} como una variable manipulada o una variable atributiva dicotómica. Éste es, por supuesto, el mismo diseño que se presenta en la figura 20.2(*a*).

La estructura del diseño 20.2 es la misma que aquella del diseño 20.1. La única diferencia es que los participantes están apareados en uno o más atributos. Sin embargo, para

FIGURA 20.3



que el diseño sea un diseño "adecuado", la aleatorización debe aparecer en escena, como se nota por la *a* unida a la *Ap* (de "apareado"). No es suficiente utilizar participantes apareados. Los miembros de cada par deben asignarse aleatoriamente a los dos grupos. Idealmente, el que un grupo sea el grupo experimental o el control debe decidirse también en forma aleatoria. En cualquier caso, cada decisión ha de tomarse lanzando una moneda o utilizando una tabla de números aleatorios; los números pares se usan para un grupo y los nones para el otro. Si existen más de dos grupos, debe utilizarse un sistema de números aleatorios.

Diseño 20.2: Grupo experimental-grupo: participantes apareados

$[Ap_a]$	X	Y	(Experimental)
	$\sim X$	Y	(Control)

Como en el diseño 20.1, es posible, aunque no siempre fácil, utilizar más de dos grupos. (La dificultad de aparear más de dos grupos se estudió anteriormente.) Sin embargo, hay ocasiones en que un diseño apareado constituye un elemento inherente a la situación de investigación. Cuando se utilizan los mismos participantes para dos o más tratamientos experimentales, o cuando se les aplica más de un ensayo a los participantes, el apareamiento es inherente a la situación. En investigación educativa, las escuelas o clases son efectivamente variables, cuando, digamos, se utilizan dos o más escuelas o clases y los tratamientos experimentales se administran en cada escuela o clase, entonces el diseño 20.2 constituye la base de la lógica del diseño. Estudie el paradigma de un diseño de escuelas en la figura 20.4. Hay que destacar el hecho de que la varianza debida a las diferencias entre escuelas —y es posible que dicha varianza sea sustancial— puede ser estimada de inmediato.

Apareamiento contra aleatorización

Aunque la aleatorización, que incluye la selección aleatoria y la asignación aleatoria, constituye el método preferido para controlar la varianza extraña, el uso del apareamiento también tiene su mérito. En varias situaciones fuera de los círculos académicos, los científicos del comportamiento no serán capaces de utilizar la aleatorización para lograr constancia entre los grupos, antes de la administración del tratamiento. En una universidad por lo común se dispone de un universo de participantes de donde seleccionar a los participantes. Los investigadores en una situación como ésta pueden darse el gusto de utilizar procedimientos aleatorios. Sin embargo, en la investigación de negocios éste quizá no sea el caso.

FIGURA 20.4

Escuelas	X_{e1} Experimental 1	X_{e2} Experimental 2	X_c Control
1			
2			
3		Medidas de Y	
4			
5			

Entre los investigadores de mercado es popular la prueba de tienda controlada, que consiste en un experimento de campo. El segundo autor ha conducido dichos estudios para varias compañías de investigación de mercado y para una cadena de tiendas de abarrotes en el sur de California. Una de las metas de la prueba de tienda controlada es ser muy discreto. Si un fabricante de productos de jabón quiere determinar los efectos de un cupón de descuento en la conducta de compra del consumidor, el fabricante no desea que el fabricante competidor de un producto similar se entere del asunto. ¿Por qué? Porque si un competidor supiera que se está realizando un estudio de investigación en una tienda, podría ir y acaparar su propio producto, contaminando así el estudio.

Para regresar al análisis sobre la aleatorización *versus* el apareamiento, muchas veces una cadena de tiendas de abarrotes o una cadena de tiendas departamentales tiene un número finito de tiendas para utilizar en un estudio. La ubicación y la clientela ejercen una enorme influencia sobre las ventas. Generalmente las ventas son la variable dependiente de dichos estudios. Con un número limitado de tiendas de donde escoger para realizar la investigación, la asignación aleatoria a menudo no funciona para igualar grupos de tiendas. Una tienda de la cadena puede hacer tres o cuatro veces el volumen de transacciones que otra y si es elegida al azar creará un gran desequilibrio en el grupo al que pertenezca, especialmente si el otro grupo no incluye una tienda similar para equilibrarlo; en pocas palabras, los grupos ya no serán iguales. Por lo tanto, la solución aquí consiste en aparear las tiendas con una base individual; un miembro del par se asigna aleatoriamente a una condición experimental y el otro miembro recibe la otra condición. Con más de dos condiciones, más tiendas tendrían que ser apareadas y luego asignarse a las condiciones de tratamiento.

En algunos estudios de ingeniería de factores humanos usando simuladores, el uso de la aleatorización en ocasiones no es factible económica ni prácticamente. Considere la prueba de dos configuraciones para un simulador. Un investigador desea saber cuál conduce a menos errores perceptuales. Los procesos de aleatorización dirían que el investigador debería asignar a los participantes aleatoriamente a las condiciones, conforme entren en el estudio. Sin embargo, cuando se requieren de tres a seis meses para cambiar la configuración del simulador, ya no es factible proceder de la manera "usual".

Un punto importante a recordar es que la aleatorización —cuando puede llevarse a cabo correcta y apropiadamente— por lo común es mejor que el apareamiento. Es quizás el único método para controlar fuentes de varianza desconocidas. Uno de los principales inconvenientes del apareamiento es que no se puede estar seguro de que se ha realizado un par exacto. Sin esa precisión, la inexactitud puede ser una explicación alternativa de por qué la variable dependiente es diferente entre las condiciones de tratamiento, después de la manipulación experimental.

que se desea tener dos o más grupos apareados por inteligencia, y que se quiere utilizar el método de apareamiento de distribución de frecuencias. Primero se necesitará una puntuación de una prueba de inteligencia para cada niño. Después se requiere crear los dos o más grupos de tal manera que los grupos tengan la misma puntuación promedio en la prueba de inteligencia, así como la misma desviación estándar y una simetría o asimetría semejante en las puntuaciones. Cada grupo sería estadísticamente igual —la media, la desviación estándar y la simetría o asimetría entre cada grupo sería estadísticamente equivalente—. Podría utilizarse una prueba estadística de hipótesis; pero el investigador requiere estar consciente de que es necesario considerar los dos tipos de error. Si más de una variable se considerara relevante para aparear a los participantes, entonces se requeriría que cada grupo de participantes tuviera las mismas medidas estadísticas en todas estas variables. El número de participantes perdidos al utilizar dicha técnica no sería tan grande como el número de pérdidas al utilizar el método de individuo por individuo, ya que cada participante adicional tan sólo tendría que contribuir para producir las medidas estadísticas apropiadas, en lugar de ser idéntico a otro participante en las variables relevantes. Por lo tanto, esta técnica es más flexible en términos de capacidad para utilizar a un participante particular.

La principal desventaja de realizar el apareamiento mediante el método de distribución de frecuencias, sucede cuando se aparea con base en más de una variable. Aquí la combinación de variables puede estar mal apareada en los diversos grupos. Si se fueran a aparear edad y tiempo de reacción, un grupo podría incluir participantes mayores con tiempos de reacción más lentos, y participantes más jóvenes con tiempos de reacción más rápidos, mientras que el otro grupo podría tener la combinación opuesta. La media y distribución de las dos variables sería equivalente; pero los participantes en cada grupo serían completamente diferentes. Tal diferencia podría afectar a la variable dependiente.

Apareamiento mediante mantener constantes las variables

Otra técnica que se utiliza para crear grupos igualados de participantes consiste en mantener constante la variable extraña o no planeada. Todos los participantes en cada grupo experimental tendrán el mismo grado o tipo de variable extraña. Si se necesita controlar la variación causada por diferencias de género, se puede mantener constante el género utilizando únicamente hombres o únicamente mujeres en el estudio. Esto tiene el efecto de aparear a todos los participantes en términos de la variable género. Este procedimiento de apareamiento crea una muestra de participantes más homogénea, debido a que solamente se utilizan participantes con cierto tipo o cantidad de la variable fortuita. Muchos proyectos de investigación de estudiantes en universidades utilizan este método, especialmente cuando el universo de participantes posee una mayoría de hombres o de mujeres. Esta técnica de mantener constantes las variables tiene por lo menos dos problemas que podrían afectar la validez del estudio. La severidad del problema se incrementa si se mantienen constantes demasiadas variables. La primera desventaja consiste en que la técnica restringe el tamaño de la población de participantes. Como consecuencia, en algunos casos resulta complicado encontrar suficientes participantes para incluir en el estudio. La investigación pionera de seccionar el cerebro, realizada por Roger Sperry, con frecuencia ha sido criticada por la restricción de los participantes utilizados en el estudio. Sus primeros trabajos incluyeron solamente pacientes epilépticos. Así, un estudio que utilice dicho método podría ser criticado por tener un sesgo en la selección.

La segunda desventaja es aún más crítica, ya que los resultados sólo pueden generalizarse al tipo de participante utilizado en el estudio. Los resultados obtenidos del estudio

de los pacientes epilépticos sólo podían generalizarse a otros pacientes epilépticos. Si alguien deseara saber si pacientes no epilépticos experimentarían los mismos cambios perceptuales, el investigador tendría que conducir un estudio similar con pacientes no epilépticos. Las conclusiones de dicho estudio podrían, en realidad, ser iguales a los obtenidos en el estudio con pacientes epilépticos, pero deben realizarse dos estudios separados. La única manera de averiguar si los resultados de un estudio pueden generalizarse a la población es replicando el estudio con participantes de diferentes características.

Apareamiento mediante la incorporación de una variable extraña al diseño de investigación

Otra forma para intentar desarrollar grupos igualados es utilizar la variable extraña como variable independiente en el diseño de investigación. Suponga que se conduce un experimento de aprendizaje con ratas y que se desean controlar los efectos del peso. La idea aquí es que el animal con mayor peso necesitará ingerir más alimento después de un periodo de privación y, por lo tanto, estará más motivada. Si se hubiese utilizado el método de mantener constante el peso, se tendrían muchos menos participantes. Al utilizar el peso como variable independiente se pueden utilizar muchos más participantes en el estudio. En términos estadísticos, un aumento en el número de participantes significa un aumento en poder y sensibilidad. Con el uso de una variable extraña como variable independiente en el diseño, se aísla una fuente de varianza sistemática y también se determina si la variable extraña tiene un efecto sobre la variable dependiente.

Sin embargo, la incorporación de una variable extraña en el diseño no debe efectuarse indiscriminadamente. Lograr que la variable extraña forme parte del diseño de investigación parece ser un excelente método de control; pero dicho método está mejor utilizado cuando existe un interés en las diferencias producidas por la variable extraña, o en la interacción entre la variable extraña y otras variables independientes. El investigador puede incluso incorporar al diseño una variable medida en una escala continua. La diferencia entre una variable extraña continua y una discreta residiría en la etapa del análisis de datos del proceso de investigación. Entonces sería preferible el uso de la regresión múltiple o del análisis de covarianza, en lugar del análisis de varianza.

Los participantes como su propio control

Puesto que cada individuo es único, resulta difícil, si no imposible, encontrar a otro individuo que fuera el par perfecto. Sin embargo, una sola persona es siempre un perfecto par para sí misma. Una de las técnicas más poderosas para lograr la igualdad y la constancia de los grupos experimentales, antes de la administración del tratamiento, consiste en utilizar a esa misma persona en cada condición del experimento. Algunos se refieren a lo anterior como el uso de participantes como su propio control. Aparte de la reactividad del experimento en sí, la posibilidad de que surja variación extraña debida a diferencias entre individuos se minimiza drásticamente. Tal método para lograr la constancia es común en algunas áreas de las ciencias del comportamiento. En psicología, el estudio de la interfase de seres humanos y máquinas (factores humanos o ingeniería humana) utiliza este método. Simon (1976) presenta varios diseños experimentales interesantes que utilizan al mismo participante en muchas condiciones de tratamiento. Sin embargo, dicho método no se ajusta a todas las aplicaciones. Algunos estudios relacionados con el aprendizaje no son elegibles, pues una persona no puede *desaprender* un problema para poderle aplicar un

método diferente después. El uso de este método requiere de mayor planeación y de una ejecución más precisa que otros métodos.

Extensiones adicionales del diseño: diseño 20.3 utilizando un pretest

El diseño 20.3 tiene muchas ventajas y se utiliza con frecuencia. Su estructura es similar a la del diseño 19.2, con dos diferencias importantes: el diseño 19.2 carece de grupo control y de aleatorización. El diseño 20.3 es similar a los diseños 20.1 y 20.2, excepto en que se añadió la situación "antes" o de pretest. Con frecuencia se utiliza para estudiar cambios. Como los diseños 20.1 y 20.2, el diseño 20.3 puede expandirse a más de dos grupos.

Diseño 20.3: grupo control antes y después (pretest-postest)

(*a*)	[*A*]	Y_b	X	Y_a	(Experimental)
		Y_b	$\sim X$	Y_a	(Control)
(*b*)	[Ap_a]	Y_b	X	Y_a	(Experimental)
		Y_b	$\sim X$	Y_a	(Control)

En el diseño 20.3(*a*) los participantes son asignados aleatoriamente al grupo experimental (línea superior) y al grupo control (línea inferior), y son sometidos a una situación de pretest en una medida de Y, la variable dependiente. Entonces, el investigador puede verificar la igualdad de los dos grupos respecto a Y. La manipulación experimental X se lleva a cabo y, después, los grupos son medidos nuevamente respecto a Y. La diferencia entre los dos grupos se prueba estadísticamente. Una característica interesante y difícil de este diseño es la naturaleza de las puntuaciones que se analizan generalmente: puntuaciones de diferencia o de cambio, $Y_d - Y_a = D$. A menos que el efecto de la manipulación experimental sea fuerte, no se recomienda el análisis de las puntuaciones de diferencia. Las puntuaciones de diferencia son considerablemente menos confiables que las puntuaciones a partir de las cuales se calculan. Una explicación clara del porqué de lo anterior la ofrecen Friedenberg (1995) y Sax (1997). Aunque existen otros problemas, aquí se analizan sólo las principales fortalezas y debilidades (para un estudio más completo al respecto, véase Campbell y Stanley, 1963). Al final del análisis se examinarán las dificultades analíticas de las puntuaciones de diferencia o de cambio.

Quizás de mayor importancia, el diseño 20.3 supera la gran debilidad del diseño 19.2, ya que brinda un grupo control contra el cual puede verificarse la diferencia, $Y_d - Y_a$. Con sólo un grupo no es posible saber si la historia, la maduración (o ambas), o la manipulación experimental X produjeron el cambio en Y. Cuando se añade un grupo control, la situación se altera radicalmente. Después de todo, si se igualan los grupos (a través de la aleatorización), los efectos de la historia y de la maduración, en caso de estar presentes, deberían presentarse en ambos grupos. Si se incrementan las edades mentales de los niños del grupo experimental, también deberían hacerlo así las edades mentales de los niños del grupo control. Entonces, si aún existe una diferencia entre las medidas de Y de los dos grupos, ésta no debe ser a causa de la historia o de la maduración; es decir, si algo afecta a los participantes del grupo experimental entre el pretest y el postest, ese algo también debería afectar a los participantes del grupo control. De manera similar, el efecto de realizar una prueba —medidas reactivas de Campbell— debe controlarse porque, si la prueba afecta a los miembros del grupo experimental, también debe afectar de forma similar a los

miembros del grupo control. (Sin embargo, existe una debilidad encubierta aquí, que se estudiará más adelante.) Ésta es la principal fuerza del diseño de grupo control-grupo experimental con pretest-postest, bien planeado y bien ejecutado.

Por otro lado, los diseños de pretest-postest tienen un aspecto problemático que disminuye tanto la validez interna como la validez externa del experimento. Esta fuente de dificultad es el pretest. Un pretest puede tener un efecto sensibilizador en los participantes. Respecto a la validez interna, por ejemplo, los participantes serían alertados sobre ciertos eventos en su ambiente, que podrían no haber notado comúnmente. Si el pretest consiste en una escala de actitud, tal vez sensibilice a los participantes respecto a los aspectos o problemas mencionados en la escala; así, cuando se administre el tratamiento X al grupo experimental, los participantes de este grupo pueden responder no tanto a la influencia tentativa (la comunicación o cualquier método que se utilice para cambiar actitudes), sino a la combinación de su sensibilidad incrementada respecto al tema y a la manipulación experimental.

Puesto que dichos efectos de interacción no son inmediatamente obvios, y como representan una amenaza para la validez externa de los experimentos, vale la pena considerarlos un poco más. Se pensaría que, puesto que tanto al grupo control como al grupo experimental se les aplica el pretest, el efecto de prueba previa, si acaso hay alguno, aseguraría la validez del experimento. Suponga que no se realiza ningún pretest, es decir, que se utiliza el diseño 20.2. Si lo demás permanece igual, una diferencia entre los grupos experimental y control después de la manipulación experimental de X puede asumirse como efecto de X. No existe razón alguna para suponer que un grupo es más sensible o que está más alerta que el otro, ya que ambos enfrentaron la situación de prueba después de X. Pero cuando se utiliza un pretest, la situación cambia. Mientras que el pretest sensibiliza a ambos grupos, puede hacer que los participantes experimentales respondan a X, completa o parcialmente, debido a la sensibilidad.

También se tiene una carencia de generalización o validez externa, ya que puede ser posible generalizar a los grupos probados antes; pero no a los grupos que no fueron probados antes. En efecto, esta situación molesta al investigador, porque: ¿quién quiere generalizar a los grupos probados antes?

Si esta debilidad se vuelve importante, ¿por qué éste es un buen diseño? Mientras que el posible efecto de interacción descrito antes repercutiría en alguna investigación, es dudoso que afecte fuertemente a la mayoría de la investigación del comportamiento, si los investigadores están conscientes de su potencial y toman precauciones adecuadas. Realizar pruebas constituye una parte normal y aceptada en muchas situaciones, especialmente en educación. Por lo tanto, resulta dudoso que los participantes de la investigación estén excesivamente sensibilizados en dichas situaciones. Aun así, pueden existir ocasiones en que resulten afectados. La regla dada por Campbell y Stanley (1963) es buena: cuando se van a utilizar procedimientos de prueba inusuales, es mejor usar diseños sin pretest.

Puntuaciones de diferencia

Observe el diseño 20.3 otra vez, particularmente los cambios entre Y_a y Y_d. Uno de los problemas más difíciles que ha abrumado e intrigado a los investigadores, a los especialistas en medición y a los estadísticos es cómo estudiar y analizar dichas puntuaciones de diferencia o de cambio. En un libro con el alcance de éste, sería imposible entrar a los problemas con detalle. El lector interesado puede leer dos excelentes libros editados: el de Harris (1963) y el de Collins y Horn (1991). No obstante, se esbozarán algunos preceptos y precauciones generales. Podría pensarse que la aplicación del análisis de varianza a las puntuaciones de

diferencia producidas por el diseño 20.3 y diseños similares, sería efectiva. Dicho análisis puede realizarse si los efectos experimentales son sustanciales. Pero las puntuaciones de diferencia, como se mencionó antes, son por lo general menos confiables que las puntuaciones a partir de las cuales se calculan. Las diferencias reales entre los grupos experimental y control quizá no sean detectables simplemente por la baja confiabilidad de las puntuaciones de diferencia. Para detectar las diferencias entre los grupos control y experimental, las puntuaciones analizadas deben ser lo suficientemente confiables para reflejar las diferencias y, así, ser detectables por medio de pruebas estadísticas. Debido a tal dificultad, investigadores como Cronbach y Furby (1970) dicen que las puntuaciones de diferencia o cambio no deben utilizarse. ¿Entonces, qué se puede hacer?

El procedimiento recomendado por lo común consiste en usar las llamadas puntuaciones residualizadas o regresionadas de ganancia, las cuales se calculan al predecir las puntuaciones del postest a partir de las puntuaciones del pretest, con base en la correlación entre el pretest y el postest, y sustrayendo después estas puntuaciones predichas de las puntuaciones del postest, para obtener las puntuaciones residuales de ganancia. (El lector no debe preocuparse si este procedimiento no está demasiado claro en este momento. Más tarde, después del estudio de la regresión y del análisis de covarianza deberá quedar más claro.) El efecto de las puntuaciones del pretest se elimina de las puntuaciones del postest, es decir, las puntuaciones residuales son puntuaciones del postest purificadas respecto a la influencia del pretest. Después se prueba la significancia de la diferencia entre las medias de estas puntuaciones. Todo lo cual puede lograrse utilizando el procedimiento descrito y una ecuación de regresión, o por medio de un análisis de covarianza.

Sin embargo, incluso el uso de puntuaciones residuales de ganancia y el análisis de covarianza no son perfectos. Si los participantes no han sido asignados aleatoriamente a los grupos experimental y control, el procedimiento no salvará la situación. Cronbach y Furby (1970) señalan que cuando los grupos difieren sistemáticamente antes del tratamiento experimental, en otras características pertinentes a la variable dependiente, la manipulación estadística no corrige tales diferencias. Sin embargo, si se utiliza un pretest, es mejor usar la asignación aleatoria y el análisis de covarianza, recordando que los resultados siempre deben tratarse con especial cuidado. Por último, el análisis de regresión múltiple proporciona la mejor solución para el problema, como se verá más adelante. Por desgracia las complejidades del diseño y del análisis estadístico pueden desanimar al estudiante de investigación, incluso al punto de hacerlo sentir desamparado. No obstante, así es la naturaleza de la investigación del comportamiento: tan sólo refleja el carácter excesivamente complejo de la realidad psicológica, sociológica y educativa, lo cual resulta frustrante y emocionante al mismo tiempo; como el matrimonio, la investigación del comportamiento es difícil y con frecuencia poco exitosa, pero no imposible. Además, es una de las mejores formas de adquirir un entendimiento confiable de nuestro mundo del comportamiento. El punto de vista de este libro es que se debe aprender y comprender lo más posible sobre lo que hacemos, que se debe tener un cuidado razonable con el diseño y el análisis, y que hay que realizar la investigación sin preocuparse demasiado sobre los aspectos analíticos. La cuestión principal es siempre el problema de investigación y el interés que se tenga en él. Ello no implica una desatención o desprecio al análisis; significa simplemente un entendimiento y cuidado razonables, y cantidades sanas tanto de optimismo como de escepticismo.

Diseño 20.4: Simulación antes-después, aleatorizado

[*A*]		X	Y_a
	Y_b		

El valor del diseño 20.4 es dudoso, aunque se considera un diseño adecuado. La demanda científica de realizar una comparación está satisfecha; hay un grupo de comparación (línea inferior). Una debilidad importante del diseño 19.3 (una versión débil del diseño 20.4) se soluciona por medio de la aleatorización. Recuerde que con el diseño 19.3 era imposible suponer de antemano que los grupos experimental y control eran equivalentes. El diseño 20.4 requiere que los participantes sean asignados aleatoriamente a ambos grupos; entonces es posible suponer que estadísticamente son iguales. Un diseño de este tipo puede utilizarse cuando existe la preocupación del efecto reactivo del pretest; o cuando, a causa de las exigencias de situaciones prácticas, no se tiene otra opción. Dicha situación sucede cuando se tiene una única oportunidad de probar un método o alguna innovación. Para probar la eficacia del método, se proporciona una línea base para juzgar el efecto de X sobre Y, aplicando un pretest a un grupo similar al grupo experimental. Entonces Y_d se prueba contra Y_a.

La validez de este diseño se desploma si los dos grupos no son seleccionados aleatoriamente a partir de la misma población, o si los participantes no son asignados aleatoriamente a los dos grupos. Además, incluso si se utiliza la aleatorización, no existe una garantía real de que ésta funcionó para igualar los dos grupos antes del tratamiento. Comparte con otros diseños similares las debilidades mencionadas, es decir, otras posibles variables pueden influir en el intervalo comprendido entre Y_a y Y_d. En otras palabras, el diseño 20.4 es superior al diseño 19.3; pero no debe utilizarse si está disponible otro diseño que se considere mejor.

Diseño 20.5: Tres grupos, antes-después

	Y_a	X	Y_d	(Experimental)
[A]	Y_a	$\sim X$	Y_d	(Control 1)
		X	Y_d	(Control 2)

El diseño 20.5 es mejor que el diseño 20.4. Además de las ventajas del diseño 20.3, proporciona una forma posible de evitar la confusión debida a los efectos interactivos del pretest. Esto se logra por medio de un segundo grupo control (tercera línea). (Parece un poco extraño tener un grupo control con X; pero el grupo de la tercera línea es realmente un grupo control.) El hecho de contar con las medidas de Y_d de este grupo, hace posible verificar el efecto de interacción. Suponga que la media del grupo experimental es significativamente mayor que la media del grupo control 1. Podría dudarse si tal diferencia se debe realmente a X; quizá se produjo por un incremento en la sensibilidad de los participantes después del pretest y por la interacción de su sensibilidad y X. Ahora observe la media de Y_d del grupo control 2. Ésta también debería ser significativamente mayor que la media del grupo control 1. Si es así, se puede suponer que el pretest no sensibilizó excesivamente a los participantes, o que X es lo suficientemente fuerte para anular un efecto de interacción entre la sensibilización y X.

Diseño 20.6: Cuatro grupos, antes-después (Solomon)

	Y_a	X	Y_d	(Experimental)
[A]	Y_a	$\sim X$	Y_d	(Control 1)
		X	Y_d	(Control 2)
		$\sim X$	Y_d	(Control 3)

El diseño propuesto por Solomon (1949) es fuerte y estéticamente satisfactorio. Posee controles potentes. En realidad, si se cambia la designación de control 2 por experimental 2, se tiene una combinación de los diseños 20.3 y 20.1, los dos mejores diseños, donde el primer diseño forma las primeras dos líneas del diseño de Solomon; y el último, las segundas dos líneas. Las virtudes de ambos se combinan en un diseño. Aunque este diseño puede tener una forma apareada, eso no se analiza aquí ni se recomienda su uso. Campbell (1957) afirma que tal diseño se ha convertido en un nuevo ideal para los científicos sociales. Aunque parece una afirmación demasiado fuerte, indica la alta estima en que se tiene al diseño.

Una de las razones de la fuerza del diseño es que la demanda de comparación queda bien satisfecha con las dos primeras líneas y con las segundas dos líneas. La aleatorización incrementa la probabilidad de la equivalencia estadística de los grupos; y la historia y la maduración están controladas por las primeras dos líneas del diseño. El efecto de interacción debido a la posible sensibilización por el pretest en los participantes está controlado por las primeras tres líneas. Al añadir la cuarta línea, se controlan los efectos temporales contemporáneos que pueden haber ocurrido entre Y_d y Y_a. Ya que los diseños 20.2 y 20.3 están combinados, se tiene la fuerza de cada uno por separado y la fuerza de la replicación porque en efecto, hay dos experimentos. Si Y_d del grupo experimental es significativamente mayor que la del grupo control 1, y la del grupo control 2 es significativamente mayor que la del grupo control 3, aunado a la consistencia de los resultados de ambos experimentos, entonces ésta es una fuerte evidencia de la validez de la hipótesis de investigación.

¿Qué defecto puede tener este ideal de diseño? Ciertamente luce bien en el papel. Parece haber solamente dos fuentes de debilidad. Una es de tipo práctico ya que es más difícil realizar dos experimentos simultáneamente, que uno; y el investigador se encuentra con la dificultad de localizar más participantes del mismo tipo.

La otra dificultad es estadística. Observe que hay una falta de balance entre los grupos. Existen cuatro grupos, pero no hay cuatro conjuntos completos de medidas. Utilizando las primeras dos líneas, es decir, con el diseño 20.3, se puede sustraer Y_a de Y_d o hacer un análisis de covarianza. Con las dos líneas se pueden probar las Y_d contra sí mismas con una prueba t o una prueba F; pero el problema reside en cómo obtener un enfoque estadístico general. Una solución es probar las Y_d de los grupos control 2 y 3 contra el promedio de las dos Y_a (las primeras dos líneas), así como también probar la significancia de la diferencia de las Y_d de las primeras dos líneas. Además, Solomon originalmente sugirió un análisis factorial de varianza de 2×2, utilizando los cuatro conjuntos de medidas de Y_d. La sugerencia de Solomon se presenta en la figura 20.5. Un estudio cuidadoso revelará que éste es un buen ejemplo de pensamiento de investigación, una excelente combinación de diseño y análisis. Con este análisis se pueden estudiar los efectos principales, X y $\sim X$, y aquellos con pretest y sin pretest. Lo que es más interesante, se puede probar la interacción de la prueba previa y X, y obtener una respuesta clara al problema anterior.

Mientras que éste y otros diseños complejos tienen fuerzas notorias, es dudoso que puedan utilizarse rutinariamente. De hecho, quizá deban reservarse para experimentos muy importantes, en los cuales, se prueben de nuevo con mayor rigor y control, hipótesis ya probadas con diseños más simples. En efecto, se recomienda que diseños como el 20.5 y el 20.6 y ciertas variantes del diseño 20.6 (que se discutirá más adelante) se reserven para

FIGURA 20.5

	X	$\sim X$
Con pretest	X_d, experimental 1	Y_d, control 1
Sin pretest	Y_d, control 2	Y_d, control 3

pruebas definitivas de hipótesis de investigación, después de que se haya realizado cierta cantidad de experimentación previa.

RESUMEN DEL CAPÍTULO

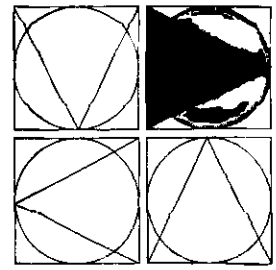
1. El diseño de un estudio es el proyecto o plan para desarrollar la investigación.
2. Un diseño es un subconjunto de un producto cartesiano cruzado de varios niveles de la variable independiente.
3. El diseño experimental es aquel donde se manipula al menos una de las variables independientes utilizadas en el estudio.
4. Los diseños no experimentales son aquellos en los cuales no hay aleatorización para igualar los grupos antes de administrar los tratamientos.
5. Generalmente el método estadístico más apropiado para los diseños experimentales es el análisis de varianza.
6. Los supuestos del análisis de varianza con frecuencia se violan en los diseños no experimentales. La regresión múltiple puede ser un método de análisis de datos más apropiado para diseños no experimentales.
7. El diseño de grupo experimental-grupo control con participantes aleatorizados (diseño 20.1) es el mejor para muchos estudios de investigación experimentales del comportamiento.
8. El diseño de cuatro grupos de Solomon (diseño 20.6) maneja varias de las preocupaciones de la investigación del comportamiento. Sin embargo, utiliza los recursos de dos estudios y quizá no resulte eficiente económicamente.
9. El diseño 20.2 es como el diseño 20.1, excepto que utiliza participantes apareados.
10. El uso de participantes apareados se vuelve útil en situaciones donde la aleatorización no funciona apropiadamente.
11. Existen varias formas de aparear participantes. La más popular es el método de individuo por individuo.
12. El apareamiento tiene problemas en cuanto a que el investigador nunca puede estar seguro de que todas las variables importantes hayan sido utilizadas en el proceso. Además, si se utilizan demasiadas variables en el apareamiento, se vuelve más difícil encontrar participantes con características iguales.
13. El diseño 20.3 utiliza un pretest; cuya aplicación es una forma para determinar si los grupos son iguales o si funcionó la aleatorización. Sin embargo, aplicar un pretest también sensibiliza a los participantes del experimento.
14. Las puntuaciones de diferencia se utilizan con frecuencia en diseños que incluyen un pretest. Sin embargo, las puntuaciones de diferencia pueden no ser confiables.
15. El diseño 20.4 es un diseño de antes-después simulado que utiliza participantes aleatorizados. El segundo grupo (control) tan sólo se mide en un pretest. El grupo experimental recibe el tratamiento y el postest.
16. El diseño 20.5 es un diseño de tres grupos antes-después. Es como el diseño 20.3, excepto por la introducción de un tercer grupo que recibe tratamiento y porque no se utiliza un pretest.

SUGERENCIAS DE ESTUDIO

1. La primera oración de este capítulo fue: "El diseño constituye una disciplina de datos." ¿Qué significa dicha frase? Justifique su respuesta.

2. Suponga que usted es un psicólogo educativo y planea probar la hipótesis de que retroalimentar con información psicológica a los maestros, incrementa el aprendizaje de los niños, al aumentar el entendimiento del maestro sobre los niños. Bosqueje un diseño de investigación ideal para probar esta hipótesis, suponiendo que usted tiene un completo dominio de la situación y suficiente dinero y asistencia. (Éstas son condiciones importantes, que se incluyen para liberar al lector de las complicaciones prácticas que tan frecuentemente comprometen los buenos diseños de investigación.) Establezca dos diseños, cada uno con aleatorización completa; que ambos diseños sigan el paradigma del diseño 20.1. En uno de ellos utilice sólo una variable independiente y un análisis de varianza de un factor. En el segundo, emplee dos variables independientes y un diseño factorial simple. ¿En qué difieren estos dos diseños respecto a sus poderes de control y en la información que producen? ¿Cuál prueba mejor la hipótesis? Explique por qué.
3. La recomendación del texto de no utilizar el análisis de varianza en investigación no experimental no aplica en gran medida al análisis de varianza de un factor ni al análisis factorial. Tampoco aplica el problema del mismo número de casos en las casillas. De hecho, en varios estudios no experimentales, el análisis de varianza de un factor se ha utilizado provechosamente. Uno de dichos estudios es el de Jones y Cook (1975). La variable independiente era la actitud hacia los afroamericanos, obviamente no manipulada. La variable dependiente era el grado de apoyo hacia una política social que afectaba a los afroamericanos.

   Se sugiere que los lectores revisen y asimilen este estudio. Usted quizás también desee realizar un análisis de varianza con los datos de la tabla 1 de los autores, utilizando el método del análisis de varianza con el uso de n, medias y desviaciones estándar, descrito anteriormente (véase anexo, capítulo 13).



CAPÍTULO 21

APLICACIONES DEL DISEÑO DE INVESTIGACIÓN: GRUPOS ALEATORIZADOS Y GRUPOS CORRELACIONADOS

Resulta difícil explicarle a alguien cómo hacer investigación. Quizás lo mejor sea asegurarse de que el principiante capte los principios y las posibilidades. Además los enfoques y las

tácticas pueden sugerirse. Para abordar un problema de investigación, el investigador debe dejar volar la mente, especular acerca de las posibilidades e incluso adivinar el patrón de los resultados. Una vez que se conocen las posibilidades, las intuiciones pueden seguirse y explorarse. Sin embargo, la intuición y la imaginación no son de mucha ayuda si se sabe poco o nada acerca de recursos técnicos. Por otro lado, la buena investigación no consiste sólo de metodología y técnica; el pensamiento intuitivo es esencial porque ayuda a los investigadores a alcanzar soluciones que no son convencionales o rutinarias. No obstante, nunca debe olvidarse que el pensamiento analítico y el pensamiento intuitivo creativo dependen del conocimiento, el entendimiento y la experiencia.

Los principales propósitos de este capítulo consisten en enriquecer e ilustrar el diseño y la discusión estadística con ejemplos reales de investigación, así como sugerir posibilidades básicas para diseñar investigación para que el estudiante finalmente resuelva problemas de investigación. El propósito general es, entonces, complementar y enriquecer análisis estadísticos y de diseño previos y más abstractos.

Diseño simple de sujetos aleatorizados

En los capítulos 13 y 14 se estudiaron e ilustraron los estadísticos del análisis de varianza simple de un factor y del análisis factorial de varianza. El diseño detrás de la discusión anterior se llama *diseño de sujetos aleatorizados*. El paradigma general del diseño (denominado como diseño 20.1) se presenta a continuación:

[*A*]	X	Y	(Experimental)
	$\sim X$	Y	(Control)

Ejemplo de investigación

La forma más simple del diseño 20.1 es un paradigma del análisis de varianza de un factor, en el cual a k grupos se les dan k tratamientos experimentales, y las k medias se comparan mediante un análisis de varianza o con pruebas separadas de significancia. Un vistazo a la parte izquierda de la figura 20.3 presenta esta forma simple del diseño 20.1 con $k = 3$. Aunque parezca extraño, no se utiliza muy a menudo; los investigadores prefieren la forma factorial del diseño 20.1 con mayor frecuencia. Más adelante se presenta un ejemplo de un factor; se utiliza la asignación aleatoria. Por desgracia algunos investigadores no reportan cómo se asignaron los participantes a los grupos o a los tratamientos. La necesidad de reportar el método de la selección de los participantes y la asignación a los grupos experimentales debe ser obvia en este momento.

Dolinski y Nawrat: miedo-luego-alivio y sumisión

Los estudios sobre sumisión han representado gran interés para los psicólogos sociales. En el capítulo 17, donde se analizó la ética al realizar investigación en ciencias del comportamiento, se mencionó la influencia del estudio de Milgrim sobre la forma en que ahora se lleva a cabo la investigación. Milgrim, si se recuerda, se interesó en saber por qué durante la Segunda Guerra Mundial los nazis obedecían órdenes y cometían actos inenarrables de brutalidad hacia otros seres humanos. En un estudio de Dolinski y Nawrat (1998), se exploró otro método para inducir a la sumisión. Éste fue un método utilizado por los nazis y los estalinistas para obligar a los prisioneros polacos a testificar en contra de

sí mismos, sus amigos o sus familias. Dolinski y Nawrat llamaron a este método "miedo-luego-alivio", el cual implica poner a un prisionero en un estado de gran ansiedad por medio de gritos y amenazas de los carceleros hacia él. Después de alcanzar el nivel de miedo deseado, los estímulos generadores de ansiedad se eliminan abruptamente; entonces, se trata amablemente al prisionero. El resultado común de este procedimiento es la intensificación de la conducta de sumisión. Dolinski y Nawrat afirman que la sumisión se debe a la reducción del miedo y no al miedo en sí. Aunque Dolinski y Nawrat utilizan un ejemplo muy extremo para ilustrar su idea, ellos también explican que el método con frecuencia se utiliza de alguna manera y forma por diadas en la vida diaria. Puede ocurrir entre padre e hijo, maestro y estudiante, y entre empleado y patrono. La policía usa tácticas similares con su rutina del "policía bueno-policía malo", que por lo común incluye a un oficial de policía ("policía malo") que regaña, grita y amenaza a un sospechoso. Cuando el sospechoso alcanza un alto nivel de ansiedad, otro oficial de policía ("policía bueno") quita al "policía malo" y le habla amable y dulcemente al prisionero. Los terroristas también utilizan este método con los rehenes.

Dolinski y Nawrat diseñaron y realizaron cuatro experimentos para probar la eficacia del método de "miedo-luego-alivio" para inducir la sumisión. Aquí se describirá uno de estos experimentos, en el cual 120 estudiantes de preparatoria voluntarios de Opole, Polonia, fueron asignados aleatoriamente a una de tres condiciones experimentales. A todos los participantes se les avisó que participarían en un estudio sobre los efectos del castigo en el aprendizaje. El grupo 1 experimentó ansiedad; se les indicó que les sería aplicado un choque eléctrico leve, no doloroso, por cada error que cometieran. Los participantes del grupo 2 experimentaron ansiedad que después fue reducida. Al inicio se les dio la misma explicación que al grupo 1, pero después se les dijo que participarían en un estudio diferente, el cual implicaba coordinación visomotora y no incluía choques eléctricos. El grupo 3 era la condición control. A estos participantes se les indicó que iban a participar en un estudio sobre coordinación visomotora. Durante el periodo de espera, antes del inicio del experimento, se le pidió a cada participante contestar un cuestionario sobre ansiedad. Después de completar el cuestionario, una estudiante cómplice del experimentador, pero que parecía estar totalmente desligada del experimento, se presentó y le pidió a cada participante unirse a una acción de caridad para un orfanatorio. A quienes obedecieron o aceptaron se les preguntó cuántas horas estaban dispuestos a trabajar para dicha causa.

La variable independiente manipulada en este estudio fue el nivel de ansiedad inducida y de alivio. Las variables dependientes fueron la sumisión, la cantidad de ansiedad y el número de horas donadas para una buena causa. Con el uso de un análisis de varianza de un factor, Dolinski y Nawrat obtuvieron un valor *F* significativo. El grupo 2, el cual experimentó ansiedad que se redujo después, tuvo la tasa más alta de sumisión y el mayor

TABLA 21.1 *Niveles de ansiedad, sumisión, número de días dispuestos para ser voluntario por ansiedad inducida y valores F (estudio de Dolinski y Nawrat)*

Condición del grupo	Ansiedad media reportada	Porcentaje de sumisión	Número medio de días como voluntario
Estudio de choque eléctrico	53.25	37.5	0.625
Estudio de choque eléctrico cambiado por estudio de coordinación visomotora	43.05	75.0	1.150
Estudio de coordinación visomotora	34.45	52.5	1.025
Valor *F*	108.9 ($p < .00001$)	6.13 ($p < .003$)	2.11 ($p > .05$)

número de días dispuestos para la caridad. El nivel de ansiedad de cada grupo resultó en la dirección esperada. El grupo 1 experimentó el mayor nivel de ansiedad, seguido por el grupo 2 y luego el grupo 3. La tabla 21.1 presenta el resumen de los datos del estudio.

Los resultados del estudio apoyaron la hipótesis de Dolinski y Nawrat de que fue el "miedo-luego-alivio", y no la emoción del miedo en sí, lo que llevó a un nivel de sumisión mayor. Crear tan sólo un estado de ansiedad en las personas no es suficiente para inducir sumisión. De hecho, tal estudio encontró que los participantes del grupo 1 (ansiedad inducida), quienes sintieron la mayor cantidad de angustia, se sometieron en menor medida que los participantes del grupo 3 (control-ansiedad baja o sin ansiedad).

Diseños factoriales

El diseño básico general continúa siendo el diseño 20.1, aunque la variación del patrón básico de grupo experimental-grupo control se altera drásticamente al agregar otros factores experimentales o variables independientes. Siguiendo una definición previa del análisis factorial de varianza, *el diseño factorial es la estructura de la investigación, en la cual se yuxtaponen dos o más variables independientes para estudiar sus efectos independientes e interactivos sobre una variable dependiente.*

Al inicio, el lector puede encontrar un poco difícil ajustar la estructura factorial dentro del paradigma general de grupo experimental-grupo control del diseño 20.1. Sin embargo, el análisis de la generalización de la noción de grupo control en el capítulo 20 quizás aclaró las relaciones entre el diseño 20.1 y los diseños factoriales. La discusión continúa ahora. Se tienen las variables independientes A y B, y la variable dependiente Y. El diseño factorial más simple, el de 2×2, tiene tres posibilidades: tanto A como B son variables activas; A es activa, B es atributiva (o a la inversa); y tanto A como B son variables atributo. (La última posibilidad, ambas variables independientes son atributo, es el caso no experimental. No obstante, como se indicó anteriormente, tal vez no sea recomendable utilizar el análisis de varianza con variables independientes no experimentales.) Como siempre, A puede dividirse en A_1 y A_2, condición experimental y control, con la variable independiente adicional B dividida en B_1 y B_2. Puesto que esta estructura ahora resulta familiar, sólo es necesario discutir uno o dos detalles del procedimiento.

El procedimiento ideal de asignación de participantes consiste en asignarlos aleatoriamente a las cuatro casillas. Si tanto A como B son variables activas, esto se vuelve factible y fácil; tan sólo se les dan a los participantes números arbitrarios de 1 a N (donde N es el número total de participantes). Después, con una tabla de números aleatorios, se anotan números de 1 a N conforme aparecen en la tabla, los cuales se colocan en cuatro grupos conforme aparecen y después se asignan los cuatro grupos de participantes a las cuatro casillas. Para estar seguros, los grupos de participantes también se asignan aleatoriamente a los tratamientos experimentales (las cuatro casillas). Los grupos se denominan como 1, 2, 3 y 4, y después se extraen estos números de una tabla de números aleatorios. Suponga que la tabla produce los números en este orden: 3, 4, 1 y 2; entonces se asigna a los participantes del grupo 3 a la casilla superior izquierda; a los participantes del grupo 4, a la casilla superior derecha, etcétera.

Con frecuencia B será una variable atributo como género, inteligencia, rendimiento, ansiedad, autopercepción o raza. La asignación de los participantes debe verse alterada. Primero, puesto que B es una variable atributo, no existe posibilidad de asignar a los participantes a B_1 y B_2 de manera aleatoria. Si B es la variable género, lo mejor que se puede hacer es asignar aleatoriamente primero a los hombres a las casillas A_1B_1 y A_2B_1, y luego a las mujeres a las casillas A_1B_2 y A_2B_2.

Diseños factoriales con más de dos variables

Con frecuencia es posible mejorar el diseño y aumentar la información obtenida de un estudio añadiendo grupos. En lugar de A_1 y A_2 y de B_1 y B_2, un experimento podría beneficiarse al utilizar A_1, A_2, A_3 y A_4; y B_1, B_2 y B_3. Los problemas prácticos y estadísticos se incrementan y algunas ocasiones se tornan bastante difíciles conforme se agregan variables. Suponga que se tiene un diseño de $3 \times 2 \times 2$ que tiene $3 \times 2 \times 2 = 12$ celdillas, cada una de las cuales debe tener por lo menos dos participantes, y de preferencia muchos más. (Es posible, pero no muy sensible, incluir solamente un participante por casilla si es posible tener más. Por supuesto que existen diseños con sólo un participante por casilla. Esto se estudia en el capítulo 22.) Si se decide que son necesarios 10 participantes por casilla, entonces deberán obtenerse y asignarse aleatoriamente $12 \times 10 = 120$ participantes. El problema se vuelve más complicado con una variable más, y también resulta más difícil la manipulación práctica de la situación de investigación. Sin embargo, la manipulación exitosa de dicho experimento permite probar varias hipótesis y brinda una gran cantidad de información. Las combinaciones de diseños de tres, cuatro y cinco variables ofrecen una amplia variedad de diseños posibles: $2 \times 5 \times 3$, $4 \times 4 \times 2$, $3 \times 2 \times 4 \times 2$, $4 \times 3 \times 2 \times 2$, etcétera.

Ejemplos de investigación de diseños factoriales

En el capítulo 14 se describieron ejemplos de diseños factoriales de dos y tres dimensiones. (Se recomienda un repaso de estos ejemplos, pues el razonamiento que subyace al diseño esencial ahora puede captarse con mayor facilidad.) Como en el capítulo 14 se presentó un número suficiente de ejemplos de diseños factoriales, los ejemplos mostrados aquí se dedican a estudios con características poco usuales o resultados muy interesantes.

Sigall y Ostrove: atractivo y crimen

A menudo se afirma que a las mujeres atractivas se les trata de forma diferente que a los hombres o que a las mujeres poco atractivas. En la mayoría de los casos, quizá, las reacciones son "favorables": las mujeres atractivas tal vez tienen una mayor probabilidad, que las mujeres no atractivas, de recibir la atención y los favores del mundo. No obstante, ¿será posible que su atractivo sea desventajoso en algunas situaciones? Sigall y Ostrove (1975) plantearon la pregunta: ¿cómo se relaciona el atractivo físico de un criminal acusado con las sentencias judiciales, y cómo la naturaleza del crimen interactúa con el atractivo? Ellos pidieron a sus participantes que asignaran sentencias, en años, a delitos de estafa y robo de acusadas atractivas, no atractivas y controles. En la tabla 21.2 se presenta el paradigma factorial del experimento, junto con los resultados. (Se evitó la descripción de muchos detalles experimentales; éstos fueron bien manejados.)

TABLA 21.2 *Sentencias medias en años de acusadas atractivas, no atractivas y control, por estafa y robo (estudio de Sigall y Ostrove)*[a]

	Condición de la acusada		
	Atractiva	**No atractiva**	**Control**
Estafa	5.45	4.35	4.35
Robo	2.80	5.20	5.10

[a] $N = 120$, 20 por casilla; F (interacción) = 4.55 ($p < .025$).

En el caso de robo, la acusada robó $2 200 en un rascacielos. En la situación de estafa, la acusada se congració con un soltero de mediana edad y lo estafó con $2 200. Observe que las condiciones de no atractiva y control no difirieron mucho entre sí. Tanto la situación *atractiva-estafa* (5.45) como la situación *atractiva-robo* (2.80) difirieron de las otras dos condiciones, ¡pero en direcciones opuestas! La situación *atractiva-estafa* recibió la mayor sentencia media: 5.45 años; mientras que la situación *atractiva-robo* recibió la menor sentencia media: 2.80 años. Los estadísticos apoyan el resumen verbal previo —la interacción fue estadísticamente significativa: la F de *atractiva-delito*, con 2 y 106 grados de libertad, fue de 4.55, $p < .025$; es decir, las acusadas atractivas tienen una ventaja sobre las acusadas no atractivas, excepto cuando sus crímenes están relacionados con su atractivo (estafa)—.

Quilici y Mayer: ejemplos, esquema y aprendizaje

¿Ayudan los ejemplos a que los estudiantes aprendan estadística? Ésta fue la pregunta básica planteada por los científicos cognitivos Quilici y Mayer (1966). En su estudio sobre la solución analítica de problemas, Quilici y Mayer examinaron sólo uno de tres procesos que definen el pensamiento analógico. Ellos se interesaron tan sólo en el proceso de reconocimiento que implica dos técnicas: 1) enfoque en las similitudes superficiales entre el ejemplo y el problema real que debe resolverse, o 2) enfoque en las similitudes estructurales.

Las similitudes superficiales tratan con los atributos compartidos de objetos en la historia del problema. En la similitud estructural lo importante son las relaciones compartidas entre objetos, tanto en el ejemplo como en el problema. Para estudiar dicho fenómeno, Quilici y Mayer utilizaron el aprendizaje de la resolución de problemas escritos en estadística. Tuvieron la impresión de que los estudiantes que aprenden la estructura de problemas estadísticos escritos serían más capaces de resolver otros problemas que enfrentaran en el futuro, al clasificarlos apropiadamente en el método de análisis estadístico correcto (por ejemplo, prueba t, correlación, etcétera). A continuación se presentan cuatro ejemplos para ilustrar las diferencias entre las similitudes superficiales y estructurales.

Ejemplo 1

Un experto en personal desea determinar si los mecanógrafos con experiencia son capaces de teclear más rápido que los mecanógrafos sin experiencia. A 20 mecanógrafos con experiencia y a 20 sin experiencia se les aplica una prueba de mecanografía. Se registra el número promedio de palabras tecleadas por minuto de cada uno de ellos.

Ejemplo 2

Un experto en personal desea determinar si la experiencia en mecanografía se relaciona con velocidades más rápidas al teclear. Se les pide a 40 mecanógrafos que informen cuántos años han trabajado como tales y se les aplica una prueba de mecanografía para determinar su número promedio de palabras tecleadas por minuto.

Ejemplo 3

Después de revisar los datos sobre el clima de los últimos 50 años, una meteoróloga afirma que la precipitación anual varía con la temperatura promedio. Para cada uno de los 50 años ella verifica la caída de lluvia anual y la temperatura promedio.

Ejemplo 4

Un decano universitario afirma que los lectores eficientes obtienen mejores calificaciones que los lectores ineficientes. Se registran las calificaciones promedio de 50

estudiantes de primer año, quienes obtuvieron una alta puntuación en una prueba de lectura de comprensión; y de 50 estudiantes de primer año que obtuvieron una baja puntuación en la misma prueba.

Si se examinan estos cuatro problemas tomados de Quilici y Mayer (1996, p. 146), el ejemplo 1 y el ejemplo 2 tendrían las mismas características superficiales; ambos tratan de mecanógrafos y de tecleo. Para resolver el ejemplo 1 se utilizaría una prueba *t*, para comparar a los mecanógrafos con experiencia con aquellos sin experiencia. No obstante, para resolver el ejemplo 2 se utilizaría una correlación, ya que la cuestión requiere de una relación entre la experiencia en mecanografía y el número promedio de palabras tecleadas por minuto. Por lo tanto, los ejemplos 1 y 2 serían estructuralmente diferentes. El ejemplo 3 también analiza la relación entre dos variables: cantidad de lluvia y temperatura. Este ejemplo tendría la misma estructura del ejemplo 2, pero una superficie diferente. Poseen la misma estructura, ya que ambos requieren del uso de una correlación para resolver el problema. El ejemplo 4 y el ejemplo 1 tienen la misma estructura; pero una superficie diferente.

Quilici y Mayer diseñaron un estudio para determinar si la experiencia con ejemplos fomentaba la construcción de esquemas estructurales. Supusieron que la exposición a problemas estadísticos escritos haría a los estudiantes más sensibles a las características estructurales que a las superficiales, en futuros problemas escritos. Los estudiantes que no son expuestos a ejemplos estadísticos escritos no exhibirían dicha conducta. Ellos también hipotetizaron que aquellos a quienes se expusiera a tres ejemplos serían capaces de exhibir la conducta en un mayor grado que aquellos que fueran expuestos a sólo un ejemplo. Estos investigadores utilizaron un diseño factorial de 3 × 2. La primera variable independiente fueron las características estructurales (prueba *t*, chi cuadrada y correlación). La segunda variable independiente fueron las características superficiales (mecanografía, clima, fatiga mental y lectura). Hubo dos variables dependientes: una puntuación de uso estructural y una puntuación de uso superficial. Los participantes fueron asignados aleatoriamente a las condiciones de tratamiento. Un análisis de varianza de dos factores confirmó su hipótesis de que aquellos expuestos a ejemplos utilizarían un esquema de base estructural; mientras que quienes no fueron expuestos a ejemplos no lo harían. Sin embargo, no hubo una diferencia estadística entre aquellos expuestos a tres ejemplos y quienes recibieron un ejemplo.

Hoyt: Conocimiento del maestro y rendimiento del alumno

Ahora se describe un estudio educativo realizado hace muchos años, que se planeó para responder una importante pregunta teórica y práctica, que ilustra claramente un diseño factorial complejo. La pregunta de investigación fue: ¿cuáles son los efectos sobre el rendimiento y las actitudes de los alumnos, si a los maestros se les informa respecto a las características de sus alumnos? El estudio de Hoyt (1955) exploró diversos aspectos de la pregunta básica y utilizó un diseño factorial para incrementar la validez tanto interna como externa de la investigación. El primer diseño se utilizó tres veces para cada uno de los tres participantes escolares; y el segundo y el tercero se usaron dos veces, una vez en cada uno de los dos sistemas escolares.

El paradigma del primer diseño se presenta en la figura 21.1. Las variables independientes fueron los tratamientos, la habilidad, el sexo y las escuelas. Los tres tratamientos autoexplicativos fueron: sin información (*N*), puntuaciones de la prueba (*P*) y puntuaciones de la prueba más otra información (*PO*). Los niveles de habilidad eran CI *alto*, *medio* y *bajo*. Las variables *género* y *escuelas* resultan obvias. Los estudiantes de octavo grado fueron asignados aleatoriamente en cuanto al género y niveles de habilidad. Ayudará a compren-

FIGURA 21.1

		N		*P*		*PO*	
		Hombre	Mujer	Hombre	Mujer	Hombre	Mujer
Escuela *A*	CI alto						
	CI medio						
	CI bajo						
Escuela *B*	CI alto						
	CI medio						
	CI bajo						

Medidas de la variable dependiente

der el diseño si se examina la forma de una tabla final de análisis de varianza del diseño. Sin embargo, antes de hacerlo debe notarse que los resultados de *rendimiento* en su mayoría fueron indeterminantes (o negativos). Las razones *F*, con una excepción, no fueron significativas. Las actitudes de los alumnos hacia los profesores, por otro lado, parecieron mejorar cuando los maestros incrementaron su conocimiento sobre los alumnos: un hallazgo interesante y potencialmente importante. La tabla del análisis de varianza se presenta en la tabla 21.3. ¡Un experimento que produce 14 pruebas! En efecto, algunas de estas pruebas no son importantes y pueden ignorarse. Las pruebas de mayor importancia (marcadas con asteriscos en la tabla) son aquellas que incluyen la variable tratamiento. La prueba más importante es entre tratamientos, el primero de los efectos principales. Quizá

TABLA 21.3 *Fuentes de varianza y grados de libertad para un diseño factorial de 3 × 3 × 2 × 2, con las variables tratamientos, habilidad, sexo y escuela (se omitieron los grados de libertad totales y dentro)*

Fuente	*gl*
Efectos principales	
Entre tratamientos*	2
Entre niveles de habilidad	2
Entre géneros	1
Entre escuelas	1
Interacciones de primer orden	
Interacción: tratamientos × habilidad	4
Interacción: tratamientos × género*	2
Interacción: tratamientos × escuela*	2
Interacción: habilidad × género	2
Interacción: habilidad × escuela	2
Interacción: género × escuela	1
Interacciones de segundo orden	
Interacción: tratamientos × habilidad × género*	4
Interacción: tratamientos × habilidad × escuela	4
Interacción: habilidad × género × escuela	2
Interacciones de tercer orden	
Interacción: tratamientos × habilidad × género × escuela	4

de igual importancia sean las interacciones que incluyen tratamientos. Tome la interacción tratamiento × sexo; si resulta significativa, entonces quiere decir que la cantidad de información que un maestro posee sobre los estudiantes ejerce una influencia en el rendimiento de estos últimos; pero los niños se ven influenciados de manera diferente que las niñas. A los niños con maestros que tienen información sobre sus alumnos les puede ir mejor que a los niños cuyos maestros no poseen dicha información; mientras que puede suceder lo contrario con las niñas, o puede no resultar ninguna diferencia en uno u otro sentido.

Las interacciones de segundo orden o triples son más difíciles de interpretar. Parece que rara vez son significativas; sin embargo, requieren un estudio especial. Las tablas de tabulación cruzadas de las medias quizás sean la mejor opción; aunque los métodos gráficos, como se analizó previamente, a menudo son ilustrativos. El lector encontrará una guía en el libro de Edward (1984) y en el manuscrito de Simon (1976).

Evaluación de los diseños de sujetos aleatorizados

Todos los diseños de sujetos aleatorizados son variantes o extensiones del diseño 20.1, el diseño básico de grupo experimental-grupo control, en el cual los participantes son asignados de forma aleatoria a los grupos experimental y control. De esta manera, incluyen las fortalezas del diseño básico, de las cuales la más importante es la característica de aleatorización y la consecuente habilidad para suponer la igualdad aproximada preexperimental de los grupos experimentales, en todas las variables independientes posibles. Se controlan la historia y la maduración ya que pasa muy poco tiempo entre la manipulación de X, y la observación y la medida de Y. No existe la posibilidad de contaminación debida al pretest.

Las otras fortalezas de estos diseños, que surgen de las múltiples variaciones posibles, son la flexibilidad y la aplicabilidad, las cuales sirven para ayudar a resolver muchos problemas en la investigación del comportamiento, puesto que parecen ajustarse particularmente bien a los tipos de problemas del diseño que surgen de problemas e hipótesis científicos, tanto sociales como educativos. El diseño de un factor, por ejemplo, incorpora cualquier número de métodos y la comprobación de métodos es una necesidad educativa importante. Las variables que constantemente necesitan control en la investigación del comportamiento —género, inteligencia, aptitud, clase social, escuelas y muchas otras— pueden incorporarse a los diseños factoriales, y así se controlan. También, con los diseños factoriales es posible realizar mezclas de variables activas y atributo —otra importante necesidad—. Sin embargo, también existen debilidades.

Una crítica consiste en que los diseños de sujetos aleatorizados no permiten pruebas de la igualdad de los grupos, como lo hacen los diseños antes-después (pretest-postest). En realidad ésta no es una crítica válida, por dos razones: 1) como se ha visto, con suficientes participantes y aleatorización se supone que los grupos son iguales; y 2) es posible verificar la igualdad de los grupos en variables diferentes de Y, la variable dependiente. Para la investigación educativa, en los expedientes escolares existe información sobre inteligencia, aptitud y rendimiento, por ejemplo. Datos pertinentes para investigación en sociología y en ciencias políticas a menudo están disponibles en expedientes de condados y de distritos electorales.

Otra debilidad es de tipo estadístico. Debe tenerse igual número de casos en las casillas de los diseños factoriales. Es posible trabajar con n desiguales; pero es insensato y representa una amenaza para la interpretación. La eliminación aleatoria de casos o el uso de métodos de casos faltantes contrarrestan pequeñas discrepancias (véase Dear, 1959;

Gleason y Staelin, 1975, dos excelentes referencias sobre la estimación de datos faltantes). Lo anterior impone una limitación sobre el uso de dichos diseños, ya que no siempre es posible tener números iguales en cada casilla. Los diseños aleatorizados de un factor no son tan delicados: los números desiguales no representan un problema difícil. Cómo ajustar y analizar datos con n desiguales constituye un problema complejo, polémico y muy discutido. Para revisar una discusión en contexto, principalmente del análisis de varianza, se recomienda consultar Snedecor y Cochran (1989). Respecto a la discusión en el contexto de la regresión múltiple, la cual representa una mejor solución del problema, véase Kerlinger y Pedhazur (1973) y Pedhazur (1996). Las discusiones de Pedhazur son detalladas y con autoridad; revisa los temas y sugiere soluciones.

En comparación con los diseños de grupos apareados, los diseños de sujetos aleatorizados por lo común son menos precisos, es decir, el término del error normalmente es mayor, si lo demás permanece igual. No obstante, es dudoso que ello sea un motivo de preocupación. En ciertos casos, en efecto lo es —por ejemplo, cuando se requiere de una prueba muy sensible para una hipótesis—. Sin embargo, en gran parte de la investigación del comportamiento tal vez sea deseable considerar como no significativo cualquier efecto que sea insuficientemente poderoso para hacerse sentir sobre y por arriba del ruido aleatorio de un diseño de sujetos aleatorizados.

De cualquier forma, entonces, éstos son diseños poderosos, flexibles, útiles y de amplia aplicación. En la opinión de los autores, son los mejores diseños disponibles, quizá los primeros a considerarse al planear el diseño de un estudio de investigación.

Grupos correlacionados

Existe un principio básico detrás de todos los diseños de grupos correlacionados: hay varianza sistemática en las medidas de la variable dependiente, debida a la correlación entre los grupos *en alguna variable relacionada con la variable dependiente.* Esta correlación y su varianza concomitante puede introducirse en las medidas y en el diseño de tres formas:

1. empleando las mismas unidades, por ejemplo participantes, en cada uno de los grupos experimentales,
2. apareando las unidades respecto a una o más variables independientes que estén relacionadas con la variable dependiente, y
3. usando más de un grupo de unidades en el diseño, como clases o escuelas.

A pesar de las aparentes diferencias entre las tres formas para introducir la correlación en las medidas de la variable dependiente, básicamente son las mismas. Ahora se examinarán las implicaciones que tiene este principio básico para el diseño, y se analizarán las formas de implementarlo.

El paradigma general

Con excepción de los diseños factoriales correlacionados y de los llamados diseños anidados, todos los paradigmas del análisis de varianza de diseños de grupos correlacionados se bosquejan fácilmente. El término *grupo* debería utilizarse para indicar conjuntos de puntuaciones; así no hay confusión cuando un experimento de ensayos repetidos se clasifica como un diseño multigrupal. El paradigma general se presenta en la figura 21.2. Para enfatizar las fuentes de varianza, se indican las medias de las columnas y de los renglones; también se incluyen las medidas individuales de la variable dependiente (Y).

Resulta útil conocer el sistema de subíndices de los símbolos utilizados en matemáticas y estadística. Una tabla rectangular de números se llama *matriz*. Los elementos de una matriz son letras y/o números. Cuando se utilizan letras, es común identificar cualquier elemento particular de la matriz con dos (en ocasiones más) subíndices. El primero de ellos indica el número de la posición del renglón; y el segundo, el número de la posición de la columna. Y_{32}, por ejemplo, indica la medida de Y en el tercer renglón y en la segunda columna; Y_{52} indica la medida de Y del quinto renglón y de la segunda columna. También se acostumbra generalizar tal sistema al añadir subíndices de letras. En este libro i simboliza cualquier número de renglón, y j cualquier número de columna. Cualquier número de la matriz se representa por Y_{ij}. Cualquier número del tercer renglón es Y_{3j}; y cualquier número de la segunda columna es Y_{i2}.

Puede notarse que existen dos fuentes de varianza sistemática: aquella debida a las columnas o los tratamientos, y aquella debida a los renglones (diferencias individuales o de unidad). El análisis de varianza debe ser de dos factores.

El lector que ya estudió el análisis sobre la varianza de correlación del capítulo 15, donde se presentaron los estadísticos y algunos de los problemas de los diseños de grupos correlacionados, no tendrá dificultad con el razonamiento de la varianza de la figura 21.2. La intención del diseño consiste en maximizar la varianza entre tratamientos, identificar la varianza entre unidades y la varianza del error (residual). El principio del maxmincon aplica aquí como en cualquier otro lado. La única diferencia, en realidad, entre los diseños de grupos correlacionados y los sujetos aleatorizados es la varianza de los renglones o de las unidades.

Unidades

Las unidades utilizadas no alteran el principio de la varianza. El término *unidad* se usa deliberadamente para enfatizar que las unidades pueden ser personas o participantes, clases, escuelas, distritos, ciudades e incluso naciones. En otras palabras, la "unidad" es un rubro generalizado que puede representar muchos tipos de entidades. La consideración importante es si las unidades —cualesquiera que sean— difieren o no entre sí; si difieren, entonces se introduce *varianza entre unidades*. En este sentido, hablar de grupos o participantes correlacionados es lo mismo que hablar de varianza entre grupos o participantes. El concepto de diferencias individuales se extiende a *diferencias* de unidades.

El valor real del diseño de grupos correlacionados, más allá de permitir al investigador aislar y estimar la varianza debida a la correlación, es la guía que permite al investigador diseñar investigaciones para aprovechar las diferencias que frecuentemente existen

FIGURA 21.2

	Tratamientos							
Unidades	X_1	X_2	X_3	.	.	.	X_k	**Renglones**
1	Y_{11}	Y_{12}	Y_{13}	.	.	.	Y_{1k}	M_1
2	Y_{21}	Y_{22}	Y_{23}	.	.	.	Y_{2k}	M_2
	Y_{31}	Y_{32}	Y_{33}	.	.	.	Y_{3k}	M_3
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
n	Y_{n1}	Y_{n2}	Y_{n3}	.	.	.	Y_{nk}	M_n
	M_{x1}	M_{x2}	M_{x3}	.	.	.	M_{xk}	(M_t)

entre las unidades. Si un estudio de investigación incluye diferentes clases de la misma escuela, éstas son una posible fuente de varianza; por lo tanto, sería sensato utilizar las "clases" como unidades en el diseño. Las diferencias bien conocidas entre escuelas son fuentes de varianza muy importantes en la investigación del comportamiento; pueden manejarse como un diseño factorial o en la forma de los diseños de este capítulo. De hecho, si se observa con detenimiento un diseño factorial con dos variables independientes, una como *escuelas*, y un diseño de grupos correlacionados con las unidades *escuelas*, en esencia se encuentra el mismo diseño. Estudie la figura 21.3; a la izquierda hay un diseño factorial y a la derecha un diseño de grupos correlacionados; sin embargo, ¡se ven iguales! Lo son respecto al principio de la varianza. (La única diferencia sería el número de puntuaciones en las casillas y el tratamiento estadístico.)

Diseño de un grupo con ensayos repetidos

En el diseño de un grupo con ensayos repetidos, como su nombre lo indica, a un grupo se le dan diferentes tratamientos en diferentes momentos. En un experimento sobre aprendizaje, el mismo grupo de participantes puede recibir varias tareas de complejidad diferente, o la manipulación experimental tal vez sea la presentación de principios de aprendizaje en órdenes distintos, por ejemplo, de simple a complejo, de complejo a simple, del todo a la parte, de la parte al todo.

Anteriormente se indicó que el mejor apareamiento posible de los participantes consiste en aparearlos consigo mismos. Las dificultades del empleo de tal solución del problema de control también se mencionó. Una de ellas se refiere a la sensibilización del pretest, el cual puede producir una interacción entre el pretest y la variable manipulada experimentalmente. Otra dificultad reside en que los participantes maduren y aprendan a través del tiempo. Un participante que ha experimentado uno o dos ensayos de una manipulación experimental y que enfrenta un tercer ensayo, ahora es una persona diferente de la que enfrentó el primer ensayo. Las situaciones experimentales difieren mucho, por supuesto. En algunas situaciones los ensayos repetidos quizá no afecten en exceso el desempeño de los participantes en ensayos posteriores; en otras situaciones posiblemente sí. El problema sobre cómo aprenden los individuos, o cómo se sensibilizan excesivamente durante un experimento, es difícil de resolver. En resumen, la *historia*, la *maduración* y la *sensibilización* son posibles debilidades de los ensayos repetidos. El efecto de regresión también es una debilidad, ya que, como se vio en un capítulo previo, los individuos con bajas puntuaciones tienden a obtener puntuaciones más altas; y los individuos con altas puntuaciones tienden a obtener puntuaciones más bajas en el postest, debido simplemente a la correlación imperfecta entre los grupos. Por supuesto, se requiere de un grupo control.

A pesar de las dificultades básicas de tiempo, habrá ocasiones en que un diseño de un grupo con ensayos repetidos sea útil. En efecto, en el análisis de datos de "tiempo", éste es

FIGURA 21.3

	Tratamientos			Tratamientos	
Escuelas	A_1	A_2	Escuelas	A_1	A_2
B_1			1		
B_2			2		
B_3			3		
Diseño factorial			Diseño de grupos correlacionados		

el diseño implícito. Si se tienen series de mediciones del crecimiento en niños, por ejemplo, los distintos momentos en que se hicieron las mediciones corresponden a los tratamientos. El paradigma del diseño es el mismo mostrado en la figura 21.2. Tan sólo se sustituyen "participantes" por "unidades" y se anotan $X_1, X_2, \ldots$ como "ensayos".

A partir de este paradigma general, es posible derivar casos especiales. El caso más simple es el diseño pretest-postest de un grupo, diseño 19.2(*a*), donde se aplicó un tratamiento experimental a un grupo de participantes, precedido de un pretest y seguido de un postest. Puesto que las debilidades de tal diseño ya se mencionaron, no es necesario ampliar la discusión. No obstante, debe notarse que este diseño, especialmente en su forma no experimental, se aproxima bastante a muchas observaciones y pensamientos de sentido común. Una persona observa prácticas educativas hoy y decide que no son buenas. Para realizar dicho juicio, uno compara implícita o explícitamente las prácticas educativas de hoy con las del pasado. De un posible número de causas, dependiendo del sesgo particular, el investigador seleccionará una o más razones por las que él considera lamentable el estado de los asuntos educativos: "educación progresiva", "educacionistas", "degeneración moral", "carencia de principios religiosos firmes", etcétera.

Diseños de dos grupos: grupo experimental-grupo control

Se trata de un diseño con dos formas, la mejor de las cuales (repetida aquí) se describió en el capítulo 20 como diseño 20.2:

$[Ap_a]$	X	Y	(Experimental)
	$\sim X$	Y	(Control)

En este diseño primero se aparea a los participantes y después se les asigna aleatoriamente a los grupos experimental y control. En la otra forma, se aparea a los participantes, pero no se les asigna a los grupos experimental y control de manera aleatoria. El último diseño se indica simplemente por medio de la eliminación del subíndice *a* (asignación aleatoria) de Ap_a, que indica el apareamiento de los sujetos y su asignación aleatoria a los grupos (descrito en el capítulo 19 como el diseño 19.4, uno de los diseños menos adecuados).

El paradigma estadístico de este caballo de batalla de los diseños se presenta en la figura 21.4. La inserción de los símbolos para las medias indica las dos fuentes de varianza

FIGURA 21.4

	Tratamientos		
Pares	X_e	X_c	
1	Y_{1e}	Y_{1c}	M_1
2	Y_{2e}	Y_{2c}	M_2
3	Y_{3e}	Y_{3c}	M_3
.	.	.	.
.	.	.	.
n	Y_{ne}	Y_{nc}	M_n
	M_e	M_c	

sistemática: *tratamientos* y *pares*, columnas y renglones. Éste contrasta claramente con los diseños aleatorizados en una sección previa de este capítulo, donde la única varianza sistemática eran los *tratamientos* o columnas.

La variante más común del diseño de dos grupos, grupo experimental-grupo control es el diseño pretest-postest de dos grupos [véase diseño 20.3(*b*)]. El paradigma estadístico del diseño y su lógica se analizarán más adelante.

Ejemplos de investigación de los diseños de grupos correlacionados

Se han publicado cientos de estudios del tipo de grupos correlacionados. Los diseños usados con mayor frecuencia son los de participantes apareados, o los mismos participantes con pretest y postest. Sin embargo, los diseños de grupos correlacionados no se limitan a dos grupos; por ejemplo, a los mismos participantes se les puede aplicar más de dos tratamientos experimentales. Los estudios que se describen a continuación se eligieron no sólo porque ilustran los diseños de grupos correlacionados, el apareamiento y los problemas de control, sino también porque son importantes histórica, psicológica o educativamente.

Estudio de transferencia del aprendizaje de Thorndike

En 1924, E. L. Thorndike publicó un notable estudio sobre el supuesto efecto de ciertas materias en la inteligencia de los estudiantes. Los estudiantes fueron apareados de acuerdo con las puntuaciones en la forma A de la medida de la variable dependiente, la *inteligencia*. Esta prueba también sirvió como un pretest. La variable independiente fue un *estudio de un año de los participantes*, en materias tales como historia, matemáticas y latín. Al final del año se les aplicó un postest, la forma B de la prueba de inteligencia. Thorndike (1924) utilizó un recurso ingenioso para separar el efecto diferencial de cada materia escolar, al aparear en la forma A de la prueba de inteligencia a aquellos alumnos que estudiaron, por ejemplo, inglés, historia, geometría y *latín*, con los alumnos que estudiaron inglés, historia, geometría y *taller*. Así, para estos dos grupos, él comparó los efectos diferenciales de *latín* y *taller*. Los incrementos en las puntuaciones finales de inteligencia se consideraron como un efecto conjunto del crecimiento y de las materias académicas estudiadas.

A pesar de sus debilidades, fue un estudio colosal. Thorndike estaba consciente de la falta de controles adecuados, como lo revela en el siguiente párrafo sobre los efectos de la selección.

> La principal razón por la cual los buenos pensadores parecen superficialmente haberlo hecho así al tomar ciertos estudios escolares es que los buenos pensadores han tomado dichos estudios... Cuando los buenos pensadores estudiaron griego y latín, tales estudios parecieron hacer buenos pensadores. Ahora que los buenos pensadores estudian física y trigonometría, éstas parecen formar buenos pensadores. Si los alumnos más capaces debieran estudiar educación física y arte dramático, entonces estas materias parecerían formar buenos pensadores (p. 98).

Thorndike señaló el camino de la investigación educativa controlada, el cual conlleva la disminución de explicaciones metafísicas y dogmáticas en la educación. Su trabajo dio un golpe a la teoría de "la frotación de la navaja" del entrenamiento mental, aquella que semejaba la mente con una navaja que podía afilarse frotándola sobre sujetos "duros".

No es fácil evaluar un estudio como éste, cuya índole e ingenuidad son impresionantes. Sin embargo, uno se pregunta sobre la adecuación de la variable dependiente, inteli-

gencia o habilidad intelectual. ¿Las materias escolares estudiadas durante un año pueden tener un gran efecto sobre la inteligencia? Además, el estudio fue no experimental. Thorndike midió la inteligencia de los estudiantes y dejó que operaran las variables independientes, *materias escolares*. Por supuesto que no era posible realizar ninguna aleatorización. Como se mencionó antes, él estaba consciente de tal debilidad en el control de su estudio, el cual es todavía un clásico que merece respeto y estudio cuidadoso, a pesar de sus debilidades en cuanto a historia y selección (se controló la maduración).

Miller y DiCara: aprendizaje de funciones autónomas

En un capítulo anterior se presentaron los datos de un estudio, del notable conjunto de estudios sobre el aprendizaje de funciones autónomas realizado por Miller y sus colegas (Miller, 1971; Miller y DiCara, 1968). Tanto los expertos como los novatos consideran que no es posible aprender y controlar las respuestas del sistema nervioso autónomo. Es decir, que respuestas glandulares y viscerales —latido cardiaco, secreción de orina y presión sanguínea, por ejemplo— se suponían más allá del "control" del individuo. Miller creía lo contrario, pues experimentalmente demostró que tales respuestas están sujetas al aprendizaje instrumental. La parte crucial de este método consistió en recompensar las respuestas viscerales cuando ocurrían. En el estudio (los datos se citaron en un capítulo previo de este libro) se recompensaba a las ratas cuando incrementaban o disminuían la secreción de orina. Se asignaron aleatoriamente 14 ratas a dos grupos llamados "ratas con incremento" y "ratas con disminución". Las ratas del primer grupo fueron recompensadas con estimulación cerebral (la cual había resultado efectiva para *incrementar* la secreción de orina); mientras que las ratas del último grupo fueron recompensadas por *disminuir* la secreción de orina durante un periodo de "entrenamiento" de 220 ensayos, en aproximadamente tres horas.

Para mostrar parte de los paradigmas experimental y analítico de este experimento, los datos antes-después de los periodos de entrenamiento de las ratas con incremento y las ratas con disminución, se incluyen en la tabla 21.4 (tomados de la tabla 1 de Miller y DiCara). Las medidas en la tabla son mililitros de orina secretada por minuto por cada 100 gramos de peso corporal. Observe que todas son cantidades muy pequeñas. El diseño de la investigación es una variante del diseño 20.3(*a*)

[*A*]	Y_a	X	Y_d	(Experimental)
	Y_a	$\sim X$	Y_d	(Control)

La diferencia es que $\sim X$, lo que en el diseño significa ausencia de tratamiento experimental para el grupo control, ahora significa recompensa por decremento en la secreción de orina. Por lo tanto, se altera el análisis usual de las medidas antes-después de los dos grupos.

El análisis se comprende mejor si se analizan los datos de la tabla 21.4, de manera diferente a como lo hicieron Miller y DiCara. (Ellos utilizaron pruebas *t*.) Aquí se realizó un análisis de varianza de dos factores (medidas repetidas) de los datos de las ratas con incremento, *antes-después* y de los datos de las ratas con disminución, *antes-después*. Las medias *antes* y *después* del grupo con incremento fueron .017 y .028, y las del grupo con disminución fueron .020 y .006. La razón *F* del grupo con incremento fue 43.875 (gl = 1.6): la *F* de las ratas con disminución fue 46.624. Ambas fueron altamente significativas. Sin embargo, las dos medias *antes*, .017 y .020, no fueron significativamente diferentes. En este caso, la comparación de las medias *después* de los dos grupos, la comparación acostumbrada con este diseño, probablemente no sea apropiada debido a que una era para el incremento y la otra para la disminución de la secreción de orina.

TABLA 21.4 *Datos de secreción de orina, ratas con incremento y ratas con disminución, entrenamiento antes-después (estudio de Miller y DiCara)*

Ratas con incremento[a]				*Ratas con disminución*[b]			
Ratas	**Antes**	**Después**	**Σ**	**Ratas**	**Antes**	**Después**	**Σ**
1	.023	.030	.053	1	.018	.007	.025
2	.014	.019	.033	2	.015	.003	.018
3	.016	.029	.045	3	.012	.005	.017
4	.018	.030	.048	4	.015	.006	.021
5	.007	.016	.023	5	.030	.009	.039
6	.026	.044	.070	6	.027	.008	.035
7	.012	.026	.038	7	.020	.003	.023
Medias	.017	.028			.020	.006	.023

[a] Incremento, antes-después: $F = 43.875$ ($p < .001$); $\omega^2 = .357$. Las medidas en la tabla son mililitros por minuto por cada 100 gramos de peso.
[b] Decremento, antes-después: $F = 46.624$ ($p < .001$); $\omega^2 = .663$.

Este estudio, con sus manipulaciones experimentales altamente controladas y sus análisis de "controles", constituye un ejemplo de la conceptualización imaginativa y del análisis competente disciplinado. El análisis anterior es un ejemplo, pero los autores del estudio hicieron mucho más. Por ejemplo, para estar más seguros de que el reforzamiento afectó únicamente la secreción de orina, compararon la frecuencia cardiaca (latidos por minuto) *antes-después*, tanto de las ratas con incremento como con disminución. Las medias fueron 367 y 412 para las ratas con incremento; y 373 y 390 para las ratas con disminución. Ninguna de las diferencias fue estadísticamente significativa. Comparaciones similares de la presión sanguínea y de otras funciones corporales tampoco fueron significativas.

Se recomienda a los estudiantes estudiar este excelente ejemplo de investigación en laboratorio hasta que comprendan claramente qué se hizo y por qué, lo cual los ayudará a aprender más sobre experimentos controlados, diseño de investigación y análisis estadístico, que la mayoría de los ejercicios en libros de texto. ¡Es un logro espléndido!

Tipper, Eissenberg y Weaver: efectos de la práctica sobre la atención selectiva

Cuando se habla de atención selectiva, algunos podrán recordar el estudio clásico de Stroop (1935), quien demostró el papel de la interferencia sobre la atención selectiva. Los estímulos irrelevantes compiten con los relevantes para lograr el control de la acción perceptual. Para aquellos que no estén familiarizados con dicho estudio, una parte memorable fue presentar a los participantes palabras como *verde* y *azul* impresas en rojo y amarillo. Luego se les pidió nombrar los colores en los que estaban escritas las palabras, pero en lugar de eso, los sujetos leían las palabras. Las personas tienen dificultad para suprimir el hábito de leer palabras aun cuando se les pide que no lo hagan. Para realizar dicha tarea de forma correcta, el participante debe concentrarse y evitar de manera consciente leer las palabras. Esta interferencia fue llamada el efecto Stroop. Desde la realización del famoso estudio de Stroop, se ha llevado a cabo un gran número de estudios sobre atención selectiva; el estudio de Tipper, Eissenberg y Weaver (1992) es uno de ellos. Este estudio es diferente, ya que discute varios aspectos respecto a numerosos estudios realizados sobre atención selectiva. Primero, Tipper y sus colaboradores hipotetizaron que cualquier experimento sobre atención selectiva que utilice participantes durante una hora o más puede estar conectando diferentes mecanismos perceptuales de los que se usan en la vida diaria. Los experimentos

de laboratorio casi siempre requieren que los participantes estén presentes durante aproximadamente una hora. En el periodo de una hora la experiencia experimental completa es aún novedosa. Es probable que la selectividad de atención se logre por medio de diferentes mecanismos conforme los estímulos se vuelven más familiares.

Tipper y sus colegas diseñaron un estudio para probar su hipótesis sobre atención selectiva utilizando un diseño dentro de sujetos. Todos los participantes experimentaron todas las condiciones del tratamiento. Ellos observaron el efecto de la interferencia sobre el tiempo de reacción y los errores. Hicieron que cada participante experimentara ambos niveles de interferencia: preparación negativa e inhibición de respuesta a través de 11 bloques o ensayos tomados durante cuatro días (efecto de la práctica). Sus resultados demostraron que hubo un efecto de interferencia ($F = 35.15$, $p < .001$) cuando se utilizó el tiempo de reacción como variable dependiente. Los tiempos de reacción fueron más largos cuando la distracción estuvo presente. También encontraron un efecto por la práctica (bloques) ($F = 9.62$, $p < .0001$) y ningún efecto de interacción. El efecto de la práctica indicó que la reacción de los participantes se torna más rápida con el incremento de la práctica. El hecho de que el efecto de interacción no fuese significativo indica que los efectos de interferencia de los estímulos irrelevantes permanecieron constantes, aun después de una práctica prolongada. Los hallazgos de Tipper y sus colegas sugieren que existen otros mecanismos de atención selectiva y que operan con diferentes niveles de experiencia.

Diseños multigrupales con grupos correlacionados

Varianza de las unidades

Mientras que es difícil aparear tres o cuatro conjuntos de participantes, y mientras que en la investigación del comportamiento por lo común no es factible ni deseable utilizar a los mismos participantes en cada uno de los grupos, hay situaciones naturales donde existen grupos correlacionados. Tales situaciones son particularmente importantes en la investigación educativa. Hasta hace poco, las varianzas debidas a las diferencias entre clases, escuelas, sistemas escolares y otras unidades "naturales" no se habían controlado bien o no habían sido utilizadas con la frecuencia deseada en el análisis de datos. Quizá la primera indicación de la importancia de este tipo de varianza fue dada en el magnífico libro de Lindquist (1940) sobre el análisis estadístico en la investigación educativa. En esta obra, Lindquist da un énfasis considerable a la varianza de las escuelas. Las escuelas, clases y otras unidades educativas tienden a diferir significativamente respecto al aprovechamiento, la inteligencia, las aptitudes y otras variables. El investigador educativo debe permanecer alerta ante estas ***diferencias de las unidades***, así como a las ***diferencias individuales***.

Considere un ejemplo obvio. Suponga que un investigador elige una muestra de cinco escuelas por su variedad y homogeneidad. La meta, por supuesto, es la validez externa: la representatividad. El investigador utiliza alumnos de las cinco escuelas y combina las medidas de las cinco para probar las diferencias entre medias en alguna variable dependiente. Al hacerlo, el investigador ignora la varianza debida a las diferencias entre las escuelas. Es entendible que las medias no difieran significativamente; la varianza de las escuelas está mezclada con la varianza del error.

Pueden surgir grandes errores por ignorar la varianza de las unidades tales como escuelas y clases. Uno de estos errores consiste en seleccionar varias escuelas y designar algunas de ellas como unidades experimentales y otras como unidades control. Aquí la varianza entre escuelas se enreda con la varianza de la variable experimental. De forma

similar, las clases, los distritos escolares y otras unidades educativas difieren y, por lo tanto, producen varianza. Las varianzas deben identificarse y controlarse, ya sea por medio de control experimental o estadístico, o de ambos.

Diseño factorial con grupos correlacionados

Los modelos factoriales pueden combinarse con la noción de unidades para producir un diseño valioso: el diseño *factorial de grupos correlacionados*, el cual es apropiado cuando las unidades son parte natural de la situación de investigación. Por ejemplo, la investigación quizá requiera la comparación de una variable antes y después de una intervención experimental, o antes y después de un evento importante. En efecto, habrá correlación entre las medidas antes-después de la variable dependiente. Otro ejemplo útil se presenta en la figura 21.5. Éste es un diseño factorial de 3×2 con cinco unidades (clases, escuelas, etcétera) en cada nivel de B_1 y B_2.

Las fortalezas y debilidades del diseño factorial con grupos correlacionados son similares a las de diseños factoriales más complejos. Las principales fortalezas son la habilidad para aislar y medir las varianzas y probar las interacciones. Observe que las dos principales fuentes de varianza, *tratamiento* (A) y *niveles* (B), así como las unidades de varianza, pueden evaluarse, es decir, es posible probar la significancia de las diferencias entre las medias de A, B y de las unidades. Además, pueden probarse tres interacciones: tratamientos por niveles, tratamientos por unidades y niveles por unidades. Si se utilizan puntuaciones individuales en las casillas en lugar de medias, entonces también puede probarse la interacción triple. Note lo importante que resulta dicha interacción, tanto teórica como prácticamente. Por ejemplo, se responderían preguntas como las siguientes: ¿los tratamientos operan de forma diferente en unidades distintas? ¿Ciertos métodos funcionan de manera distinta en diferentes niveles de inteligencia? ¿Con diferentes sexos? ¿Con niños de distintos niveles socioeconómicos? El estudiante avanzado deseará saber cómo manejar unidades (escuelas, clases, etcétera) y unidades de varianza en diseños factoriales. Una guía detallada se encuentra en Edwards (1984) y en Kirk (1995). El tema es difícil, incluso los nombres de los diseños se vuelven complejos: bloques aleatorizados, tratamientos anidados, diseños

FIGURA 21.5

			Métodos (tratamiento)		
		Unidades	A_1	A_2	A_3
	B_1	1			
		2			
		3			
		4			
		5			
Niveles (dispositivos, tipos, etcétera)			Medias de Y o medidas		
	B_2	1			
		2			
		3			
		4			
		5			

de diagrama dividido. Sin embargo, tales diseños son poderosos: combinan las virtudes de los diseños factoriales y de los diseños con grupos correlacionados. Cuando se requiera, Edwards y Kirk serán una buena guía. Además, se sugiere solicitar ayuda de alguien que entienda tanto de estadística como de investigación del comportamiento. Es absurdo utilizar programas computacionales sólo porque sus nombres parezcan apropiados o porque estén disponibles. También lo es buscar ayuda analítica del personal de informática; no es posible esperar que ellos conozcan y entiendan, por ejemplo, del análisis factorial de varianza, ya que ése no es su campo. Se tratará más sobre análisis computacional en capítulos posteriores.

Suedfeld y Rank: líderes revolucionarios y complejidad conceptual

Suedfeld y Rank (1976) probaron la intrigante noción de que los líderes revolucionarios existosos —Lenin, Cromwell y Jefferson, entre otros— son conceptualmente simples en sus discursos públicos *antes* de la revolución y conceptualmente complejos *después* de la misma. Los líderes revolucionarios no exitosos, por el otro lado, no difieren en su complejidad conceptual antes y después de la revolución. El problema se presta para un diseño factorial y para un análisis de medidas repetidas. El diseño y los datos sobre la complejidad conceptual se muestran en la tabla 21.5. Puede verse que los líderes exitosos se tornaron conceptualmente más complejos —de 1.67 a 3.65— pero los líderes no exitosos no cambiaron mucho —de 2.37 a 2.21—. La razón F de la interacción fue 12.37, significativa al nivel .005. La hipótesis fue apoyada.

Deben aclararse algunos puntos aquí. Primero, observe la combinación efectiva del diseño factorial y de las medidas repetidas. Cuando la combinación es apropiada, como en este caso, es bastante efectiva principalmente porque deja de lado, por así decirlo, la varianza en las medidas de la variable dependiente debidas a las diferencias individuales (o de grupo o de bloque). Por lo tanto, el término del error es menor y más capaz de evaluar la significancia estadística de las diferencias entre las medias. En segundo lugar, dicho estudio fue no experimental: no se manipuló ninguna variable experimental. En tercer lugar, y lo más importante, el interés intrínseco y la significancia del problema de investigación y su teoría; y la ingenuidad de medir y utilizar la complejidad conceptual como variable para "explicar" el éxito de los líderes revolucionarios ensombrece posibles puntos metodológicamente cuestionables. La frase anterior, por ejemplo, quizá sea incongruente con el uso de las variables en este estudio. Suedfeld y Rank analizaron medidas de la variable independiente, complejidad conceptual; pero la hipótesis bajo estudio en realidad era: si hay complejidad conceptual (después de la revolución), entonces habrá liderazgo exitoso. Sin embargo, con un problema de investigación de tan imponente interés y con una variable de tal importancia (complejidad conceptual) medida con gran imaginación y competencia, ¿quién quiere objetar?

TABLA 21.5 *Diseño factorial con medidas repetidas: líderes revolucionarios (estudio de Suedfeld y Rank)*[a]

	Antes de tomar el poder	Después de tomar el poder	
Éxito	1.67	3.65	2.66
Fracaso	2.37	2.22	2.30
	1.96	3.05	

[a] Las medidas en la tabla son medias de complejidad conceptual. F de la interacción = 12.37 ($p < .005$).

Perrine, Lisle y Tucker: ofrecimiento de ayuda y disposición para buscar apoyo

Los maestros en todos los niveles de educación utilizan un programa sobre la asignatura para introducir a los estudiantes al curso. ¿Qué y cuántas características del programa tienen el mayor impacto en los estudiantes, incluso antes de que empiece la instrucción en el salón de clases? Perrine, Lisle y Tucker (1995) realizaron un estudio para saber si el ofrecimiento de ayuda en el programa del instructor anima a los estudiantes universitarios, de diferentes edades, a buscar ayuda de sus instructores. De acuerdo con Perrine y sus colaboradores, éste es el primer estudio que explora el uso del apoyo social, por parte de instructores universitarios, en beneficio de los estudiantes. Perrine y sus colaboradores también estudiaron el efecto del tamaño de la clase en la disposición de los estudiantes para buscar ayuda. El estudio utilizó 104 estudiantes de licenciatura, de los cuales 82 eran mujeres y 22 eran hombres. Se pidió a cada participante leer una descripción de dos clases de psicología; las descripciones incluían afirmaciones realizadas por los instructores de cada clase en los programas de la asignatura. En la descripción se manipuló el tamaño de la clase, incluyendo 15, 45 o 150 estudiantes. El curso era descrito como demandante de mucho trabajo, pero digno de disfrutarse. También animaba a los estudiantes a no retrasarse en las lecturas ni en las tareas. Las dos afirmaciones separadas de los instructores consistieron en una que demostraba apoyo y otra que permanecía neutral. En la declaración de apoyo se animaba al estudiante a acercarse al instructor para pedir ayuda si alguna vez encontraba problemas en la clase; el neutral no incluía dicho comentario. Cada participante leyó ambas descripciones, después de lo cual, el participante respondía preguntas acerca de su disposición a buscar ayuda del instructor por seis posibles problemas académicos encontrados en la clase: 1) dificultades para entender un libro de texto, 2) baja calificación en el primer examen, 3) problemas al escuchar la exposición del instructor, 4) habilidades de estudio inefectivas para el curso, 5) planes para abandonar el curso y 6) dificultades para entender un tema importante. El participante utilizaba una escala de evaluación de 6 puntos que iba desde 0 = definitivamente no, hasta 6 = definitivamente sí.

El diseño fue un diseño factorial de 3 × 2 × 2 (tamaño de la clase × afirmaciones del discurso × edad del estudiante). El diseño contenía una variable independiente manipulada (activa), una variable independiente medida (atributo) y una variable independiente dentro de sujetos (correlacionada). El tamaño de la clase era la variable independiente manipulada y aleatorizada. La edad del estudiante era la variable independiente medida y el comentario del programa fue la variable independiente correlacionada. El uso del análisis de varianza apropiado (generalmente conocido como ANOVA mezclado, cuando al menos una varia-

TABLA 21.6 *Medias y valores F de las diferencias de los comentarios del programa y diferencias de edad (estudio de Perrine, Lisle y Tucker)*

	Programa			*Edad*		
Problema académico	**Con apoyo**	**Neutral**	***F***	**Mayores**	**Menores**	***F***
Dificultades para entender un libro de texto	4.7	3.7	76.08**	4.8	4.1	5.48*
Baja calificación en el primer examen	4.8	4.0	49.89**	5.2	4.3	7.64*
Problemas al escuchar la exposición del instructor	4.4	3.8	36.05**	4.4	4.0	1.01
Habilidades de estudio inefectivas para el curso	4.7	3.6	79.57**	4.8	4.0	6.32*
Planes para abandonar el curso	4.9	3.8	61.80**	4.8	4.3	2.18
Dificultades para entender un tema importante	5.3	4.2	82.97**	5.3	4.6	7.69*

* $p < .05$
** $p < .01$

ble independiente es entre sujetos, y al menos otra es dentro de sujetos) reveló que los participantes expresaron significativamente mayor disposición para buscar ayuda del instructor cuando la declaración de apoyo aparecía en el programa de la asignatura, que cuando sólo aparecía el comentario neutral. Los estudiantes más jóvenes (menores de 25 años) expresaron menor disposición que los estudiantes mayores. También hubo una interacción de *edad* × *programa* ($F = 4.85$, $p < .05$) que fue significativa. La respuesta al ofrecimiento de ayuda fue diferente entre los grupos de edades. Los comentarios afectaron menos a los estudiantes más jóvenes que a los mayores. El tamaño de la clase no pareció ser un factor significativo respecto a si los estudiantes estaban dispuestos o no a solicitar ayuda. La tabla 21.6 presenta el resumen estadístico del estudio.

Análisis de covarianza

La invención de Ronald Fisher del análisis de covarianza fue un evento importante en la metodología de la investigación del comportamiento. Constituye un uso creativo de los principios de la varianza, comunes al diseño experimental y a la teoría de la correlación y la regresión —que se estudiarán más adelante en el libro— para ayudar a resolver un antiguo problema del control.

El *análisis de covarianza* es una forma de análisis de varianza que prueba la significancia de las diferencias entre las medias de los grupos experimentales, después de tomar en cuenta las diferencias iniciales entre los grupos y la correlación de las medidas iniciales y las medidas de la variable dependiente. Es decir, el análisis de covarianza analiza las diferencias entre los grupos experimentales sobre Y, la variable dependiente, después de tomar en cuenta ya sean las diferencias iniciales entre los grupos sobre Y (pretest) o las diferencias entre los grupos en alguna(s) variable(s) independiente(s) potencial(es), X, correlacionadas sustancialmente con Y, la variable dependiente. La medida utilizada como variable control —el pretest o variable pertinente— se llama un *covariable*.

El lector debe ser precavido al utilizar el análisis de covarianza; es particularmente sensible a las violaciones de sus supuestos. El mal uso potencial de este método fue de tanta preocupación que la revista *Biometrics*, en 1957, dedicó un ejemplar completo a ello. Elashoff (1969) escribió un artículo importante para los investigadores educativos respecto al uso de este método. El consenso es que generalmente no es buena idea utilizarlo para diseños de investigación no experimentales.

Clark y Walberg: reforzamiento masivo y rendimiento en lectura

No tiene mucho caso describir los procedimientos y cálculos estadísticos del análisis de covarianza. Primero, porque en su forma convencional son complejos y difíciles de seguir; segundo, aquí sólo se desea mostrar el significado y propósito del método; tercero y más importante, existe una forma más fácil de hacer lo mismo que el análisis de covarianza hace. Más adelante en el libro se verá que el análisis de covarianza es un caso especial de regresión múltiple y es mucho más fácil realizarlo con las técnicas de la regresión múltiple. Para dar al lector una idea de lo que se logra con el análisis de covarianza, se estudiará un efectivo del procedimiento en estudios educativos y psicológicos.

Clark y Walberg (1968) pensaron que sus participantes, quienes posiblemente abandonarían la escuela pues su rendimiento era deficiente, necesitaban mucho más reforzamiento (ánimo, recompensa, etcétera) que los participantes que se desempeñaban bien. Por lo tanto, utilizaron reforzamiento masivo con su grupo experimental de participantes y reforzamiento moderado con su grupo control de participantes. Puesto que su variable dependiente, *rendimiento en lectura*, está altamente correlacionada con la *inteligencia*, tam-

TABLA 21.7 *Paradigma del análisis de covarianza (estudio de Clark y Walberg)*

	Experimental (reforzamiento masivo)		Control (reforzamiento moderado)	
	X (Inteligencia)	*Y* (Lectura)	*X* (Inteligencia)	*Y* (Lectura)
Medias	92.05	31.62	90.73	26.86

bién necesitaban controlar la inteligencia. Un análisis de varianza de un factor de las medias del rendimiento en lectura, de los grupos experimental y control, produjo una *F* de 9.52, significativa al nivel .01, lo cual apoyó su creencia. No obstante, es posible que la diferencia entre los grupos experimental y control se debiera a la inteligencia más que al reforzamiento. Es decir, aunque los sujetos fueron asignados aleatoriamente al grupo experimental, una diferencia inicial en la inteligencia, a favor del grupo experimental, pudo haber sido suficiente para volver la media de lectura del grupo experimental significativamente mayor que la media de lectura del grupo control, ya que la inteligencia está altamente correlacionada con la lectura. Con la asignación aleatoria es poco probable que suceda, pero puede ocurrir. Para controlar esta posibilidad, Clark y Walberg utilizaron el análisis de covarianza.

Estudie la tabla 21.7 que presenta un bosquejo del diseño y del análisis. Las medias de las puntuaciones de *X* y de *Y*, como reportaron Clark y Walberg, aparecen al final de la tabla. La medidas de *Y* son la parte más importante; resultaron significativamente diferentes. Aunque es dudoso que el análisis de covarianza cambie estos resultados, es posible que la diferencia entre las medias de *X*, 92.05 y 90.73 haya inclinado las balanzas estadísticas, en la prueba de la diferencia entre las medias de *Y*, a favor del grupo experimental. La prueba *F* del análisis de covarianza, que utiliza las sumas de cuadrados y los cuadrados medios de *Y* libres de la influencia de *X*, fue significativa al nivel .01: $F = 7.90$. Así, las puntuaciones promedio de lectura de los grupos experimental y control difirieron significativamente, después de ajustarlas y controlarlas respecto a la inteligencia.

Diseño y análisis de investigación: observaciones concluyentes

Cuatro objetivos principales guiaron la organización y la preparación de la parte seis de este libro. El primero fue familiarizar al estudiante respecto a los principales diseños de investigación. Al hacerlo se esperaba que se ampliaran los conceptos estrechamente circunscritos sobre la realización de investigación con, digamos, sólo un grupo experimental y sólo un grupo control; o con participantes apareados o con un grupo, antes y después. El segundo objetivo fue brindar un sentido de la estructura equilibrada de los buenos diseños de investigación, para desarrollar una idea sensible por la arquitectura del diseño. El diseño debe ser formal, así como funcional (con la y ajustada a los problemas de investigación que se busca resolver). El tercer objetivo consistió en ayudar al lector a entender la lógica de la investigación experimental y los diferentes diseños. Los diseños de investigación son rutas alternativas hacia el mismo destino: planteamientos válidos y confiables de las relaciones entre variables. Algunos diseños, si pueden llevarse a la práctica, generan planteamientos relacionales más fuertes que otros diseños.

En cierto sentido, el cuarto objetivo de la parte seis —ayudar al estudiante a comprender la relación entre el diseño de investigación y la estadística— es el más difícil de lograr.

La estadística es, en un sentido, la disciplina técnica del manejo de la varianza; y, como se ha visto, uno de los propósitos básicos del diseño consiste en proporcionar control de las varianzas sistemática y del error. Ésta es la lógica utilizada para tratar la estadística en tanto detalle en la parte cuatro y en la parte cinco, antes de considerar el diseño en la parte seis. Fisher (1951, p. 3) expresa dicha idea de forma sucinta cuando dice: "El procedimiento estadístico y el diseño experimental son sólo dos aspectos diferentes de un todo, y ese todo comprende todos los requerimientos lógicos del proceso completo de adición al conocimiento natural por medio de la experimentación."

Un diseño bien concebido no es garantía de la validez de los hallazgos de investigación. Los diseños elegantes y bien adaptados a los problemas de investigación aun pueden resultar en conclusiones erróneas o distorsionadas. Sin embargo, las oportunidades de llegar a conclusiones precisas y válidas son mayores con diseños sólidos que con aquellos que no lo son. Esto es relativamente seguro: si un diseño es inadecuado, no es factible llegar a conclusiones claras. Si, por ejemplo, se utiliza un diseño de dos grupos con sujetos apareados cuando el problema de investigación demanda lógicamente un diseño factorial, o si se utiliza un diseño factorial cuando la naturaleza de la situación de investigación requiere de un diseño de grupos correlacionados, ninguna cantidad de manipulación interpretativa o estadística puede incrementar la confianza en las conclusiones de dicha investigación.

Fisher (1951) dijo la última palabra a este respecto. En el primer capítulo de su libro, *The Design of Experiments*, afirma:

> Si el diseño de un experimento resulta inadecuado, cualquier método de interpretación que lo convierta en decisivo es inadecuado también. Es verdad que existe una enorme cantidad de procedimientos experimentales que son bien diseñados, que pueden conducir a conclusiones decisivas; pero en otras ocasiones pueden fallar en hacerlo; en tales casos, si de hecho se sacan conclusiones decisivas cuando no estén justificadas, podemos afirmar que la falla está por entero en la interpretación, no en el diseño. Pero la falla de la interpretación... reside en pasar por alto los rasgos característicos del diseño, lo que conduce a que el resultado algunas veces no sea concluyente, o concluyente en algunos aspectos pero no en todos. Comprender correctamente un aspecto del problema es entender el otro (p. 3).

Anexo computacional

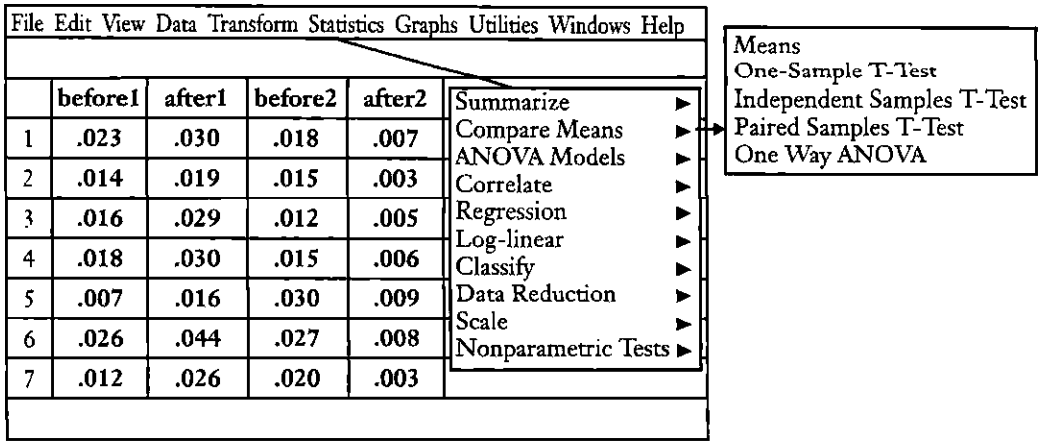
Los diseños aleatorizados pueden analizarse con pruebas *t* de muestras independientes o análisis de varianza. La organización y análisis del SPSS se incluyeron en los capítulos 13 y 14. Aquí se explicará cómo utilizar el SPSS para llevar a cabo análisis estadísticos de datos de diseños de grupos correlacionados (medidas repetidas). Para el análisis se utilizarán los datos de Miller y DiCara (1968) presentados en la tabla 21.4.

Siguiendo las instrucciones previamente establecidas sobre el ingreso de los datos, usted ingresará los datos en el SPSS, de tal manera que la hoja de datos del SPSS resultante se vea como la que aparece en la figura 21.6.

Recuerde la discusión previa, la meta era comparar ratas que presentaban un aumento en la secreción de orina con ratas cuya secreción de orina disminuía. El volumen de la secreción de orina de las ratas que mostraron un incremento es "before1" y "after1". Las variables "before2" y "after2" sirven para representar las secreciones antes y después de las ratas que mostraron una disminución.

Para que el SPSS realice el análisis apropiado para los datos presentados en la tabla 21.4 y en la figura 21.6, señale y haga clic en la opción "Statistics". Esta acción presentará un menú del cual usted debe escoger "Compare Means". Después de hacer clic en "Com-

FIGURA 21.6 *Datos de Miller y DiCara en el SPSS*



pare Means", se despliega otro menú, del cual debe elegir "Paired Sampled T- Test". Después de esta elección, aparece una nueva pantalla (véase figura 21.7).

Aquí pueden probarse simultáneamente dos pruebas *t* de muestras dependientes, siguiendo los siguientes pasos:

1. Resalte la variable "after1" (señale y haga clic en ella).
2. Resalte la variable "before1".
3. Haga clic en el botón de la flecha.
4. Resalte la variable "after2".
5. Resalte la variable "antes2".
6. Haga clic en el botón de la flecha.

FIGURA 21.7 *Pantalla del SPSS utilizada para especificar las variables para el análisis*

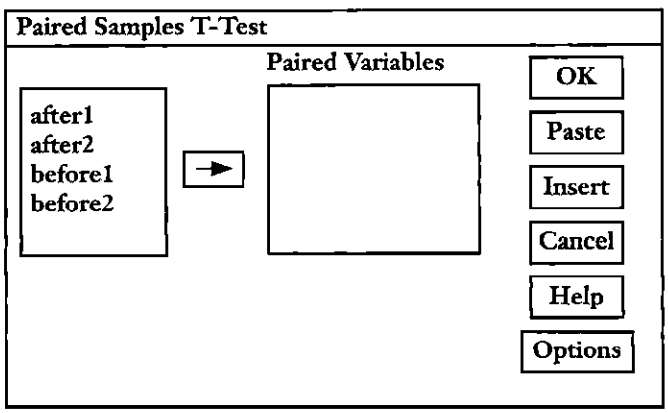
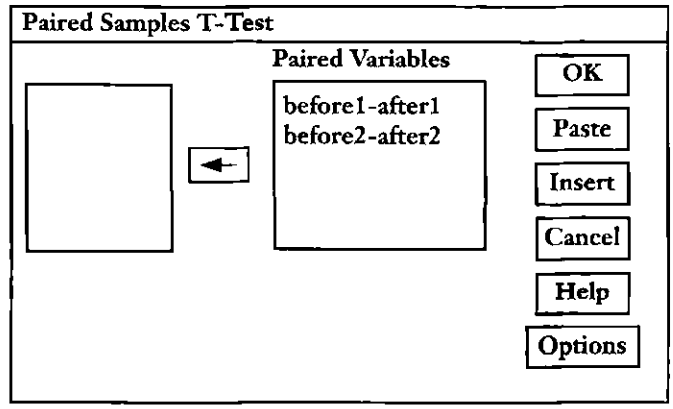


FIGURA 21.8 *Preparación para el análisis del SPSS*



En este punto usted verá que el SPSS ha formado dos ecuaciones de diferencia, que aparecen en el lado derecho del cuadro. Esto se observa en la figura 21.8. Cuando haga clic en el botón "OK", el SPSS realizará el análisis y mostrará la tabla de resultados. Una versión abreviada de la tabla de resultados se presenta en la figura 2.9.

Este análisis se realizó utilizando la prueba *t* del SPSS. Usted también puede realizar el mismo análisis con el comando "General Linear Model" del SPSS.

RESUMEN DEL CAPÍTULO

1. Los diseños de sujetos aleatorizados son los diseños preferentes para la investigación del comportamiento.
2. Los diseños de sujetos aleatorizados son verdaderos experimentos con variables independientes activas, manipuladas.
3. El método estadístico generalmente utilizado para analizar datos de diseños de sujetos aleatorizados es el análisis de varianza.
4. Los diseños de sujetos aleatorizados por lo común requieren de un número (*N*) grande de participantes para lograr la precisión deseada.

FIGURA 21.9 *Resultados del SPSS*

		Paired Differences Mean	Std. Deviation	t	df	Sig. (2-tailed)
Pair 1 Increase	BEFORE1-AFTER1	−.00111	.00445	−6.624	6	.001
Pair 2 Decrease	BEFORE2-AFTER2	.00137	.00531	6.828	6	.000

5. Los diseños de sujetos correlacionados generalmente incluyen
   a) el uso de los mismos participantes en cada condición de tratamiento
   b) aparear a los participantes en una o más variables independientes relacionadas con la variable dependiente.
   c) El empleo de más de un grupo de participantes (por ejemplo, salones de clase).
6. Las unidades pueden ser diferentes tipos de entidades. En la investigación psicológica, las unidades por lo general son personas o animales.
7. Los diseños de sujetos correlacionados incluyen al diseño de un grupo con ensayos (medidas) repetidos.
8. El diseño 20.2 es el diseño a usar cuando los participantes son apareados y fueron asignados aleatoriamente a los grupos de tratamiento.
9. Una covariable es una variable independiente potencial utilizada para ajustar las diferencias individuales entre los grupos, que *no* se deben al tratamiento. Los pretest son las covariables más comunes.
10. El análisis de covarianza es un método de sujetos correlacionados del análisis estadístico. Una covariable ajusta la variable dependiente y, después. los valores ajustados se utilizan en un análisis de varianza. La regresión múltiple es otro método estadístico que sirve para tal propósito.

SUGERENCIAS DE ESTUDIO

1. Al estudiar diseño de investigación resulta útil realizar análisis de varianza, tantos como sea posible: análisis simples de un factor y análisis factoriales de dos variables, quizá incluso un análisis de tres variables. Por medio de esta práctica estadística usted logrará un mejor entendimiento de los diseños. También puede asignar nombres de variables a sus "datos", en lugar de trabajar únicamente con números. A continuación se incluyen algunas sugerencias para realizar proyectos con números aleatorios.
   a) Obtenga tres grupos de números aleatorios del 0 al 9. Asigne nombres a las variables independiente y dependiente. Formule una hipótesis y tradúzcala a lenguaje de diseño estadístico. Realice un análisis de varianza de un factor. Interprete.
   b) Repita el paso 1 *a*) utilizando cinco grupos de números.
   c) Ahora, sume 2 a cada uno de los datos en uno de sus grupos y reste 2 a cada uno de los datos de otro grupo. Repita el análisis estadístico.
   d) Extraiga cuatro grupos de números aleatorios, con 10 números en cada uno. Ordénelos aleatoriamente en un diseño factorial de 2×2. Realice un análisis de varianza factorial.
   e) Produzca un sesgo en los números de las dos casillas derechas al sumar 3 a cada número. Repita el análisis. Compare los resultados con los del inciso *d*).
   f) Produzca un sesgo en los números de los datos del inciso *d*) de la siguiente forma: sume 2 a cada uno de los números de las casillas superior izquierda e inferior derecha. Repita el análisis e interprete.
2. Regrese al capítulo 14, a las sugerencias de estudio 2 y 3. Trabaje ambos ejemplos de nuevo. ¿Son más fáciles para usted ahora?
3. Suponga que usted es el director de una escuela primaria. Algunos de los maestros de cuarto y quinto grado desean prescindir de los libros de trabajo. Al director general no le gusta la idea pero está dispuesto a permitirle a usted probar la idea de que los libros de trabajo no hacen mucha diferencia. (Uno de los maestros incluso sugi-

TABLA 21.8 *Datos hipotéticos (medias) de un experimento factorial ficticio*

	Métodos			
	A_1	A_2	A_3	
Hombre	45	45	36	42
Mujer	35	39	40	38
	40	42	38	

rió que los libros de texto pueden traer efectos nocivos tanto en los maestros como en los alumnos.) Para probar la eficacia de dichos libros, establezca dos planes y diseños de investigación: de un factor y otro factorial. Considere las variables *rendimiento*, *inteligencia* y *género*. También podría considerar la actitud de los maestros hacia los libros de trabajo como una posible variable independiente.

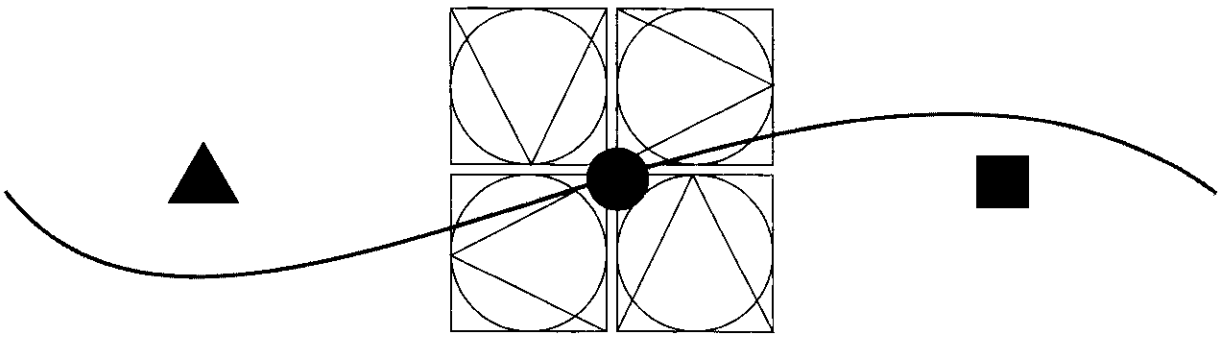
4. Suponga que se realizó una investigación que utilizó *métodos* y *género* como variables independientes y *logro* como variable dependiente, y sus resultados son los que se reportan en la tabla 21.8. Los números en las casillas son las medias ficticias. Las razones *F* de *métodos* y *género* no son significativas. La razón de *F* en la interacción es significativa al nivel .01. Interprete los resultados estadística y sustantivamente. Para hacer esto último, asigne nombres a los tres métodos.
5. Aunque difícil y en ocasiones frustrante, no existe un sustituto para la lectura y el estudio de reportes de investigación originales. En éste y en capítulos previos se han citado y resumido numerosos estudios que utilizan un diseño factorial y el análisis de varianza. Seleccione y lea dos de esos estudios e intente resumir alguno. Critique ambos estudios respecto a la adecuación del diseño y la realización de la investigación (con lo mejor de su conocimiento y habilidades actuales). Enfóquese particularmente en la adecuación del diseño para responder la(s) pregunta(s) de investigación.
6. Se realizó un análisis de varianza de dos factores (medidas repetidas) con los datos de Miller y DiCara sobre las ratas con incremento, en la tabla 21.4, con algunos de los datos reportados en la tabla: ω^2 (omega al cuadrado de Hays) fue de .357; ω^2 para los datos de las ratas con disminución fue de .633. ¿Qué significan estos coeficientes? ¿Por qué calcularlos?
7. Kolb (1965), quien basó su estudio en el trabajo sobresaliente de McClelland sobre motivación de logro, realizó un experimento fascinante con jóvenes de secundaria con bajo logro y alta inteligencia. De 57 jóvenes, asignó a 20 aleatoriamente a un programa de entrenamiento donde, a través de distintos medios, se les "enseñó" motivación de logro a los jóvenes (un intento por crear una necesidad de logro en los jóvenes). A los jóvenes se les aplicó un pretest de motivación de logro en el verano, el cual se aplicó otra vez seis meses después. Las *puntuaciones* medias de cambio fueron, para los grupos experimental y control, 6.72 y –.34, respectivamente. Éstas fueron significativas al nivel .005.
   a) Comente sobre el uso de puntuaciones de cambio. ¿Su uso debilita la fe que usted tiene en la significancia estadística de los resultados?
   b) ¿Pueden otros factores, diferentes del entrenamiento experimental, haber inducido el cambio? Si así es, ¿cuáles serían esos factores?
8. Para evitar que el estudiante crea que sólo se analizan medidas continuas y que el análisis de varianza sólo se utiliza en experimentos psicológicos y educativos, lea el estudio de Freedman, Wallington y Bless (1967) sobre la culpa y la sumisión. Había

un grupo experimental (sujetos inducidos a mentir) y un grupo control. La variable dependiente se midió viendo si un participante obedecía o no a una solicitud de ayuda. Los resultados fueron reportados en tablas de frecuencias de tabulación cruzada. Lea el estudio y, después de estudiar el diseño y los resultados de los autores, diseñe uno de los tres experimentos de otra forma. Introduzca otra variable independiente, por ejemplo. Suponga que se sabía que había grandes diferencias individuales en la sumisión. ¿Cómo puede controlarse lo anterior? Asigne nombre y describa dos tipos de diseño para hacerlo.

9. En un estudio donde el entrenamiento en las complejidades de los estímulos artísticos afectó la actitud hacia la música, entre otras cuestiones, Renner (1970) utilizó un análisis de covarianza, donde la covariable eran las medidas de una escala diseñada para medir la actitud hacia la música. Éste fue el pretest. Hubo tres grupos experimentales. Estructure el diseño a partir de esta breve descripción. ¿Por qué Renner utilizó la escala de actitud hacia la música como pretest? ¿Por qué utilizó un análisis de covarianza? (Nota: vale la pena leer el reporte original. El estudio, que en parte trata sobre creatividad, es creativo en sí.)
10. En un estudio significativo del efecto de la educación en artes liberales sobre la formación de conceptos complejos, Winter y McClelland (1978) encontraron que la diferencia entre los estudiantes de primer y segundo año de una universidad de artes liberales, respecto a la medida de la formación de conceptos complejos, fue estadísticamente significativa ($M_{\text{primer año}} = 2.00, M_{\text{segundo año}} = 1.22; t = 3.76; (p < .001)$. Como se dieron cuenta de que se necesitaba una comparación, también probaron las diferencias de medias similares en una universidad pedagógica y una universidad comunitaria. Ninguna de estas diferencias resultó estadísticamente significativa. ¿Por qué Winter y McClelland probaron la relación en la universidad pedagógica y en la universidad comunitaria? Se sugiere que los estudiantes encuentren y lean el reporte original —vale la pena su estudio— y realicen un análisis de varianza de las n, medias y desviaciones estándar reportadas, utilizando el método descrito en el capítulo 13 (anexo).
11. Una virtud del análisis de covarianza, rara vez mencionada en los textos, es que pueden calcularse tres estimados de la correlación entre X y Y. Éstos son: (i) la r total sobre todas las puntuaciones; (ii) la r entre grupos, que es la r entre las medias de X y de Y; (iii) y la r dentro de grupos, la r calculada a partir de un promedio de las r entre X y Y dentro de k grupos. La r dentro de grupos es el "mejor" estimado de la r "verdadera" entre X y Y. ¿Por qué es esto así?
[Sugerencia: ¿Puede una r total, aquella que se calcula generalmente en la práctica, inflarse o desinflarse por la varianza entre grupos?]

PARTE SIETE

TIPOS DE INVESTIGACIÓN

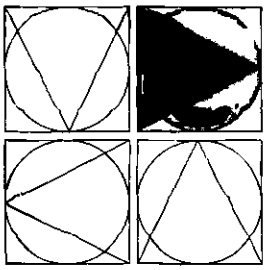


Capítulo 22
DISEÑOS DE INVESTIGACIÓN CUASI-EXPERIMENTALES Y CON $N = 1$

Capítulo 23
INVESTIGACIÓN NO EXPERIMENTAL

Capítulo 24
EXPERIMENTOS DE LABORATORIO, EXPERIMENTOS DE CAMPO Y ESTUDIOS DE CAMPO

Capítulo 25
INVESTIGACIÓN POR ENCUESTA



CAPÍTULO 22

DISEÑOS DE INVESTIGACIÓN CUASI-EXPERIMENTALES Y CON $N = 1$

En capítulos previos se estableció y enfatizó que una de las principales metas de la ciencia consiste en encontrar relaciones causales. En las ciencias del comportamiento, el experimento verdadero es la técnica más fuerte utilizada para alcanzar dicha meta. Cuando el experimento verdadero se estructura y ejecuta correctamente, proporciona al investigador una proposición de *causa y efecto* respecto a la relación entre X (variable independiente) y Y (variable dependiente). Esta se considera generalmente la forma más elevada de experimentación. No obstante, existen problemas de investigación en las ciencias del comportamiento, y especialmente en la investigación educativa, que no pueden estudiarse utilizando un diseño experimental verdadero; es decir, los diseños 20.1 a 20.6 y algunas de sus variantes, revisadas en los capítulos 20 y 21, no pueden utilizarse. Faltan uno o más

de los componentes de un experimento verdadero o se han debilitado, ya sea por la naturaleza del estudio o por una planeación pobre. El debilitamiento de los componentes del experimento verdadero constituye el tema que se analizará en el presente capítulo. Se examinarán dos tipos de diseños de investigación donde se ven comprometidos uno o más de los componentes del experimento verdadero. El primer tipo son los llamados diseños *cuasi-experimentales* y el segundo tipo son los conocidos como diseños *de un solo sujeto* o $N = 1$.

Diseños comprometidos, también conocidos como diseños cuasi-experimentales

Es posible, y de hecho necesario, utilizar diseños que estén comprometidos con la experimentación verdadera. Recuerde que la experimentación verdadera requiere por lo menos de dos grupos, uno que reciba un tratamiento experimental, y otro que no lo reciba o que lo reciba de forma diferente. El experimento verdadero requiere la manipulación de por lo menos una variable independiente, la asignación aleatoria de los participantes a los grupos y la asignación aleatoria del tratamiento a los grupos. Cuando falta uno o más de estos prerrequisitos por cualquier razón, se tiene un *diseño comprometido*. Los diseños comprometidos se conocen popularmente como diseños cuasi-experimentales. Se les llama *cuasi* porque dicho término significa "casi" o "tipo de". Cook y Campbell (1979) presentan dos principales clasificaciones del diseño cuasi-experimental. El primero se llama "diseño de grupo control no equivalente"; el segundo es el "diseño de series interrumpidas". Numerosos estudios de investigación que se realizan fuera del laboratorio podrían caer en una de tales categorías. Muchos estudios de investigación de mercado tienen la forma de diseños cuasi-experimentales. Con frecuencia se le pide a un investigador que "diseñe" y analice los datos de un estudio sin planeación. Por ejemplo, un comprador de abarrotes decide abastecer una nueva marca de alimento para bebés. Sus superiores se preguntan posteriormente si tal movimiento fue rentable; entonces el comprador consulta a un investigador de mercado para determinar lo que puede hacerse para demostrar si la decisión fue rentable o no. Dicho análisis no tendría la mejor selección ni asignación aleatorias, sólo consistiría de datos tomados a través del tiempo. Además, otros anuncios o la estación del año podrían influir en las ventas del alimento para bebés. El único componente que asemeja un experimento verdadero es el hecho de que se manipuló la variable independiente. No todas las tiendas recibieron ese producto en particular. Con tales problemas, el investigador optaría por el uso de diseños de investigación cuasi-experimentales o comprometidos.

Diseño de grupo control no equivalente

Quizás el diseño cuasi-experimental más utilizado es el de grupo experimental-grupo control, en el cual no se tiene mucha seguridad de que los grupos experimental y control sean equivalentes. Algunos autores como Cook y Campbell (1979), Christensen (1977), Ray (1977), y Graziano y Raulin (1993) se refieren a él como diseño de grupo control no equivalente. Cook y Campbell presentan ocho variaciones de este diseño, que ellos consideran "interpretables":

- diseños de grupo control sin tratamiento
- diseños de variables dependientes no equivalentes

diseños de grupo con retiro del tratamiento
diseños de tratamiento repetido
diseños de grupo control no equivalente con reversión del tratamiento
diseños cohorte
diseños sólo con postest
diseños de continuidad de regresión

En este libro se analizará en detalle sólo uno de ellos. Es el que tiene mayor posibilidad de ocurrir en alguna forma y variación en la literatura de investigación. Para un estudio más detallado de estos ocho tipos de diseños de grupo control no equivalente, se recomienda leer a Cook y Campbell (1979).

Diseño de grupo control sin tratamiento

La estructura del diseño de grupo control sin tratamiento ya se consideró en el diseño 20.3. Cook y Campbell (1979) se refieren a éste como el diseño de grupo control sin tratamiento con pretest y postest. La forma comprometida es como sigue:

Diseño 22.1: Diseño de grupo control sin tratamiento

Y_a	X	Y_d	(Experimental)
Y_a	$\sim X$	Y_d	(Control)

La diferencia entre el diseño 20.3 y el diseño 22.1 es marcada. En el diseño 22.1 no hay una asignación aleatorizada de los participantes a los grupos como en el 20.3(*a*), ni hay apareamiento de los participantes y luego asignación aleatoria como en el 20.3(*b*). Por lo tanto, el diseño 22.1 está sujeto a las debilidades debidas a la posible falta de equivalencia entre los grupos en variables distintas a X. Por lo común los investigadores sufren para establecer equivalencia por otros medios y, dependiendo del grado en que sean exitosos al hacerlo, el diseño será válido, lo cual se logra en formas que se analizarán a continuación.

En ocasiones es difícil o imposible igualar grupos por medio de selección o asignación aleatorias, o por medio del apareamiento. ¿Debe entonces renunciarse a llevar a cabo la investigación? Por ningún motivo. Deben realizarse todos los esfuerzos posibles para 1) seleccionar y 2) asignar aleatoriamente. Si ambas cuestiones no son posibles, quizá se puedan lograr el apareamiento y la asignación aleatoria. Si el apareamiento y la asignación aleatoria no son posibles, por lo menos debe hacerse el esfuerzo de utilizar muestras que provengan de la misma población o muestras que sean lo más similares posibles. Los tratamientos experimentales deben asignarse aleatoriamente y después debe verificarse la similitud de los grupos, utilizando cualquier información disponible (sexo, edad, clase social, etcétera). La equivalencia de los grupos puede verificarse utilizando las medias y las desviaciones estándar de los pretest: las pruebas-*t* y las pruebas-*F* sirven para este fin. Las distribuciones también deben verificarse. Aunque no se alcanza la seguridad ofrecida por la aleatorización, si todos estos aspectos resultan satisfactorios, entonces se puede continuar con un estudio, sabiendo por lo menos que no existe evidencia conocida en contra del supuesto de equivalencia.

Estas precauciones incrementan las posibilidades de conseguir validez interna. Aún existen dificultades, todas las cuales están subordinadas a una dificultad principal: la *selección*. Estas otras dificultades no se estudiarán aquí; para un análisis más detallado, véase Campbell y Stanley (1963), o Cook y Campbell (1979).

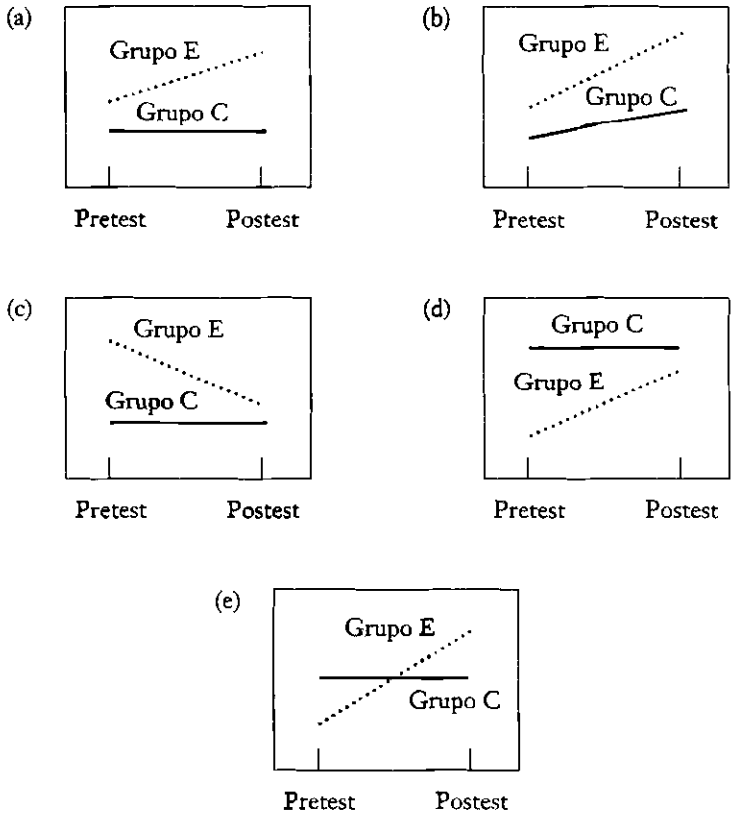
La selección constituye uno de los problemas más difíciles y complicados de la investigación del comportamiento. Puesto que sus aspectos se verán con detalle en el capítulo 23 que se trata sobre la investigación no experimental, aquí solamente se incluirá una breve descripción. Una de las razones importantes del énfasis en la selección y asignación aleatorias es evitar las dificultades de la selección. Cuando se integra a los participantes a los grupos con bases extrañas a los propósitos de investigación, a esto se le llama "selección" o, alternativamente, "autoselección". Considere un ejemplo común: suponga que los voluntarios se utilizan en el grupo experimental y otros participantes sirven como controles. Si los voluntarios difieren en una característica relacionada con Y, la variable dependiente, la diferencia última entre los grupos experimental y control quizá se deba a dicha característica, más que a X, la variable independiente (tratamiento). Los voluntarios pueden ser más (o menos) inteligentes que los no voluntarios. Si se realizara un experimento con cierto tipo de aprendizaje como variable dependiente, los voluntarios obviamente se desempeñarían mejor en Y debido a una inteligencia superior, a pesar de la semejanza inicial de los dos grupos en el pretest. Note que si se hubieran utilizado sólo voluntarios que se hubiesen asignado aleatoriamente a los grupos experimental y control, la dificultad de selección se disminuiría. Sin embargo, la validez externa o representatividad habría disminuido.

Cook y Campbell (1979) afirman que aun en los casos muy extremos es posible extraer conclusiones sólidas si se consideran y justifican todas las amenazas en contra de la validez. Sin el beneficio de la asignación aleatoria, deben llevarse a cabo intentos con otros medios para eliminar hipótesis rivales. Aquí se considera únicamente el diseño que utiliza el pretest debido a que éste ofrece información útil respecto a la efectividad de la variable independiente sobre la variable dependiente. El pretest puede proporcionar datos respecto a la igualdad entre los grupos, antes de la administración del tratamiento al grupo experimental.

Otro ejemplo más frecuente en investigación educativa consiste en tomar algunos grupos escolares para el grupo experimental y otros para el grupo control. Si se selecciona un número bastante grande de grupos, y se asignan aleatoriamente a los grupos experimental y control, entonces no hay mucho problema; pero si no se asignan aleatoriamente, algunos de ellos pueden asignarse a sí mismos a los grupos experimentales, y estos grupos quizá tengan características que los predisponen a tener puntuaciones medias de Y más altas que los otros grupos. Por ejemplo, sus maestros pueden estar más alertas, ser más inteligentes y más agresivos. Las características interaccionan con la selección para producir, independientemente de X, puntuaciones más altas para el grupo experimental que para el grupo control Y. En otras palabras, algo que influya en el proceso de selección (por ejemplo, participantes voluntarios), también influye en las medidas de la variable dependiente. Esto sucede aunque el pretest muestre que los grupos son iguales o similares respecto a la variable dependiente. La manipulación de X es "efectiva" debido a la selección o autoselección; pero no es efectiva por sí misma. Además, en ocasiones un investigador educativo necesita recibir la aprobación del distrito escolar para realizar la investigación. En ocasiones el distrito asignará las escuelas y los grupos que el investigador pueda usar.

Un estudio clásico de Sanford y Hemphill (1952), reportado por Campbell y Stanley (1963), utilizó este diseño. Este estudio se condujo en la U.S. Naval Academy en Annapolis, con el fin de saber si un curso de psicología en el currículum incrementaba la confianza de los estudiantes (guardias marinos) en las situaciones sociales. Los guardias marinos de segundo año fueron el primer grupo de estudiantes en tomar el curso de psicología. El grupo comparativo o control lo conformó la clase de tercer año, quienes no habían tomado el curso durante su segundo año. Se administró un cuestionario de situaciones sociales a ambas clases al inicio del año académico y al final del año. Los resultados demostraron

FIGURA 22.1 *Cinco posibles resultados del diseño de grupo control no equivalente*[a]



[a] E = grupo experimental y C = grupo control

un incremento en las puntuaciones de confianza en la clase de segundo año, de 43.26 a 51.42. La clase de tercer año también mostró un incremento; pero éste fue considerablemente menor, con un cambio de 55.80 a 56.78. A partir de estos datos se podría concluir que tomar el curso de psicología tuvo un efecto de incremento en la confianza de los sujetos en situaciones sociales. Sin embargo, también son posibles otras explicaciones. Una podría explicar que las mayores ganancias logradas por la clase de segundo año fueron el resultado de algún desarrollo maduracional que tiene su mayor crecimiento en el segundo año, y un menor crecimiento en el tercer año. Si dicho proceso existe, el mayor incremento en las puntuaciones de la clase de segundo año se hubiera dado aun si los guardias marinos no hubieran tomado la clase de psicología. El hecho de que la clase de segundo año iniciara con una puntuación más baja que la clase de tercer año, podría indicar que estos estudiantes no habían alcanzado todavía un nivel equivalente al de la clase de tercer año. Además, las puntuaciones del final del año de la clase de segundo año no fueron equivalentes a las puntuaciones iniciales de la clase de tercer año. Un mejor y más fuerte diseño consistiría en crear dos grupos equivalentes de la clase de segundo año, a

través de la selección aleatoria, e impartir la clase de psicología aleatoriamente a un solo grupo.

Resultados posibles de tal diseño se presentan en la figura 22.1. Existe la posibilidad de una interpretación diferente de la causalidad, según el resultado que obtenga el investigador. En la mayoría de los casos la amenaza más probable contra la validez interna sería la interacción selección-maduración. Quizá se recuerde que dicha interacción ocurre cuando 1) dos grupos son diferentes desde el inicio, de acuerdo a las medidas; 2) uno de los grupos experimenta mayores cambios diferenciales, como tornarse más experimentado, más preciso, más cansado, etcétera, que el otro grupo. La diferencia posterior al tratamiento, de acuerdo al postest, no puede atribuirse exactamente al tratamiento por sí mismo.

En la figura 22.1(a) existen tres amenazas posibles contra la validez interna. Como antes se mencionó, la amenaza con mayor prevalencia es la interacción selección-maduración. Para el resultado en la figura 22.1(a), Cook y Campbell (1979) afirman que hay cuatro explicaciones alternativas.

La primera es la interacción selección-maduración. Digamos que el estudio implica la comparación de dos estrategias o métodos de solución de problemas. El grupo E posee mayor inteligencia que el grupo C. El grupo E tiene puntuaciones mayores en el pretest que el grupo C. El grupo E muestra un incremento en las puntuaciones del postest después del tratamiento. El grupo C presenta poco o ningún cambio. Quizá parezca que el tratamiento que recibe el grupo E es superior al tratamiento recibido por el grupo C. Sin embargo, con la interacción selección-maduración, el incremento del grupo E puede deberse a su mayor nivel de inteligencia. Con un nivel más alto de inteligencia, tales participantes quizá pueden procesar más o quizá crezcan más rápido que los del grupo C.

Una segunda explicación se refiere a la instrumentación. La escala utilizada para medir la variable dependiente tal vez sea más sensible en ciertos niveles que en otros. Como ejemplo considere los percentiles, los cuales tienen una ventaja sobre las puntuaciones en bruto, pues transmiten un significado directo sin otras piezas de información. No obstante, los percentiles son transformaciones no lineales de las puntuaciones en bruto. En una distribución normal, los cambios en las puntuaciones en bruto cercanas al centro de la distribución reflejan cambios percentilares más grandes que en las colas. Un cambio de sólo 2 o 3 puntos en la escala de puntuación en bruto puede reflejar un cambio percentilar de 10 puntos cerca del centro de la distribución. Éste no sería el caso al considerar las colas de la distribución normal. Se necesitaría un cambio de 15 puntos de la puntuación en bruto para tener un incremento de 10 puntos percentilares en la cola de la distribución. Por lo tanto, las mediciones de los percentiles del grupo C pueden cambiar poco debido a que las mediciones no son lo suficientemente sensibles para detectar los cambios [en las colas]. Sin embargo, el grupo E mostrará una cantidad mayor de cambio, ya que su percentil está en la parte más sensible de la escala de medición.

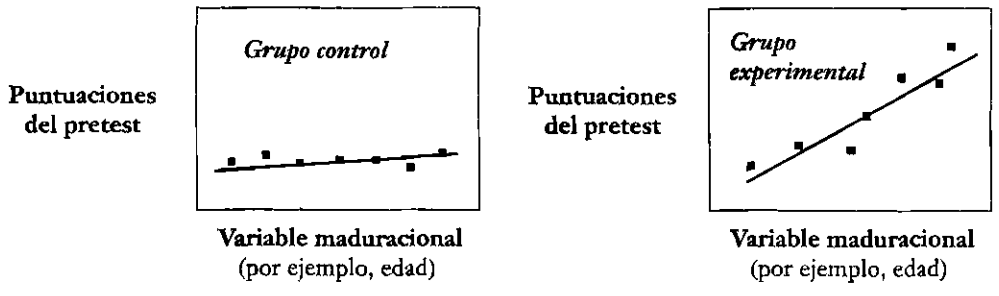
La tercera explicación es la regresión estadística. Digamos que los dos grupos, E y C, en realidad provienen de diferentes poblaciones, y que el grupo C es el grupo de interés. El investigador desea introducir un plan educativo para ayudar a incrementar el funcionamiento intelectual de estos participantes, quienes son seleccionados porque generalmente obtienen puntuaciones bajas en pruebas de inteligencia. El investigador crea un grupo comparativo o control a partir de estudiantes con puntuaciones normales. Este grupo se representa como grupo E en la figura 22.1(a). Estos estudiantes estarían en el extremo bajo de la escala de puntuaciones de la prueba; pero no tan abajo como el grupo C. Si ésta es la situación, entonces la regresión estadística constituye una explicación alternativa viable. El incremento de las puntuaciones del grupo E se deberían a su selección con base en las puntuaciones extremas. En el postest sus puntuaciones aumentarían pues estarían acercándose a la línea base de la población.

La cuarta explicación se centra en la interacción entre la historia y la selección. Cook y Campbell (1979) se refieren a lo anterior como el efecto de la historia local. En dicha situación, algo diferente a la variable independiente afectará a uno de los grupos (grupo E) y no al otro grupo (grupo C). Suponga que un investigador de mercado desea determinar la efectividad de un anuncio de condimento para sopa. Se reúnen datos sobre las ventas antes y después de la introducción del anuncio. Se utilizan dos grupos de diferentes regiones del país: un grupo es del sur de California y el otro es del oeste medio. En este caso, el crecimiento en las ventas observado en uno de los grupos (E) pudo no deberse necesariamente al anuncio. Ambos grupos pueden tener conductas de compra similares durante la primavera y el verano; es decir, que no hay una alta necesidad de los condimentos para sopa. No obstante, conforme se acerca la temporada de otoño, la venta de condimentos para sopa puede aumentar en el grupo del medio oeste. En el sur de California, donde las temperaturas son considerablemente más cálidas durante todo el año, la demanda de condimentos para sopa permanecería bastante constante. Por lo tanto, la explicación del incremento en las ventas en el medio oeste sería la estación del año y no el anuncio.

Todas las amenazas mencionadas respecto a la figura 22.1(a) también resultan verdaderas para la figura 22.1(b). Mientras que en la figura 22.1(a) uno de los grupos (grupo C) permanece constante, en la figura 22.1(b) ambos grupos experimentan un incremento del pretest al postest. La interacción selección-maduración aún es una posibilidad ya que, por definición, los grupos están creciendo (o disminuyendo) a diferente ritmo, pues el grupo de puntuaciones más bajas (grupo C) progresa a un ritmo menor que el grupo de altas puntuaciones (grupo E). Para determinar si la selección y la maduración juegan un papel importante en los resultados, Cook y Campbell (1979) recomiendan dos métodos. El primero implica observar únicamente los datos del grupo experimental (grupo E). Si la varianza dentro de grupos del postest es considerablemente mayor que la varianza dentro de grupos del postest, entonces hay evidencia de una explicación por la interacción selección-maduración. El segundo método consiste en trazar dos gráficas y la línea de regresión asociada con cada gráfica. Una gráfica es para el grupo experimental (grupo E). Las puntuaciones del pretest se grafican contra la variable de maduración, que puede ser la edad o la experiencia. La segunda gráfica sería igual, excepto que sería para el grupo control (grupo C). Si las pendientes de la línea de regresión de cada gráfica difieren entre sí, entonces existe evidencia de un ritmo promedio de crecimiento diferencial, lo cual significaría la posibilidad de una interacción selección-maduración (véase figura 22.2).

El resultado que se presenta en la figura 22.1(c) se encuentra con mayor frecuencia en los estudios de psicología clínica. Se supone que el tratamiento resultará en una disminución de una conducta indeseable. Como los dos resultados previos, éste también es susceptible de la interacción selección-maduración, de la regresión estadística, de la

FIGURA 22.2 *Comparación del grupo experimental y el grupo control*



instrumentación y de los efectos de la historia local. En este resultado, la diferencia entre los grupos experimental y control es muy dramática en el pretest, pero después del tratamiento los grupos se acercan entre sí.

El cuarto resultado se presenta en la figura 22.1(d). Éste difiere de los tres previos en que el grupo control (grupo C) inicia más alto que el grupo experimental (grupo E) y permanece más alto incluso en el postest. Sin embargo, el grupo E mostró un mayor incremento del pretest al postest. La regresión estadística sería una amenaza si los participantes del grupo E fuesen seleccionados con base en su puntuación extremadamente baja. Sin embargo, Cook y Campbell (1979) afirman que la amenaza de la selección-maduración puede excluirse, ya que este efecto generalmente resulta en un ritmo de crecimiento más lento en aquellos con puntuaciones bajas, y en un ritmo de crecimiento más rápido en aquellos con puntuaciones altas. Aquí los participantes con bajas puntuaciones muestran mayor ganancia en las puntuaciones que los participantes con puntuaciones altas. Esta evidencia confiere apoyo a la efectividad de la condición de tratamiento recibida por el grupo E. Lo que no puede eliminarse fácilmente son las amenazas de la instrumentación y de la historia local, vistas en los tres resultados previos de los diseños de grupo control no equivalente.

Con el resultado final mostrado en la figura 22.1(e), las medias de los grupos experimental (grupo E) y control (grupo C) son significativamente diferentes entre sí, tanto en el pretest como en el postest. No obstante, las diferencias resultan en dirección opuesta en el pretest, respecto al postest. Las líneas se cruzan entre sí. El grupo E inicia abajo, pero después supera al grupo C, que inicialmente tuvo puntuaciones altas. Cook y Campbell (1979) consideraron este resultado más interpretable que los cuatro anteriores. La instrumentación o la escalación se descarta, ya que ninguna transformación de las puntuaciones podría suprimir o reducir este cruce o efecto de interacción. La regresión estadística se vuelve insostenible porque es extremadamente raro que una puntuación baja tenga suficiente regresión para superar una puntuación inicialmente alta. Además de un efecto muy complicado de interacción de selección-maduración, dicho patrón no se asemeja a las amenazas por selección-maduración. La maduración, por ejemplo, generalmente no inicia diferente, se cruza y luego continúa en la dirección opuesta. Por lo tanto, el resultado de la figura 22.1(e) parece ser el más fuerte y debe permitir al investigador hacer proposiciones causales respecto al tratamiento. Sin embargo, Cook y Campbell advierten que los investigadores no deben planear el desarrollo de la investigación cuasi-experimental con la esperanza de obtener este resultado. En definitiva el diseño de un estudio de grupo control no equivalente debe realizarse con cuidado y precaución.

Ejemplos de investigación

Nelson, Hall y Walsh-Bowers: diseño de grupo control no equivalente

La investigación realizada por Nelson, Hall y Walsh-Bowers (1997) afirma específicamente que ***utilizaron un diseño de grupo control no equivalente*** para comparar los efectos a largo plazo de los apartamentos de apoyo (AA), hogares de grupo (HG) y los hogares de alojamiento y cuidado (HAC) para residentes psiquiátricos. Los apartamentos de apoyo y los hogares de grupo son manejados por organizaciones no lucrativas; los hogares de alojamiento y cuidado sí se manejan con fines lucrativos. El objetivo principal fue comparar los dos grupos de intervención: los apartamentos de apoyo y los hogares de grupo. Los autores no pudieron asignar aleatoriamente a los participantes a los diferentes hogares. Nelson *et al.*, realizaron su mayor esfuerzo para aparear a los residentes; pero existían algunas diferencias significativas en la composición de los grupos que los condujeron a

utilizar el diseño de grupo control no equivalente. Con este diseño decidieron usar a residentes de HAC como grupo comparativo. No fueron capaces de corregir, por medio del apareamiento, las siguientes variables que pudieron tener un efecto sobre las variables dependientes: 1) Los grupos AA y HG tendían a ser más jóvenes que el grupo HAC (33 años contra 45 años) y tenían menos tiempo de residencia (2.5 años contra 39 años). 2) Los residentes de AA y HG tenían un mayor nivel de educación que los del grupo de HAC. Nelson y sus colaboradores encontraron una diferencia significativa entre estos grupos en esas variables. Aunque el género no resultó significativo, había más hombres que mujeres en los grupos de AA y HG; y más mujeres que hombres en el grupo HAC.

Nelson *et al.* proponen que las diferencias que encontraron entre estos tres grupos en las medidas del postest, quizá se deban al problema de selección y no al tipo de instalación para el cuidado.

Chapman y McCauley: cuasi-experimento

En este estudio, Chapman y McCauley (1993) examinaron el desarrollo de la carrera de estudiantes graduados que solicitaron una beca para graduados de la National Science Foundation (NSF) Graduate Fellowship Award. Aunque quizás se puede pensar que este estudio es no experimental, Chapman y McCauley consideraron que podía clasificarse como cuasi-experimental; debe entenderse por qué. Al comparar a los ganadores de la beca con los no ganadores, la elección de ganadores no se realizó exactamente al azar. El estudio no consideró a los solicitantes del grupo de calidad 1. Los solicitantes del grupo 1 estaban dentro del 5% más alto y todos recibieron becas. Los solicitantes al NSF del grupo de calidad 2 conformaron el siguiente 10% y fueron considerados como un grupo altamente homogéneo. Las becas fueron otorgadas aproximadamente a la mitad del grupo homogéneo de solicitantes, por medio de un procedimiento que Chapman y McCauley consideran como una asignación aproximadamente aleatoria de la beca o de la mención honorífica. Se asignó a los estudiantes con respecto al potencial académico. Chapman y McCauley consideraron que las diferencias en el desempeño entre los solicitantes del grupo de calidad 2, donde algunos estudiantes fueron adjudicados con una beca NSF y otros no, podrían revelar el efecto de las expectativas positivas asociadas con esta prestigiosa beca.

Los resultados mostraron que quienes recibieron la beca NSF tenían mayores posibilidades de terminar el doctorado. Sin embargo, Chapman y McCauley no encontraron un efecto confiable de la beca sobre el logro de estatus dentro de la facultad o sobre el solicitar o recibir un reconocimiento del NSF o una beca de investigación de los institutos nacionales de salud. Parece que las expectativas positivas asociadas con esta prestigiosa beca tienen alguna influencia en las escuelas de posgrado; pero no la tienen en los logros posteriores a la escuela de posgrado.

Diseños de tiempo

Variantes importantes del diseño cuasi-experimental básico son los diseños de tiempo. La forma del diseño 20.6 puede alterarse para incluir un lapso de tiempo:

Y_a	X	Y_d	
Y_a	$\sim X$	Y_d	
	X		Y_d
	$\sim X$		Y_d

Las Y_2 de la tercera y cuarta líneas son observaciones de la variable dependiente en cualquier fecha posterior específica. Dicha alteración, por supuesto, cambia el propósito del diseño y puede causar que se pierdan algunas de las virtudes del diseño 20.6. Es posible, si se tiene el tiempo, la paciencia y los recursos, mantener todos los beneficios anteriores y aun extender el tiempo añadiendo dos o más grupos al propio diseño 20.6.

Un problema común de investigación, especialmente en estudios sobre el desarrollo y crecimiento de los niños, incluye el estudio de individuos y de grupos utilizando el tiempo como variable. Éstos son estudios longitudinales de los participantes, con frecuencia niños, en diferentes puntos de tiempo. Un diseño entre muchos podría ser:

Diseño 22.2: Un diseño de tiempo longitudinal (también conocido como diseño de series de tiempo interrumpidas)

Y_1	Y_2	Y_3	Y_4	X	Y_5	Y_6	Y_7	Y_8

Observe la similitud con el diseño 19.2, donde un grupo es comparado consigo mismo. El uso del diseño 22.2 permite evitar una de las dificultades del diseño 19.2; su empleo hace posible separar los efectos reactivos de medición del efecto de X. Permite determinar si las mediciones tienen un efecto reactivo y si X fue lo suficientemente fuerte para superar tal efecto. El efecto reactivo debe mostrarse a sí mismo al comparar Y_3 con Y_4; lo cual puede contrastarse con Y_5. Si existe un incremento en Y_5 por encima del incremento en Y_4, a partir de Y_3, puede atribuirse a X. Un argumento similar se aplica para la maduración y la historia.

Una dificultad con los estudios longitudinales o de tiempo, especialmente con niños, es el crecimiento o el aprendizaje que ocurre de manera natural a través del tiempo: los niños no detienen su crecimiento ni su aprendizaje para conveniencia de la investigación. A mayor periodo, mayor será el problema. En otras palabras, el tiempo en sí mismo es una variable. Con un diseño como el 20.2, $Y_a X Y_d$, la variable tiempo puede confundir a X, la variable independiente experimental. Si existe una diferencia significativa entre Y_a y Y_d, no es posible decir si X o una "variable" de tiempo provocó el cambio. Pero con el diseño 22.2 se tienen otras medidas de Y y, por lo tanto, una línea base con la cual comparar el cambio en Y, presumiblemente debido a X.

Un método para determinar si el tratamiento experimental tuvo un efecto consiste en observar una gráfica de los datos a través del tiempo. Caporaso (1973) presenta varios patrones de conducta adicionales posibles, los cuales se obtienen de datos de series de tiempo. Ya sea que un cambio significativo en la conducta venga o no después de la introducción de la condición de tratamiento, esto se determina por medio de una prueba de significancia. La prueba estadística más utilizada es ARIMA (promedio autorregresivo, integrado y móvil), desarrollada por Box y Jenkins (1970) (véase también Gottman, 1981). Este método consiste en determinar si el patrón de medidas postrespuesta difiere del patrón de medidas prerrespuesta. El uso de dicho análisis estadístico requiere la disponibilidad de muchos puntos de datos.

El análisis estadístico de medidas de tiempo representa un problema complicado y especial: las pruebas comunes de significancia que se aplican para medidas de tiempo pueden generar resultados falsos. Una razón de esto es que tales datos tienden a ser altamente variables y es fácil interpretar equívocamente cambios que no se deben a X, como si lo fueran. Es decir, con datos de tiempo las puntuaciones individuales y medias tienden a moverse bastante. Es fácil caer en la trampa de considerar uno de estos cambios como "significativo", en especial si va de acuerdo con la hipótesis. Si es posible suponer legítimamente que otras influencias, diferentes a X —ambas aleatorias y sistemáticas— son

uniformes sobre todas las series de Y, entonces el problema estadístico puede resolverse. Pero dicho supuesto puede ser, y con frecuencia es injustificado.

El investigador que explora estudios de tiempo debe estudiar con especial cuidado los problemas estadísticos y consultar a un especialista en estadística. Para el practicante la complejidad estadística resulta desafortunada, ya que quizá desmotive la realización de estudios prácticos necesarios. Puesto que los diseños longitudinales de un solo grupo se adaptan particularmente bien con la investigación de clase individual, se recomienda que en estudios longitudinales de métodos o estudios de niños en situaciones educativas, el análisis sea confinado al trazado de gráficas de resultados, y a su interpretación cualitativa. No obstante, las pruebas cruciales, especialmente aquellas para estudios publicables, deben reforzarse con pruebas estadísticas.

Diseño de series de tiempo múltiples

El diseño de series de tiempo múltiples es una extensión del diseño de series de tiempo interrumpidas. Con el diseño de series de tiempo interrumpidas solamente se utilizó un grupo de participantes; como resultado, las explicaciones alternativas pueden provenir de un efecto de la historia. El diseño de series de tiempo múltiples tiene la ventaja de que elimina el efecto de la historia al incluir un grupo control compuesto de un grupo equivalente de participantes —o por lo menos comparable— que no recibe la condición de tratamiento. Lo anterior se presenta en el diseño 22.3, donde un grupo experimental recibe la condición de tratamiento y el grupo control no. Como consecuencia, el diseño ofrece un mayor grado de control sobre las fuentes de explicaciones alternativas o hipótesis rivales. Los efectos de la historia, por ejemplo, se controlan debido a que ejercerían la misma influencia en el grupo experimental que en el grupo control.

Diseño 22.3: Un diseño de series de tiempo múltiples

Y_1	Y_2	Y_3	Y_4	X	Y_5	Y_6	Y_7	Y_8	(Experimental)
Y_1	Y_2	Y_3	Y_4		Y_5	Y_6	Y_7	Y_8	(Control)

Existen, naturalmente, otras variaciones posibles del diseño 22.2, además del diseño 22.3. Una variación importante consiste en añadir uno o más grupos control; otra sería añadir más observaciones de tiempo. Otra más consistiría en agregar más X, más intervenciones experimentales (véase Gottman, 1981; Gottman, McFall y Barnett, 1969; Campbell y Stanley, 1963).

Diseños experimentales de un solo sujeto

La mayoría de la investigación del comportamiento actual implica el uso de participantes. Sin embargo, existen otros métodos. En esta sección se analizan las estrategias para lograr control en los experimentos por medio del uso de uno o de unos pocos participantes. Estos diseños de un solo sujeto algunas veces se llaman diseños $N = 1$. Los diseños de un solo sujeto son una extensión del diseño de series de tiempo interrumpidas. Mientras las series de tiempo interrumpidas generalmente observan un grupo de individuos a través del tiempo (por ejemplo, niños), el estudio de un solo sujeto utiliza únicamente un participante o, cuando mucho, pocos participantes. Aun cuando se utilicen pocos participantes, cada uno es estudiado individual y extensamente; éstos también se llamarán

diseños o estudios de un solo sujeto. Aunque tengan diferentes nombres, todos comparten las siguientes características:

- Solamente se utilizan uno o pocos participantes en el estudio.
- Cada sujeto participa en varios ensayos (medidas repetidas), lo cual es similar a los diseños dentro de participantes que se describieron en el capítulo 21.
- Los procedimientos de aleatorización (por ejemplo, asignación aleatoria y/o selección aleatoria) se utilizan en muy raras ocasiones. En su lugar, las mediciones repetidas o intervalos de tiempo se asignan aleatoriamente a las diferentes condiciones de tratamiento.

Estos diseños observan el comportamiento del organismo antes del tratamiento experimental y utilizan las observaciones como una medida de la línea base. Las observaciones realizadas después del tratamiento se comparan, posteriormente, con las observaciones de la línea base; el participante sirve como su propio control. Estos diseños por lo común se aplican en investigación escolar, clínica y de asesoría. Se utilizan para evaluar los efectos de intervenciones conductuales a través del tiempo. Este tipo de investigación es popular entre quienes realizan experimentos sobre aprendizaje operante o modificación conductual.

La investigación con participantes únicos no es nueva, como lo ilustró Gustav Fechner, quien desarrolló la disciplina de la psicofísica en la década de 1860, utilizando solamente dos participantes: él y su cuñado. A Fechner se le acredita como el inventor de los métodos psicofísicos básicos que aun hoy en día se utilizan para medir los umbrales sensoriales. Fechner ejerció una fuerte influencia en Hermann Ebbinghaus, conocido por su trabajo experimental sobre la memoria. Ebbinghaus también se utilizó a sí mismo como sujeto. Wilhelm Wundt, considerado como el fundador del primer laboratorio psicológico en 1879, condujo experimentos donde medía varias respuestas psicológicas y de conducta en participantes individuales. I. P. Pavlov realizó su trabajo pionero sobre el condicionamiento instrumental utilizando perros de forma individual. La lista de psicólogos que han utilizado participantes únicos es extensa, la mayoría de los cuales lo hicieron antes de 1930 y antes del advenimiento del trabajo de R. A. Fisher y William Sealy Gossett sobre estadística moderna.

Los científicos del comportamiento que realizaron investigación antes del desarrollo de la estadística moderna intentaron resolver el problema de la confiabilidad y de la validez llevando a cabo extensas observaciones y réplica frecuentes de los resultados. Éste es un procedimiento tradicional utilizado por los investigadores que conducen experimentos con un solo sujeto. El supuesto es que los participantes individuales son, en esencia, equivalentes y que se requiere estudiar participantes adicionales tan sólo para asegurarse de que el sujeto original estaba dentro de la norma.

La popularidad del trabajo de Fisher sobre el análisis de varianza, y el estudio de Gossett sobre la prueba *t* de Student, abrieron el camino para la metodología de investigación orientada a grupos. Algunos afirman que estos trabajos fueron tan populares que la tradición del sujeto único casi se extinguió. De hecho, incluso en el mundo actual, existen criterios de contratación en las principales universidades, que dependen de si el candidato es un científico orientado hacia la investigación de grupos o un investigador orientado hacia los diseños de participantes únicos. A pesar de la popularidad de los métodos de Fisher y de la investigación orientada a grupos, algunos psicólogos continúan trabajando en la tradición del sujeto único. El más notable fue Burrus Frederick Skinner. Skinner se abstiene de utilizar estadística inferencial; no recomienda el uso de estadística inferencial compleja. En cambio considera que es posible demostrar la eficacia del tratamiento al graficar las acciones del comportamiento del organismo a través del tiempo; él llamó a

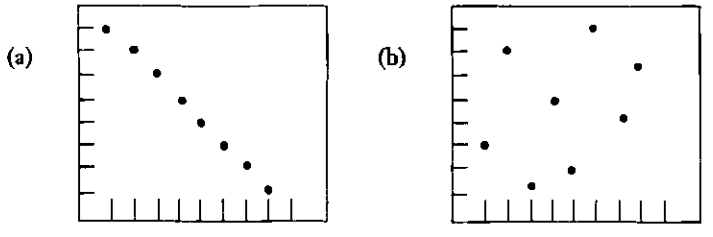
esto el registro acumulativo. Algunos, como E. L. Thorndike, le llaman "curva de aprendizaje". Skinner considera que es más útil estudiar un animal durante 1 000 horas, que estudiar 1 000 animales durante una hora. En su libro clásico, Sidman (1960) describe la filosofía de investigación de Skinner y hace una clara distinción entre el método de investigación de un solo sujeto y el método de investigación de grupo. El primero supone que la varianza del comportamiento del sujeto está dictada por la situación; como resultado, esta varianza puede eliminarse a través de un control experimental cuidadoso. El segundo método, en cambio, supone que la mayoría de la variabilidad es inherente y puede controlarse y analizarse estadísticamente.

Algunas ventajas de los estudios de un solo sujeto

La investigación orientada hacia grupos generalmente incluye el cálculo de la media o de alguna otra medida promedio o de tendencia central; pero los promedios pueden confundir. Observe los gráficos a) y b) en la figura 22.3 —ambos tienen exactamente los mismos valores—. Si se calculara la media de los datos de cada figura, se encontraría que son exactamente iguales. Incluso, si se calcularan la desviación estándar o la varianza, resultaría que las dos medidas de variación son exactamente las mismas. No obstante, una inspección visual de los datos muestra que el gráfico 22.3 a), exhibe una tendencia, mientras que el gráfico 22.3 b) no la muestra. De hecho, el gráfico 22.3 b) presenta lo que parece ser un patrón aleatorio. El método de un solo sujeto no tiene este problema, ya que se estudia a un participante de forma extensa a través del tiempo. El registro acumulativo de ese participante muestra el desempeño real de dicho participante.

Uno de los principales problemas del uso de muestras grandes es que la significancia estadística se logra por medio de diferencias muy pequeñas. Con la estadística inferencial, una muestra grande tenderá a reducir la cantidad de varianza del error. Tome la prueba t como ejemplo. Aun cuando la diferencia entre las medias permanezca igual, el incremento en el tamaño de la muestra tenderá a disminuir el error estándar. Con la reducción del error estándar, el valor de t se vuelve más grande, incrementando así su posibilidad de significancia estadística. Sin embargo, la significancia estadística y la significancia práctica son dos cuestiones diferentes. El experimento puede tener poca significancia práctica aun cuando tenga enorme significancia estadística. Simon (1987) criticó el uso indiscriminado de grupos grandes de participantes, pues los considera un desperdicio e incapaces de producir información útil. Simon recomienda el uso de experimentos de rastreo para encontrar las variables independientes con el mayor efecto sobre la variable dependiente; éstas serían las variables poderosas que producen grandes efectos. Simon no apoya exactamente los diseños de un solo sujeto; él recomienda el uso de diseños bien construidos, con el número necesario de participantes para encontrar los mayores efectos. Simon (1976) se refiere a esto como los "diseños económicos multifactoriales". Los investigadores de un solo sujeto, por otro lado, favorecen el incremento del tamaño del efecto en lugar de

FIGURA 22.3 *Comparación del grupo experimental y el grupo control*



intentar reducir la varianza del error, pues consideran que esto puede realizarse por medio de un control más rígido del experimento.

En la misma línea, los diseños de un solo sujeto tienen la ventaja, sobre los diseños orientados a grupos, de que con sólo unos cuantos participantes los investigadores pueden probar diferentes tratamientos. En otras palabras, determinan la eficacia o la no eficacia de una intervención de tratamiento sin emplear un número grande de participantes, lo cual puede resultar costoso.

Con los estudios de un solo sujeto, el investigador puede evitar algunos de los problemas éticos que enfrentan los investigadores orientados a grupos. Uno de dichos problemas éticos se refiere al grupo control, en el cual algunas situaciones no recibe ningún tratamiento real. Aunque en la mayoría de los estudios realizados hoy en día a los participantes del grupo control no se les daña en forma alguna, aún existen algunas cuestiones éticas. Considérese como ejemplo el estudio de Gould y Clum (1995) que buscaba determinar si la autoayuda, con mínimo contacto con el terapeuta, es efectiva en el tratamiento del trastorno de pánico. Todos los participantes de este estudio sufrían de ataques de pánico y fueron asignados aleatoriamente tanto al grupo experimental como al grupo control. El grupo experimental recibió material de autoayuda. El grupo control "no recibió tratamiento durante el transcurso del experimento" (p. 536). En su lugar, al grupo control se le indicó que estaban en la lista de espera para el tratamiento.

En el estudio de cierto tipo de individuos el tamaño de la población es pequeño y, por lo tanto, sería difícil adecuar el muestreo u obtener suficientes participantes para el estudio. De hecho, el estudio de Strube (1991) indica que incluso el muestreo aleatorio tiende a fallar cuando se utilizan muestras pequeñas. Si no se tienen disponibles suficientes participantes con ciertas características para el estudio, entonces el investigador puede considerar el uso de diseños de un solo sujeto, en lugar de abandonar el estudio. Simon (1987) cita el estudio que intentaron realizar Adelson y Williams, en 1954, sobre los parámetros de entrenamiento importantes en la educación de los pilotos. El estudio fue abandonado pues había demasiadas variables que considerar y no se tenían suficientes participantes. Simon señaló que el estudio pudo haberse realizado, pero no con el uso de la metodología tradicional orientada hacia grupos.

Algunas desventajas del diseño de un solo sujeto

Los estudios de un solo sujeto no están exentos de problemas o limitaciones. Algunas de ellas se harán más notorias cuando se analicen los tipos de diseños de un solo sujeto. Uno de los problemas más generales del paradigma de un solo sujeto es la validez externa. Algunos encuentran difícil creer que los hallazgos de un estudio que utilice un sujeto (o quizá tres o cuatro) puedan generalizarse a la población entera.

Con ensayos repetidos en un participante puede cuestionarse si el tratamiento sería igualmente eficaz para un participante que no ha experimentado tratamientos previos. Si se habla de un tratamiento terapéutico, entonces la eficacia quizá radique en la acumulación de sesiones, en lugar de una sola sesión. La persona que está en el ensayo *enésimo* puede ser una persona muy diferente de la que se encuentra en el primer ensayo. Es aquí donde la investigación orientada hacia grupos puede eliminar este problema; el tratamiento se aplica a cada persona solamente una vez.

Los estudios de un solo sujeto son quizás aún más sensibles a las aberraciones por parte del experimentador y del participante. Dichos estudios son eficaces sólo cuando el investigador puede evitar sesgos y el participante está motivado y coopera. El investigador puede mostrar una tendencia a observar sólo ciertos efectos y a ignorar otros. Se analizó el caso de Blondlot en este libro; él era el único científico capaz de ver los "rayos-N". No era tanto cuestión de que fuera un fraude, sino de que estaba sesgado hacia ver algo que no

estaba ahí. Un investigador que hace investigación con un solo sujeto se ve afectado más de esta forma que el investigador orientado a grupos, y requiere desarrollar un sistema de verificación y balances para evitar esta dificultad.

Numerosas investigaciones requieren, por naturaleza, seguir métodos orientados a grupos y, como tales, serían impropios para diseños de un solo sujeto. Por ejemplo, para estudiar el comportamiento de miembros del jurado se requeriría el uso de grupos y la influencia de la dinámica de grupos. Antes se analizó la investigación alrededor del pensamiento grupal. El estudio de este importante fenómeno fue mejor realizado con grupos, ya que fue el grupo como un todo el que mostró dicho fenómeno.

Algunos paradigmas de la investigación de un solo sujeto

La línea base estable: una meta importante

En un diseño orientado hacia grupos, un grupo de participantes se compara con otro; o se compara un grupo de participantes que recibe una condición con el mismo conjunto de participantes que reciben una condición diferente. Se supone que los grupos son iguales antes de la aplicación del tratamiento de tal manera que, si la variable dependiente difiere después del tratamiento, se pueden asociar tales diferencias con el tratamiento. La determinación de un tratamiento eficaz se realiza al comparar estadísticamente la diferencia entre los dos grupos respecto a alguna variable resultante. Sin embargo, cuando se utiliza un solo sujeto, debe emplearse una táctica diferente. En la situación de un sujeto es necesario comparar el comportamiento que ocurre antes con el comportamiento que ocurre después de la introducción de una intervención experimental. El comportamiento previo a la intervención del tratamiento debe medirse durante un periodo lo suficientemente grande para poder obtener una línea base estable. Dicha línea base o nivel operante es importante porque se compara con el comportamiento posterior. Si la línea base varía de manera considerable, entonces sería más difícil evaluar cualquier cambio confiable en el comportamiento después de la intervención. El problema de la línea base con los diseños de un solo sujeto es importante. Para encontrar una descripción completa del problema y de sus posibles soluciones, se debe consultar a Barlow y Hersen (1984). Otra excelente referencia es Kazdin (1982).

Un ejemplo donde las medidas de línea base son muy importantes es el uso de un polígrafo (detector de mentiras). Aquí, el operador obtiene mediciones fisiológicas de la persona (sospechoso). El operador formula ciertas preguntas al sospechoso, cuya respuesta se sabe cierta (nombre, color de ojos, lugar de nacimiento, etcétera). Las respuestas emitidas se registran y se utilizan como medida de línea base para las respuestas honestas. Se toma otra línea base para respuestas conocidas como falsas: se le indica al sospechoso mentir deliberadamente a las preguntas planteadas. Después del establecimiento de estas dos líneas base, se plantea la pregunta de importancia (v.g., ¿cometió usted el crimen?) y se compara con las dos líneas base. Si la respuesta en el polígrafo se asemeja a la línea base de mentir, entonces se considera que el sospechoso mintió.

Diseños que utilizan el retiro del tratamiento

El diseño ABA

El diseño ABA incluye tres grandes pasos. El primero consiste en establecer una línea base estable (A). En el segundo paso (B) se aplica la intervención experimental al participante.

Si el tratamiento es efectivo, habrá una respuesta diferente a la de la línea base. Para determinar si la intervención del tratamiento causó el cambio en el comportamiento, el investigador lleva a cabo el paso tres: un regreso a la línea base (A). El tercer paso se requiere porque no se sabe cuál habría sido la tasa de respuesta si el participante no recibiera tratamiento. También se necesita saber si el cambio en la respuesta se debió a la intervención del tratamiento o a algo más.

Un problema importante del diseño ABA es que el efecto de la intervención puede no ser completamente reversible. Si el tratamiento implicó una cirugía, donde se removió el hipotálamo o se seccionó el cuerpo calloso, sería imposible revertir estos procedimientos. Un método de aprendizaje que provoque algún cambio permanente en el comportamiento del participante no sería reversible.

Existen también algunas consideraciones éticas respecto a regresar al paciente al estado original, si tal estado fuese un comportamiento indeseable (Tingstrom, 1996). Los experimentos en modificación conductual rara vez regresan al participante a la línea base. Este regreso a la línea base se llama condición de retiro. Para beneficiar a los participantes, se reintroduce el tratamiento. El diseño ABAB hace esto.

Repetición de tratamientos (diseño ABAB)

Existen dos versiones del diseño ABAB. El primero se describió brevemente en la sección anterior. El diseño ABAB es igual al diseño ABA, excepto que el tratamiento se reintroduce al participante y éste deja el estudio después de lograr cierto nivel benéfico. La repetición del tratamiento también proporciona al experimentador información adicional sobre la fortaleza de la intervención del tratamiento. El hecho de demostrar que la intervención del tratamiento puede llevar al participante al nivel de beneficio previo, después de regresar a la persona a la línea base, da fuerza a la afirmación de que el tratamiento causó el cambio en el comportamiento; es decir, brinda evidencia de validez interna. El diseño ABAB esencialmente produce el efecto experimental dos veces.

La segunda variante del diseño ABAB es el llamado diseño de *tratamientos alternantes*. En esta variante no se toma la línea base. A y B en este diseño son dos tratamientos diferentes que se alternan aleatoriamente. El objetivo de este diseño consiste en evaluar la eficacia relativa de las dos intervenciones de tratamiento. A y B pueden ser dos métodos diferentes para controlar la alimentación en exceso. Cada tratamiento se aplica al participante en diferentes momentos. Después de un periodo, un método puede emerger como más efectivo que el otro. La ventaja que tiene este diseño sobre el primer diseño ABAB es que no se requiere obtener una línea base y el participante no está sujeto a procedimientos de retiro. Puesto que este método implica la comparación de dos conjuntos de series de datos, algunos lo llaman diseño entre series.

Existen otras variantes interesantes del diseño ABAB, donde no se lleva a cabo el retiro del tratamiento. McGuigan (1996) lo llama el diseño ABCB. En la tercera fase de este diseño, el organismo recibe una condición "placebo". La condición placebo es esencialmente un método diferente.

Los diseños de un solo sujeto se diferencian de los diseños de grupo en que sólo permiten que el investigador varíe una variable a la vez. El investigador no sería capaz de determinar qué variable o qué combinación de ellas causó los cambios en la respuesta, si dos o más variables se alteraron simultáneamente. Lo mejor que cualquiera puede hacer es afirmar que la combinación de las variables condujo al cambio. Sin embargo, el investigador será incapaz de determinar cuál o qué tanto de cada una; si hay dos variables, llamadas B y C, y la línea base es A, entonces una posible secuencia de presentación de las condiciones sería A-B-A-B-BC-B-BC. En dicha secuencia cada condición es precedida y procedida por la misma condición una vez por lo menos, con una sola variable cambiando a la vez.

El diseño A-B-A-B-BC-B-BC con frecuencia se denomina un diseño de interacción. Sin embargo, no están presentes todas las combinaciones posibles de B y C. La condición C nunca ocurre sola (A representa la ausencia de B y C). Esta interacción difiere de las interacciones analizadas en el capítulo sobre diseños factoriales. Lo que se prueba con este procedimiento es si C se añade o no al efecto de B.

En un experimento de aprendizaje que utilice este diseño se podría examinar el efecto de elogiar a un estudiante por dar la respuesta correcta (C) a una pregunta sobre geografía, aunado a un punto meritorio (B). Si se descubre que el elogio, junto con el punto meritorio tienen un mayor efecto que el punto meritorio solo, se tiene información útil para diseñar una situación de aprendizaje para éste y otros estudiantes; pero no se conocerá el efecto singular del elogio. Utilizar únicamente el elogio podría haber sido tan efectivo como el punto meritorio más el elogio; o quizá emplear tan sólo el elogio hubiera tenido poco o ningún efecto. Sin embargo, es posible evaluar el elogio extendiendo el diseño de un solo sujeto: la secuencia A-B-A-B-BC-B-BC-C-BC. Pero extender un experimento de un solo sujeto de este tipo acarrea otros problemas; por ejemplo, un sujeto puede experimentar fatiga o perder el interés. Como resultado, una sesión demasiado larga quizá no produzca información útil, aunque el diseño parezca correcto.

Un ejemplo de investigación

Powell y Nelson: ejemplo de un diseño ABAB

Este estudio de Powell y Nelson (1997) incluyó un participante, Evan, un niño de 7 años de edad, diagnosticado con trastorno por déficit de atención con hiperactividad (TDAH). Evan recibía 15 mg de Ritalin® al día. La mayor parte de su comportamiento en el salón de clases se consideraba indeseable; también tenía relaciones pobres con sus compañeros y no comprendía su trabajo escolar. Las conductas indeseables incluían falta de obediencia, abandono de su escritorio, molestar a otros, iniciar las actividades a destiempo y no realizar su trabajo. Los datos se recolectaron a través de las interacciones entre Evan y su maestro.

El tratamiento consistió en permitir a Evan elegir las materias en las cuales deseaba trabajar. Había dos condiciones: elección y no elección. Los datos de la línea base fueron recolectados durante la fase de no elección; a Evan se le impartió la misma materia que al resto de la clase. Durante las fases de elección el maestro le presentó a Evan tres materias diferentes y él eligió una. Las opciones eran materias idénticas en longitud y dificultad y sólo variaban respecto a su contenido. A Evan no se le dio la misma opción de materias dos veces.

Powell y Nelson utilizaron un diseño ABAB para evaluar los efectos de la toma de decisiones sobre la conducta indeseable de Evan. Los resultados mostraron que durante la condición de elección disminuyó el número de conductas indeseables. Dicho estudio apoyó la eficacia de la toma de decisiones como técnica control de antecedente. Tales resultados sugieren que los educadores que intentan manejar la conducta de los estudiantes en clase pueden utilizar procedimientos de elección.

Uso de líneas base múltiples

Existe una forma de investigación de un solo sujeto que emplea más de una línea base. Se establecen varias líneas base antes de aplicar el tratamiento al participante. Estos tipos de estudios se llaman *estudios de líneas base múltiples*. Existen tres clases de diseños de investigación de líneas base múltiples: a través de conductas, a través de participantes y a través de escenarios.

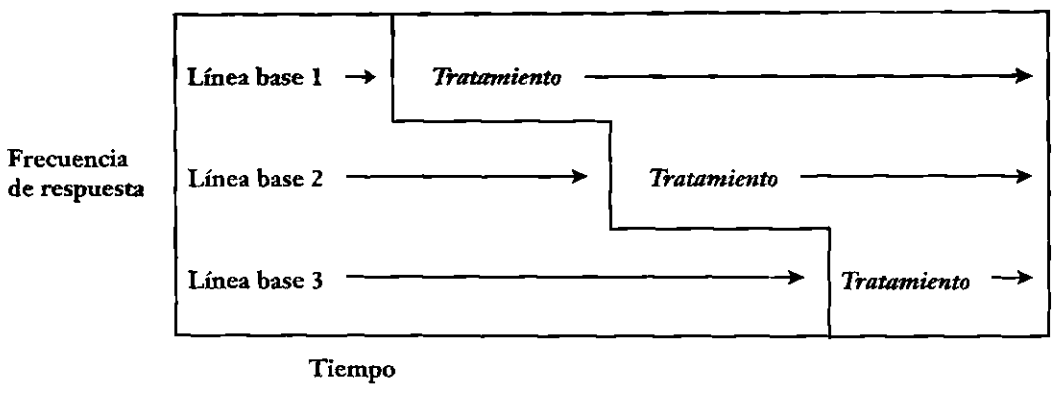
El uso de líneas base múltiples constituye otro método para demostrar la eficacia de un tratamiento sobre el cambio en el comportamiento. Existe un patrón común para la implementación de las tres clases de este diseño, cuyo patrón se muestra en la figura 22.4.

Con las líneas base múltiples *a través de conductas*, las intervenciones de tratamiento para cada conducta diferente se introducen en diferentes momentos. En la figura 22.4, cada línea base sería de una conducta diferente. En el caso de un niño autista, la línea base 1 podría ser golpear su cabeza contra la pared. La línea base 2 tal vez sea hablar constantemente en diferentes tonos añadiendo ruidos. La línea base 3 sería golpear a otros. Se establecen las tres líneas base para saber si el cambio en las conductas coincide con la intervención del tratamiento. Si una de las conductas cambia, mientras las otras permanecen constantes o estables en la línea base, el investigador podría afirmar que el tratamiento resultó eficaz con esa conducta específica. Después de que pase cierto periodo, se aplica el mismo tratamiento a la segunda conducta indeseable (línea base 2). Cada conducta subsecuente se somete al tratamiento en el mismo procedimiento paso a paso. Si la intervención del tratamiento es eficaz al cambiar la tasa de respuesta de cada conducta, entonces es posible afirmar que el tratamiento fue eficaz.

Una consideración importante con esta clase particular de diseño de líneas base múltiples, es que parte del supuesto de que las respuestas de cada conducta son independientes de las respuestas de otras conductas. La intervención puede considerarse eficaz si existe tal independencia. Si las respuestas están correlacionadas de alguna manera, entonces la interpretación de los resultados se torna más difícil.

En el diseño de líneas base múltiples *a través de participantes* se aplica el mismo tratamiento en series a la misma conducta de distintos individuos en el *mismo ambiente*. Ahora cada línea base en la figura 22.4 representa un participante diferente. Cada participante recibe el mismo tratamiento para la misma conducta, en el mismo ambiente. El estudio de Tingstrom, Marlow, Edwards, Kelshaw y Olmi (1997) constituye un ejemplo de un estudio de líneas base múltiples a través de participantes. El paquete de entrenamiento de obediencia es la intervención del tratamiento. La intervención utiliza *tiempo dentro* (contacto físico y elogio verbal) y *tiempo fuera* (un procedimiento coercitivo) para aumentar la tasa de obediencia del estudiante hacia las instrucciones del maestro. La conducta de interés es la obediencia a las instrucciones del maestro. El ambiente es el salón de clases. Los participantes de este estudio fueron tres estudiantes —A, B y C— quienes habían mostrado conducta de no obediencia. Los tres estudiantes presentaban trastornos de articulación

FIGURA 22.4 *Formato general del diseño de líneas base múltiples*



y de lenguaje. El diseño del estudio se adhirió a las siguientes fases de intervención: línea base, sólo tiempo dentro, tiempo dentro y tiempo fuera combinados, y seguimiento. Los estudiantes B y C permanecieron en la línea base mientras se implementaba la fase de sólo tiempo dentro para el estudiante A. Cuando A mostró un cambio en la obediencia, se implementó la fase sólo tiempo dentro para B, mientras que C permanecía en línea base. Cuando B mostró un cambio en la obediencia, se implementó sólo tiempo dentro para C. Tingstrom y sus colaboradores fueron capaces de demostrar la efectividad de la intervención combinada de tiempo dentro y tiempo fuera para incrementar la obediencia.

En el diseño de líneas base múltiples *a través de escenarios*, el mismo tratamiento se aplica a diferentes participantes, quienes se encuentran en *diferentes escenarios*. En este diseño, cada línea base en la figura 22.4 representa a un participante diferente en un ambiente diferente. El tratamiento y la conducta bajo estudio serían los mismos. Aquí es posible tener tres pacientes diferentes, cada uno residente de un tipo distinto de instalación de cuidado psiquiátrico, como las estudiadas por Nelson y sus colaboradores, que se analizaron previamente en este capítulo. En dicho estudio Nelson, Hall y Walsh-Bowers (1997) compararon los efectos a largo plazo de apartamentos de apoyo (AA), hogares de grupo (HG) y hogares de alojamiento y cuidado (HAC).

RESUMEN DEL CAPÍTULO

1. Los experimentos verdaderos son aquellos en que el experimentador puede seleccionar a los participantes de manera aleatoria, asignar a los participantes a las condiciones de tratamiento aleatoriamente y controlar la manipulación de la variable independiente. El diseño cuasi-experimental carece de una o más de estas características.
2. Cook y Campbell (1979) analizan ocho variantes del diseño de grupo control no equivalente. El que cubre este libro es el diseño de grupo control sin tratamiento. Se analizan cinco resultados diferentes en términos de la validez interna.
3. Los diseños de series de tiempo son diseños longitudinales que incluyen mediciones repetidas de las mismas variables dependientes en diferentes intervalos de tiempo fijos. Casi siempre en cierto momento se introduce la intervención del tratamiento.
4. La selección y la interacción entre la selección y la maduración son dos explicaciones alternativas que cubren los resultados obtenidos a partir de los diseños cuasi-experimentales.
5. Los experimentos que utilizan participantes únicos no son nuevos. Los investigadores pioneros de la psicología experimental utilizaron diseños de un solo sujeto.
6. Los investigadores consideran que en los diseños de un solo sujeto la variabilidad de la situación se elimina con el control experimental adecuado.
7. Los investigadores orientados a grupos consideran que la variabilidad puede analizarse estadísticamente.
8. La investigación de un solo sujeto posee varias ventajas sobre la investigación de grupos, en términos de flexibilidad y ética. Sin embargo, carece de credibilidad de la validez externa.
9. Los efectos pequeños, pero estadísticamente significativos, encontrados en la investigación de grupos pueden tener poca significancia clínica o práctica, y haberse introducido artificialmente por tamaños grandes de muestras. Cuando esto sucede, el tamaño del efecto es pequeño. La investigación de un solo sujeto se concentra en el tamaño del efecto y no en el tamaño de la muestra.

10. El establecimiento de una línea base estable constituye una de las tareas más importantes en la investigación de un solo sujeto.
11. Al establecimiento de una línea base, seguido por la administración de un tratamiento y, después, por el retiro del tratamiento, se le llama diseño ABA.
12. Un problema importante del diseño ABA es que tal vez el tratamiento sea irreversible, dejando al participante en el estado de mejoría, en lugar de regresar a la persona al estado indeseable original.
13. Una variación del diseño ABA es el diseño ABAB, donde se reinstaura el estado de mejoría del participante.
14. En un estudio de un solo sujeto se puede variar únicamente una variable independiente a la vez.
15. El llamado diseño de interacción no permite probar una interacción del tipo definido previamente en los diseños factoriales. Solamente examina dos variables en conjunto.
16. Existen tres tipos de diseños de líneas base múltiples. En cada caso, la intervención se introduce en diferentes momentos para diferentes conductas, participantes o escenarios. Si los cambios en las conductas coinciden con la introducción del tratamiento, esto proporciona evidencia de que el tratamiento es eficaz.

SUGERENCIAS DE ESTUDIO

1. Revise cada uno de los siguientes estudios y determine cuáles son diseños cuasiexperimentales, de grupo control no equivalente y de un solo sujeto.

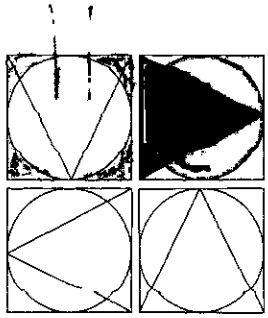
   Adkins, V. K. y Matthews, R. M. (1997). Prompted voiding to reduce incontinence in community-dwelling older adults. *Journal of Applied Behavior Analysis, 30,* 153–156.

   Lee, M. J. y Tingstrom, D. H. (1994). A group math intervention: The modification of cover, copy, and compare for group application. *Psychology in the Schools, 31,* 133–145.

   Streufert, S., Satish, U., Pogash, R., Roache, J. y Severs, W. (1997). Excess coffee consumption in simulated complex work settings: Detriment or facilitation of performance? *Journal of Applied Psychology, 82,* 774–782.

2. ¿Por qué es necesaria una medida de línea base en los diseños de un solo sujeto?
3. ¿Los datos de diseños de un solo sujeto deben analizarse estadísticamente? Explique por qué.
4. Dé un ejemplo donde deba utilizarse un diseño de un solo sujeto. También cite una situación de investigación donde resulte más apropiado un diseño de grupo.
5. Un estudiante universitario desea realizar un estudio de series de tiempo sobre los efectos de la luna llena en pacientes psiquiátricos. ¿Qué variable dependiente debe utilizar? ¿Dónde debe buscar para obtener datos para un estudio de este tipo?
6. ¿Los diseños de un solo sujeto son aplicables a la investigación médica? ¿Deben enseñarse diseños de un solo sujeto a los estudiantes de escuelas de medicina? Lea el siguiente artículo:

   Bryson-Brockmann, W. y Roll, D. (1996). Single-case experimental designs in medical education: An innovative research method. *Academic Medicine, 71,* 78-85.



CAPÍTULO 23

INVESTIGACIÓN NO EXPERIMENTAL

Entre las falacias prevalecientes, una de las más peligrosas para la ciencia es la conocida como *post hoc, ergo propter hoc*: después de esto, por lo tanto, causado por esto. Es posible bromear al decir con un toque de seriedad, "si me llevo mi paraguas, no lloverá". Incluso se podría decir seriamente que los delincuentes son delincuentes debido a la falta de disciplina en las escuelas, o que la educación religiosa hace a los niños más virtuosos. Es fácil asumir que una cuestión causa otra, simplemente porque ocurre antes de la otra y porque se tiene una amplia gama de "causas" posibles. Entonces, también muchas explicaciones parecen frecuentemente plausibles. Por ejemplo, es fácil creer que el aprendizaje de los niños mejora porque se instituye una nueva práctica educativa o porque se enseña de cierta manera. Se asume que la mejoría en su aprendizaje se debió al nuevo

método de ortografía, a la institución de los procesos de grupo en la situación dentro del salón de clases, a una disciplina severa y a mayor cantidad de tarea (o poca disciplina y menos tarea). En raras ocasiones se considera que los niños aprenderán algo si se les da la oportunidad de aprender.

El científico social y el científico educativo enfrentan a menudo el problema de la falacia *post hoc*. El sociólogo que busca las causas de la delincuencia sabe que debe tenerse extremo cuidado al estudiar el problema. Las condiciones de pobreza, los hogares destruidos, las cantidades de plomo en las tuberías de agua, la carencia de amor —todas y cada una— son causas posibles de la delincuencia. El psicólogo que busca las raíces de la personalidad adulta enfrenta un problema aún más sutil: rasgos heredados, prácticas de crianza, influencias educativas, personalidad de los padres y circunstancias ambientales; todas son explicaciones plausibles. El científico educativo, con la meta de entender las bases de logro del éxito escolar, enfrenta también un gran número de posibilidades razonables: inteligencia, aptitud, motivación, ambiente familiar, personalidad del maestro, personalidad del alumno y métodos de enseñanza.

El peligro del supuesto *post hoc* es que puede, y con frecuencia lo hace, conducir a interpretaciones erróneas y confusas de los datos de investigación, cuyo efecto es particularmente serio cuando los científicos tienen poco o ningún control sobre el tiempo y sobre las variables independientes. Cuando buscan explicar un fenómeno que ya ha ocurrido, los científicos se ven confrontados con el desagradable hecho de que no tienen un control real de las causas posibles. Por lo tanto, deben elaborar un curso de acción de investigación, que difiera en ejecución e interpretación del que siguen los científicos que realizan experimentación.

Definición

La investigación no experimental es la búsqueda empírica y sistemática en la que el científico no posee control directo de las variables independientes, debido a que sus manifestaciones ya han ocurrido o a que son inherentemente no manipulables. Se hacen inferencias sobre las relaciones entre las variables, sin intervención directa, de la variación concomitante de las variables independiente y dependiente.

Suponga que un investigador tiene interés en la relación del sexo con la creatividad de los niños. Mide la creatividad de una muestra de niños y niñas, y prueba la significancia de la diferencia entre las medias de los dos sexos. La media de los niños es significativamente mayor que la media de las niñas. Una conclusión es que los niños son más creativos que las niñas y quizá no sea una conclusión válida. La relación existe, es verdad; sin embargo, con esta sola evidencia la conclusión resulta dudosa. Surgiría una pregunta: ¿es la relación demostrada realmente entre el sexo y la creatividad? Como existen otras variables que están correlacionadas con el sexo, una o más de ellas pudieron haber generado la diferencia entre las puntuaciones de creatividad de los dos sexos.

Diferencia básica entre la investigación experimental y la no experimental

La base de la estructura sobre la que la ciencia experimental opera es simple. Se hipotetiza: si x, entonces y; si frustración, entonces agresión. Dependiendo de las circunstancias y de la predilección personal en el diseño de investigación, se utiliza un método para manipular o medir x. Entonces se observa y para observar si ocurre una variación concomitante, es

decir, la variación esperada o que se predice a partir de la variación de x. Si esto sucede, es evidencia de la validez de la proposición $x \rightarrow y$, "si x entonces y". Considere que aquí se predice y a partir de una x controlada. Para lograr mayor control, se puede utilizar el principio de aleatorización y manipulación activa de x, y se puede asumir que, siendo lo demás igual, y varía como resultado de la manipulación de x.

Por otra parte, en la investigación no experimental se observa y, y también se observan una o varias x. Éstas se observan ya sea antes, después o concomitantemente a la observación de y. No hay diferencia en la lógica básica. Es posible demostrar que la estructura del argumento y su validez *lógica* son iguales en la investigación experimental y en la no experimental. Además, el propósito básico de ambas es el mismo: establecer la validez *empírica* de las llamadas proposiciones condicionales de la forma: si p, entonces q. La diferencia esencial es el control directo de p, la variable independiente. En la investigación experimental p puede manipularse, lo cual es más bien un "control" directo. Cuando Clark y Walberg (1968) pidieron a unos maestros que dieran reforzamiento masivo a un grupo de participantes y a otros maestros que dieran reforzamiento moderado a otro grupo, estaban manipulando o controlando directamente la variable reforzamiento. De la misma forma, cuando Dolinski y Nawrat (1998) sometieron a un grupo a estrés (ansiedad), sometieron a otro grupo a estrés y después lo redujeron, e incluyeron un tercer grupo con poco o ningún estrés, estaban manipulando directamente la variable ansiedad. Además, los participantes pueden asignarse aleatoriamente a los grupos experimentales.

En la investigación no experimental el control *directo* no es posible: tampoco puede utilizarse la manipulación experimental ni la asignación aleatoria. Existen dos diferencias esenciales entre los modelos experimental y no experimental. A causa de la carencia de control relativo de x y de otras posibles x, la "verdad" de la relación hipotetizada entre x y y no puede sostenerse con la misma confianza que en la situación experimental. Básicamente la investigación no experimental tiene, por así decirlo, una debilidad inherente: la carencia de control de las variables independientes.

La diferencia más importante entre la investigación experimental y la investigación no experimental es, entonces, el *control*. En los experimentos, los investigadores tienen, por lo menos, control manipulativo: por lo menos tienen una variable activa. Si un experimento es un "verdadero" experimento, también puede ejercerse control por medio de la aleatorización. Es posible asignar participantes aleatoriamente a los grupos, o asignar los tratamientos aleatoriamente a los grupos. En la situación de investigación no experimental, este tipo de control de las variables independientes no es posible. Los investigadores deben tomar las cosas como son e intentar entenderlas.

Considere un caso bien conocido. Cuando se pinta la piel de las ratas con sustancias cancerígenas (x), se controlan adecuadamente otras variables, y las ratas finalmente desarrollan un carcinoma (y), el argumento es contundente, ya que se está controlando x (y otras x teóricamente posibles), y se predice y. Cuando se encuentran casos de cáncer pulmonar (y) y se regresa a revisar en la posible multiplicidad de causas ($x_1, x_2, \ldots, x_n$), y se elige el hábito de fumar cigarrillos (digamos x_3) como la causa, se está en una situación más difícil y ambigua. Ninguna situación resulta segura, por supuesto; ambas son probabilísticas. Pero en el caso experimental se puede estar considerablemente más seguro, si se mantuvieron "constantes el resto de las condiciones", de que la proposición si x, entonces y, es válida empíricamente. Sin embargo, en el caso no experimental no se pisa tierra tan firme, ya que no puede afirmarse con mucha certeza, "que se mantuvieron constantes el resto de las condiciones". No es posible controlar las variables independientes por medio de la manipulación o de la aleatorización. En pocas palabras, la probabilidad de que x esté "realmente" relacionada con y es mayor en la situación experimental que en la situación no experimental porque el control de x es mayor.

Autoselección e investigación no experimental

En un mundo ideal de la investigación del comportamiento, la obtención de muestras aleatorias de participantes, así como la asignación aleatoria de participantes y tratamientos a los grupos, siempre sería posible. No obstante, en el mundo real, ni una, ni dos e incluso ninguna de estas tres posibilidades existe. Es posible seleccionar participantes al azar, tanto en la investigación experimental como en la no experimental. Pero no es posible, en la investigación no experimental, asignar a los participantes aleatoriamente a los grupos o asignar los tratamientos aleatoriamente a los grupos. Los participantes pueden "asignarse a sí mismos" a los grupos. Es posible que se "seleccionen a sí mismos" en los grupos con base en características diferentes de aquellas que interesan al investigador. Los participantes y los tratamientos llegan como si ya hubieran sido asignados a los grupos.

La *autoselección* ocurre cuando los miembros de los grupos estudiados están en los grupos, en parte, porque poseen rasgos o características diferentes, ajenas al problema de investigación; características que posiblemente influyan o estén relacionadas con las variables del problema de investigación. Algunos ejemplos de autoselección ayudarán a una mejor comprensión.

En la bien conocida investigación sobre el tabaquismo y el cáncer, se estudiaron los hábitos de fumar de un gran número de personas. Este gran grupo se dividió en aquellos que padecían cáncer pulmonar —o que habían muerto por su causa— y aquellos que no lo padecían. Por lo tanto, la variable dependiente era la presencia o la ausencia del cáncer. Los investigadores exploraron los antecedentes de los participantes para determinar si fumaban cigarrillos, y si así era, cuántos. Fumar cigarrillos era la variable independiente. Los investigadores encontraron que la incidencia del cáncer pulmonar se incrementaba de acuerdo con el número de cigarrillos fumados por día. También encontraron que la incidencia fue menor en el caso de los fumadores moderados y de los no fumadores. Llegaron a la conclusión de que el hábito de fumar cigarrillos "causa" cáncer pulmonar. Esta conclusión puede o no resultar cierta; pero los investigadores no pueden llegar a esta conclusión, aunque afirmen que existe una relación estadísticamente significativa entre las variables. Observe que los investigadores científicos cuidadosos generalmente no utilizan el término *causa* a menos que el estudio haya sido realizado bajo las más estrictas condiciones. La palabra *causa* se utiliza aquí para ilustrar cómo los medios de comunicación masiva a menudo interpretan los hallazgos científicos que sugieren causalidad.

Los investigadores científicos tampoco pueden establecer una conexión causal porque existen muchas otras variables; y una de ellas o alguna combinación de ellas, que pudo causar el cáncer. Además ellos no controlaron otras variables independientes posibles. *No pueden* controlarlas excepto por medio de la comprobación de hipótesis alternativas, un procedimiento que se explicará más adelante. Aun cuando ellos también estudian "grupos control" de personas que no padecen cáncer, quizá esté operando la autoselección. Tal vez, por ejemplo, los hombres tensos y ansiosos están condenados a padecer cáncer pulmonar si se casan con mujeres altas. Puede suceder que este tipo de hombre también fume gran cantidad de cigarrillos. No es el tabaquismo lo que lo mata —él se mata a sí mismo al estar tenso y ansioso— y posiblemente por el hecho de casarse con una mujer alta. Tales hombres son seleccionados para formar parte de la muestra por los investigadores sólo porque fuman cigarrillos; sin embargo, ellos se seleccionan a sí mismos para la muestra porque comúnmente poseen un temperamento concomitante al tabaquismo.

La autoselección constituye un aspecto sutil. Existen dos tipos: 1) autoselección para *muestras* y 2) autoselección para *grupos comparativos*. Esto último sucede cuando se selecciona a los participantes porque pertenecen a un grupo o a otro: cáncer y sin cáncer, universitario y no universitario, bajo rendimiento o sin bajo rendimiento. Es decir, son

seleccionados *porque* poseen la variable dependiente en mayor o menor grado. La autoselección para muestras ocurre cuando se selecciona a los participantes de forma no aleatoria para la muestra.

Lo esencial del tema es que cuando la *asignación* no es aleatoria, siempre existe un resquicio para que otras variables se inmiscuyan. Cuando se colocan participantes dentro de los grupos, en el caso anterior o en casos similares, o cuando se "colocan ellos mismos" dentro de los grupos, con base en una variable, es posible que otra variable (o variables) correlacionada(s) con esta variable, sean la base "real" de la relación. El estudio no experimental común utiliza grupos que muestran diferencias respecto a la variable dependiente. En ciertos estudios de tipo longitudinal, los grupos se diferencian primero con base en la variable independiente. Pero ambos casos son básicamente iguales, ya que la pertenencia al grupo, con base en una *variable*, siempre conlleva la selección.

Por ejemplo, es posible seleccionar aleatoriamente a universitarios de primer año y después seguirlos para determinar la relación entre inteligencia y el éxito en la universidad. Los estudiantes se seleccionaron a sí mismos dentro de la universidad, por así decirlo. Una o más de las características que llevan consigo a la universidad, además de la inteligencia —nivel socioeconómico, motivación, antecedentes familiares— pueden ser los principales determinantes del éxito universitario. El hecho de iniciar con la variable independiente, en este caso la inteligencia, no cambia la naturaleza autoselectiva de la situación de investigación. En términos de muestreo, los estudiantes se seleccionaron a sí mismos para la universidad, lo cual sería un factor importante si se estudiaran alumnos universitarios y no universitarios. Pero si el interés radica tal sólo en el éxito y en el no éxito de estudiantes *universitarios*, entonces la autoselección para la universidad se vuelve irrelevante; mientras que la autoselección para los grupos de éxito y de no éxito resulta crucial. El hecho de medir la inteligencia de los estudiantes al entrar a la universidad, y seguirlos hacia el éxito y hacia el no éxito, no cambia ni el problema de selección ni el carácter no experimental de la investigación. En suma, los estudiantes se seleccionaron a sí mismos dentro de la universidad y se seleccionaron a sí mismos para tener o no éxito en la universidad.

Investigación no experimental a gran escala

Como siempre, los ejemplos de investigación ayudan a comprender la naturaleza de la investigación no experimental. En lugar de resumir únicamente estudios individuales, como se ha hecho hasta ahora, se describirán tanto estudios individuales como conjuntos de estudios centrados en algún fenómeno o variable de interés. La investigación no experimental del comportamiento con frecuencia se enfoca en grandes problemas de importancia social y humana: clase social, procesos políticos, segregación y disgregación, actitudes públicas, rendimiento escolar, por ejemplo. La importancia —*relevancia* es la palabra de moda— del tema de dichos estudios no debe oscurecer la comprensión de su carácter no experimental. Sin embargo, aunque la investigación no experimental tiene debilidades inherentes, eso no significa que la investigación experimental sea más importante. Como se mencionó antes, el experimento es uno de los grandes inventos de todos los tiempos, un ideal de control al que todos aspiran. Ello no significa que los experimentos sean necesariamente "mejores" que los estudios no experimentales. Por otro lado, la investigación no experimental no siempre es "mejor" que la investigación experimental porque su contenido y variables parezcan ser socialmente importantes. ¡Esto sería como decir que la investigación psicológica es "mejor" que la investigación sociológica, debido a que los psicólogos utilizan con mayor frecuencia un modelo experimental y los sociólogos un modelo no experimental!

Determinantes del rendimiento escolar

Una gran preocupación de los investigadores educativos ha sido la búsqueda de los factores determinantes del rendimiento escolar. ¿Qué factores conducen al éxito del rendimiento en la escuela? La inteligencia es un factor importante, por supuesto. Mientras que la inteligencia medida, especialmente la habilidad verbal, explica una gran proporción de la varianza del rendimiento, existen muchas otras variables, tanto psicológicas como sociológicas: sexo, raza, clase social, aptitud, características ambientales, condiciones de la escuela y habilidades del maestro, antecedentes familiares, métodos de enseñanza. El estudio del rendimiento está caracterizado tanto por modelos experimentales como no experimentales. Aquí nos ocupamos de estos últimos, ya que ilustran claramente los problemas de la investigación no experimental.

En 1966 se publicó el ahora famoso reporte Coleman (Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld y York, 1966). Como su título indica (*Equality of Educational Opportunity*), fue un intento a gran escala para responder la pregunta: ¿ofrecen las escuelas estadounidenses iguales oportunidades educativas a todos los niños? Sin embargo, de igual importancia era la pregunta sobre la relación entre el rendimiento de los estudiantes y el tipo de escuela a la que asistían. Dicho estudio representó un esfuerzo masivo y admirable por responder estas preguntas (y otras). Su hallazgo más famoso y controvertido fue que las diferencias entre las escuelas explican sólo una pequeña fracción de las diferencias en el rendimiento escolar. La mayoría de la varianza del rendimiento estaba explicada por aquello que los niños llevaban consigo a la escuela. Hubo mucho que cuestionar respecto a la metodología y conclusiones del estudio. De hecho, sus consecuencias aún persisten. Algunos de los distritos escolares más importantes de Estados Unidos utilizaron el reporte como justificación para implementar ciertas políticas educativas controvertidas, tales como encomendar demasiadas ocupaciones a los niños.

La principal variable dependiente en este estudio era el rendimiento verbal. Hubo, sin embargo, más de 100 variables independientes. Los autores utilizaron procedimientos multivariados relativamente sofisticados para analizar los datos. Gran parte del foco de los problemas analíticos, la interpretación de los resultados y las críticas subsecuentes recaen en la naturaleza no experimental de la investigación.

La controvertida conclusión mencionada antes, sobre la importancia relativa de las variables de antecedentes del hogar y de las variables escolares, depende de un método completamente confiable y válido de evaluación de los impactos relativos de las diferentes variables. En la investigación experimental se está más a salvo al hacer conclusiones comparativas, ya que las variables independientes no están correlacionadas. Sin embargo, en el mundo educativo real las variables están correlacionadas, lo cual provoca que sea difícil determinar su contribución única al rendimiento. Aunque existen métodos estadísticos para manejar dichos problemas, ningún método puede decir, sin ambigüedades, que X_1 influye en Y en tal o cual medida porque la influencia real puede ser x_2, que influye tanto en x_1 como en y. La interpretación "correcta" de los hallazgos de *igualdad*, y de estudios similares, resulta siempre insostenible. Aunque existen métodos analíticos poderosos que se utilizan con datos no experimentales, las respuestas inequívocas a las preguntas sobre los factores determinantes del rendimiento son inalcanzables.

Diferencias del estilo de respuesta entre estudiantes del este asiático y estadounidenses

Esta investigación realizada por Chen, Lee y Stevenson (1995) fue un estudio de gran escala que abarcó cuatro países, cuatro culturas y dos continentes. El interés principal del

estudio se centró en el uso de escalas de medición, las cuales son básicas en la investigación de las ciencias del comportamiento. Sin embargo, ¿existen diferencias culturales respecto a la manera en que ciertos grupos étnicos responden las preguntas con el uso de una escala de medición? Chen y sus colaboradores deseaban determinar si existía o no una diferencia en el estilo de respuesta entre los asiáticos del este y los estadounidenses. Estos investigadores recolectaron datos de estudiantes: 944 japoneses, 1 357 taiwaneses, 687 canadienses y 2 174 del medio oeste y de la costa este de Estados Unidos. Las comparaciones en su estudio incluyeron las diferencias entre dos culturas (del este asiático contra estadounidense) y las diferencias entre los dos grupos representativos dentro de cada cultura (Estados Unidos contra Canadá, y Japón contra Taiwán). El cuestionario administrado a dichos estudiantes incluía reactivos sobre ideas, valores, actitudes, creencias y autoevaluaciones, en relación con la escuela y la vida cotidiana. Se utilizó una escala tipo Likert de 7 puntos, donde el 7 normalmente indicaba "más" o "fuertemente de acuerdo"; el 1 fue utilizado para "sin importancia", "menos" o "fuertemente en desacuerdo". Los resultados demostraron que en el área de individualismo y colectivismo hubo diferencias altamente significativas entre las dos culturas. También encontraron una relación entre el apoyo al individualismo y el uso de valores extremos en la escala.

Este estudio está clasificado también como no experimental ya que no hay una variable independiente manipulada. La variable independiente de este estudio era la cultura, y ésta no fue manipulada. El estudio mostró una diferencia en el estilo de respuesta entre diferentes culturas. Independientemente del muestreo de este estudio, éste es no experimental. Por lo tanto, no se puede afirmar de manera explícita que si usted es de esta cultura, entonces responderá de tal o cual forma en las escalas de medición.

Investigación no experimental a menor escala

Ilustrar estudios o series de estudios de investigación no experimental sobre el comportamiento no es fácil, pues existen demasiados; no obstante, algunos satisfacen los criterios personales de los autores sobre la solidez metodológica y el interés sustantivo. Se escogieron los siguientes estudios por tres razones: 1) los autores consideran que cada uno representa un enfoque único, original e interesante para un importante problema sociológico, psicológico o educativo; 2) cada uno de ellos contribuye significativamente al conocimiento científico, y 3) cada uno de ellos es no experimental.

Cochran y Mays: sexo, mentiras y VIH

Éste es el estudio clásico, frecuentemente citado y mencionado respecto a la diferencia en la conducta sexual entre hombres y mujeres. Cochran y Mays (1990) encontraron que aconsejar a los adultos jóvenes y a los adolescentes sobre las precauciones que deben tomar para protegerse del virus de inmunodeficiencia humana (VIH) puede resultar inútil. Por ejemplo, un consejo es que el individuo pregunte a la persona con quien va a salir, sobre su historia de riesgos, antes de decidir si se involucra o no en una relación sexual. Sin embargo, Cochran y Mays encontraron que la gente joven tiende a mentir respecto a su historia sexual. En una muestra de 665 estudiantes (entre 18 y 25 años de edad) en el sur de California, 196 hombres sexualmente experimentados y 226 mujeres sexualmente experimentadas reportaron que mentían para lograr relaciones sexuales, lo cual significa que más del 63% de la muestra afirmó haber mentido en el pasado, para tener sexo con quien salía. Se encontró que los hombres mentían significativamente más seguido que las mujeres. Además, tanto hombres como mujeres indicaron que volverían a engañar a su compañero o compañera de citas; los hombres estaban más dispuestos a hacer esto que las

TABLA 23.1 *Porcentaje de respuestas a preguntas sobre sexo y deshonestidad para hombres y mujeres (estudio de Cochran y Mays)*

Pregunta	Hombres	Mujeres
Ha mentido para tener relaciones sexuales	34	10
Mintió sobre el control en la eyaculación o la posibilidad de un embarazo	38	14
Mentiría respecto a tener resultados negativos de anticuerpos VIH	20	4
Mentiría sobre el control de la eyaculación o la posibilidad de un embarazo	29	2
Subestimaría el número de parejas previas	47	42

mujeres. En la tabla 23.1 se presenta un resumen del análisis realizado por Cochran y Mays sobre el cuestionario de 18 páginas respecto a la conducta sexual, la reducción del riesgo del VIH y del engaño en las citas. La comparación entre hombres y mujeres incluye la variable independiente medida o atributiva *sexo*. Dicho estudio posee gran significancia respecto a aconsejar a la gente joven sobre prácticas de sexo seguro. La implicación es que no se puede confiar en la palabra de la persona con quien se tiene una cita. Aunque los datos señalan fuertemente la disposición de los hombres a mentir para obtener favores sexuales, no es posible asumir automáticamente que todos los hombres mentirán sobre su historia sexual.

Elbert: problemas de lectura y del lenguaje escrito en niños con déficit de atención

El fenómeno conocido como trastorno por déficit de atención con hiperactividad (TDAH) es actualmente un área común de investigación psicológica y educativa. Los niños que la padecen por lo común exhiben escasa autorregulación de conductas y pobre desempeño escolar (generalmente de 1 a 1.5 unidades de desviación estándar por debajo de las puntuaciones de los niños normales). Conforme la investigación en esta área maduraba, los investigadores volcaron su interés a cuestiones más específicas que la comparación de los niños con TDAH con niños normales. Un área de interés son las subclases o subgrupos de TDAH, especialmente el trastorno por déficit de atención (TDA). Dichos estudios normalmente comparaban niños con déficit de atención e hiperactividad (TDA + H) contra niños con déficit de atención sin hiperactividad (TDA – H).

Uno de ellos fue realizado por Elbert (1993), quien deseaba determinar si estos dos subgrupos de TDAH (TDA + H y TDA – H) diferían respecto al rendimiento. El rendimiento se evaluaba a través de pruebas estandarizadas de lectura, ortografía y lenguaje escrito. Elbert también buscaba determinar si existía una interacción entre el género (hombre y mujer), la edad (6 a 7 años, 8 a 9 años y 10 a 12 años) y el tipo de subgrupo (TDA + H y TDA – H). El estudio utilizó los datos de 115 niños cuyas edades iban de los 6 a los 12 años. Cada niño era clasificado con TDA + H o TDA – H, por medio de evaluaciones objetivas de maestros y por los lineamientos establecidos por Barkley (1990). Note que aquí no hay una variable independiente manipulada. La inclusión en el grupo no se determinó a través de un proceso aleatorio. La naturaleza no experimental de tal estudio resultó en grupos de muy distintos tamaños. El grupo TDA + H quedó integrado por 83 niños; y el grupo TDA – H, por 32 niños. También había más hombres (86) que mujeres (29). Sin embargo, Elbert realizó numerosos análisis para verificar la igualdad de los grupos respecto a las variables edad, nivel de calificaciones, nivel de educación de la madre y CI. Las pruebas estadísticas entre TDA + H y TDA – H, con estas variables, no fueron significativas.

Los resultados en las puntuaciones de lectura mostraron un desempeño más pobre del grupo de niños con TDA + H (97), que en el grupo de TDA – H (97). Elbert también encontró una interacción significativa de género y edad, respecto a las puntuaciones de lectura. Pruebas *post hoc* demostraron que las niñas del grupo de edad media tuvieron un peor desempeño que los niños. Las pruebas estadísticas realizadas con las puntuaciones en cuanto a ortografía y lenguaje escrito no mostraron algún efecto entre los grupos de edad, género o tipo de subgrupo. No obstante, se encontró un efecto de interacción significativo del género y la edad. De nuevo, las niñas del grupo de edad media tuvieron el desempeño más pobre. Elbert descubrió además que una subprueba de ortografía y lenguaje escrito (ortografía escrita por dicción) fue la habilidad más limitada en ambos subgrupos de TDA. Observe la naturaleza no experimental del estudio de Elbert. No hubo variable independiente manipulada ni aleatorización.

Con estos estudios no experimentales de respaldo, ahora es posible discutir y evaluar la investigación no experimental, en general. Sin embargo, es necesaria una discusión evaluativa, con un cuestionamiento más sistemático de la comprobación de hipótesis alternativas, una de las características más importantes de la investigación científica.

Comprobación de hipótesis alternativas

La mayoría de las investigaciones se inician con las hipótesis; y después se prueban las implicaciones empíricas de tales hipótesis. Aunque las hipótesis se "confirman" de la manera descrita en capítulos anteriores, también se pueden "confirmar" y "desmentir" las hipótesis bajo estudio al tratar de demostrar qué hipótesis alternativas posibles son apoyadas o no. Primero, considere variables independientes alternativas como antecedentes de una variable dependiente. El razonamiento es el mismo. Por ejemplo, si se dice "variables independientes alternativas", en efecto se están proponiendo hipótesis o explicaciones alternativas de una variable dependiente.

En los estudios no experimentales, aunque no es posible confiar en la "verdad" de una proposición "si x, entonces y", como en un experimento verdadero, sí es posible establecer y probar hipótesis alternativas o "control". (Por supuesto, las hipótesis alternativas también pueden comprobarse, y de hecho así ocurre, en estudios experimentales.) Este procedimiento ha sido formalizado y explicado por Platt (1964), quien recibió la influencia de Chamberlin (1890; 1965). Platt le llama "inferencia fuerte". Chamberlin denomina apropiadamente al procedimiento "el método de trabajar hipótesis múltiples" y describe cómo pueden evitarse los propios "afectos intelectuales". Chamberlin (p. 756) asevera:

> El esfuerzo radica en sacar a la luz cada explicación racional de nuevos fenómenos e intentar desarrollar cada hipótesis sostenible, respetando su causa e historia. Así, el investigador se convierte en padre de una familia de hipótesis; y, por su relación parental con todas, el investigador tiene prohibido mostrar indebidamente su afecto a una de ellas.

Para revisar el desarrollo histórico de las hipótesis alternativas, véase Cowles (1989).

Sean x_1, x_2 y x_3 tres variables independientes alternativas, y sea y la variable dependiente, el fenómeno a "explicarse" con la proposición "si x, entonces y". Suponga que x_1, x_2 y x_3 agotan todas las posibilidades. Esta suposición no puede hacerse en realidad en la investigación científica, ya que resulta prácticamente imposible agotar todas las posibilidades. Aun así, aquí se supondrá esto por razones didácticas.

Un investigador tiene evidencia de que x_1 y y están sustancialmente relacionadas, y teniendo razones para creer que x_1 es el factor determinante, x_2 y x_3 se mantienen constantes. El supuesto aquí es que uno de los tres factores, x_1 o x_2 o x_3, es la "verdadera" variable independiente. De nuevo, note el supuesto. Puede ser ninguna o alguna combinación de

las tres. Suponga que el investigador tiene éxito en eliminar x_2; es decir, se demuestra que x_2 no está relacionada con y. Si el investigador también tiene éxito en eliminar x_3, entonces la conclusión es que x_1 es la variable independiente influyente. Puesto que las hipótesis alternativas o "control" no fueron justificadas, entonces se refuerza la hipótesis original.

De forma similar, es posible probar variables *dependientes* alternativas, lo que también implica hipótesis alternativas; se cambian las alternativas a la variable dependiente. Alper, Blane y Abrams (1955) lo ilustran en un estudio sobre las diferentes reacciones de niños de clase media y clase baja a pintar con los dedos, como consecuencia de diferentes prácticas de crianza del niño. La pregunta general planteada era: ¿las diferencias en la clase social en las prácticas de entrenamiento de niños resultarán en diferencias de personalidad entre clases? La teoría subyacente requería que hubiese diferencias en las reacciones ante el hecho de pintar con los dedos. Alper y sus colaboradores pensaron que los niños de clase media reaccionarían de manera distinta que los niños de clase baja a 16 variables diferentes, *cuando se utilizaran pinturas dactilares*: aceptación de la tarea, lavarse, etcétera. Las reacciones fueron significativamente diferentes respecto a la mayoría de las variables. En un "experimento control" se siguió el mismo procedimiento *utilizando crayolas en lugar de pinturas dactilares*. Los dos grupos no difirieron significativamente en ninguna de las 11 variables medidas. Éste fue un sorprendente contraste con los resultados de la pintura dactilar. El estudio fue no experimental ya que no era posible manipular la variable independiente, y porque los niños llegaron al estudio con sus reacciones ya construidas. El uso de un estudio control fue ingenioso y crucial. ¡Imagínese la consternación del investigador si las diferencias entre los dos grupos, en la tarea con crayolas, hubiesen resultado significativas!

Ahora considere el estudio clásico de Sarnoff, Lighthall, Waite, Davidson y Sarason (1958) que predijo que los niños ingleses y estadounidenses diferirían significativamente respecto a la ansiedad al contestar una prueba, pero no respecto a la ansiedad en general. La hipótesis fue delineada cuidadosamente: si se toma el examen eleven-plus, entonces resultará ansiedad de prueba. (El examen eleven-plus se aplica a los estudiantes ingleses a la edad de 11 años, como ayuda para determinar sus futuros educativos.) Puesto que era posible que hubiera otras variables independientes que causaran las diferencias entre los niños ingleses y estadounidenses, respecto a la ansiedad de prueba, evidentemente los investigadores quisieron excluir, por lo menos, a algunos de los principales competidores. Esto lo lograron apareando con cuidado las muestras: ellos quizás pensaron que la diferencia en la ansiedad de prueba podría deberse a una diferencia en la ansiedad en general, ya que la medida de la ansiedad de prueba obviamente debe reflejar alguna ansiedad en general. Si el resultado fuera éste, no se sustentaría la hipótesis principal. Por lo tanto, Sarnoff y sus colegas, además de comprobar la relación entre el examen y la ansiedad de prueba, también probaron la relación entre la examinación y la ansiedad general. La hipótesis de que los niños ingleses tendrían puntuaciones de ansiedad de prueba mayores que los estadunidenses fue confirmada por los datos. También encontraron que no hubo diferencias significativas entre los dos países respecto a la ansiedad en general, y que las niñas mostraban un nivel de ansiedad de prueba mayor que los niños, en ambos países. Se encontró que la ansiedad de prueba estaba correlacionada positivamente con el nivel de calificaciones.

Aunque el método de comprobación de hipótesis alternativas es importante en toda investigación, resulta fundamental en los estudios no experimentales, ya que es una de las únicas formas de controlar las variables independientes de dicha investigación. Al carecer de la posibilidad de la aleatorización y de la manipulación, los investigadores no experimentales, quizás más que los experimentales, deben ser muy sensibles a las posibilidades de prueba de hipótesis alternativas.

Evaluación de la investigación no experimental

El lector puede sentir, a causa de la discusión precedente, que la investigación no experimental es inferior que la experimental; pero esta conclusión sería injustificada. Es fácil afirmar que la investigación experimental es "mejor" que la investigación no experimental, o que la investigación experimental tiende a ser "trivial", o que la no experimental es "meramente correlacional". Dichas afirmaciones son, en sí mismas, simplificaciones excesivas. Lo que el estudiante de investigación necesita es una comprensión balanceada de las fortalezas y debilidades de ambos tipos de investigación. Estar comprometido inequívocamente con la investigación experimental o con la no experimental llega a convertirse en una visión limitada.

Limitaciones de la interpretación no experimental

La investigación no experimental posee tres grandes debilidades, dos de las cuales ya se analizaron en detalle: 1) la incapacidad de manipular variables independientes, 2) la falta de poder para aleatorizar y 3) el riesgo de realizar interpretaciones inadecuadas. En otras palabras, comparada con la investigación experimental, manteniendo lo demás igual, la investigación no experimental carece de control; tal carencia es la base de la tercera debilidad: el riesgo de la interpretación inadecuada.

El peligro de llegar a interpretaciones inadecuadas y erróneas en la investigación no experimental surge, en parte, de la posibilidad de muchas explicaciones para eventos complejos. Es fácil aceptar la primera y más obvia interpretación de una relación establecida, especialmente si se trabaja sin hipótesis que guíen la investigación. La investigación que no está guiada por hipótesis, o la investigación que se realiza "para averiguar cosas", con frecuencia es no experimental. La investigación experimental tiende más a basarse en hipótesis establecidas cuidadosamente.

Las hipótesis son predicciones de la forma si-entonces. En un experimento de investigación, la predicción se hace a partir de una x bien controlada, a una y. Si la predicción resulta verdadera, entonces se está relativamente a salvo al plantear la proposición condicional: "si x, entonces y". No obstante, en un estudio no experimental bajo las mismas condiciones, se está considerablemente menos a salvo al establecer la proposición condicional, por las razones discutidas anteriormente. Las protecciones cuidadosas son más importantes en el último caso, especialmente en la selección y comprobación de hipótesis alternativas, tales como la predicha carencia de relación entre el examen eleven-plus y la ansiedad en general, en el estudio de Sarnoff. Una relación predicha (o no predicha) en investigación no experimental puede resultar bastante espuria; aunque su plausibilidad y conformidad con la preconcepción puede volverla fácil de aceptar. Éste es un peligro en la investigación experimental, pero es menos peligroso que en la investigación no experimental, pues una situación experimental resulta mucho más fácil de controlar.

La investigación no experimental que se realiza sin hipótesis y sin predicciones, es decir, aquella en la cual los datos simplemente se recolectan y luego se interpretan, es aún más peligrosa por su capacidad de generar confusión. Si es posible, se localizan las diferencias o correlaciones significativas, y luego se interpretan. El segundo autor de este libro ha visto a estudiantes de posgrado recolectar una gran cantidad de datos sin hipótesis, y luego utilizar un programa de cómputo para realizar cualquier análisis posible, con la esperanza de encontrar significancia estadística en alguna parte. Después de encontrar diferencias significativas, se desarrollan las hipótesis que se adecuen al análisis. Para ilustrar el problema, suponga que un educador decide estudiar los factores que conducen a un bajo rendimiento académico. Se selecciona un grupo de sujetos con bajo rendimiento y un grupo de

personas con rendimiento normal. A cada grupo se le aplica una batería de pruebas. Después, se calculan las medias de las pruebas de los dos grupos, y las diferencias entre las medias se analizan mediante pruebas *t*. Digamos que de 12 diferencias, tres son significativas. Entonces, el investigador concluye que las personas con bajo rendimiento y las personas con rendimiento normal difieren respecto a las variables medidas con esas tres pruebas. Armado con estos análisis, el investigador ahora se siente deseoso de señalar a otros qué aspectos caracterizan a aquellos con bajo rendimiento. Puesto que las tres pruebas parecen medir inseguridad, entonces la causa del bajo rendimiento es la inseguridad.

Cuando están guiados por hipótesis, la credibilidad de los resultados de los estudios, como el descrito anteriormente, puede incrementarse; pero los resultados permanecen débiles porque se generan por el azar: debido únicamente al azar, uno o dos resultados de muchas pruebas estadísticas pueden ser significativos; sobre todo, la plausibilidad puede confundir. Una explicación plausible con frecuencia parece irresistible —¡aunque muy equivocada!—. Por ejemplo, parece muy obvio que los conservadores y los liberales son opositores; sin embargo, la evidencia de investigación parece indicar que no es así (Kerlinger, 1967, 1980, 1984). Otra dificultad es que las explicaciones plausibles, una vez halladas y creídas, a menudo son difíciles de probar. Según Merton (1949), las explicaciones *post factum* no conducen, por sí mismas, a la nulificación, ya que son demasiado flexibles. Cualesquiera que sean las observaciones, indica, pueden encontrarse nuevas interpretaciones que "se adecuen a los hechos" (pp. 90-91).

El valor de la investigación no experimental

A pesar de sus debilidades, en psicología, sociología y educación debe realizarse gran cantidad de investigación no experimental tan sólo porque muchos problemas de investigación no se prestan al cuestionamiento experimental. Una breve reflexión respecto a algunas de las variables importantes en investigación del comportamiento —inteligencia, aptitud, antecedentes familiares, rendimiento, clase social, rigidez, etnocentrismo— mostrará que no son manipulables. La búsqueda controlada es posible, por supuesto, aunque no la experimentación verdadera.

Incluso puede decirse que la investigación no experimental es más importante que la investigación experimental; ésta no es, por supuesto, una observación metodológica; más bien significa que la mayoría de los problemas de investigación científica social y educativa no conducen por sí mismos a la experimentación, aunque muchos de ellos conducen a la búsqueda controlada del tipo no experimental. Considere los estudios de Piaget sobre el pensamiento de los niños; los estudios de Adorno, Frenkel-Brunswik, Levinson y Sanford sobre el autoritarismo; el muy importante estudio de *Equality of Educational Opportunity*, y el estudio de Cochran y May sobre las mentiras y la práctica del sexo seguro. Considere además la influencia del estudio no experimental respecto del tabaquismo y de los problemas de salud: condujo a la creación de una legislación específica respecto de poner advertencias impresas en los propios productos. Si se hiciera un registro de estudios firmes e importantes en las ciencias del comportamiento y educación, es posible que los estudios no experimentales superaran en número y calidad a los estudios experimentales.

Conclusiones

Los estudiantes de investigación difieren ampliamente en sus puntos de vista respecto a los valores relativos de la investigación experimental y de la no experimental. Están aquellos

que exaltan la investigación experimental y subestiman la no experimental y aquellos que critican la discutida estrechez y la carencia de "realidad" de los experimentos, especialmente los de laboratorio. Dichas críticas, sobre todo en educación, enfatizan el valor y la relevancia de la investigación no experimental en situaciones "de la vida real" y "naturales" (como repaso consulte a Keith, 1988). Una posición racional parece obvia. Si es posible, utilice la experimentación porque, *si el resto de las cosas permanecen iguales*, es posible interpretar los resultados de la mayoría de los experimentos con gran confianza en que las proposiciones de la forma "si p, entonces q" son lo que se dice que son. También parece deseable probar las proposiciones "si p, entonces q" en otros escenarios. Se buscaría evidencia no experimental de la validez empírica de la propia hipótesis. Así, si es posible, las proposiciones condicionales deben estudiarse utilizando tanto el modelo experimental como el no experimental. Algunos estudios de investigación no experimental son impresionantes y convincentes. ¡Pero cuánto más impresionante y convincente sería si conclusiones similares surgieran de experimentos bien conducidos! De forma inversa, cuánto más convincentes son las conclusiones experimentales cuando se basan en investigación no experimental bien conducida.

La réplica es siempre deseable, incluso necesaria. Un aspecto importante a remarcar es que la réplica de la investigación no sólo significa repetición de los mismos estudios en las mismas situaciones. Podría y debería significar la comprobación de las implicaciones empíricas de la teoría —interpretando el término "teoría" en un sentido amplio— en situaciones similares y diferentes, y experimental y no experimentalmente. Es más fácil pedir extensiones de la investigación del laboratorio hacia el campo; pero los investigadores deben intentar concebir la comprobación experimental de proposiciones surgidas de manera no experimental. Por supuesto, esto es más difícil y rara vez se hace. Lo importante aquí es que *debe* concebirse y, cuando sea posible, hacerse.

El adoptar una posición firme de que la investigación experimental y no experimental es el único camino al cielo de la investigación es un asunto dogmático. Quizás sea muy difícil, incluso imposible en muchos casos, realizar tanto investigación experimental como no experimental en un mismo problema. ¿Será posible manipular experimentalmente la variable género de Cochran y May, o la variable cultural de Chen, Lee y Stevenson, por ejemplo? Por supuesto que difícil no quiere decir imposible. La cuestión aquí es que las posibilidades experimentales y las no experimentales deben ser exploradas y explotadas cuando sea posible hacerlo. Además, no debe asumirse de manera inmediata que no es posible realizar investigación de manera distinta a como se ha hecho. No existe un solo camino metodológico hacia la validez científica; existen muchos. Los caminos deben elegirse por su adecuación a los problemas bajo estudio. Sin embargo, esto no quiere decir que no se pueda explotar un modelo que difiere de aquello a lo que se está acostumbrado.

Por alguna extraña razón, quizás la creencia espuria en la supuesta certeza de la ciencia, cuando la gente —incluyendo a los científicos— piensa en la ciencia y en la investigación científica, se considera erróneamente que tan sólo existe una forma "correcta" de acercamiento y de hacer investigación. Muy raras veces se comete dicho error en la música, el arte o en la construcción de una casa. También la ciencia tiene muchos caminos, y los modelos experimental y no experimental son dos de ellos. Ninguno de ellos es correcto o erróneo; sino que son diferentes. La tarea aquí ha sido tratar de entender las diferencias y sus consecuencias. Sin embargo, todavía falta mucho para finalizar el tema. Es probable que se logre más comprensión antes de finalizar. Cuando se piense en los diferentes puntos de vista de los métodos experimentales y los no experimentales, debe considerarse la máxima china que afirma que "existen muchos caminos hacia la cima de la montaña, pero la vista ahí siempre es la misma".

RESUMEN DEL CAPÍTULO

1. Cuando se conducen de forma correcta los estudios no experimentales son tan valiosos como los experimentales.
2. Un ingrediente para un buen estudio no experimental es el desarrollo de hipótesis antes de iniciar el estudio.
3. La réplica sirve para incrementar la credibilidad de los resultados obtenidos a partir de estudios no experimentales.
4. La investigación no experimental se define como aquella que no posee una variable independiente activa.
5. La diferencia más importante entre los métodos experimental y no experimental es el control.
6. La autoselección de los participantes es un problema importante de los estudios no experimentales.
7. Existe un gran número de estudios no experimentales realizados y publicados en las ciencias del comportamiento.
8. Algunos estudios no experimentales —como el que relacionó el tabaquismo con los problemas de salud— han tenido una gran influencia.
9. Existen tres debilidades principales en la investigación no experimental: (i) las variables independientes no pueden ser manipuladas, (ii) la carencia de aleatorización, y (iii) el riesgo de la interpretación inadecuada.
10. La eliminación paso a paso de las hipótesis alternativas es una forma de llegar a la variable que posiblemente "causa" los cambios en la variable dependiente.
11. Desarrollos relativamente nuevos en estudios no experimentales incluyen el modelo matemático de "causa y efecto". Estos modelos, en realidad, no implican causa y efecto.

SUGERENCIAS DE ESTUDIO

1. Un psicólogo social planea investigar los factores que subyacen al antisemitismo. Él considera que las personas que tuvieron padres autoritarios y una educación autoritaria tienden a ser antisemitas. ¿Un proyecto de investigación diseñado para probar esta hipótesis sería experimental o no experimental? Explique.
2. Un psicólogo educativo decide probar la hipótesis de que la inteligencia y la motivación son los principales factores determinantes del éxito escolar. ¿Esta investigación sería principalmente experimental o no experimental? Argumente.
3. Un investigador está interesado en la relación entre la percepción del papel y los valores sociales.
   a) ¿Cuál es la variable independiente? ¿Y la variable dependiente?
   b) Cualquiera que haya sido su juicio, ¿puede invertir de forma justificada las variables?
   c) ¿Cree usted que un proyecto de investigación diseñado para investigar este problema sería básicamente experimental o no experimental?
   d) ¿Puede el investigador hacer dos investigaciones, una experimental y otra no experimental, ambas diseñadas para probar la misma hipótesis?
   e) Si su respuesta para *d)* fue "Sí", ¿serían las mismas variables en los dos problemas? Suponiendo que las relaciones en ambas investigaciones fuesen significativas, ¿las conclusiones serían sustancialmente las mismas?

TABLA 23.2 *Países con alta y baja motivación de logro, cuya producción de energía eléctrica estuvo por arriba o por debajo de las expectativas (estudio de McClelland)*[a]

	Por arriba de las expectativas	Por debajo de las expectativas
Alta motivación de logro	13 (65%)	7 (35%)
Baja motivación de logro	5 (26%)	14 (70%)

[a] Los datos en las casillas son el número de países que, por ejemplo, tuvieron alta motivación de logro y cuya producción de electricidad estuvo por arriba de las expectativas (13). Los índices en los paréntesis son los porcentajes.

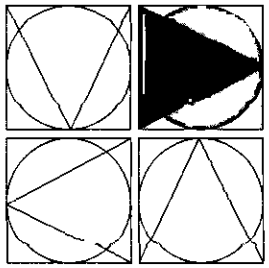
¿Apoyan estos resultados la hipótesis de McClelland? (Sugerencia: Calcule χ^2 y C, como en el capítulo 10. Utilice los porcentajes para ayudarse a interpretar la tabla.)

[Respuesta: $\chi^2 = 5.87$, $gl = 1$ ($p < .05$); C = .36. Sí, la hipótesis se mantiene.]

4. En las sugerencias de estudio del capítulo 2 se presentaron una serie de problemas e hipótesis. Tome cada uno de estos problemas e hipótesis y decida si la investigación diseñada para explorar los problemas y para probar las hipótesis serían básicamente experimentales o no experimentales. ¿Puede alguno de los problemas e hipótesis tratarse de ambas maneras?
5. McClelland (1961) presenta datos sobre la producción de energía eléctrica durante los años 1952-1958 de países con alta motivación de logro y baja motivación de logro. Al contar el número de países en cada una de las cuatro casillas se obtienen los datos que se muestran en la tabla 23.2.
6. El estudiante emprendedor quizás desee dar un paso decisivo hacia un pensamiento estimulante, provocativo, controvertido e importante. El famoso reporte del Club de Roma de Meadows, Meadows, Randers y Behrens (1974) molestó a algunos observadores, alarmó casi a todos los que lo han leído y ofendió a todos. Con el uso de variables sociales importantes —recursos naturales, contaminación, población, por ejemplo— y sus interacciones complejas, se ha pronosticado un desastre final para las ciudades y el mundo. La investigación en que se basan las conclusiones es totalmente no experimental. Lea este reporte. ¿Considera que el carácter no experimental de la investigación disminuye su credibilidad?
7. Lea uno de (o todos) los siguientes estudios. Todos son no experimentales. Escriba las razones por las que usted piensa que son no experimentales con base en los puntos resaltados en este capítulo.

Goodman, S. H. y Emory, E. K. (1992). Perinatal complications in births to low socioeconomic status schizophrenic and depressed women. *Journal of Abnormal Psychology, 101*, 225-229.

Koniak-Griffin, D. y Brecht, M. (1995). Linkages between sexual risk taking, substance use, and AIDS knowledge among pregnant adolescents and young mothers. *Nursing Research, 44*, 340-346.



CAPÍTULO 24

EXPERIMENTOS DE LABORATORIO, EXPERIMENTOS DE CAMPO Y ESTUDIOS DE CAMPO

La investigación científica social puede dividirse en cuatro grandes categorías: experimentos de laboratorio, experimentos de campo, estudios de campo e investigación mediante encuestas. Esta clasificación surge de dos fuentes: la distinción entre investigación experimental y no experimental, y la distinción entre investigación de laboratorio y de campo. Este capítulo le debe mucho al material de tres libros: 1) Festinger y Katz (1953),

2) Taylor y Bodgan (1998) y 3) Padgett (1998). Aunque la publicación de Festinger y Katz tiene más de 45 años, continúa siendo una fuente valiosa sobre muchos aspectos de la metodología de investigación del comportamiento. Este capítulo inicia con la presentación de ejemplos del experimento de laboratorio, dos experimentos de campo y un estudio de campo. Esto es para que el lector perciba los principales componentes de cada método y las diferencias entre cada uno.

Experimento de laboratorio: estudios de Miller del aprendizaje de respuestas viscerales

Una serie de brillantes experimentos realizados por Miller (1969, 1971) ha trastornado una creencia mantenida por largo tiempo: que el aprendizaje ocurre tan sólo con respuestas voluntarias, y que el sistema autónomo involuntario está sujeto únicamente al condicionamiento clásico. Esto, en efecto, indica que respuestas como mover la mano y hablar pueden controlarse y, por lo tanto, enseñarse; pero que las respuestas involuntarias, como la frecuencia cardiaca, las contracciones intestinales y la presión sanguínea no pueden tenerse bajo control instrumental y, por lo tanto, no pueden "enseñarse". Para entender los estudios de Miller es necesario definir ciertos términos psicológicos. En el *condicionamiento clásico* un estímulo neutral, inherentemente incapaz de producir cierta respuesta, se vuelve capaz de hacerlo al ser asociado repetidamente con un estímulo inherentemente capaz de producirla. El ejemplo más famoso es el del perro de Pavlov que salivaba ante el sonido de un metrónomo, que había sido asociado repetidamente con carne en polvo. En el *condicionamiento instrumental* u *operante*, dar un reforzamiento a un organismo, inmediatamente después de haber dado una respuesta, produce un incremento de la respuesta. Recompense una respuesta y ésta se repetirá. Se enseña que las respuestas o conductas voluntarias son superiores, quizás porque están bajo el control del individuo; mientras que las respuestas involuntarias son inferiores porque no son controladas. Se ha considerado que las respuestas involuntarias pueden modificarse únicamente por medio del condicionamiento clásico y no por medio del condicionamiento instrumental. En otras palabras, la posibilidad de "enseñar" al corazón, estómago o a la sangre es remota, ya que las situaciones del condicionamiento clásico son difíciles de lograr. No obstante, si se sujeta a los órganos al condicionamiento instrumental, pueden ponerse bajo control experimental y pueden ser "enseñados"; y ellos pueden "aprender".

El trabajo de Miller ha demostrado que, a través de condicionamiento instrumental, la frecuencia cardiaca es susceptible de modificarse, las contracciones estomacales llegan a alterarse e inclusive la formación de orina ¡puede incrementarse o disminuirse! Tal descubrimiento tiene una importancia teórica y práctica enorme. Para mostrar la naturaleza de los experimentos de laboratorio se tomará uno de los experimentos más interesantes y creativos de Miller.

La idea del experimento resulta simple: recompensar a un grupo de ratas cuando su frecuencia cardiaca aumente, y recompensar a otro grupo cuando su frecuencia cardiaca disminuya. Éste es un ejemplo directo del diseño de dos grupos analizado anteriormente. El gran problema de Miller era el control. Existen muchas otras causas del cambio en la frecuencia cardiaca, por ejemplo, el esfuerzo muscular. Para controlar dichas variables extrañas, Miller y su colega (Trowill) paralizaron a las ratas con curare. Pero si las ratas estaban paralizadas, ¿qué podría utilizarse como recompensa? Decidieron utilizar estimulación eléctrica directa en el cerebro. La variable dependiente, frecuencia cardiaca, se registraba de manera continua con el electrocardiógrafo. Cuando ocurría un pequeño

cambio en la frecuencia cardiaca (en la dirección "correcta": incremento para un grupo y decremento para el otro), se le aplicaba un impulso eléctrico al animal en un centro de recompensa de su cerebro (véase también Olds y Fobes, 1981, una investigación cerebral que demuestra que una pequeña estimulación eléctrica de cierta parte del cerebro actúa como recompensa). Esto se continuó hasta que los animales estuvieron "entrenados".

Los incrementos y decrementos de la frecuencia cardiaca fueron confiables estadísticamente, pero pequeños: sólo 5% en cada dirección. Por ello, Miller y otro colega (DiCara) utilizaron la técnica conocida como moldeamiento que, en este caso, implicó recompensar primero pequeños cambios y después exigir cambios crecientes en la frecuencia para obtener las recompensas. Esto incrementó los cambios en la frecuencia cardiaca a un promedio de 20% en cada dirección. Además, investigación posterior, donde se utilizó el escape de un choque eléctrico leve como reforzamiento, demostró que los animales recordaron lo que habían aprendido y que "diferenciaron" las respuestas cardiacas de otras respuestas.

Miller ha tenido éxito al "entrenar" otras respuestas involuntarias: contracción intestinal, formación de orina y presión sanguínea, por ejemplo. En resumen, las respuestas viscerales *pueden aprenderse* y *pueden moldearse*; ¿pero puede utilizarse este método con personas? Miller dice que él considera que las personas son tan inteligentes como las ratas, pero que no ha sido completamente probado todavía. Aunque el uso de curare podría presentar dificultades, Miller dice que puede hipnotizarse a la gente.

Un experimento de campo: el estudio de Rind y Bordia sobre los efectos del agradecimiento de un mesero y la personalización en las propinas de los restaurantes

¿La práctica común entre los meseros de escribir "gracias" al reverso de la cuenta y entregarla, de tal manera que el comensal vea la gratitud del mesero, produce mayores propinas? Si en realidad es así, entonces el mesero se beneficia con esta acción a un costo extremadamente bajo. Rind y Bordia (1995) realizaron este experimento de campo para determinar la efectividad de utilizar dicha técnica: específicamente, escribir "gracias" y personalizar la interacción mesero-comensal añadiendo el nombre del mesero. El estudio fue realizado en un elegante restaurante de Filadelfia durante el periodo del almuerzo, por cinco días. Participaron 51 comensales en el estudio. Todas las meseras eran del sexo femenino. La variable independiente era la *impresión* y consistió de tres niveles: 1) la parte posterior de la cuenta no contenía nada, 2) la palabra "gracias" escrita a mano, o 3) la palabra "gracias" más el nombre de pila de la mesera. Rind y Bordia hipotetizaron, con base en la teoría del manejo de la impresión, que el añadir el agradecimiento escrito y personalizado conduciría a propinas más altas, que si la mesera no escribía nada al reverso de la cuenta. También hipotetizaron que la personalización de la cuenta conduciría a propinas más altas que sin la personalización. Cada nivel o condición de la variable independiente se determinó aleatoriamente para cada grupo de comensales. Antes de entregar la cuenta al comensal, la mesera tomaba aleatoriamente de su bolsillo una de tres monedas (fechadas 1981, 1982 y 1983). Si tomaba la moneda de 1981, la mesera no escribía nada al reverso de la cuenta. Si seleccionaba la moneda de 1982, anotaba "gracias" al reverso de la cuenta. Si la moneda elegida era la de 1983, escribía "gracias" al reverso de la cuenta, pero también añadía su nombre. Se registró el tamaño de la propina, el tamaño de la cuenta, el tamaño del grupo de comensales y la forma de pago, para cada grupo de comensales. Los resultados de dicho estudio demostraron que agregar la palabra "gracias"

a la cuenta producía propinas significativamente mayores que cuando no se escribía nada en la cuenta (18% de las cuentas contra 16.3%). No hubo diferencias significativas entre el agradecimiento escrito y el agradecimiento escrito más la personalización. Rind y Bordia mencionan que existen teorías encontradas respecto al porqué de sus hallazgos. Sin embargo, a partir de sus resultados, parece que esta práctica es benéfica para el mesero. Los investigadores también señalan las limitaciones de su experimento. Una de ellas es que su elección de conducir el estudio en un restaurante elegante puede generar resultados diferentes que si el estudio se hubiese realizado en un restaurante popular. El uso de mujeres únicamente abre la posibilidad de que los clientes pudieran tratar a los meseros hombres de manera diferente.

Un estudio de campo: el estudio de Bennington College realizado por Newcomb

Newcomb (1943) llevó a cabo uno de los estudios clásicos más importantes respecto a la influencia del ambiente universitario sobre los estudiantes. En él examinó al cuerpo entero de estudiantes del Bennington College (cerca de 600 mujeres jóvenes), de 1935 a 1939. Una faceta inusual del estudio fue el intento de Newcomb de explicar la influencia tanto de los factores personales como de los sociales sobre los cambios de actitud de los estudiantes. Aunque también se probaron otras hipótesis, la hipótesis principal del estudio de Bennington fue que los nuevos estudiantes coincidirían respecto a las normas del grupo universitario, y que cuanto más se integraran a la comunidad universitaria, mayor sería el cambio en sus actitudes sociales.

Newcomb utilizó varias escalas de actitud de papel y lápiz, informes escritos sobre los estudiantes y entrevistas individuales. El estudio fue longitudinal y no experimental. La variable independiente, que aunque no fue fácil de categorizar, puede decirse que fueron las normas sociales del Bennington College. La variable dependiente fueron las actitudes sociales y ciertos comportamientos de las estudiantes.

Newcomb encontró cambios significativos en las actitudes entre las estudiantes de nuevo ingreso, por un lado, y las estudiantes de niveles intermedios y de último año, por el otro. Los cambios se dieron hacia un menor conservadurismo en una variedad de aspectos sociales. Por ejemplo, las preferencias políticas de los estudiantes intermedios y del último año, en la elección presidencial de 1936, fueron mucho menos conservadoras que las de hombres jóvenes y estudiantes de segundo año. De 50 estudiantes intermedios y del último año, 15% preferían a Landon (republicano), mientras que de 52 estudiantes de nuevo ingreso, 62% preferían a Landon. Los porcentajes de preferencia por Roosevelt (demócrata) fueron 54% y 29%. Las puntuaciones medias de todas las estudiantes de los cuatro años, en una escala diseñada para medir conservadurismo político y económico, fueron: estudiantes de nuevo ingreso, 74.2; estudiantes de segundo año, 69.4; intermedios, 65.9, y estudiantes del último año, 62.4. Evidentemente la universidad afectó las actitudes de las estudiantes.

Newcomb planteó una pregunta de "control": ¿habrían cambiado estas actitudes en otras universidades? Para responder esta pregunta, administró sus medidas de conservadurismo a estudiantes del Williams College y del Skidmore College. Las puntuaciones medias comparativas de los estudiantes de Skidmore —desde los estudiantes de recién ingreso hasta los de último año— fueron 79.9, 78.1, 77.0 y 74.1. Parece que los estudiantes de Skidmore (y Williams) no cambiaron tanto, ni tan consistentemente a través del tiempo, como lo hicieron los de Bennington.

Newcomb, Koenig, Flacks y Warwick (1967) reportaron un estudio de seguimiento de los estudiantes del Bennington College, después de 25 años. Encontraron que los cambios habían perdurado y que la influencia de Bennington era persistente.

Características y criterios de los experimentos de laboratorio, experimentos de campo y estudios de campo

Un *experimento de laboratorio* es una investigación en la que la varianza de todas, o de casi todas, las posibles variables independientes influyentes, sin pertinencia al problema de investigación inmediato, se mantienen al mínimo. Esto se logra aislando la investigación en una situación física separada de la rutina de la vida ordinaria, y por medio de manipular una o más variables independientes bajo condiciones rigurosamente específicas, operacionalizadas y controladas.

Fortalezas y debilidades de los experimentos de laboratorio

El experimento de laboratorio tiene la virtud inherente de la posibilidad de un control relativamente completo. El experimento de laboratorio puede, y frecuentemente lo hace, aislar la situación de investigación de la vida fuera del laboratorio, al eliminar las muchas influencias extrañas que lleguen a afectar las variables independiente y dependiente.

Además del control de la situación, los experimentos de laboratorio generalmente se sirven de la asignación aleatoria y manipulan una o más variables independientes. Existen otros aspectos del control de laboratorio: en la mayoría de los casos, el investigador logra un alto grado de especificidad en las definiciones operacionales de las variables. Las relativamente crudas definiciones operacionales de las situaciones de campo, tales como muchas de las que se asocian con la medición de valores, actitudes, aptitudes y características de personalidad, no saturan al experimentador; aunque el problema de definición nunca resulta simple. El experimento de Miller (1969, 1971) constituye un buen ejemplo. Las definiciones operacionales del reforzamiento y del cambio en la frecuencia cardiaca son precisas y altamente objetivas.

Muy relacionada con la fortaleza operacional está la precisión de los experimentos de laboratorio. Preciso significa exacto, definido y no ambiguo. Las mediciones precisas se efectúan con instrumentos de precisión. En términos de varianza, cuanto más preciso sea un procedimiento experimental, menor será la varianza del error. Cuanto más exacto o preciso sea un instrumento de medición, mayor certeza se tendrá de que las medidas obtenidas no varían mucho de sus valores "verdaderos".

Los resultados precisos de laboratorio se logran principalmente por medio de la manipulación y medición controladas, en un ambiente donde se eliminaron las posibles condiciones "contaminantes". Los reportes de investigación de experimentos de laboratorio generalmente especifican en detalle la manera en que se realizaron las manipulaciones y los medios utilizados para controlar las condiciones del ambiente bajo las cuales se efectuaron. Al especificar de manera exacta las condiciones del experimento, se reduce el riesgo de que los participantes puedan responder erróneamente y, por lo tanto, de introducir varianza aleatoria a la situación experimental. El experimento de Miller constituye un modelo de precisión experimental de laboratorio.

La mayor debilidad del experimento de laboratorio probablemente es la carencia de fortaleza de las variables independientes. Puesto que las situaciones de laboratorio son, después de todo, situaciones creadas para propósitos especiales, puede decirse que los

efectos de las manipulaciones experimentales son generalmente débiles. Los incrementos y los decrementos en la frecuencia cardiaca por medio del reforzamiento eléctrico del cerebro fueron, aunque asombrosas, relativamente pequeñas. Compare esto con los efectos relativamente grandes de las variables independientes en situaciones reales. En el estudio Bennington, por ejemplo, la comunidad universitaria aparentemente tuvo un efecto masivo. En la investigación de laboratorio sobre la conformidad, generalmente se producen tan sólo pequeños efectos por la presión grupal sobre los individuos. Confronte lo anterior con el efecto relativamente fuerte de la mayoría de un gran grupo sobre un miembro individual de un grupo, en una situación de la vida real. El miembro del consejo de educación que sabe que cierta acción deseable va en contra de los deseos de la mayoría de sus colegas, y quizás de la mayoría de la comunidad, se encuentra bajo mucha presión para coincidir con la norma.

Una razón de la preocupación respecto a la precisión del laboratorio y de la estadística refinada es la debilidad de los efectos de laboratorio. Detectar una diferencia significativa en el laboratorio requiere de situaciones y medidas con el menor ruido debido al azar, y de pruebas estadísticas precisas y sensibles que muestren relaciones y diferencias significativas cuando existan.

Otra debilidad es un producto de la primera: la artificialidad de la situación de investigación experimental. En realidad, es difícil saber si la artificialidad es una debilidad o simplemente una característica neutral de las situaciones experimentales de laboratorio. Cuando una situación de investigación se idea deliberadamente para excluir las muchas distracciones del ambiente, quizá resulte ilógico etiquetar a la situación con un término que exprese en parte el resultado que se busca. La crítica sobre la artificialidad no proviene de los experimentadores que saben que las situaciones experimentales son artificiales; sino proviene de los individuos que no comprenden los propósitos de los experimentos de laboratorio.

La tentación de interpretar de forma incorrecta los resultados de los experimentos de laboratorio es enorme. Mientras que los resultados de Miller son considerados altamente significativos por los científicos sociales, sólo pueden ser extrapolados de manera tentativa a situaciones fuera del laboratorio. Resultados similares pueden obtenerse en situaciones de la vida real, y existe evidencia de que así sucede en algunos casos; pero esto no es así necesariamente. La relación debe siempre probarse una y otra vez bajo situaciones que no sean de laboratorio. Por ejemplo, la investigación de Miller tendrá que realizarse cuidadosa y precavidamente con seres humanos en hospitales e inclusive en escuelas.

A pesar de que los experimentos de laboratorio tienen una validez interna relativamente alta, carecen de validez externa. Antes se planteó la pregunta: ¿hizo realmente *X*, la manipulación experimental, una diferencia significativa? A mayor confianza en la "verdad" de las relaciones descubiertas en un estudio de investigación, mayor será la validez interna del estudio. Cuando se descubre una relación en un experimento de laboratorio bien realizado, por lo común se tiene bastante confianza en él, pues se ha ejercido el máximo control posible de la variable independiente y de otras posibles variables independientes extrañas. Cuando Miller "descubrió" que las respuestas viscerales podían aprenderse y moldearse, quizás estuvo relativamente seguro de la "verdad" de la relación entre el reforzamiento y la respuesta visceral en el laboratorio. Él consiguió un alto grado de control y de validez interna.

Puede decirse que si se estudia un problema utilizando experimentos de campo, *quizá* se encontrará alguna relación. Ésta es una cuestión empírica, no especulativa; la relación que se desea probar debe ponerse en la situación donde se quiere generalizar. Si un investigador encuentra que los individuos coinciden respecto a las normas grupales en el laboratorio, ¿ocurrirá el mismo fenómeno o uno similar en los grupos comunitarios, facultades

y cuerpos legislativos? Esta carencia de validez externa forma la base de las objeciones de muchos educadores respecto a los estudios con animales de las teorías de aprendizaje. Sus objeciones son válidas únicamente si un experimentador generaliza, a partir del comportamiento y aprendizaje de animales de laboratorio, al comportamiento y aprendizaje de los niños. Sin embargo, los experimentalistas capaces rara vez cometen un error de este tipo —ellos saben que el laboratorio constituye un ambiente restringido—.

Propósitos del experimento de laboratorio

Los experimentos de laboratorio tienen tres propósitos relacionados. Primero, son un medio para estudiar las relaciones bajo condiciones "puras" y no contaminadas. Los experimentadores se plantean las siguientes preguntas: ¿está x relacionada con y? ¿Cómo se relaciona con y? ¿Qué tan fuerte es la relación? ¿Bajo qué circunstancias cambia la relación? Ellos buscan escribir ecuaciones de la forma $y = f(x)$, hacer predicciones con base en la función, y ver qué tan bien y bajo qué condiciones se lleva a cabo la función.

Un segundo propósito debe mencionarse en conjunción con el primer propósito: en un inicio hay que comprobar las predicciones derivadas de la teoría y, después, aquellas derivadas de otras investigaciones.

Un tercer propósito de los experimentos de laboratorio consiste en refinar las teorías y la hipótesis, para formular hipótesis relacionadas con otras hipótesis probadas experimental o no experimentalmente y, quizás lo más importante, ayudar a la construcción de sistemas teóricos. Éste fue uno de los principales propósitos de Miller. Aunque algunos experimentos de laboratorio se realizan sin dicho propósito, la mayoría de los experimentos de laboratorio están, por supuesto, orientados a la teoría.

Entonces, el propósito de los experimentos de laboratorio consiste en probar hipótesis derivadas de la teoría, estudiar las interrelaciones precisas de variables y su operación, y controlar la varianza bajo condiciones de investigación que no estén contaminadas por la operación de variables extrañas. Como tal, el experimento de laboratorio es uno de los más grandes inventos de todos los tiempos. A pesar de que existen debilidades, lo son sólo en un sentido realmente irrelevante. Admitiendo la carencia de representatividad (validez externa), el experimento de laboratorio bien realizado aún cumple el prerrequisito fundamental de cualquier investigación: la validez interna.

El experimento de campo

Un experimento de campo consiste en un estudio de investigación realizado en una situación real, donde una o más variables independientes son manipuladas por el experimentador bajo condiciones tan cuidadosamente controladas como la situación lo permita. El contraste entre el experimento de laboratorio y el experimento de campo no es grande: las diferencias son principalmente cuestiones de grado. Algunas veces resulta difícil etiquetar un estudio particular como "experimento de laboratorio" o como "experimento de campo". En tanto que el experimento de laboratorio tiene un control máximo, la mayoría de los experimentos de campo deben operar con menos control, un factor que a menudo constituye una severa limitante.

Fortalezas y debilidades de los experimentos de campo

Los experimentos de campo poseen valores que los recomiendan especialmente para los psicólogos sociales, sociólogos y educadores, a causa de que se ajustan de forma admirable

a muchos de los problemas sociales y educativos de interés para la psicología social, la sociología y la educación. Puesto que las variables independientes se manipulan y se utiliza la aleatorización, puede cumplirse el criterio de control —por lo menos teóricamente—.

No obstante, el control de la situación experimental de campo rara vez resulta tan rígido como el del laboratorio. Esto implica tanto una fortaleza como una debilidad. En un experimento de campo, aunque el investigador tiene el poder de la manipulación, siempre se enfrenta con la desagradable posibilidad de que las variables independientes estén contaminadas por variables ambientales no controladas. Se enfatiza este punto pues la necesidad de controlar variables independientes extrañas es particularmente crítica en los experimentos de campo. El experimento de laboratorio se lleva a cabo en una situación altamente controlada; mientras que el experimento de campo se realiza en una situación natural y frecuentemente laxa. Por lo tanto, una de las preocupaciones principales del experimentador de campo consiste en tratar de hacer que la situación de investigación se aproxime lo más posible a las condiciones del experimento de laboratorio. Por supuesto que con frecuencia ésta es una meta difícil de lograr, pero si la situación de investigación puede mantenerse rígida, entonces el experimento de campo es poderoso porque, en general, es posible tener mayor confianza en que las relaciones en realidad sean lo que se dice que son.

Como compensación para la disminución del control, el experimento de campo posee dos o tres virtudes únicas. Las variables de un experimento de campo por lo común tienen un efecto más poderoso que las de los experimentos de laboratorio. Los efectos de los experimentos de campo a menudo son lo suficientemente fuertes para penetrar las distracciones de las situaciones experimentales. El principio es: *cuanto más realista sea la situación de investigación, más fuertes serán las variables.* Ésta es una de las ventajas de realizar investigación en ambientes educativos. En su mayoría, la investigación en los ambientes educativos es similar a las actividades educativas y, por lo tanto, no debe verse necesariamente como algo especial y separado de la vida escolar. A pesar de la petición de muchos educadores porque se realice investigación educativa más realista, no existe virtud especial en el realismo por sí mismo. El realismo tan sólo incrementa la fortaleza de las variables; también contribuye a la validez externa, ya que a mayor realismo de la situación, habrá mayor posibilidad de que las generalizaciones a otras situaciones resulten más validas.

Otra virtud de los experimentos de campo es que son apropiados para estudiar influencias, procesos y cambios sociales y psicológicos complejos, en situaciones similares a la vida real. Glick, DeMorest y Hotze (1988), por ejemplo, estudiaron los efectos de la pertenecia al grupo, del espacio personal y de las solicitudes de ayuda sobre la ansiedad interpersonal y la obediencia de los individuos. Los investigadores realizaron su estudio en una plaza comercial, utilizando compradores reales como participantes. Schmitt, Dube y Leclerc (1992) estudiaron un problema similar sobre el espacio personal, examinando intrusiones en filas de espera. Estos investigadores condujeron tres experimentos de laboratorio y uno de campo, como un intento para determinar si las reacciones conductuales a la intrusión se basan en intereses personales o sociales. Jaffe (1991) realizó un experimento de campo sobre anuncios comerciales dirigidos a las mujeres. En dicha investigación los participantes evaluaron un anuncio impreso que incluía mujeres en diferentes situaciones. La situación utilizada era la de la mujer tradicional (nutriente y orientada a la familia) o la mujer moderna (exitosa en su carrera y en la familia). Rabinowitz, Colmar, Elgie, Hale, Niss, Sharp y Sinclito (1993) estudiaron la conducta compleja de los cajeros en tiendas de recuerdos de viaje que atienden a turistas. Los investigadores deseaban saber si el mal manejo del dinero se debía a la deshonestidad, la indiferencia o al descuido. El estudio de Wogalter y Young (1991) sobre la eficacia de advertencias verbales o impresas en el manejo de sustancias peligrosas, o en el piso resbaloso de una plaza comercial, son útiles para

quienes se preocupan de los aspectos de seguridad en ambientes industriales o comerciales. Wogalter y Young hicieron dos estudios de laboratorio y un experimento de campo para demostrar que la combinación de advertencias verbales e impresas era lo más eficaz para producir la conducta de obediencia en la gente. Todos estos estudios utilizaron manipulación experimental en los participantes de la vida real, en ambientes reales.

Los experimentos de laboratorio son adecuados principalmente para comprobar aspectos de teorías; mientras que los experimentos de campo son adecuados tanto para comprobar hipótesis derivadas de teorías, como para encontrar respuestas a problemas prácticos. Los experimentos sobre métodos educativos, generalmente con propósitos prácticos, con frecuencia buscan cuál de dos o tres métodos es el mejor para lograr cierto propósito. La investigación industrial y la investigación del consumidor dependen en gran parte de los experimentos de campo. Por otro lado, gran parte de la investigación en psicología social es básicamente teórica. El estudio de Schmitt y sus colaboradores (1992), mencionado anteriormente, probó dos teorías sobre las reacciones conductuales de las personas que, formadas en filas, experimentan una intrusión. Las dos teorías probadas fueron la teoría del ultraje moral y la teoría del costo individual. Los experimentos de campo de Glick y colaboradores (1988) y de Jaffe (1991), también estuvieron orientados a la teoría.

La flexibilidad y la aplicabilidad a una gran variedad de problemas constituyen rasgos importantes de los experimentos de campo. Las únicas dos limitantes incluyen el que sea posible o no la manipulación de una o más variables independientes, y al que las exigencias prácticas de la situación de investigación sean tales que el experimento de campo pueda realizarse sobre el problema particular bajo estudio. Superar estas dos dificultades no resulta fácil. Cuando puede hacerse, un amplio rango de problemas prácticos y teóricos se abren a la experimentación.

Como se indicó antes, las principales debilidades de los experimentos de campo son de índole práctica. La manipulación de las variables independientes y la aleatorización son quizás los dos problemas más importantes; son particularmente agudos en la investigación de ambientes escolares. La manipulación, aunque muy posible, a menudo puede no ser practicable porque, por ejemplo, los padres objetan cuando sus hijos, quienes fueron asignados aleatoriamente a un grupo control, no obtendrán un tratamiento experimental deseado. O quizá haya objeciones a un tratamiento experimental porque prive a los niños de alguna gratificación o los ubique en situaciones conflictivas.

No existe una razón real de por qué la aleatorización no pueda ser utilizada en los experimentos de campo. Sin embargo, con frecuencia se enfrentan dificultades. La mala voluntad para separar grupos de clases o para permitir que los niños sean asignados aleatoriamente a grupos experimentales son algunos ejemplos. Aun cuando la asignación aleatoria sea posible y permitida, la variable independiente puede empañarse seriamente a causa de que los efectos de los tratamientos no puedan aislarse de otros efectos. Por ejemplo, los maestros y los niños pueden discutir sobre lo que está sucediendo durante el curso del experimento. Para prevenir dicho oscurecimiento de las variables, el experimentador debería explicar a los administradores y maestros la necesidad de la asignación aleatoria y del control cuidadoso.

Una característica del campo experimental de naturaleza diferente es una debilidad en algunos experimentos, y en otros representa una fortaleza. Los investigadores de campo deben ser, por lo menos en cierto grado, operadores hábiles en lo referente a las habilidades sociales. Deben ser capaces de trabajar, hablar y convencer a la gente de la importancia y necesidad de su investigación. Deben estar preparados para pasar muchas horas, inclusive días y semanas, en paciente discusión con gente responsable de la situación institucional o comunitaria en la que van a trabajar. Por ejemplo, si van a trabajar en un sistema escolar rural, deben tener conocimientos sobre los problemas educativos rura-

les y generales, y sobre el sistema rural particular que desean estudiar. Algunos investigadores se tornan impacientes con estos aspectos preliminares, debido a que están ansiosos por realizar su trabajo de investigación. Encuentran difícil dedicar el tiempo y el esfuerzo necesarios en la mayoría de las situaciones prácticas. Otros disfrutan la socialización inevitable que acompaña a la investigación de campo. En French (1953) se presentan buenos consejos para el manejo de este aspecto de las situaciones de campo.

Un obstáculo importante para el buen diseño y que generalmente parece que se pasa por alto, es la actitud del investigador. Por ejemplo, la planeación de la investigación educativa parece caracterizarse con frecuencia por una actitud negativa resumida por afirmaciones tales como "eso no puede realizarse en escuelas", "los administradores y maestros no lo permitirán" y "no es posible hacer experimentos sobre este problema en esa situación". Iniciar con actitudes como éstas compromete cualquier buen diseño de investigación desde antes de que inicie la investigación. Si un diseño de investigación requiere de la asignación aleatoria de maestros a las clases, y si la falta de dicha asignación pone seriamente en riesgo la validez interna del estudio propuesto, debe realizarse cualquier esfuerzo para asignar aleatoriamente a los maestros. Los educadores que planean una investigación parecen suponer que los administradores o los maestros no permitirán el empleo de la asignación aleatoria. Sin embargo, esta suposición no es necesariamente correcta.

A menudo el consentimiento y cooperación de los maestros y administradores puede obtenerse si se utiliza un método apropiado, con orientación adecuada y precisa, y si se ofrecen explicaciones de las razones del uso de métodos experimentales específicos. Los puntos a enfatizar son los siguientes: diseñar investigaciones para obtener respuestas válidas a las preguntas de investigación; entonces, si es necesario hacer posible el experimento, se modifica el diseño "ideal". Con imaginación, paciencia y cortesía, muchos de los problemas prácticos de la implementación de un diseño de investigación pueden resolverse de manera satisfactoria.

Otra de las debilidades inherentes a las situaciones experimentales de campo es la falta de precisión. En el experimento de laboratorio es posible lograr un alto grado de precisión o exactitud, de tal manera que los problemas de medición y control de laboratorio generalmente son más simples que los de los experimentos de campo. En situaciones reales, siempre existe una gran cantidad de ruido sistemático y aleatorio. Para medir el efecto de una variable independiente sobre una variable dependiente en un experimento de campo, no es necesario únicamente maximizar la varianza de la variable manipulada ni de cualesquiera variables asignadas, sino también medir la variable dependiente de la manera más precisa posible. Pero en las situaciones reales, como en escuelas y grupos comunitarios, abundan las variables independientes extrañas. Además, las medidas de las variables dependientes, por desgracia, algunas veces no son lo suficientemente sensibles para recoger los mensajes de las variables independientes. En otras palabras, las medidas de la variable dependiente a menudo son tan inadecuadas que no pueden captar toda la varianza que las variables independientes han producido.

Estudios de campo

Los *estudios de campo* son investigaciones científicas no experimentales que buscan descubrir las relaciones e interacciones entre variables sociológicas, psicológicas y educativas en estructuras sociales reales. En este libro cualquier estudio científico (grande o pequeño), que busque relaciones de manera sistemática y que pruebe hipótesis, que sea no experimental y que se realice en situaciones de la vida (por ejemplo, comunidades, escuelas, fábricas, organizaciones e instituciones) será considerado un estudio de campo.

El investigador de un estudio de campo busca primero una situación social o institucional, y después estudia las relaciones entre las actitudes, valores, percepciones y conductas de individuos y grupos en dicha situación. El investigador del estudio de campo por lo común no manipula variables independientes. Antes de discutir y evaluar los diferentes tipos de estudios de campo sería útil considerar algunos ejemplos. Ya se han examinado estudios de campo en capítulos previos y en este capítulo se revisó el estudio Bennington de Newcomb. Ahora se examinarán brevemente dos estudios de campo más pequeños.

Anderson, Warner y Spencer (1984) estudiaron el sesgo por exageración de solicitantes de empleo. Los participantes de tal estudio eran solicitantes reales de puestos en el estado de Colorado. Los solicitantes de un empleo a menudo declaran tener más experiencia y más conocimientos de los que en realidad poseen. Para medir el grado de esta exageración, Anderson y sus colaboradores inventaron una actividad inexistente y les preguntaron a los solicitantes qué tanta experiencia tenían en dicha actividad. Los resultados demostraron que cerca de la mitad de los solicitantes declararon tener experiencia en una o más actividades inexistentes. Aquellos solicitantes que declararon tener gran experiencia en actividades inexistentes también exageraron su habilidad en tareas reales. Este estudio de campo ofrece información importante para aquellos involucrados en la toma de decisiones de contratación. Se trató de un estudio de campo pues no hubo una variable independiente manipulada. Se mencionaron actividades reales y falsas en un cuestionario y se les pidió a los participantes indicar la cantidad de experiencia que tenían en cada tarea, utilizando una escala de 4 puntos. Observe que dicho estudio no fue realizado en el laboratorio, y que utilizó participantes que no sospechaban la situación.

La investigación de campo realizada por Tom y Lucey (1997) estudió el tiempo de espera en las cajas de los supermercados y la satisfacción del cliente con el cajero y con la tienda. Estos investigadores estudiaron a cajeros rápidos y lentos durante periodos concurridos y no concurridos de las operaciones de la tienda. Los investigadores registraron los tiempos de espera de cada cliente y también entrevistaron al cliente cuando salía de la tienda. Los resultados demostraron que generalmente los clientes estaban más satisfechos respecto a la tienda y al cajero cuando el tiempo de espera percibido era más corto. Sin embargo, Tom y Lucey notaron que éste no era siempre el caso. En una de las dos tiendas utilizadas para el estudio, encontraron que algunos clientes reportaron mayor satisfacción con los cajeros lentos. Un cuestionamiento posterior reveló que los cajeros eran más lentos debido a que se tomaban el tiempo para dar al cliente una atención más personal.

Observe que los problemas de estos estudios de campo fueron atacados de manera no experimental: ni la aleatorización ni la manipulación experimental eran posibles. En el estudio de Jones y Cook, los datos fueron recolectados directamente de los estudiantes en dos universidades. En el estudio de Tom y Lucey sólo se utilizaron dos tiendas de abarrotes. Ninguno de estos estudios tuvo aleatorización o una variable independiente activa; no obstante, ambos fueron capaces de proporcionar información útil.

Tipos de estudios de campo

Katz (1953) dividió los estudios de campo en dos grandes tipos: *exploratorios* y de comprobación de hipótesis. El primer tipo, dice Katz, *busca lo que es*, en lugar de *predecir relaciones que se buscan*. El célebre estudio *Equality of Educational Opportunity*, citado en el capítulo 23, ejemplifica este tipo de estudio de campo. Los estudios exploratorios tienen tres propósitos: descubrir variables significativas en la situación de campo, descubrir relaciones entre variables y establecer las bases para una comprobación de hipótesis posterior, más sistemática y rigurosa.

Hasta este punto, en el libro se ha enfatizado el uso y la comprobación de hipótesis. Sin embargo, resulta conveniente reconocer que existen actividades preliminares a la comprobación de hipótesis en la investigación científica. Para lograr la meta deseada de la comprobación de hipótesis, con frecuencia debe realizarse investigación metodológica y de medición preliminar. Parte del mejor trabajo del siglo XX se ha realizado en esta área. Un ejemplo es el que lleva a cabo el analista factorial que está preocupado por el descubrimiento, aislamiento, especificación y medición de las dimensiones subyacentes del rendimiento, inteligencia, aptitudes, actitudes, situaciones y características de personalidad.

El segundo subtipo de estudios exploratorios de campo —investigación cuya meta es descubrir o revelar relaciones— es indispensable para el avance científico en las ciencias sociales. Es necesario conocer, por ejemplo, los correlatos de las variables. De hecho, el significado científico de un constructo surge de las relaciones que tienen con otros constructos. Suponga que no se tiene conocimiento científico del constructo "inteligencia"; no se conoce nada sobre sus causas o concomitantes. Por ejemplo, considere que no se sabe nada absolutamente sobre la relación entre inteligencia y rendimiento. Es concebible que se realice un estudio de campo en situaciones escolares. Podría observarse cuidadosamente un número de niños y niñas considerados inteligentes o no inteligentes por los maestros (aunque aquí se introduce contaminación ya que los maestros deben juzgar la inteligencia, por lo menos en parte, por el rendimiento). Quizá se observe que un gran número de los niños "más inteligentes" proviene de hogares de los niveles socioeconómicos más altos; que en clase resuelven problemas más rápido que otros niños; que poseen un vocabulario más amplio, etcétera. Ahora se tienen algunas pistas sobre la naturaleza de la inteligencia, de manera que es posible intentar construir una medida simple de inteligencia. Note que aquí la "definición" de inteligencia surge de lo que los supuestos niños inteligentes y no inteligentes hacen. Un procedimiento similar puede llevarse a cabo con la variable "rendimiento".

Fortalezas y debilidades de los estudios de campo

Los estudios de campo son fuertes en su realismo, significancia, fortaleza de las variables, orientación teórica y calidad heurística. La varianza de muchas variables en escenarios de campo reales es grande, especialmente comparada con la varianza de las variables de experimentos de laboratorio. Considere el contraste entre el impacto de normas sociales en un experimento de laboratorio como el de Sherif (1963), y el impacto de estas normas en una comunidad donde, por ejemplo, se aprueban ciertas acciones de los maestros y se desaprueban otras. Considere además la diferencia entre el estudio de la cohesión en el laboratorio, donde se le pregunta a los participantes, por ejemplo, si desean permanecer en un grupo (medida de cohesión), y el estudio de la cohesión del profesorado de una escuela, donde la permanencia en el grupo es parte esencial del futuro profesional de la persona. Compare la atmósfera de grupo en el estudio del Bennington College y la de un experimento de campo donde los instructores universitarios, que juegan diferentes papeles, crean diferentes atmósferas. Variables tales como clase social, prejuicio, conservadurismo, cohesión y clima social llegan a tener efectos fuertes en estos estudios. Sin embargo, la fortaleza de las variables no es una bendición pura. En una situación de campo por lo común existe tanto ruido en el canal que aunque los efectos sean fuertes y la varianza sea grande, para el experimentador no resulta fácil separar las variables.

El realismo de los estudios de campo es obvio. De todos los tipos de estudios, éstos son los que se asemejan más a la vida real; no puede haber una queja de artificialidad aquí. (Las observaciones sobre el realismo en los experimentos de campo se aplican, *a fortiori*, al realismo de los estudios de campo.)

Los estudios de campo son altamente heurísticos y *ad hoc*. Una de las dificultades de investigación de un estudio de campo es mantenerlo dentro de los límites del problema. Las hipótesis a menudo se le ocurren al investigador de inmediato debido a que el campo es rico en su potencial de descubrimiento. Por ejemplo, quizá se desee probar la hipótesis de que las actitudes sociales de los miembros del consejo de educación es un factor determinante de las decisiones políticas del consejo de educación. No obstante, después de empezar a reunir los datos, surgen muchos conceptos interesantes que pueden desviar el curso de la investigación.

A pesar de estas fortalezas, el estudio de campo es un primo científicamente débil de los experimentos de laboratorio y de campo. Su debilidad más seria, por supuesto, es su carácter no experimental. Por lo tanto, las proposiciones de relaciones son más débiles de lo que son en la investigación experimental. Para complicar las cosas, la situación de campo casi siempre tiene una gran cantidad de variables y de varianza. Piense en las muchas variables independientes posibles que pueden elegirse como determinantes de la delincuencia o del rendimiento escolar. En un estudio experimental dichas variables pueden ser controladas en gran parte, pero en un estudio de campo deben ser controladas, de alguna manera, con medios más indirectos y menos satisfactorios.

Otra debilidad metodológica es la carencia de precisión en la medición de variables de campo. Naturalmente, el problema de la precisión es más agudo en los estudios de campo que en los experimentos de campo. La dificultad encontrada por Astin (1968) para medir el ambiente universitario es uno de muchos ejemplos similares. Por ejemplo, se midió el ambiente administrativo a través de las percepciones de los estudiantes sobre los aspectos del ambiente. Mucha de la falta de precisión se debe a la mayor complejidad de las situaciones de campo.

Los estudios de organizaciones, por ejemplo, en su mayoría son estudios de campo, y la medición de las variables organizacionales ilustran bien las dificultades. "Efectividad organizacional" parece tan complejo como "efectividad del maestro". Para un análisis más profundo e ilustrativo véase Katz y Kahn (capítulo 8, 1978). Vale la pena una cuidadosa lectura y estudio de este excelente libro.

Otra debilidad de los estudios de campo incluye problemas prácticos: viabilidad, costo, muestreo y tiempo. Tales dificultades en realidad son debilidades *potenciales* —ninguna es necesariamente una debilidad real—. Las preguntas más obvias que se plantean son: ¿puede realizarse el estudio con las facilidades de que dispone el investigador? ¿Pueden medirse las variables? ¿Costará demasiado? ¿Requerirá de demasiado tiempo y esfuerzo? ¿Serán cooperadores los participantes? ¿Es posible el muestreo aleatorio? Cualquiera que contemple la posibilidad de un estudio de campo debe plantear y responder estas preguntas. Al diseñar la investigación es importante no subestimar las enormes cantidades de tiempo, energía y habilidad necesarios para la realización exitosa de la mayoría de los estudios de campo. El investigador de campo necesita ser un vendedor, administrador y empresario, a la vez que investigador.

Investigación cualitativa

Un área dentro de los estudios de campo es la investigación cualitativa. Hasta ahora, se ha hablado exclusivamente de investigación cuantitativa. Los estudios de campo con un énfasis cuantitativo poseen los problemas mencionados en la última sección. Sin embargo, la investigación cualitativa es diferente, pues no se basa en el uso de números o mediciones. Esta área de investigación cualitativa va creciendo en interés principalmente debido a que los investigadores se han dado cuenta de que no todos los estudios pueden o deben ser

cuantificados. Existen áreas de investigación donde los métodos cuantitativos no son capaces de captar adecuadamente la información. Por ejemplo, la investigación cuantitativa sería incapaz de captar información valiosa que sirviera para entender las experiencias de vida de pacientes renales que están bajo diálisis. La investigación cuantitativa puede proporcionar a los doctores y enfermeras información sobre la relación entre factores clínicos (tales como nutrición) y medidas de resultados (tales como tasas de supervivencia); pero no pueden indicar lo que el paciente en diálisis experimenta. Es la descripción de estas experiencias lo que permite el desarrollo de mejores programas de rehabilitación. El término "investigación cualitativa" se utiliza aquí para referirse a la investigación social y conductual basada en observaciones de campo discretas que se analizan sin utilizar números o estadística. Anteriormente se mencionó que quienes están involucrados con el aprendizaje operante o investigación skinneriana tampoco están interesados en el uso de la estadística inferencial; no obstante, se apartan de la investigación cualitativa, ya que sí utilizan números y mediciones. Los participantes de la investigación cualitativa pueden no estar conscientes de ser observados o estudiados. Varía el grado en que el participante se involucra en el proceso de investigación. A diferencia de la investigación de un solo sujeto o de series de tiempo, el participante no está consciente de que se estén haciendo mediciones. Dooley (1995) presenta un ejemplo sobresaliente de investigación cualitativa con el estudio de la teoría de la disonancia cognoscitiva. Dooley cita la investigación de Festinger (1956), quien estudió a la gente que predijo el fin del mundo pero que no vio su predicción convertirse en realidad. Este tipo de investigación requiere una metodología que no sea cuantitativa y que sea discreta. Sería muy difícil hacer que estas personas, que pertenecen a una secta, vinieran a un laboratorio de una universidad para ser estudiados. El investigador no tiene oportunidad de estudiar con eficacia a estas personas que acaban de experimentar disonancia cognoscitiva, pidiéndoles que completen un cuestionario o que participen en una entrevista estructurada. En su lugar, el investigador debe ser lo más discreto posible: tendría que actuar como alguien curioso o preocupado, e inclusive podría unirse a la secta como observador y encontrar la información requerida, de forma no amenazante. Se estudia a los participantes sin que ellos noten que están siendo estudiados. Sin embargo, Festinger realizó la investigación hace muchos años (1956). En el ambiente actual sería demasiado peligroso para los investigadores unirse a una secta con el propósito de estudiarla. ¿Por qué? En años recientes, especialmente en 1997, todos los miembros de una secta llamada Heaven's Gate se suicidaron por la llegada del cometa Hale-Bopp. Los miembros masculinos de esta secta estuvieron sujetos a severas alteraciones físico-quirúrgicas. Existe también una cantidad de sectas poderosas que utilizan métodos de programación rigurosos y drogas hipnóticas con sus miembros, para mantenerlos bajo control. Por lo tanto, aunque el ejemplo de Dooley es una buena ilustración de la investigación cualitativa, los autores de este libro de texto no recomiendan a alguien interesado en realizar investigación cualitativa sobre las sectas, unirse o convertirse en miembro de una de ellas.

Sería un poco más seguro considerar un estudio realizado por Rosenhan (1973), quien estaba interesado en la forma en que los hospitales psiquiátricos efectuaban diagnósticos especializados y en cómo serían las experiencias de un paciente psiquiátrico. Rosenhan pidió a ocho de sus cómplices que actuaran como pacientes psiquiátricos que sufrían de alucinaciones. Cada uno de estos seudopacientes fue admitido en diferentes hospitales. Durante su estancia, los seudopacientes nunca exhibieron algún síntoma. Los cómplices de Rosenhan realizaron observaciones sobre las condiciones hospitalarias, sobre cómo eran tratados y sobre el comportamiento del personal y de otros pacientes. Rosenhan reportó que el personal del hospital nunca supo que los seudopacientes no estaban enfermos.

No obstante, también existen estudios de investigación cualitativa donde el participante sabe que está participando en un estudio. En estos casos, el investigador necesita

desarrollar un alto nivel de empatía con los participantes. Por ejemplo, Jones (1998) utilizó el modelo cualitativo para estudiar a una cultura única (las bandas de adolescentes) en la sociedad estadounidense. Poco se ha reportado acerca de las bandas, con excepción de estadísticas. Se sabe poco sobre las dinámicas dentro de las bandas y sobre las diferencias entre algunos tipos de bandas. Jones tuvo que pasar una gran cantidad de tiempo en prisiones y centros de detención, entrevistando a miembros de bandas. Las experiencias de estar en una banda, las dinámicas entre sus miembros, sus sistemas de valores y cómo estos miembros de la sociedad estadounidense dan significado a sus vidas, satisfacen la meta de la metodología de la investigación cualitativa, la cual, como el estudio de Jones, resulta adecuada para estudiar experiencias de vida complejas.

La investigación cualitativa constituye un estudio de campo porque se realiza en el campo donde los participantes se comportan de manera natural. Heppner, Kivlighan y Wampold (1992) se refieren a la investigación cualitativa como naturalista-etnográfica o fenomenológica. Heppner y sus colaboradores presentan cuatro diferencias entre la investigación cualitativa y la cuantitativa (resumidas en la tabla 24.1).

La investigación cualitativa posee varias ventajas sobre la investigación cuantitativa. La primera utiliza observación directa y entrevistas semiestructuradas en escenarios del mundo real. El investigador busca transacciones e interacciones sociales entre la gente y los eventos. El proceso de recolección de datos resulta menos estructurado que en la investigación cuantitativa. El investigador puede hacer una serie de ajustes durante las observaciones; inclusive puede desarrollar nuevas hipótesis durante el proceso de investigación. La investigación cualitativa es naturalista, participativa e interpretativa.

La investigación cuantitativa rara vez se desvía del plan de investigación. La investigación cualitativa, por otro lado, es muy flexible. Esto ha provocado cierta crítica contra la investigación cualitativa. Algunos consideran que la investigación cualitativa sufre de algunos de los mismos problemas de validez, inherentes a los diseños de un solo sujeto. Otra área vulnerable es el sesgo del experimentador. El investigador cualitativo debe ser sumamente cuidadoso para evitar percibir las situaciones con un sesgo personal. No obstante, los investigadores cualitativos sostienen que el involucramiento discreto y la mezcla natural del observador con el ambiente reduce la cantidad de interrupción en el escenario y en el grupo de estudio. Después de un corto periodo, los participantes regresan a su forma normal de comportamiento y ya no muestran una *fachada*. El observador bien entrenado puede obtener percepciones del comportamiento de los participantes desde distintos pun-

TABLA 24.1 *Cuatro diferencias entre la investigación cuantitativa y la investigación cualitativa (Heppner, Kivlighan y Wampold)*

Cuantitativa	Cualitativa
Emana de la tradición post-positivista; los principales constituyentes son los objetos físicos y los procesos	Emana de la perspectiva fenomenológica; enfatiza eventos mentales internos como la unidad básica de la existencia
Asume que el conocimiento proviene de observaciones del mundo físico	El conocimiento se construye activamente y proviene del examen de los constructos internos de las personas
El investigador infiere, con base en observaciones directas o derivadas de observaciones directas	El investigador se basa en esquemas observacionales externos e intenta mantener intacta la perspectiva de los participantes
La meta consiste en describir causa y efecto	Intenta describir las formas en que la gente da significado al comportamiento

tos de vista. Si se realiza de manera apropiada, los datos recolectados por medio de la investigación cualitativa llegan a generar más información y menos variabilidad espuria que otros métodos de investigación. Quizá las dos visiones de la ciencia presentadas por Sampson (1991) en el capítulo 1 de este libro, incluyan las diferencias entre la investigación cuantitativa y cualitativa. En la investigación cualitativa la determinación del tamaño de la muestra puede realizarse cerca del final del estudio, en lugar de hacerlo al inicio, lo cual no es tan importante para el investigador cualitativo. Una regla de la investigación cualitativa es que a mayor número de entrevistas con cada participante, habrá menor necesidad de tener más participantes.

El diseño de la investigación cualitativa generalmente utiliza un observador discreto o un observador participante. Como observador discreto, el investigador realiza observaciones pasivas e intenta evitar responder al participante de cualquier manera. No se manipulan variables; el investigador sólo deja que los eventos naturales ocurran. Si el investigador deseara ver si la presencia de otra persona en el cuarto de baño afecta la disposición de alguien para lavarse las manos, entonces el investigador debe esperar y observar el comportamiento de la gente cuando haya otra persona en el cuarto de baño, y cuando no haya otra persona. En la investigación cuantitativa, el investigador utilizaría un cómplice para alterar la situación (véase Pedersen, Keithly y Brady, 1986). En la situación participante-observador, el investigador se vuelve parte del ambiente en estudio. Una característica de la forma participante-observador es que el investigador puede ver el efecto de manipular su propio comportamiento; de esta manera, en ocasiones los estudios de investigación cualitativa asemejan experimentos naturales.

Uno de los estudios de investigación cualitativa más famosos es el trabajo de Margaret Mead, quien estudió la cultura de Samoa. Dichos estudios no sólo se basan en observaciones personales sino que también requieren frecuentemente del reclutamiento de informantes. Estudios tales como los realizados para saber cómo era la vida de inmigrantes de primera generación, que llegaron a Estados Unidos en la primera parte del siglo XX, pueden ser estudios cualitativos. Los investigadores entrevistan a una cantidad de inmigrantes de primera generación y desarrollan historias de vida. Con suficientes historias de vida que muestren patrones de comportamiento similares, es posible desarrollar una descripción de cómo era la vida de quienes vivieron en esa época. Para mejores resultados, las entrevistas son videograbadas. El proceso de entrevista se conduce de tal manera y con tal duración que permite al informante adaptarse al entrevistador y al aparato de grabación. Es parte del plan de la investigación cualitativa elegir cuidadosamente al entrevistador para lograr la mejor combinación con el informante.

Puesto que los diarios, las grabaciones y las descripciones se obtienen de la gente estudiada en su ambiente natural, las cuestiones éticas son muy importantes; en particular, la confidencialidad de los registros y de la información debe mantenerse estrictamente segura. Hertz e Imber (1993) manifestaron que la investigación en ciencias sociales tiende a concentrarse en aquellos con menor poder (por ejemplo, animales, estudiantes universitarios) puesto que son de fácil acceso; mientras que los individuos poderosos no lo son (por ejemplo, políticos, ejecutivos corporativos, administradores escolares). Es poco probable que se le permita a un estudiante hacer un estudio de caso con el rector de una universidad. Por lo tanto, algunos estudios de investigación cualitativa incluyen el uso del engaño, lo cual es un tema que requiere de una revisión y justificación de cada caso.

Una excelente referencia sobre investigación cualitativa es la de Taylor y Bogdan (1998). Este libro está ya en su tercera edición y proporciona detalles claros del diseño, recolección de datos y el reporte final de la investigación cualitativa. Otra referencia muy útil sobre la investigación cualitativa es Cresswell (1998), quien señala que existen cinco tradiciones diferentes dentro de la investigación cualitativa. Él compara y critica la biografía, la

fenomenología, la teoría básica, la etnografía y el estudio de caso. Taylor y Bogdan, y Cresswell proporcionan ejemplos detallados de investigación cualitativa. Una excelente referencia sobre el uso de la metodología de investigación cualitativa para el estudio de pacientes con insuficiencia renal es *The Renal Rehabilitation Report*,[1] publicado por el Life Options Rehabilitation Advisory Council. En el volumen de julio/agosto de 1998 de esta publicación, se presenta una comparación entre el modelo tradicional y el modelo cualitativo. Tal artículo explica las razones por las que los métodos cualitativos son científicamente recomendables. Mientras Cresswell compara cinco diferentes tradiciones dentro de la investigación cualitativa, el artículo proporciona las descripciones de siete áreas diferentes. Entre las categorías se encuentran la investigación feminista, la investigación de acción y la investigación de evaluación cualitativa. La investigación feminista se enfoca en la mejora de las necesidades, intereses, experiencias y metas de las mujeres. La investigación de acción implica el esfuerzo conjunto del investigador y del participante para lograr un cambio. La investigación de evaluación cualitativa trata con historias y estudios de caso.

A pesar de que se presentó una visión positiva de los métodos de investigación cualitativa, no todos los individuos mantienen la misma opinión. La mayoría de las ciencias del comportamiento —especialmente la psicología— se han mostrado a favor del modelo cuantitativo. Existen algunos, como Sampson (1991) y Phillips (1973), quienes han afirmado que la cuantificación no es un método apropiado para todas las situaciones de investigación. Ya se analizaron brevemente los beneficios de los métodos cuantitativos; sin embargo, al mismo tiempo, se mencionó que son incapaces de responder ciertas preguntas respecto a la cultura o a ciertas formas de vida. Este conflicto entre la metodología de la investigación cualitativa y la metodología cuantitativa se encuentra bien documentado en la literatura (véase Cook y Reichardt, 1979; Padgett, 1998). Existen fieles defensores de ambas posturas; sin embargo, los investigadores cuantitativos, como Cook y Reichardt, han discutido los temas y presentado algunas ideas para combinar ambas, en lugar de separarlas. Ellos hablan acerca de la posibilidad de un estudio de investigación que combine elementos cuantitativos y cualitativos. De hecho, Padgett (1998) describe tres formas para realizar tanto investigación cuantitativa como cualitativa en un estudio. La combinación de los dos métodos —cuantitativo y cualitativo— se llama *investigación multimétodo*.

Según Padgett, la primera de las tres formas de hacer investigación multimétodo consiste en iniciar la investigación de manera cualitativa y terminarla de manera cuantitativa. El método cualitativo sirve para explorar e identificar las ideas, hipótesis y variables de interés para el investigador. Esto se haría por medio de observación directa, entrevistas o grupos de enfoque. Los conceptos derivados de la porción cualitativa del estudio pueden, entonces, estudiarse a través del uso de métodos cuantitativos y de la comprobación de hipótesis. La generalización de los conceptos y de las hipótesis, probados a través de la investigación cuantitativa, puede proporcionar mayor credibilidad al obtener un mejor vínculo con el mundo real. Los métodos cualitativos proporcionarían ese vínculo.

La segunda forma de hacer investigación multimétodo es utilizar el método cuantitativo primero, seguido por el método cualitativo. Los resultados de la porción cuantitativa del estudio se usan como el punto de inicio de la porción cualitativa. Padgett considera que muchos estudios cuantitativos podrían beneficiarse de un análisis cuantitativo de los resultados. Los métodos cualitativos pueden ayudar a proporcionar entendimiento e información respecto a las preguntas que no fueron respondidas o no podían responderse por medio del estudio cuantitativo. Por ejemplo, en estudios cuantitativos que utilizan el

[1] Una copia de este reporte está disponible en el Life Options Rehabilitation Rosource Center al (800) 468-7777. Los autores desean agradecer al doctor Abdul Abukurah por proporcionarnos una copia de esta publicación.

análisis de regresión múltiple (que se cubre en un capítulo posterior de este libro), el investigador a menudo se queda con un cierto porcentaje de varianza injustificada. Por ejemplo, un estudio de investigación que reporta un coeficiente de correlación de .48, entre las puntuaciones del Graduate Record Examination (GRE) y el éxito en los estudios de posgrado, está indicando que únicamente el 23% de la varianza total del éxito en los estudios de posgrado se explica por las puntuaciones del GRE. Esto también indica que el 77% no está justificado. En este punto, a través del uso de los métodos cualitativos, se puede iniciar el proceso de determinar qué otras variables estarían involucradas. Esto puede, a su vez, conducir a otro estudio cuantitativo que incluya aquellas variables encontradas en la porción cualitativa del estudio.

La tercera forma descrita por Padgett difiere ligeramente de las dos primeras, ya que tiene una división temporal más definida; es decir, después de completar un método, continúa el otro. En la tercera forma de investigación multimétodo, ambos métodos, cualitativo y cuantitativo, se utilizan simultáneamente. Dichos métodos pueden tener un método más dominante que el otro. Cuando ello sucede, un método —el menos dominante— se "anida" dentro del otro —el dominante—. Padgett informa que existen más estudios con esta naturaleza "anidada" que la integración verdadera de los dos métodos en el estudio. En el caso en que los investigadores continúan un resultado cuantitativo con un hallazgo cualitativo, se dice que los métodos cualitativos complementan pero no alteran el modelo cuantitativo del estudio. En el caso opuesto, donde el método cualitativo es el dominante, el investigador realiza una encuesta o entrevista; pero utiliza escalas e instrumentos de medición estandarizados en el proceso, lo cual incluye el uso de escalas de tipo Likert y datos de censo para complementar los datos obtenidos de las entrevistas intensivas. Aquí, los datos cuantitativos no se entrometen dentro de la naturaleza inductiva y holística de los métodos cualitativos.

A pesar de que el uso conjunto de los métodos cualitativos y cuantitativos resulta promisorio, existen todavía algunas dudas en las mentes de muchos. Cook y Reichardt (1979), por ejemplo, señalan algunos de los obstáculos que enfrenta la investigación multimétodo. Los obstáculos que señalan se relacionan principalmente con la economía y el entrenamiento. Un estudio con los esfuerzos conjuntos de los métodos cualitativo y cuantitativo puede ser costoso en términos de tiempo y dinero. Aun cuando la investigación multimétodo requiera de fe y vigilancia, Padgett (1998) considera que la investigación multimétodo vale el costo y el esfuerzo.

Anexo

El paradigma experimental holístico

El paradigma experimental holístico proporciona un medio económico de cuantificación empírica de las relaciones complejas entre los factores críticos que afectan el desempeño humano, sobre tareas operacionales individuales. El modelo produce una ecuación de orden requerido por la mayoría de los factores potencialmente críticos relacionados con la tarea, la gente, el equipo y el tiempo, a través de sus rangos operacionales efectivos. Al combinarse con varias técnicas de reducción del sesgo, el modelo holístico mejora materialmente la precisión predictiva de los resultados experimentales y produce información más generalizable de lo que permite el modelo de pocos factores a la vez.

Contrario a los alegatos efectuados en muchos de los libros de texto de las ciencias del comportamiento actuales, respecto a los experimentos factoriales grandes, el modelo es extremadamente económico. De hecho, con el uso de diseños fraccionales de manera

secuencial, resulta mucho menos costoso llevar a cabo experimentos megafactoriales,[2] de lo que sería obtener la información respecto al mismo número de factores en una serie de experimentos pequeños. Los experimentos megafactoriales grandes tampoco son meras extensiones de experimentos más pequeños; se realizan de manera diferente. Requieren de menos supuestos de los tipos presentes en los experimentos de pocos factores, y en el paradigma holístico los pocos supuestos que se hacen son tentativos y se probarán eventualmente conforme el experimento progrese, y se modifican conforme se necesite. Mientras que la metodología se adapta a los problemas que involucren factores cuantitativos y el modelo del ANOVA, muchos de sus principios pueden utilizarse en el camino, en investigación científica del comportamiento. El modelo es heurístico, pragmático y empírico.

El paradigma experimental holístico conforma una metodología completa, que integra un conjunto de principios, una estrategia y un cuerpo de técnicas que proporcionan respuestas cuantitativas con un sesgo mínimo a preguntas complejas del comportamiento. El sesgo se define como la diferencia entre los estimados del desempeño, basada en los resultados y realizaciones experimentales, obtenidas bajo condiciones operacionales. La estrategia básica en este modelo holístico fue tomada de la metodología G.E.P. de superficie de respuesta de Box (Box, 1954), modificada para ajustarla a los problemas especiales que se encuentran en los experimentos del comportamiento. Sin embargo, el modelo holístico no depende de un diseño experimental o técnica estadística específica; por el contrario, las disposiciones estadísticas juegan un papel mucho menor que en los experimentos tradicionales.

El modelo holístico enfatiza una planeación pre-experimental y una fase de exploración, como un punto para verificar las condiciones que afecten de forma negativa la conducción del experimento y el desempeño de los operadores. Al mismo tiempo, los factores experimentales relevantes se seleccionan con base en el conocimiento actual del investigador y con pruebas preliminares. Después, con el uso de la estrategia de Box, utilizando una secuencia de diseños factoriales fraccionales, esos factores se estudian al nivel más bajo de resolución, una ecuación de primer orden. A continuación, si se encuentra que este modelo no se ajusta adecuadamente a los datos empíricos, se recolecta otro grupo de datos para expandir el orden de la ecuación y se realiza otra prueba. Puesto que la mayor parte del desempeño humano puede ser aproximado adecuadamente por no más que un polinomio de tercer orden, por lo común se concluye este modelo iterativo, después de tomar mucho menos datos de los que se requerirían para llenar el diseño factorial.

Las técnicas más importantes empleadas en este modelo fueron desarrolladas principalmente en los años treinta y sesenta. Se desarrollaron nuevas técnicas para robustecer la recolección secuencial de datos de cientos de condiciones experimentales, para las tendencias y la transferencia intraserial o efectos "remanentes", sin compensación o aleatorización excesivas. Otras técnicas empleadas en el modelo holístico incluyen transformaciones gráficas y análisis gráfico de datos.

Un análisis crítico del modelo tradicional de la experimentación del comportamiento revela que muchos de sus ritos se han tornado sacrosantos, deificados en algo que no son. Conforme este libro pasa a impresión, se desafía uno de los iconos de la ciencia del comportamiento: la prueba de la significancia estadística (véase Harlow, Mulaik y Steiger, 1997), como ya ha sucedido a menudo durante más de 30 años (véase capítulo 1 de Bakan, 1973).

[2] El doctor Charles Simon acuñó este término para evitar una confusión con la palabra "multifactorial", a la que los escritores de libros de texto con frecuencia se refieren como experimentos de 2, 3 y 4 factores. La primera definición de "mega" es "grande" y "megafactor" implica un número mucho mayor de factores de aquellos que tradicionalmente se han utilizado.

Otras llamadas reglas de investigación científica han dictado experimentos que producen resultados de poco o ningún valor duradero y algunas ocasiones con conclusiones totalmente incorrectas. Esto sucede, por ejemplo, cuando factores críticos no incluidos en el experimento se mantienen constantes. La elección de los valores constantes en los que se mantienen dichos factores, llegan a alterar el nivel de dificultad de la tarea y alterar marcadamente los resultados generales. La aleatorización no garantiza evitar el sesgo ni garantiza la "validez interna", y únicamente debe utilizarse después de haber agotado todos los controles sistemáticos conocidos.

Un procedimiento, recomendado frecuentemente por los tradicionalistas, para incrementar la "validez externa" o generalización de los resultados del experimento consiste en realizar unos cuantos estudios con parámetros modestamente diferentes, después de completar el experimento principal. Éste es un costoso modelo de ensayo y error.

La generalización se logra con mayor precisión y menor costo en el modelo holístico al incluir todos los factores relevantes identificables en el plan experimental original.

El paradigma experimental holístico fue desarrollado por Charles W. Simon durante los pasados 30 años, apoyado principalmente por los departamentos de investigación de la fuerza aérea, la marina y el ejército de Estados Unidos. En los años setenta se impartieron seminarios a grupos de investigación industriales y militares. Hasta la fecha, no está disponible un reporte consolidado, sino únicamente numerosos reportes de las diferentes técnicas, a menudo aislados unos de otros, y no necesariamente actualizados con desarrollos recientes. Un libro está actualmente en preparación.[3]

RESUMEN DEL CAPÍTULO

1. Se comparan y contrastan los experimentos de laboratorio, los experimentos de campo y los estudios de campo.
2. El experimento de laboratorio posee la mayor validez interna, pero tiene debilidad en su validez externa.
3. Los experimentos de laboratorio por lo común muestran variables con un efecto pequeño; mientras que los estudios de campo y los experimentos de campo muestran variables con efectos grandes.
4. Aunque los experimentos de campo tienen variables que muestran efectos grandes, con frecuencia dicho efecto es enmascarado por otras variables siendo difícil superar este obstáculo.
5. Los experimentos de laboratorio tienen una gran orientación teórica, y se diseñan para probar una teoría general.
6. Los experimentos de campo y los estudios de campo tienen una mayor orientación aplicada, y buscan responder una pregunta específica sobre fenómenos observables.
7. Los experimentos de campo difieren de los experimentos de laboratorio, ya que los primeros no poseen los controles estrictos hallados en la investigación de laboratorio.
8. Los experimentos de campo intentan conducir un estudio del tipo del laboratorio en un ambiente del mundo real, con el uso de participantes del mundo real. Casi siempre hay una variable independiente activa.
9. Los estudios de campo son estudios no experimentales realizados en el mundo real. Por lo común no hay una variable independiente activa.

[3] El Dr. Charles W. Simon preparó esta descripción del modelo holístico.

10. La meta de los estudios de campo es descubrir las relaciones e interacciones entre un número de variables sociales y del comportamiento.
11. La mayoría de la investigación de las ciencias del comportamiento y de las ciencias sociales tiene una orientación cuantitativa. A un estudio de campo orientado cuantitativamente a menudo se le llama encuesta o investigación epidemiológica. A un estudio de campo orientado cualitativamente se le denomina investigación cualitativa o naturalista-etnográfica.
12. La investigación cualitativa incluye métodos con observadores discretos o con observadores participantes.
13. En el método del observador discreto, éste se mezcla con el ambiente y no tiene contacto con los participantes. El método del observador participante requiere que el observador se convierta en miembro del grupo en estudio.
14. Los métodos cualitativos son adecuados para estudiar experiencias humanas poco conocidas o complejas. La investigación cualitativa complementa la investigación cuantitativa y no pretende suplantarla.

SUGERENCIAS DE ESTUDIO

1. ¿Dónde es más probable el uso del análisis factorial de varianza, en los experimentos de laboratorio, en los experimentos de campo o en los estudios de campo? Explique.
2. En el capítulo 15 se describió un estudio sobre los efectos comparativos de la marihuana y el alcohol. Suponga que dicho estudio es un experimento de laboratorio. ¿Eso limita su utilidad y generalización? ¿Diferiría un estudio como éste, respecto a la generalización, de un experimento de laboratorio sobre frustración y agresión?
3. A continuación se presenta una lista de estudios. Algunos se resumen en capítulos previos y otros no. Consulte estos estudios y después clasifique cada uno como experimento de laboratorio, experimento de campo o estudio de campo. Explique por qué categoriza cada estudio de esa manera.

   Henemann, H. G. (1977). Impact of test information and applicant sex on applicant evaluation in a selection simulation. *Journal of Applied Psychology, 62*, 524-526.

   Johnson, C. B., Stockdale, M. S. y Saal, F. E. (1991). Persistence of men's misperceptions of friendly cues across a variety of interpersonal encounters. *Psychology of Women Quarterly, 15*, 463-475.

   McKay, J. R., Alterman, A. I., McLellan, T., Snider, E. C. y O'Brien, C. P. (1995). Effect of random versus nonrandom assignment in a comparison of inpatient and day hospital rehabilitation for male alcoholics. *Journal of Consulting and Clinical Psychology, 63*, 70-78.

   Reinholtz, R. K. y Muehlenhard, C. L. (1995). Genital perceptions and sexual activity in a college population. *Journal of Sex Research, 32*, 155-165.

   Wansink, B., Kent, R. J. y Hoch, S. J. (1998). An anchoring and adjustment model of purchase quantity decisions. *Journal of Marketing Research, 35*, 71-81.

   Wilson, F. L. (1996). Patient education materials nurses use in community health. *Western Journal of Nursing Research, 18*, 195-205.

4. "El experimento es uno de los más grandes inventos del último siglo." ¿Concuerda usted con esta afirmación? Si es así, mencione las razones para ello: ¿por qué es correcta la afirmación (si, de hecho, es correcta)? Si no está de acuerdo, explique por

qué no. Antes de formular juicios rápidos, lea y pondere las referencias que se ofrecen en las sugerencias de estudio número 6, abajo.

5. Por desgracia ha habido mucha crítica desinformada sobre los experimentos. Antes de realizar juicios racionales sobre cualquier fenómeno complejo se debe conocer primero sobre lo que se está hablando y, segundo, se debe conocer la naturaleza y el propósito del fenómeno que se critica. Para ayudarle a obtener conclusiones racionales acerca del experimento y de la experimentación, se ofrecen las siguientes referencias como lectura previa.

Berkowitz, L. y Donnerstein, E. (1982). External validity is more than skin deep: Some answers to criticisms of experiments. *American Psychologist, 3*, 245-257. (Una respuesta profunda a la crítica sobre la carencia de validez externa de los experimentos.)

Kaplan, A. (1964). *The conduct of inquiry*. San Francisco, California: Chandler. (El capítulo IV, llamado "Experiment", parece incluir la observación más controlada.)

6. Aquellos que deseen conocer más sobre el paradigma experimental holístico pueden consultar algunas de las primeras publicaciones de Charles W. Simon, que brindan la filosofía subyacente al modelo. Tome en cuenta que desde la época en que fueron escritos, se ha refinado el modelo y se le han incorporado nuevas técnicas.

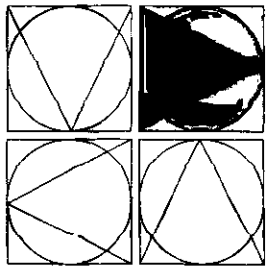
Simon, C. W. (1976). Analysis of human factors engineering experiments: Characteristics, results and applications. Westlake Village, California: Canyon Research Group, Inc., Tech. Rep. No. CWS-02-767, 104 pp. (AD A038-184).

Simon, C. W. (1978). New research paradigm for applied experimental psychology: A system approach. Westlake Village, California: Canyon Research Group, Inc., Tech. Rep. No. CWS-04-77A, 123 pp. (AD A 056-984).

Simon, C. W. (1987). Will egg-sucking ever become a science? *Human Factors Society Bulletin, 30*, 1-4.

Simon, C. W. y Roscoe, S. N. (1984). Application of a multifactor approach to transfer of learning research. *Human Factors, 26*, 591-612.

Westra, D. P., Simon, C. W., Collyer, S. C. y Chambers, W. S. (1982). *(Simulator design features for carrier landings: I. Performance experiments.* NAVTRAEQUIPCEN. (78-C-0060-7): 64 pp. (AD A122-064).



CAPÍTULO 25

INVESTIGACIÓN POR ENCUESTA

La investigación por encuesta estudia poblaciones (o universos) grandes o pequeñas, por medio de la selección y estudio de muestras tomadas de la población, para descubrir la incidencia, distribución e interrelaciones relativas de variables sociológicas y psicológicas. Como tal, la investigación por encuesta puede clasificarse como estudios de campo con una orientación cuantitativa. Algunos la consideran una variación del diseño de investigación correlacional. Este capítulo se concentra en el empleo de la investigación por encuesta en la investigación científica y rechaza las llamadas encuestas de estatus, las cuales tienen una meta diferente de la investigación por encuesta: su meta es conocer el *status quo*, en lugar de estudiar las relaciones entre variables; así como examinar la situación actual de algunas características poblacionales. Las investigaciones por encuesta se utilizaban ya en 1830 en Gran Bretaña, para estudiar las condiciones laborales de niños y adultos durante la Revolución Industrial. La investigación por encuesta en las ciencias sociales y del comportamiento es más moderna, pues es un desarrollo del siglo XX.

No existe la intención de ir en contra de las encuestas de estatus, ya que son útiles e, incluso, indispensables. La intención es destacar la importancia y utilidad de la investigación por encuesta en el estudio científico de problemas social y educativamente significativos. El trabajo de los encuestadores sobre opinión pública, tal como Gallup y Roper, no será examinado. Si se desea revisar una buena explicación sobre encuestas y sus tipos, véase Parten (1950), en el capítulo 1. (Aunque antiguo, este libro aún es valioso.) Un libro un poco más nuevo, sobre la forma en que se utilizan las encuestas para captar la opinión

pública en Estados Unidos, es el de Wheeler (1976). El texto clásico, utilizado durante muchos años, hasta que recientemente dejó de editarse, es el de Warwick y Lininger (1975); tiene la ventaja de haber sido guiado por las ideas y práctica del Survey Research Center de la University of Michigan. También posee la ventaja de tener un énfasis transcultural. Orlich (1978) presenta el método y el procedimiento de la investigación por encuesta de una forma muy directa. Inclusive explica cómo diseñar y secuenciar los reactivos de una encuesta. Algunas de las publicaciones más recientes sobre investigación por encuesta son las realizadas por Alreck (1994), Babbie (1990), Suskie (1996) y Weisberg (1996).

Las encuestas incluidas en la definición anterior con frecuencia se denominan *encuestas muestrales*, quizá debido a que la investigación por encuesta se desarrolló como una actividad de investigación separada, junto con el desarrollo y mejoramiento de los procedimientos de muestreo. La investigación por encuesta se considera como una rama de la investigación científica social, la cual se distingue inmediatamente de la encuesta de estatus. Sus procedimientos y métodos han sido desarrollados principalmente por psicólogos, sociólogos, economistas, científicos políticos y estadísticos (véase Campbell y Katona, 1953). Estos individuos han puesto un sello de rigor científico en la investigación por encuesta y, en el proceso, han influido profundamente en las ciencias sociales.

La definición también liga poblaciones y muestras. Los investigadores de encuestas están interesados en la evaluación precisa de las características de poblaciones completas de personas. Ellos desean saber, por ejemplo, cuántas personas en Estados Unidos votaron por un candidato republicano, y la relación entre dicha votación y variables como sexo, raza, preferencia religiosa y otras similares. Buscan conocer la relación entre las actitudes hacia la educación y el apoyo público a los presupuestos escolares.

Sin embargo, sólo en raras ocasiones los investigadores por encuestas estudian poblaciones completas; más bien estudian *muestras* obtenidas de poblaciones. A partir de las muestras ellos infieren las características de la población o universo definidos. El estudio de muestras, a partir de las cuales se pueden realizar inferencias sobre poblaciones, es necesario debido a las dificultades para estudiar poblaciones completas. Las muestras aleatorias pueden, frecuentemente, generar la misma información que un censo (una enumeración y estudio de una población entera), a un costo mucho menor, con mayor eficiencia y, algunas veces, ¡con mayor precisión!

Algunas encuestas intentan determinar la incidencia, distribución e interrelaciones entre variables sociológicas y psicológicas y, al hacerlo, generalmente se enfocan en la gente, los factores vitales de la gente, y sus creencias, opiniones, actitudes, motivaciones y comportamiento. La naturaleza científica social de la investigación por encuesta se revela por la naturaleza de sus variables, que pueden clasificarse como hechos, opiniones y actitudes sociológicas. Los *hechos sociológicos* son atributos de los individuos que surgen de su pertenencia a grupos sociales: sexo, ingreso, afiliaciones políticas y religiosas, nivel socioeconómico, educación, edad, gasto para vivir, ocupación, raza, etcétera. El segundo tipo de variable es psicológica e incluye *opiniones* y *actitudes*, por un lado, y *comportamiento* por el otro. Los investigadores por encuestas no se interesan únicamente en las relaciones entre variables sociológicas; tienden a interesarse más en lo que la gente piensa y hace, así como en las relaciones entre variables sociológicas y psicológicas. El estudio sobre la calidad de vida en Estados Unidos, realizado por el Survey Research Center de la University of Michigan, por ejemplo, ofrece datos deprimentes sobre la relación entre la raza y los sentimientos de confianza en la gente, una variable sociológica y psicológica (los datos se presentan en la tabla 25.1). La relación es sustancial. En efecto, las personas afroamericanas confían menos en la gente, que las personas blancas. Como Campbell, Converse y Rodgers (1976) afirman (p. 455), "aquellas personas que han tenido menos éxito en sus encuentros con la sociedad tienen menos razones para tener confianza en ella".

TABLA 25.1 *Relación entre raza y confianza en la gente (en porcentajes) (estudio de Campbell y colaboradores)*[a]

	Poca confianza	Mucha confianza
Afroamericanos	72	28
Americanos blancos	38	62

[a] $N = 2\,070$.

Por supuesto, los investigadores por encuesta también estudian las relaciones entre variables psicológicas; sin embargo, las relaciones de la investigación por encuesta ocurren entre variables sociológicas y psicológicas: entre educación y tolerancia, entre raza y autoestima y entre educación y sentido de eficacia política.

Tipos de encuestas

Las encuestas pueden ser clasificadas convenientemente de acuerdo con los siguientes métodos para obtener información: entrevista personal, cuestionario enviado por correo, por panel y por teléfono. De éstas, la entrevista personal eclipsa, por mucho, a las otras, y quizás sea la herramienta más poderosa y útil de la investigación social científica por encuesta. Estos tipos de encuestas se describirán brevemente; en capítulos posteriores, cuando se revisen los métodos de recolección de datos, se estudiará con profundidad la entrevista personal.

Entrevistas e inventarios

La mejor investigación por encuesta utiliza la entrevista personal como método principal para obtener información. Esto se logra, en parte, por la construcción cuidadosa y laboriosa de un inventario o cuestionario. Se utilizará el término "inventario" pues tiene un significado claro: es el instrumento utilizado para reunir información de encuesta, a través de una entrevista personal. El término "cuestionario" ha sido utilizado para nombrar instrumentos personales de entrevista e instrumentos de actitudes o de personalidad. Estos últimos se llaman "escalas" en este libro. La información del inventario incluye información factual, opiniones y actitudes, y razones del comportamiento, de opiniones y de actitudes. Los inventarios de entrevista son difíciles de construir, consumen mucho tiempo y son relativamente costosos; pero ningún otro método proporciona la información que ellos ofrecen.

La *información factual* obtenida en encuestas incluye los llamados datos sociológicos que se mencionaron antes: género, estado civil, educación, ingreso, preferencia política, preferencia religiosa y otros similares. Dicha información resulta indispensable, ya que sirve para estudiar las relaciones entre variables y para verificar la adecuación de las muestras. Estos datos, que se anotan en una "carátula", se denominan "información de la carátula". Esta información, o al menos parte de ella, casi siempre se obtiene al inicio de la entrevista. La mayoría de ella es de carácter neutral y ayuda al entrevistador a establecer empatía con el entrevistado. Las preguntas de naturaleza más personal, tales como los hábitos personales y el ingreso, y preguntas más difíciles de responder, tales como el nivel de conocimiento o habilidad del entrevistado, pueden reservarse para un cuestionamiento posterior, quizás para el final del inventario. Saber cuál es el momento adecuado debe ser necesariamente una cuestión de juicio y experiencia (véase Warwick y Lininger, 1975).

Otros tipos de información factual incluyen lo que los entrevistados saben sobre el tema de investigación, lo que ellos hicieron en el pasado, lo que están haciendo ahora y lo que pretenden hacer en el futuro. Después de todo, a menos que se observe de manera directa, todos los datos sobre el comportamiento de los entrevistados deben provenir de ellos o de otras personas. En este especial sentido, todo el comportamiento pasado, presente y futuro puede clasificarse bajo el "hecho" de comportamiento, aun cuando el comportamiento tan sólo sea una intención. Un aspecto importante de dichas cuestiones factuales es que el entrevistador presumiblemente conoce bastante sobre las acciones y comportamientos personales. Si el entrevistado dice que votó por una emisión de bonos escolares, puede considerarse que la afirmación es verdadera —a menos de que exista evidencia contundente de lo contrario—. De forma similar, puede creérsele al entrevistado, quizás con mayor reserva (puesto que el hecho aún no ha sucedido), si manifiesta su intención de votar por la emisión de los bonos escolares.

Las creencias, opiniones, actitudes y sentimientos que tienen los entrevistados acerca de los objetos cognitivos son de gran importancia, quizás aún más importantes desde un punto de vista científico social. *Objetos cognitivos* es una expresión que indica el objeto de una actitud. Casi cualquier cosa puede ser el objeto de una actitud, pero generalmente el término se reserva para "objetos" sociales importantes, por ejemplo, grupos (religiosos, raciales y educativos) e instituciones (escuela, matrimonio y partidos políticos). Un término más general y quizá más adecuado, aunque no de uso general, es *referente*. Muchos de los objetos cognitivos de la investigación por encuesta pueden no ser de interés para los investigadores: inversiones, ciertos productos comerciales, candidatos políticos y otros similares. Otros objetos cognitivos tal vez sean más interesantes: las Naciones Unidas, la Suprema Corte, las prácticas educativas, la integración, el comportamiento sexual, el subsidio federal a la educación, los estudiantes universitarios y el movimiento feminista.

La entrevista personal puede ayudar a conocer las razones del entrevistado para hacer o creer algo. Cuando se le pregunta a la gente las razones de las acciones, intenciones o creencias, llegan a responder que han hecho algo, intentado hacer algo o que se sienten de ciertas formas acerca de algo. Pueden decir que afiliaciones grupales, lealtades o ciertos eventos han influido en ellos; o pueden haber escuchado acerca de temas bajo investigación por los medios de comunicación masiva. Por ejemplo, un entrevistado masculino puede declarar que antes se oponía al subsidio federal para la educación a causa de que él y su partido político siempre se habían opuesto a la interferencia gubernamental. No obstante, ahora apoya la ayuda federal porque ha leído mucho acerca del problema en periódicos y revistas, y ha llegado a la conclusión de que el subsidio federal beneficiará a la educación del país.

Los deseos, valores y necesidades de un entrevistado influyen en sus actitudes y acciones. Al explicar por qué está de acuerdo con el subsidio federal a la educación, quizá manifieste que sus propias aspiraciones educativas se vieron truncadas y que siempre ha abogado por una mayor educación; o quizá señale que su grupo religioso tiene un fuerte compromiso con la educación infantil como parte de sus estructura de valores. Si el individuo bajo estudio ha revelado en forma concreta sus valores, deseos y necesidades —y llega a expresarlos de manera verbal— la entrevista personal resulta muy valiosa.

Otros tipos de investigación por encuesta

El siguiente tipo importante de la investigación por encuesta es el *panel*. Se selecciona una muestra de participantes y se les entrevista; después se entrevistan de nuevo y se estudian posteriormente. La técnica del panel permite que el investigador estudie cambios en

ambigua y la especificación del problema de investigación; así como el análisis e interpretación de los datos.

En el espacio limitado de una sección de un capítulo es, de hecho, imposible analizar adecuadamente la metodología de la investigación por encuesta. Por lo tanto, se incluirán sólo aquellas partes de la metodología relacionadas con los propósitos de este libro: el diseño de la encuesta o del estudio, el llamado plan o gráfica de flujo de los investigadores por encuestas, la verificación de la confiabilidad y la validez de la muestra, y los métodos de recolección de datos. (Tanto el muestreo como el análisis ya se estudiaron en capítulos previos.)

Los investigadores de encuestas utilizan un *plan* o *gráfica de flujo* para bosquejar el diseño y las implementaciones subsecuentes de una encuesta. El plan de flujo inicia con los objetivos de la encuesta, enlista cada paso a seguir y termina con el reporte final. Primero se establecen los problemas generales y específicos a resolver, de la forma más cuidadosa y completa posible. Puesto que, en principio, no hay nada muy diferente aquí del análisis de los problemas e hipótesis del capítulo 2, puede omitirse un estudio detallado y presentar un ejemplo hipotético simple. Un investigador educativo ha sido comisionado por un consejo de educación para estudiar las actitudes de los miembros de la comunidad hacia el sistema escolar. Al discutir el problema general con el consejo y con los administradores del sistema escolar, el investigador nota que hay varios problemas más específicos, tales como: ¿la actitud de los miembros de la comunidad se ve afectada por el hecho de tener a sus hijos en la escuela? ¿Su nivel educativo afecta sus actitudes?

Uno de los trabajos más importantes del investigador consiste en especificar y aclarar el problema. Para hacerlo bien, el investigador no debe preguntar a las personas únicamente lo que piensan de las escuelas, aunque ésta pueda ser una buena manera de empezar, en caso de no saber mucho sobre el tema. Deben plantearse preguntas específicas, orientadas a las diferentes facetas del problema. Cada una de estas preguntas debe incluirse en el inventario de la entrevista. Algunos investigadores de encuesta diseñan tablas para el análisis de los datos en este momento, para aclarar el problema de investigación y para guiar la construcción de las preguntas de la entrevista. Ya que dicho procedimiento es recomendable, se diseñará una tabla para demostrar cómo puede utilizarse para especificar los objetivos y las preguntas de la encuesta.

Considere la pregunta: ¿está relacionada la actitud con el nivel educativo? La pregunta requiere que la "actitud" y el "nivel educativo" sean definidos operacionalmente. Las actitudes negativas y positivas se inferirán a partir de las respuestas a las preguntas y reactivos del inventario. Si, en respuesta a una pregunta tan vaga como "en general, ¿qué piensa de nuestro sistema escolar?", el encuestado responde: "es uno de los mejores en esta área", puede inferirse que tiene una actitud positiva hacia las escuelas. Naturalmente, una pregunta no será suficiente; deben utilizarse preguntas relacionadas. Resulta bastante fácil obtener una definición de "nivel educativo". Se decide utilizar tres niveles: 1) licenciatura

FIGURA 25.1

	Actitud positiva	Actitud negativa
Licenciatura inconclusa		
Preparatoria terminada		
Preparatoria inconclusa		

inconclusa, 2) preparatoria terminada y 3) preparatoria inconclusa. El paradigma del análisis podría verse como el de la figura 25.1.

La virtud de paradigmas como éste es que el investigador puede determinar de inmediato si el problema específico se ha planteado con claridad y si está relacionado con el problema general. También proporciona cierta noción sobre cuántos encuestados se necesitarán para llenar adecuadamente las casillas de la tabla; también ofrece lineamientos para la codificación y el análisis. Además, como Katz (1953, pp. 80-81) afirma:

> Al realizar la tarea mecánica de diseñar dichas tablas, los investigadores seguramente descubrirán las complejidades de una variable que necesita medidas y calificaciones más detalladas de las hipótesis, en relación con las condiciones especiales.

El próximo paso en plan de flujo es la muestra y el plan del muestreo. El muestreo es demasiado complejo para discutirlo aquí en detalle, por lo que sólo se bosquejan las principales ideas. Para un tratamiento más detallado del tema, véase los capítulos 8 y 12 en esta obra, y el capítulo 5 del libro de Warwick y Lininger (1975). El ejemplo detallado del muestreo multietapas de área de Warwick y Lininger resulta especialmente útil. El *muestreo por área* es el tipo más utilizado en la investigación por encuesta. Primero deben definirse aleatoriamente las áreas grandes que van a muestrearse. Esto es equivalente a la partición del universo y al muestreo aleatorio de las casillas de la partición, las cuales pueden ser áreas delineadas por la cuadrícula de mapas o por fotografías aéreas de países, distritos escolares o manzanas de ciudades. Después se pueden obtener muestras de subáreas, a partir de las áreas grandes ya obtenidas. Finalmente, se eligen todos los individuos o familias o muestras aleatorias de individuos y familias.

En primer lugar, debe definirse el universo que va a estudiarse y del cual se obtendrán las muestras. ¿Se incluyen a todos los ciudadanos que viven en la comunidad? ¿A los líderes comunitarios? ¿A los ciudadanos que pagan impuestos escolares? ¿A quienes tienen hijos en edad escolar? Una vez definido el universo, se decide cómo se obtendrá la muestra y cuántos casos se elegirán. En la mejor investigación por encuesta se utilizan muestras aleatorias. Algunas ocasiones se utilizan las muestras *por cuota* en lugar de muestras aleatorias, debido a que estas últimas son costosas y más difíciles de llevar a cabo. En una muestra por cuota (o control por cuota), presumiblemente se logra la "representatividad" por medio de la asignación por cuotas a los entrevistadores —tantos hombres y mujeres, tantos blancos y afroamericanos, etcétera—. Aunque el muestreo por cuota puede lograr representatividad, carece de las virtudes del muestreo aleatorio —y, por lo tanto, debe evitarse—.

El siguiente gran paso en una encuesta es la construcción del inventario de entrevista **y de otros instrumentos de medición que serán utilizados.** Ésta es una tarea laboriosa y difícil que no tiene ninguna semejanza con los cuestionarios armados en forma apresurada en forma por los principiantes. La tarea principal consiste en traducir la pregunta de investigación en un instrumento de medición y en otros instrumentos construidos para la encuesta. Uno de los problemas del estudio podría ser, por ejemplo: ¿qué relación tienen las actitudes permisivas y restrictivas hacia la disciplina de los niños con los sistemas educativos locales? De entre las posibles preguntas para evaluar las actitudes permisivas y restrictivas, una sería: ¿cómo considera que debe disciplinarse a los niños? Después de completar bocetos de inventarios de entrevistas y de otros instrumentos, se prueban con antelación en una muestra pequeña representativa del universo. Después se revisan y se les da la forma final.

Los pasos descritos anteriormente constituyen la primera gran parte de cualquier encuesta. Después de que el investigador ha desarrollado el instrumento y ha determinado a qué población va a medir, también necesita decidir si los datos se recolectarán con el uso de un diseño transversal o con un diseño longitudinal. El diseño longitudinal implica la

Verificación de los datos obtenidos mediante encuestas

La investigación por encuesta posee una ventaja única entre los métodos científicos en ciencias sociales: con frecuencia es posible verificar la validez de los datos de la encuesta. Puede entrevistarse de nuevo a alguno de los entrevistados y comparar los resultados de ambos cuestionarios. Se ha encontrado que la confiabilidad de los reactivos de aspectos personales, como edad e ingreso, es alta. La confiabilidad de las respuestas de actitud es más difícil de determinar debido a que una respuesta modificada puede significar una actitud modificada. La confiabilidad de las respuestas promedio es más alta que la confiabilidad de las respuestas individuales. Por fortuna, el investigador generalmente está más interesado en promedios o medidas de grupo, que en respuestas individuales.

Una forma para verificar la validez de un instrumento de medición consiste en utilizar un criterio externo. Los resultados se comparan con un criterio externo, presumiblemente válido. Por ejemplo, un entrevistado declara haber votado en la última elección de miembros del consejo escolar. Es posible descubrir si esto es verdad o no, verificando los archivos de registro y de votación. Por lo común no se verifica el comportamiento individual, puesto que es difícil obtener información acerca de individuos; aunque la información grupal está más disponible. Esta información se utiliza para probar, hasta cierto grado, la validez de la muestra de la encuesta y de las respuestas.

Un buen ejemplo de una verificación externa de datos de encuesta es el uso de información sobre el último censo. Esto es particularmente útil en encuestas a gran escala; aunque también ayuda con encuestas más pequeñas. Es posible comparar proporciones de hombres y mujeres, razas, niveles educativos, edades, etcétera, de la muestra, con el censo de cada país. Por ejemplo, en el estudio de Verba y Nie (1972), sobre participación política, los autores informan ciertas comparaciones de este tipo. Los estimados de su muestra son precisos: sólo uno de ellos, *edad* 20-34, se desvía de los estimados del censo en más de 2 por ciento, lo cual constituye evidencia sólida de lo adecuado de la muestra. Para asegurarse, su muestra fue grande (> 2 500), pero también se han encontrado muestras pequeñas que son bastante precisas. En un estudio en Detroit, realizado por la University of Michigan, en 1952, la muestra fue de sólo 735, pero los estimados de la muestra estuvieron cerca de los del censo de 1950. Campbell y Katona (1953) analizan los métodos de verificación de la validez y confiabilidad de las muestras. Warwick y Lininger (1975) presentan tablas de errores de muestreo, con una explicación respecto a su significado y su uso estadístico. Ahí se observa, por ejemplo, que los porcentajes reportados entre 20 y 80, de una muestra de 700, tienen un error estándar de 4. ¡Para reducir el error estándar a 2, se requiere de una muestra de 3 000!

Los investigadores del SIDA, el doctor Vickie Mays y la doctora Susan Cochran, de la University of California en Los Ángeles, tienen una forma ingeniosa para verificar algunas de las respuestas en sus encuestas. Ellos incluyen en los cuestionarios reactivos que son específicos para ciertos grupos de personas. Por ejemplo, incluyen algunas preguntas que sólo incumben a hombres homosexuales, y otras preguntas dirigidas sólo a hombres heterosexuales. Posteriormente, después de que los datos se recolectan, se codifican y se traducen, de forma que la computadora pueda leerlos, se utiliza un programa estadístico de computadora para realizar una serie de tablas cruzadas de las preguntas respecto a la variable *preferencia sexual*. Los homosexuales que respondieron a las preguntas que eran sólo para hombres heterosexuales, o los heterosexuales que respondieron las preguntas para los homosexuales, se consideran como datos mal codificados, o capturados de manera incorrecta en la computadora. Entonces la forma de respuesta real puede recuperarse de los archivos y volverse a analizar para corregir los datos. Puesto que la realización de una encuesta significativa es costosa, cada cuestionario de cada participante es importante. A

diferencia de otras áreas de investigación, donde se recomendaría eliminar datos de participantes, la investigación por encuesta no puede hacer esto y aun así esperar obtener la información más precisa.

Tres estudios

Se han realizado muchas encuestas, tanto malas como buenas. Los estudiantes probablemente no se interesarían en la mayoría de ellas, ya que son sólo un poco más que intentos refinados para obtener información simple: estudios sobre la elección presidencial, sobre plantas industriales, etcétera. Sin embargo, existen encuestas de considerable interés y significancia —incluso muy grande— para los científicos del comportamiento. Tres de estos estudios se resumen a continuación.

Verba y Nie: participación política en Estados Unidos

Verba y Nie (1972) se preguntaron, entre otras cuestiones, cómo la participación política de los ciudadanos en una democracia influye en los procesos gubernamentales. Ellos entrevistaron a más de 2 500 residentes en Estados Unidos, en 200 lugares en 1967, los cuales fueron seleccionados por medio de un procedimiento de muestreo de probabilidad por área. (Sus comparaciones muestra-censo revelaron un alto acuerdo en general.) El principal hallazgo fue que la participación de los ciudadanos influye, de hecho, en los líderes políticos; pero son los ciudadanos más acaudalados, los más educados y con un estatus más alto en general, quienes más influyen con su participación. Los autores señalan que aunque los estadounidenses no se caracterizan por su ideología basada en la clase, el nivel social sí se relaciona con la participación. El estudio se caracteriza especialmente por su medición y metodología analítica sofisticadas, y por un hallazgo desconcertante importante. En capítulos posteriores se retomará el estudio y su metodología.

Docter y Prince: una encuesta de travestis masculinos

Docter y Prince (1997) reportan que una de las últimas encuestas más importantes publicadas sobre hombres travestis se llevó a cabo en 1972. El estudio de Docter y Prince (1997) utilizó el mismo instrumento de encuesta, usado en 1972, para medir a los travestis en 1992. Los investigadores añadieron algunas preguntas adicionales respecto al travestismo y a la excitación sexual. Uno de los objetivos consistía en evaluar si habían ocurrido cambios desde la encuesta de 1972. La razón por la que estos investigadores consideraron que quizá ocurrieron cambios, se centra en la discriminación del travestismo en algunas áreas de Estados Unidos, en la mayor exposición que han tenido los travestis y los transexuales en los medios de comunicación masiva y en el crecimiento de grupos y organizaciones nacionales de apoyo. Docter y Prince compararon las dos muestras encuestadas en, por lo menos, seis dimensiones: 1) factores demográficos, de la infancia y familiares; 2) orientación sexual y comportamiento sexual; 3) identidad de género; 4) comportamiento del papel de género; 5) planes futuros de vivir completamente como mujer, y 6) confianza en los servicios de asesoría o de salud mental. Docter y Prince utilizan el término "travesti" para definir a los hombres biológicos que ocasionalmente se visten con ropa de mujer, que, sin embargo, no buscan una reasignación de sexo. Un transgenérico es aquel que vive continuamente en el papel del género opuesto a su sexo biológico, sin procedimientos de reasignación de sexo; un transexual es quien ya tuvo una reasignación de sexo. Docter y Prince encuestaron a 1 032 travestis autodefinidos, en edades de 20 a 80 años. Biológicamente todos eran

hombres. La población de la muestra fueron voluntarios de todo Estados Unidos, quienes respondieron a la solicitud de participantes para investigación en reuniones de clubes, convenciones y publicaciones para travestis. La muestra de 1992 representó una base más amplia de travestis que la muestra de 1972: la de 1972 consistió principalmente de lectores de publicaciones para travestis; en cambio, la muestra de 1992 estuvo compuesta de lectores de un número de diferentes publicaciones y de miembros de clubes de travestis. La comparación entre las dos muestras demostró que hubo algunos cambios entre los travestis de 1972 y los de 1992. En particular, en la muestra de 1992 existían más participantes interesados en vivir como mujeres todo el tiempo. También hubo más participantes en la muestra de 1992 que tenían una identidad de género preferida, que fue igual para hombre que para mujer, con respecto a la muestra de 1972. Docter y Prince (1997) documentan las diferencias en el método de muestreo entre las dos muestras y señalan las limitaciones de la muestra más reciente, en comparación con la muestra antigua. Ésta es la naturaleza de la investigación por encuesta: ciertas cuestiones cambian a través del tiempo, y vuelven más difícil obtener exactamente el mismo ambiente de investigación, de un periodo a otro.

Sue, Fujino, Hu, Takeuchi y Zane: servicios comunitarios de salud para minorías étnicas

Este estudio (1991) no se ajusta exactamente a lo que se llamaría investigación por encuesta. Tales investigadores no diseñaron la encuesta para el estudio, ni recolectaron los datos para el mismo. En su lugar, utilizaron los datos suministrados por el Automated Information System (AIS), mantenido por el Departamento de Salud Mental del condado de Los Ángeles. Estos datos fueron utilizados por la agencia del gobierno con el propósito de la administración del sistema, recaudación fiscal, manejo clínico e investigación. Todos los pacientes eran receptores de consulta externa. El estudio califica como investigación por encuesta debido a que se trata de un estudio de campo —cuantitativo y epidemiológico— que reunió información que describe las relaciones entre variables dentro del conjunto de datos. Dicho tipo de investigación por encuesta está basado en la búsqueda de registros (Isaac y Michael, 1987). Sue y sus colaboradores utilizaron los datos para responder algunas preguntas respecto a los servicios de salud mental de cuatro grupos étnicos: afroamericanos, asiaticoamericanos, latinos y estadounidenses blancos. La muestra del AIS consistió de 7 136 asiaticoamericanos, 47 220 afroamericanos, 58 844 latinos y 99 036 estadounidenses blancos. El conjunto de datos originales cubrió un periodo de 15 años. Sue y sus colaboradores utilizaron únicamente el último periodo de cinco años. En una comunicación personal, Sue informó al segundo autor de este libro (HBL) que él y su equipo dedicaron gran cantidad de tiempo, esfuerzo y dinero en la reorganización de los datos, para que pudieran ser sujetos de su investigación. La hipótesis que los investigadores comprobaron fue que los pacientes que fueron apareados tanto étnicamente como por género con un terapeuta, tendrían una mayor mejoría en su salud mental. La medida de salud mental fue la escala de evaluación global (EEG). Las variables dependientes en dicho estudio fueron los abandonos del tratamiento, el número promedio de sesiones de tratamiento y los resultados del tratamiento. Los resultados del estudio mostraron en todos los grupos, con excepción de los afroamericanos, menores posibilidades de abandono del tratamiento cuando los pacientes estaban apareados étnicamente con el terapeuta. Cuando se aparearon por género, tan sólo los asiaticoamericanos y los estadounidenses blancos demostraron una menor posibilidad de abandono del tratamiento. Tal hallazgo señala un ingrediente importante para la prevención del abandono de los pacientes, en instituciones públicas de salud mental. Al analizar los datos del AIS, Sue y sus colaboradores encontraron, que sólo una tercera parte de los pacientes étnicos fueron tratados por terapeutas de la misma etnia; mientras que el 75 por ciento de los estadounidenses blancos fueron tratados por terapeutas blancos. Respecto al número promedio de sesiones de tratamiento, todos los grupos que

incluían un apareamiento paciente-terapeuta tuvieron un número promedio mayor de sesiones de tratamiento. No obstante, respecto a la variable de apareamiento por género, únicamente los mexicoamericanos y los estadounidenses blancos mostraron un número mayor de sesiones de tratamiento. El EEG fue utilizado para medir el resultado del tratamiento. No hubo un efecto del apareamiento por género. Respecto al apareamiento étnico, sólo los mexicoamericanos tuvieron puntuaciones más altas en el EEG en el momento de la conclusión del tratamiento.

Hall, Kaplan y Lee (1994) hallaron patrones similares, al utilizar la misma base de datos, pero observando únicamente a los pacientes que eran niños. Encontraron que los niños más pequeños tenían una mayor mejoría cuando eran apareados con terapeutas similares en las áreas de etnicidad e idioma. Esto puede atribuirse al hecho de que el idioma, en niños bilingües más pequeños, aún no está bien desarrollado, lo cual resulta en la necesidad de un terapeuta que pueda satisfacer sus requerimientos culturales y de idioma. Otro estudio originado a partir de esta importante base de datos es el de Russell, Fujino, Sue, Cheung y Snowden (1996).

Aplicaciones de la investigación por encuesta en educación

Estos estudios demuestran con claridad la aplicabilidad de la investigación por encuesta y su metodología en sociología, psicología social, trabajo social, psicología clínica y ciencias políticas. El fuerte énfasis de la investigación por encuesta en las muestras representativas, el diseño general, el plan de investigación y las entrevistas a expertos con el uso de inventarios de encuesta construidos cuidadosa y competentemente, ha sido, y continuará siendo, una influencia benéfica en la investigación del comportamiento. A pesar de su evidente valor potencial en todos los campos de la investigación del comportamiento, la investigación por encuesta no ha sido utilizada en tan amplia medida en donde parecería tener un gran valor teórico y práctico, es decir, en la educación. Su utilidad distintiva en la educación y en la investigación educativa parece haber sido lentamente percibida. No obstante, una revisión de la literatura actual muestra que la situación quizás esté cambiando. Por lo tanto, se dedica esta sección a la aplicación de la investigación por encuesta a la educación y a los problemas educativos.

Obviamente, la investigación por encuesta constituye una herramienta útil para indagar hechos en la educación. Un administrador, un consejo educativo o un grupo de maestros aprenden mucho acerca de un sistema escolar o de una comunidad sin contactar a cada niño, cada maestro ni cada ciudadano. En resumen, los métodos de muestreo desarrollados en la investigación por encuesta resultan de mucha utilidad. Es poco satisfactorio depender de las llamadas muestras representativas que son relativamente de ensayo y error, basadas en juicios "expertos". Tampoco es necesario reunir datos en poblaciones completas; las muestras son suficientes para muchos propósitos.

La mayor parte de la investigación en educación se conduce utilizando muestras no aleatorias, relativamente pequeñas. Si las hipótesis se sostienen, pueden comprobarse después con muestras aleatorias de poblaciones y, si nuevamente son apoyadas, entonces pueden generalizarse los resultados a poblaciones de escuelas, niños y gente ordinaria. En otras palabras, la investigación por encuesta puede usarse para probar hipótesis que ya han sido probadas en situaciones más limitadas, lo cual resulta en un incremento de la validez externa.

La investigación por encuesta parece ajustarse de manera ideal a algunos de los aspectos controvertidos más importantes en educación. Por ejemplo, su habilidad para manejar problemas "difíciles" como la integración y la clausura de escuelas, a través de entrevistas

cuidadosas y prudentes, la ubica en los primeros lugares de la lista de los métodos de investigación para dichos problemas. Las entrevistas de muestras aleatorias de ciudadanos y maestros de distritos escolares, al inicio de un programa de educación especial o de educación para superdotados, o tras la experiencia de la clausura probable de ciertas escuelas primarias a causa de la disminución de inscripciones, pueden ofrecer información valiosa respecto a sus preocupaciones y temores, de tal forma que se tomen medidas apropiadas para informarles y para disminuir sus temores. El efecto de estas medidas, por supuesto, se estudia también.

La investigación por encuesta quizá se adapta más a la obtención de hechos personales y sociales, creencias y actitudes. Es significativo que, aunque se dicen y escriben cientos de miles de palabras sobre educación y sobre lo que se supone que la gente piensa acerca de ella, existe poca información confiable sobre el tema. Simplemente no se conocen las actitudes que tiene la gente hacia la educación. Es necesario depender de futuros escritores y de los llamados expertos, para obtener dicha información. Los consejos educativos con frecuencia dependen de administradores y de líderes locales para que les digan lo que la gente piensa. Algunas de las preguntas que se plantean y, que, posiblemente, se contestan con el uso de la investigación por encuesta son: ¿apoyará la comunidad un mayor presupuesto el próximo año? ¿Qué pensarán respecto a la división de los distritos escolares? ¿Cómo reaccionarán los padres ante la asignación de tareas a los niños para lograr suprimir la segregación racial? ¿Cuál es el currículum actual? ¿Cuál es la tasa de abandono de los estudiantes de posgrado? ¿En qué grado copian más en los exámenes los estudiantes de la escuela de medicina? ¿Los niños con diversos antecedentes culturales que viven en Israel difieren respecto a sus temores? Un antiguo y sobresaliente ejemplo de investigación por encuesta en educación es el estudio de Gross, Mason y McEachern (1958), el cual constituye una lectura indispensable para los administradores educativos y para los miembros de los consejos educativos.

Es alentador que en los pasados 12 años se hayan realizado más estudios en ambientes educativos. Considere, por ejemplo, el estudio de Stile, Kitano, Kelley y Lecrone (1993), quienes llevaron a cabo una encuesta nacional sobre lo que está sucediendo en programas preescolares y de jardín de niños para alumnos superdotados. Su encuesta examinó escuelas en todas las entidades, en Estados Unidos. Reportaron que tan sólo 29 de los 50 estados (58 por ciento) y un territorio tenían programas para niños superdotados. Dichos programas abarcaron un poco más de 2 655 distritos escolares. Sólo 16 estados muestran que tienen programas en el nivel de jardín de niños para aprendices superdotados, que provienen de familias en desventaja. Aunque el estudio de Stile y colaboradores parece, de alguna manera, una encuesta sobre estatus, sí señala cómo se ve "el panorama completo" de los programas educativos para superdotados en Estados Unidos y sus territorios. También señala el tipo de fondos utilizados para los programas con superdotados.

El estudio de Cooke, Sims y Peyrefitte (1995) brinda información que no se había publicado antes, respecto al abandono escolar de estudiantes de posgrado. Se conoce mucho acerca del abandono escolar de estudiantes de licenciatura; pero muy poco sobre los estudiantes de posgrado. Por lo común, el muestreo de estudiantes de posgrado no es tan abundante como el de los estudiantes de licenciatura. En dicho estudio los investigadores reunieron datos de 230 estudiantes de posgrado inscritos en programas de negocios, ingeniería, administración pública y educación. Se eligieron estos programas debido a que se contaba con mayores cantidades de estudiantes de minorías étnicas inscritos en ellos. El instrumento de encuesta se envió por correo a los participantes a principios de 1992; una encuesta de seguimiento se envió 18 meses después. Las dos encuestas fueron utilizadas para determinar si podía predecirse el abandono después de 18 meses. Los resultados demostraron que las minorías étnicas tenían una mayor intención de renunciar a los estu-

dios de posgrado, y que estaban menos satisfechos con los estudios de posgrado que aquellos que no pertenecían a tales minorías étnicas. Sin embargo, aunque estas diferencias existían, no se encontró que estuvieran relacionadas con el abandono escolar. El abandono estuvo más relacionado con las variables necesidad de logro, compromiso afectivo y si el programa de posgrado cumplía las propias expectativas.

Little y Lee (1995) realizaron un estudio sobre todos los programas escolares de posgrado en psicología, a través de todo el territorio de Estados Unidos. Su propósito era determinar la cantidad de entrenamiento que recibían los estudiantes de posgrado, en las áreas de métodos de investigación y estadística. Entre las muchas comparaciones, está la que se hizo entre los programas que otorgaban doctorados y los que no lo hacían. Little y Lee no estaban interesados únicamente en la cantidad de cursos, sino también en su contenido y en el empleo de programas estadísticos de computadora. Se enviaron por correo un total de 181 encuestas a los programas certificados por la National Association of School Psychologists (NASP) y a la American Psychological Association (APA), así como a aquellos anotados en la *Petersen's Guide to Graduate Education*. De éstos, se obtuvieron 101 encuestas útiles. Los resultados no mostraron diferencias significativas dentro de los programas predoctorales y doctorales, respecto a la cantidad de cursos sobre estadística y diseño de investigación. Sin embargo, se encontraron diferencias al comparar los programas predoctorales con los de doctorado, los cuales, por lo general, requerían el doble de cursos de estadística y de diseño de investigación, que los programas predoctorales. Little y Lee proporcionan información valiosa que puede utilizarse en programas de posgrado ya existentes, o nuevos, en psicología escolar, para ajustar o desarrollar su currículum.

Baldwin, Daugherty, Rowley y Schwarz (1996) enviaron una encuesta a 3 975 estudiantes de segundo año que acudían a 31 escuelas de medicina; 2 459 (62 por ciento) completaron el cuestionario de encuesta. La encuesta se llevó a cabo para determinar el grado de comportamiento y actitud fraudulentos en los exámenes. El 39 por ciento de los encuestados afirmaron haber visto por lo menos un incidente en donde se hiciera trampa en un examen. Cerca de dos terceras partes de la muestra declaran haber escuchado que los estudiantes cometen fraude en los exámenes. El hecho fraudulento se dividió en categorías: 1) obtener información previa respecto al examen, 2) copiar las respuestas de otro estudiante durante el examen y 3) intercambiar respuestas durante el examen. El 82 por ciento de los estudiantes que declararon haber hecho trampa en la escuela de medicina también afirmaron haberla hecho en la escuela, antes de ingresar a la facultad de medicina. Aproximadamente el 5 por ciento de los estudiantes reportaron haber hecho trampa alguna vez durante los primeros dos años de la escuela de medicina. Más hombres que mujeres afirmaron haber realizado dichas prácticas en los exámenes.

Ventajas y desventajas de la investigación por encuesta

La investigación por encuesta tiene la ventaja de una visión amplia: se obtiene gran cantidad de información a partir de una población grande. Permite estudiar una población grande o un sistema escolar grande, a mucho menor costo que el que generaría un censo. Mientras que las encuestas tienden a ser más costosas que los experimentos de laboratorio y de campo, y aun que los estudios de campo, por la cantidad y calidad de la información que brindan, resultan económicas. Además, es posible utilizar las instalaciones y el personal educativo para reducir los costos de la investigación.

La información de la investigación por encuesta es precisa —con cierto error de muestreo, por supuesto—. La precisión de las muestras obtenidas apropiadamente es, con

frecuencia, sorprendente incluso para los expertos en el campo. Una muestra de 600 a 700 individuos o familias brinda un retrato notablemente preciso de una comunidad: sus valores, actitudes y creencias.

Aunadas a estas ventajas, están las debilidades y desventajas inevitables. Una primera desventaja es que la información de las encuestas generalmente no profundiza mucho debajo de la superficie. Por lo común, la visión de la información buscada se enfatiza a expensas de la profundidad. No obstante, ésta parece constituir una debilidad que no es necesariamente inherente al método. Los estudios de Verba y Nie (1972), y de Smith y Garner (1976; véase también Garner y Smith, 1977) muestran que es posible ir considerablemente a mayor profundidad de las opiniones superficiales. Smith y Garner diseñaron un procedimiento para acompañar un cuestionario bien diseñado, que les permitió penetrar en el comportamiento homosexual de atletas universitarios. En lugar de aplicar un instrumento de encuesta una vez, ellos lo aplicaron por lo menos tres veces para verificar la consistencia de las respuestas. También desarrollaron otros medios de verificación de las respuestas de los atletas, respecto a un tema muy sensible y emplearon medios totalmente inofensivos para recolectar sus datos. Smith y Garner obtuvieron información útil respecto a un tema altamente emocional. A pesar de estos ejemplos sobre la profundidad de la información de la investigación por encuesta, ésta parece adaptarse mejor a la investigación extensiva, que a la intensiva. Otros tipos de investigación quizás están mejor adaptados para una exploración más profunda de algunas relaciones.

La segunda desventaja es de tipo práctico. La investigación por encuesta demanda mucho tiempo, energía y dinero. En una encuesta grande, pueden pasar varios meses antes de que una sola hipótesis llegue a probarse. El muestreo y el desarrollo de buenos inventarios son operaciones importantes. Las entrevistas requieren habilidad, tiempo y dinero. Las encuestas a menor escala pueden evitar dichos problemas en cierto grado.

Cualquier investigación que utilice muestreo está sujeta, naturalmente, al error de muestreo. Aunque es verdad que la información de las encuestas ha mostrado ser relativamente precisa, siempre existe una probabilidad de 20 o 100 de que ocurra un error —más serio de lo que podrían causar fluctuaciones pequeñas por azar. La probabilidad de dicho error puede disminuirse al crear verificaciones de seguridad en un estudio— si se incluyen comparaciones con datos de censos u otra información externa, y por medio del muestreo independiente de la misma población.

Una debilidad potencial, más que real, de este método es que la entrevista de encuesta saca temporalmente al entrevistado de su propio contexto social, lo que puede invalidar los resultados de la encuesta. La entrevista es un suceso especial en la vida ordinaria del entrevistado. Tal vez esa diferencia haga que el entrevistado hable e interactúe con el entrevistador de una manera poco natural. Por ejemplo, una madre, al ser interrogada acerca de las prácticas del cuidado de sus hijos, quizá dé respuestas que revelen métodos que a ella le gustaría utilizar, en lugar de los que en realidad utiliza. Los entrevistadores tienen la oportunidad de limitar los efectos de sacar a los entrevistados de su contexto social, por medio de un manejo experto, en especial con una manera propia y por medio de expresar y plantear las preguntas de manera cuidadosa (véase Cannell y Kahn, 1968).

La investigación por encuesta también requiere de bastante conocimiento y sofisticación sobre investigación. El investigador por encuesta competente debe saber sobre muestreo, sobre la construcción de preguntas e inventarios, la forma de realizar entrevistas, el análisis de datos y otros aspectos técnicos de la entrevista. Dicho conocimiento es difícil de adquirir; pocos investigadores alcanzan este tipo y cantidad de experiencia. Conforme se va apreciando el valor de la investigación por encuesta, tanto de gran escala como de pequeña escala, puede anticiparse que dicho conocimiento y experiencia serán considerados, por lo menos de forma mínima, necesarios para los investigadores.

Meta-análisis

En el momento de escribir estas líneas va en aumento el número de estudios de investigación reportados que utilizan el meta-análisis. El estudiante que lea atentamente la literatura, tiene probabilidad de encontrarse con un estudio que utilice el meta-análisis. Pero, ¿qué es el *meta-análisis*, y por qué se incluye dentro de la investigación por encuesta? Bueno, muchos escritores de libros de texto han tenido problemas para ubicar estos métodos dentro de capítulos de temas específicos. Robert Rosenthal, una de las principales autoridades sobre dicho tema, ubicó el tema del meta-análisis en el apéndice del libro que escribió junto con Ralph Rosnow sobre investigación del comportamiento (Rosnow y Rosenthal, 1996). Algunos autores lo han integrado dentro de capítulos que tratan sobre estadística. Aquí no es diferente. Los autores perciben este método como perteneciente a la investigación por encuesta. Aunque no se diseña ningún cuestionario ni se planea ninguna muestra, sí implica la búsqueda de datos previamente recolectados. Tales datos provienen de la literatura de investigación. Podría decirse que es un tipo de encuesta de la literatura. El meta-análisis es de naturaleza cuantitativa y no experimental. Algunos autores, como Mann (1990), se han referido a éste como un experimento natural. El propósito del meta-análisis es buscar en la literatura un tema específico que contenga un gran número de estudios. Algunos de dichos estudios pueden coincidir entre sí de alguna manera. Si es así, producen una convergencia de conocimiento, y ese conocimiento se vuelve útil en la toma de decisiones. Por ejemplo, si existe un efecto de la preparación para el Scholastic Aptitud Test (SAT), entonces todos los estudios realizados sobre el tema deben tener hallazgos básicos similares. El meta-análisis implica tomar todos estos estudios de forma colectiva, para determinar si un hallazgo similar se encuentra una y otra vez bajo situaciones diferentes. La meta es ser capaz de establecer algún tipo de ley general del comportamiento. A diferencia de los estudios de investigación "regulares", que tienen a los participantes individuales o grupos de participantes como unidad de medición, el meta-análisis utiliza los estudios individuales como unidades de medición. Los resultados de estos estudios de investigación son resumidos por medio del uso de medidas del tamaño del efecto, similares a la ETA cuadrada (η^2) o a la omega cuadrada (ω^2), que se analizaron en un capítulo previo y que se utilizan para estudios de investigación individual. En el meta-análisis, el tamaño del efecto se mide utilizando un estadístico *d*. Gran parte de la investigación meta-analítica que se utiliza hoy, reporta de manera tabulada los diferentes estudios, el tamaño de la muestra y el tamaño del efecto. La tabla 25.2 presenta una adaptación del estudio de Scogin y McElreath (1994) sobre la efectividad de la intervención psicológica en la depresión de adultos mayores.

Para determinar el efecto general del fenómeno bajo estudio, Rosenthal (1978) ofrece un procedimiento estadístico para calcular el tamaño del efecto combinado. Rosenthal

TABLA 25.2 *Tabla meta-analítica del tamaño de la muestra y del tamaño del efecto (de Scogin y McElreath)*

Estudio	Tamaño de la muestra	Tamaño del efecto
1	31	.41
2	36	.00
3	84	.97
4	61	.70
5	28	.82
6	20	.28

toma, en esencia, los valores de p[1] de cada estudio, encuentra la puntuación estándar para cada valor de p, y la utiliza en la fórmula:

$$Z_{general} = \frac{\sum_{i=1}^{n} Z_i}{\sqrt{n}}$$

Después se determina la probabilidad de este valor Z, a través del uso de la tabla de la distribución normal (véase apéndice B). Esto indicará al investigador si el efecto combinado general de los estudios es estadísticamente significativo o no. Por lo tanto, en un meta-análisis el investigador puede encontrar una gran cantidad de estudios respecto a un fenómeno particular, que no fueron estadísticamente significativos. Sin embargo, al combinarse, es posible lograr la significancia estadística; por ejemplo, un investigador encuentra cuatro estudios que tienen los siguientes valores p: .25, .32, .04, .19, para pruebas de una cola. Sus valores Z correspondientes son .69, .47, 1.75 y .50, respectivamente. Observe que sólo el último de estos estudios fue estadísticamente significativo. El valor Z general para este ejemplo sería:

$$Z_{general} = \frac{0.69 + 0.47 + 1.75 + 0.50}{\sqrt{4}} = \frac{3.41}{2} = 1.72$$

El valor Z general es significativo al nivel 0.0427. Rosenthal (1978) muestra nueve formas de reunir resultados de estudios para crear un estadístico general.

El meta-análisis no debe confundirse con otros dos métodos similares: la réplica y el análisis con diferentes modelos o métodos. En la réplica se utilizan la misma metodología y los mismos datos recolectados, de una muestra diferente. La meta de los estudios de réplica es establecer la confiabilidad de los resultados en la misma situación. En el modelo del uso de diferentes métodos, se recolectan los mismos datos de una muestra diferente o se utilizan los datos originales; sin embargo, en este método se utilizan diferentes métodos. El objetivo aquí es indagar qué tan robustos fueron los hallazgos originales. Se hace un esfuerzo por encontrar el modelo que mejor ajuste para realizar predicciones o para tomar decisiones. Ésta puede ser considerada una forma de "extracción de datos", o investigación de métodos múltiples. El meta-análisis esencialmente combina estos dos, la búsqueda de métodos diferentes y de datos diferentes. El objetivo del meta-análisis es generalizar los resultados a situaciones nuevas.

El desarrollo del meta-análisis se acredita a Glass (1976). Smith y Glass (1977) demostraron el meta-análisis en una búsqueda a través de la literatura psicológica, realizada para determinar la eficacia de la psicoterapia. Encontraron cerca de 400 estudios que ofrecían información sobre psicoterapia relevante para su meta. Smith y Glass fueron capaces de sintetizar los resultados de cada uno de estos estudios, para establecer una conclusión general acerca de la eficacia de la psicoterapia. Además, pudieron comparar la eficacia relativa de varios métodos distintos de tratamiento dentro de la psicoterapia. Por lo tanto, el meta-análisis es capaz de responder gran cantidad de preguntas prácticas y de investigación que la investigación individual no puede lograr. Por ejemplo, Blumenthal (1998)

[1] El valor p es otra forma de expresión de la probabilidad de un error tipo I. Generalmente, los estudios reportan los resultados como $p < .05$ (estadísticamente significativa) o $p > .05$ (estadísticamente no significativa). No obstante, en años recientes, con la disponibilidad de las computadoras de alta velocidad y de los programas de computadora, el valor p se calcula directamente.

utiliza un meta-análisis para responder preguntas respecto a las diferencias de género en la percepción del acoso sexual. Este estudio es muy significativo en el área de casos legales y de la corte, o en políticas legales. La mayor parte de los estudios sobre el acoso sexual han incluido participantes a quienes se les presenta una o dos escenas breves sobre un incidente, y después se les plantea una serie de preguntas respecto a la situación. Mientras que la mayoría de los estudios presentan resultados en donde los hombres y las mujeres difieren en su percepción del acoso sexual, la magnitud de los hallazgos ha variado. Inclusive existen algunos estudios que no encontraron diferencias significativas. El estudio de Blumenthal examina la literatura sobre este tema y determina, de manera sistemática, cuál es la situación general en tal aspecto. El estudio de Blumenthal utiliza búsquedas por computadora de estudios con palabras clave tales como "acoso sexual", "percepción" y "diferencias por género". El advenimiento de las búsquedas por computadora ha facilitado el crecimiento de los estudios meta-analíticos.

Antes del desarrollo del meta-análisis, los investigadores se basaban en artículos que aparecían en publicaciones especializadas de revisión, tales como *Psychological Review, American Psychologist, Psychological Bulletin, Harvard Education Review* y el *Annual Review of Psychology*, para encontrar resúmenes de investigación que hubiesen sido realizados en una cierta área. Los escritores de las revisiones eran elegidos, por lo general, debido a que se les consideraba expertos en esa área. A pesar de que los revisores hacían grandes esfuerzos para presentar los datos de forma objetiva, era inevitable cierto nivel de subjetividad. Mann (1990) presenta algunos ejemplos de revisiones subjetivas, realizadas con el modelo tradicional en la ciencia médica. Mann afirma que siempre existe la posibilidad de que ciertos elementos importantes puedan pasarse por alto al hacer una de estas revisiones tradicionales. El meta-análisis proporciona una metodología que suplementa estas revisiones y que satisface una necesidad crítica en la ciencia. Esa necesidad es la resolución de hallazgos de investigación conflictivos. Simon (1987) considera que el meta-análisis no podría resolver el conflicto por completo. Basa su argumento en la premisa de que los estudios meta-analíticos no toman en cuenta suficientes variables independientes. Si se considera el problema planteado por Adelson y Williams, reportado en Simon (1987), el meta-análisis no sería capaz de responder la pregunta sobre cuál de las 34 variables independientes posibles ejerce el mayor efecto sobre el desempeño del piloto.

Para ilustrar algunas de las áreas adecuadas para el meta-análisis, se citarán algunos de los estudios que se han realizado. Scogin y McElreath (1994) realizaron un meta-análisis de 17 estudios respecto a la eficacia de los tratamientos psicosociales para la depresión de adultos mayores. Estos 17 estudios cumplieron el criterio de los estudios que tuvieron una condición de control. Los investigadores buscaron en la literatura artículos relacionados con el tema, publicados entre 1975 y 1990. El promedio del tamaño del efecto encontrado por estos investigadores, resultó estadísticamente significativo e indicó que aquellos que recibieron tratamiento psicosocial estaban más saludables que quienes no recibieron dicho tratamiento. Verhaeghan y DeMeersman (1998) realizaron un meta-análisis de estudios que comparaban adultos mayores y adultos jóvenes sobre el efecto de interferencia de Stroop. El efecto de Stroop ya fue mencionado en un capítulo anterior. Los participantes del estudio nombraban el color de la tinta con que estaba escrita la palabra, en lugar de la palabra misma. Es decir, con la palabra *amarillo* escrita con tinta color verde, ellos tendían a decir "verde", cuando se les pedía que leyeran la palabra. Verhaeghan y DeMeersman encontraron 20 estudios en una búsqueda de literatura por computadora. Los hallazgos de Verhaeghan y DeMeersman demostraron que el efecto de Stroop no se vio afectado por la edad. Hellman (1997) utilizó un meta-análisis para estudiar la relación entre la satisfacción laboral y la intención de dejar el trabajo. Hellman identificó 38 estudios en la búsqueda de la literatura en un periodo de 13 años. La relación entre estas dos variables se encontró,

de forma consistente estadísticamente significativa y en dirección negativa. Es decir, a mayor satisfacción, menor posibilidad de dejar el trabajo; o a menor satisfacción, mayores intentos por dejarlo.

El meta-análisis es un método que puede resumir los resultados de muchos estudios realizados respecto a la misma o similar área temática. No requiere que los estudios estén replicados de manera exacta. Además, posee el apoyo de por lo menos un índice cuantitativo —tamaño promedio del efecto— como ayuda en la evaluación. Además, los índices del tamaño del efecto también pueden compararse estadísticamente entre sí. Sin embargo, existe por lo menos un problema que se ha asociado con el meta-análisis: el "problema del cajón de archivo", el cual surge del hecho de que los editores de las revistas generalmente no aceptan publicar artículos que tengan resultados "no significativos". Es decir, estudios donde no se rechaza la hipótesis nula o estudios que no sean estadísticamente significativos al místico nivel $\alpha = .05$. Barber (1976) lo llama el "efecto negativo". Dichos estudios son considerados "impublicables". Por lo tanto, el meta-análisis, que generalmente se realiza para revisar la literatura de investigación y encontrar artículos publicados del área de interés, contendrá sólo el análisis y las conclusiones extraídas a partir de estudios que sean "estadísticamente significativos". Mientras tanto, los investigadores pueden haber almacenado en sus "cajones de archivo" los estudios de investigación que fracasaron en la producción de resultados significativos. Si un cajón de archivo de este tipo existe, con un gran número de hallazgos no significativos, entonces los meta-análisis reportados por los investigadores pueden ser exageradamente optimistas. Para contrarrestar este problema, algunos investigadores han desarrollado tablas y métodos para dar al investigador alguna idea de la cantidad de tolerancia o distorsión que pudiera estar presente (véase Bradley y Gupta, 1997; Sharpe, 1997). Sin embargo, Rosenthal y Rubin (1978) han desarrollado una fórmula estadística para determinar cuántos estudios negativos se necesitarían para echar por tierra la conclusión extraída con el uso de estudios positivos en un meta-análisis. Rosenthal y Rubin han mostrado que por 345 estudios publicados, se necesitarían 65 123 estudios no publicados, que mostraran un efecto negativo. Sin embargo, como Light y Pillemer (1984) han señalado, existe una diferencia importante entre 50 estudios sin efecto no publicados, y 50 000 estudios sin efecto no publicados, aun cuando ambos estén por debajo de los 65 123 estipulados por Rosenthal y Rubin.

El análisis estadístico del meta-análisis puede ser bastante complejo. Se hizo una breve mención acerca del tamaño del efecto calculado. Sin embargo, el análisis va más lejos que éste. Existen numerosos programas de cómputo disponibles para manejar los cálculos (véase Johnson, 1993; Mullen, 1993).

RESUMEN DE CAPÍTULO

1. La investigación por encuesta es un tipo de estudio de campo cuantitativo.
2. La investigación por encuesta intenta encontrar relaciones entre variables sociológicas y psicológicas.
3. La investigación por encuesta es un desarrollo del siglo XX.
4. El enfoque general de la investigación por encuesta es la gente.
5. Las entrevistas, los inventarios, los paneles y las encuestas por teléfono y por correo constituyen diferentes tipos de encuestas.
6. El tipo de encuesta que produce la mejor información es la entrevista. Las encuestas por correo son las que contienen la mayor cantidad de problemas.
7. La investigación por encuesta puede obtener un amplio rango de información, pero no proporciona información profunda. Es más extensiva que intensiva.

8. La metodología de la investigación por encuesta incluye un "plan de flujo". Este plan bosqueja el diseño y la implementación de una encuesta.
9. La construcción del cuestionario o encuesta es una de las partes importantes del plan. Otra parte importante es el plan de muestreo (por ejemplo, ¿a quién se va a muestrear y cómo se llevará a cabo el muestreo?).
10. La recolección de datos es con frecuencia una tarea laboriosa. Si se utilizan entrevistas, entonces el entrevistador necesita ser entrenado apropiadamente.
11. Convertir los datos a una forma que la computadora pueda leer es otra gran tarea en la investigación por encuesta. Esto también incluye el análisis de los datos.
12. La investigación por encuesta puede ser costosa en términos del tiempo, dinero y trabajo. En una encuesta grande, los hallazgos no son accesibles rápidamente antes de finalizar el estudio.
13. El meta-análisis es una forma de investigación por encuesta. La investigación experimental generalmente utiliza un participante individual como unidad de medición. En el meta-análisis, los estudios individuales son, por sí mismos, la unidad de la medición.
14. El meta-análisis implica recolectar una cantidad de estudios sobre un tema similar y resumir los hallazgos. La meta es definir algunas leyes generales de comportamiento.

SUGERENCIAS DE ESTUDIO

1. A continuación se muestran varios buenos ejemplos de investigación por encuesta; algunos son artículos y otros son libros. Elija uno de ellos. Si elige un libro, lea el primer capítulo para aprender sobre el problema del estudio. Después vaya a la sección técnica (si existe una), para ver cómo se realizaron el muestreo y la entrevista. (La mayoría de los estudios de investigación por encuesta publicados contienen dicha sección.) Intente determinar las variables principales y sus relaciones. Dentro de los corchetes se incluyen resúmenes del contenido.

Cai, D. y You, M. (1998). An ergonomic approach to public squatting-type toilet design. *Applied Ergonomics, 29*, 147-153. [Un estudio sobre el diseño de un tipo de retrete público.]

Glock, C. y Stark, R. (1966). *Christian beliefs and anti-Semitism.* Nueva York: Harper y Row. [Religión y prejuicio.]

Lortie, D. (1975). *Schoolteacher: A sociological study.* Chicago: University of Chicago Press. [Un estudio valioso y revelador sobre los maestros.]

MacDonald, S. Wells, S. y Lothian, S. (1998). Comparison of lifestyle and substance use factors related to accidental injuries at work, home, and recreational events. *Accident Analysis and Prevention, 30*, 21-27.

Miller, W. y Levitin, T. (1976). *Leadership and change: The new politics and the American electorate.* Cambridge, Massachusetts.: Winthrop. [La "nueva izquierda" y la "minoría silenciosa". Con base en los datos de 25 años del Centro de Investigación por Encuesta.]

Murray, A. (1998). The home and school background of young drivers involved in traffic accidents. *Accident Analysis and Prevention, 30*, 169-182. [Investiga la relación entre los antecedentes del hogar y escolares, y los datos sobre accidentes de más de 4 000 conductores masculinos y femeninos, de edades entre 16 y 22 años.

Oates, G. L. (1997). Self-esteem enhancement through fertility? Socioeconomic prospects, gender, and mutual influence. *American Sociological Review, 62*, 965-973. [Determina si tener hijos influye o no en la autoestima de las personas.]

2. Rensis Likert fue un científico social sobresaliente, un pionero metodológico en la investigación por encuesta y fundador del Instituto de Investigación Social de la University of Michigan (de la cual forma parte el Centro de Investigación por Encuesta). Dos de sus colegas, Seashore y Katz (1982) escribieron un obituario en el que describieron las contribuciones de Likert. Se sugiere que los estudiantes lean el obituario, el cual es virtualmente una explicación del nacimiento y crecimiento de aspectos metodológicos importantes de la investigación por encuesta, así como una descripción interesante de las contribuciones de este creativo y competente individuo.
3. Lea una de las siguientes referencias sobre el método del meta-análisis.

Light, R. J. y Pillemer, D. B. (1984). *Summing up: The science of reviewing research.* Cambridge, Massachusetts: Harvard University Press.

Farley, J. U. y Lehmann, D. R. (1986). *Meta-analysis in marketing: Generalization of response models.* Lexington, Massachusetts: Lexington Books.

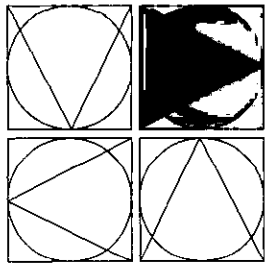
Plucker, J. A. (1997). Debunking the myth of the "highly significant" result: Effect sizes in gifted education research. *Roeper Review, 20*, 122-126.

Rosenthal, R. (1984). *Meta-analytic procedures for social research.* Thousand Oaks, California: Sage.

Sharpe, D. (1997). Of apples and oranges, file drawers and garbage: Why validity issues in meta-analysis will not go away. *Clinical Psychology Review, 17*, 881-901.

4. ¿Alguna vez ha sufrido de dolor de cabeza? Encontrará interesante el siguiente artículo.

McCrory, D. C. y Hasselblad, V. (1997). Cranial electrostimulation for headache: Meta-analysis. *Journal of Nervous and Mental Disease, 185*, 766-767.



CAPÍTULO 26

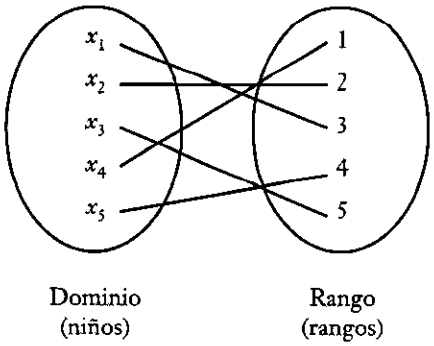
FUNDAMENTOS DE MEDICIÓN

- DEFINICIÓN DE MEDICIÓN
- ISOMORFISMO ENTRE MEDICIÓN Y "REALIDAD"
- PROPIEDADES, CONSTRUCTOS E INDICADORES DE OBJETOS
- NIVELES DE MEDICIÓN Y ESCALACIÓN
  - **Clasificación y enumeración**
  - **Medición nominal**
  - **Medición ordinal**
  - **Medición de intervalo (escalas)**
  - **Medición de razón (escalas)**
- COMPARACIÓN DE ESCALAS: CONSIDERACIONES PRÁCTICAS Y ESTADÍSTICOS

La medición es una de las piedras angulares de la investigación. Cualquier cuantificación de eventos, objetos, lugares y cosas involucra medición. Janda (1998) expresa acertadamente, en el prefacio de su libro, que la medición es fundamental para todas las áreas de la psicología y las ciencias sociales. Todos los procedimientos estadísticos descritos en este libro dependen de la medición. La mayoría de los métodos de recolección de datos, que eventualmente requieren algún tipo de cuantificación, se basan en la medición. Stevens (1951, 1968) afirma que "en su sentido más amplio, la medición es la asignación de valores numéricos a objetos o eventos, de acuerdo con ciertas reglas". La definición de Stevens expresa, de forma sucinta, la naturaleza básica de la medición. Para entenderla, sin embargo, se requiere definir y explicar cada término importante —tarea a la cual se dedica este capítulo—.

Suponga que se le pide a un juez que se pare a siete pies de distancia de un grupo de estudiantes, que observe a los estudiantes y que después estime el grado en que cada uno de ellos posee cinco atributos: simpatía, fuerza de carácter, personalidad, habilidad musical e inteligencia. Las estimaciones deben expresarse numéricamente con una escala de números del 1 al 5, donde el 1 indica una muy pequeña cantidad de la característica en cuestión, y 5 indica una gran cantidad de la misma. En otras palabras, con sólo observar a los estudiantes, el juez debe evaluar qué tan "simpáticos" son, qué tan "fuertes" son sus caracteres, etcétera, utilizando los números 1, 2, 3, 4 y 5 para indicar la cantidad de cada característica que ellos poseen.

FIGURA 26.2



la regla expresada en el párrafo anterior: "si una persona es mujer, asígnele un 1; si es hombre, asígnele un 0". Suponga que 0 y 1 forman el conjunto llamado B; entonces $B = \{0, 1\}$. El diagrama de medición se presenta en la figura 26.1.

Este procedimiento es igual al que se utilizó en el capítulo 5, cuando se discutió sobre relaciones y funciones. En efecto, la medición es una relación. Puesto que a cada miembro de A, el dominio, solamente se le asigna uno y solamente un objeto de B, el rango, la relación constituye una función. ¿Esto significa, entonces, que todos los procedimientos de medición son funciones? Sí, lo son, siempre que los objetos medidos sean considerados el dominio, y los valores numéricos a los que se asignen, o sobre los que se representen, sean considerados el rango.

Aquí hay otra forma de reunir los conceptos de conjunto, relación, función y medición. Recuerde que una relación es un conjunto de pares ordenados; también una función lo es. Entonces, cualquier procedimiento de medición establece un conjunto de pares ordenados, donde el primer miembro de cada par es el objeto medido, y el segundo miembro es el valor numérico asignado al objeto, en concordancia con la regla de medición, cualquiera que ésta sea. Así, ahora es posible escribir una ecuación general para cualquier procedimiento de medición:

$$f = \{(x, y);\ x = \text{cualquier objeto, y } y = \text{un valor numérico}\}$$

que se lee: "la función, f, o la regla de correspondencia, es igual al conjunto de pares ordenados (x, y), de tal manera que x es un objeto, y cada y correspondiente es un valor numérico". Se trata de una regla general que es adecuada para cualquier caso de medición.

Ahora se citará un ejemplo para hacer más concreto el análisis. Los eventos a medir, las x, son cinco niños. Los valores numéricos son los rangos 1, 2, 3, 4 y 5. Suponga que f es una regla que le indica a un maestro lo siguiente: "dé el rango 1 al niño que tenga la mayor motivación para hacer trabajo escolar. Dé el rango 2 al niño que tiene la siguiente mayor motivación para hacer trabajo escolar, y así sucesivamente, hasta el rango 5, el cual debe asignarse al niño que tenga la menor motivación para hacer trabajo escolar". La medición o la función aparece en la figura 26.2.

Observe que f, la regla de correspondencia, quizás habría sido: "si un niño tiene alta motivación para el trabajo escolar, déle un 1; pero si un niño tiene baja motivación para el trabajo escolar, déle un 0". Entonces, el rango sería {0, 1}. Ello tan sólo significa que el conjunto de cinco niños se ha dividido en dos subconjuntos, y a cada uno de ellos se les

asignará, por medio de *f*, los valores numéricos 0 y 1. Un diagrama de esto es similar a la figura 26.1, donde el conjunto *A* es el dominio y el conjunto *B* es el rango.

Regresando a las reglas, aquí es donde la evaluación entra en escena. Las reglas pueden ser "buenas" o "malas". Con reglas "buenas" se tiene una medición "buena" o acertada, si lo demás permanece igual. Con reglas "malas" se tiene una medición "mala" o pobre. Muchas cosas son fáciles de medir a causa de que las reglas son fáciles de elaborar y de seguir. Por ejemplo, medir el sexo resulta fácil, ya que varios criterios simples y bastante claros sirven para determinar el sexo y para indicar al investigador cuándo asignar 1 y cuándo asignar 0. También es fácil medir otras características humanas, tales como color de cabello, color de ojos, estatura o peso. Por desgracia, la mayoría de las características humanas son mucho más difíciles de medir, principalmente porque es difícil idear reglas claras que sean "buenas". No obstante, siempre deben tenerse reglas de algún tipo para medir cualquier cosa.

Isomorfismo entre medición y "realidad"

Como se ha visto, la medición puede ser un asunto sin sentido. ¿Cómo evitar esto? La definición de conjuntos de objetos a medir, la definición de conjuntos numéricos a partir de los cuales se asignan valores numéricos a los objetos que se miden, y las reglas de asignación o correspondencia, deben ligarse con la "realidad". Cuando se mide la dureza de objetos, hay poca dificultad. Si una sustancia *a* puede rayar a *b* (y no a la inversa), entonces *a* es más dura que *b*. De la misma forma, si *a* puede rayar a *b*, y *b* puede rayar a *c*, entonces (probablemente), *a* puede rayar a *c*. Éstas son cuestiones empíricas que son fáciles de comprobar, de tal manera que puede encontrarse un orden de rango de la dureza. Es posible medir la dureza de un conjunto de objetos por medio de unas cuantas pruebas de rayado, asignando valores numéricos para indicar el grado de dureza. Se afirma que el procedimiento de medición y el sistema de números son *isomórficos* a la realidad.

Isomorfismo significa identidad o similitud de forma. Las preguntas planteadas son: ¿este conjunto de objetos es isomórfico a aquel conjunto de objetos? ¿Los dos conjuntos son iguales o similares en algún aspecto formal? ¿Los procedimientos de medición utilizados tienen alguna correspondencia racional y empírica con la "realidad"?

Para demostrar la naturaleza del isomorfismo, es posible utilizar la idea de la correspondencia de conjuntos de objetos. Quizá se desea medir la *persistencia* de siete individuos. Suponga, también, que existe un ser omnisciente, que conoce la cantidad exacta de persistencia que cada individuo posee; es decir, conoce los valores "verdaderos" de persistencia de cada individuo. (Considere que *persistencia* ha sido definida adecuadamente.) Sin embargo, *usted*, quien mide, no conoce estos valores "verdaderos". Es necesario que usted *evalúe* la persistencia de los individuos de alguna forma falible, y usted piensa que ya encontró dicha forma. Por ejemplo, usted evaluaría la persistencia dándoles a los individuos una tarea que realizar y registrando el tiempo total que cada uno requiera para completarla, o puede anotar el número total de veces que el individuo intenta realizar la tarea antes de dirigirse a otra actividad (Feather, 1962). Usted utiliza su método y mide la persistencia de los individuos. Resultan, digamos, los siguientes siete valores: 6, 6, 4, 3, 3, 2, 1. Ahora, el ser omnisciente conoce los valores "verdaderos", que son 8, 5, 2, 4, 3, 3, 1. Este conjunto de valores es la "realidad". La correspondencia de su conjunto con la "realidad" se presenta en la figura 26.3.

En dos casos, usted ha evaluado los valores "verdaderos" de forma exacta; y ha "fallado" en todos los demás. Sin embargo, sólo una de estas "fallas" es seria, y hay una correspondencia bastante buena entre los dos órdenes de rango de los valores. Note, además,

dificultades surgen principalmente sobre el desacuerdo de los estadísticos que pueden utilizarse legítimamente para los diferentes niveles de medición. La posición de Stevens y la definición de medición citada anteriormente es una perspectiva amplia que, con relajación liberal, se sigue en este texto. Una posición más restrictiva —pero defendible— requiere que las diferencias entre las medidas puedan interpretarse como *diferencias cuantitativas de la propiedad medida*. En la perspectiva de algunos expertos, "cuantitativo" significa que una diferencia de magnitud entre dos valores de atributo representa una diferencia cuantitativa correspondiente en los atributos (véase Jones, 1971, pp. 335-355). Estrictamente hablando, esta visión excluye como *medición* a las escalas nominales y ordinales, las cuales se definirán en la siguiente sección de este capítulo. Los autores de este libro consideran que la experiencia real de medición en las ciencias del comportamiento y en la educación justifica una posición más relajada. Nuevamente, esto no tiene una importancia considerable, en caso de que el estudiante *entienda* las ideas generales presentadas. Se recomienda que el estudiante más avanzado lea los capítulos 1 y 2 de Torgerson (1958), y el capítulo 1 de Nunnally (1978); ambas referencias ofrecen buenas presentaciones. Comrey (1950, 1976) y Michell (1990) han influido de manera importante en la orientación que el segundo autor da a este capítulo. Comrey (1976) presenta un ensayo revelador sobre el problema fundamental de la medición en las ciencias sociales y del comportamiento. Un tratado más antiguo y excelente que ha ejercido gran influencia en esta obra es el de Guilford (1954). El estudiante curioso disfrutará la colección de artículos sobre la controversia publicada en el capítulo 2 de un libro editado por Kirk (1972). Los lectores que tengan intenciones de realizar investigación y que siempre se enfrentarán con problemas de medición deben leer cuidadosa y repetidamente las excelentes presentaciones que Nunnally (1978) o Nunnally y Bernstein (1994) hacen de los problemas y de su solución.

En el siguiente análisis, primero se considera el problema científico fundamental y de medición de la clasificación y la enumeración.

Clasificación y enumeración

El primer y más elemental paso en cualquier procedimiento de medición consiste en definir los objetos del universo de información. Suponga que U, el conjunto universal, se define como todos los alumnos de primer año de cierta preparatoria. A continuación, deben definirse las propiedades de los objetos de U. Todas las mediciones requieren que U se separe en, por lo menos, dos subconjuntos. La forma más elemental de medición sería clasificar o categorizar todos los objetos como poseedores o no de alguna característica. Considere que dicha característica es la condición masculina. Se separa U en hombres y no hombres, u hombres y mujeres. Éstos, por supuesto, son dos *subconjuntos* de U, o *particiones* de U. (Recuerde que partir un conjunto consiste en separarlo en subconjuntos que sean *mutuamente excluyentes* y *exhaustivos*; es decir, cada objeto del conjunto debe asignarse a uno y solamente un subconjunto, y que todos los objetos del conjunto de U deben asignarse de esta manera.)

Lo que se ha hecho es clasificar los objetos de interés. Se han ubicado en categorías: se han partido. La simpleza obvia de este procedimiento parece provocar dificultad a los estudiantes. La gente pasa gran parte de su vida categorizando cosas, eventos y personas. La vida no podría continuar sin dicha categorización, aunque asociar el proceso con la medición parece difícil de lograr.

Después de encontrar un método de clasificación, se tiene como efecto una regla que indica cuáles objetos de U van dentro de qué clases, subconjuntos o particiones. Se utiliza la regla y los objetos del conjunto se ubican en los subconjuntos. Aquí están los niños; acá

las niñas. Fácil. Aquí están los niños de clase media; acá los niños de clase trabajadora. No tan fácil, pero tampoco demasiado difícil. Aquí están los delincuentes; acá los no delincuentes. Más difícil. Aquí están los destacados; acá los mediocres, y más allá los lerdos. Mucho más difícil. Aquí están quienes son creativos; acá quienes no son creativos. Muchísimo más difícil.

Después de que los objetos del universo se han clasificado dentro de subconjuntos designados, es posible contar a los miembros de los conjuntos. En caso de dicotomía, la regla de conteo fue expresada en el capítulo 4. Si un miembro de U posee la característica en cuestión, por ejemplo, condición masculina, entonces se asigna 1. Si el miembro no posee la característica, entonces se asigna 0 (véase figura 26.1). Cuando los miembros del conjunto se cuentan de esta manera, todos los objetos de un subconjunto se consideran iguales entre sí, y desiguales respecto a los miembros de otros subconjuntos.

Existen cuatro niveles generales de medición: nominal, ordinal, de intervalo y de razón. Estos cuatro niveles conducen a cuatro tipos de escalas. Algunos escritores sobre el tema aceptan únicamente la medición ordinal, de intervalo y de razón; mientras que otros afirman que los cuatro pertenecen a la familia de la medición. Comrey y Lee (1995) consideran que la escala nominal constituye una forma de medición. Sin embargo, ésta no es tan cuantitativa como la ordinal, la de intervalo y la de razón. Es decir, los números utilizados en la medición nominal son sólo etiquetas numéricas ligadas a categorías predefinidas. No es necesario ser tan exigentes respecto a esto mientras se comprendan las características de las diferentes escalas y niveles.

Medición nominal

Las reglas utilizadas para asignar valores numéricos a los objetos definen el tipo de escala y el nivel de medición. El nivel más bajo de medición es el *nominal* (véase el análisis previo sobre categorización). Los números asignados a los objetos son valores numéricos que no tienen un significado numérico; no pueden ordenarse o sumarse. Son *etiquetas*, parecidas a las letras que se utilizan para nombrar conjuntos. Si a grupos o individuos se les asigna 1, 2, 3, tales valores numéricos son simplemente nombres. Por ejemplo, a los jugadores de beisbol y de futbol se les asignan este tipo de números; a los teléfonos también. A los grupos se les pueden asignar las etiquetas I, II y III o A_1, A_2 y A_3. Utilizamos medición nominal en nuestro pensamiento y vida cotidianos. Identificamos a otros como "hombres", "mujeres", "protestantes", "australianos", etcétera. De cualquier manera, los símbolos asignados a objetos, o mejor dicho, a conjuntos de objetos, constituyen escalas nominales. Algunos expertos no creen que esto sea medición, como se indicó previamente. Pero dicha exclusión de la medición nominal no permitiría que muchos de los procedimientos de investigación en ciencias sociales fuesen llamados medición. Puesto que se satisface la definición de medición y como los miembros de los conjuntos etiquetados pueden contarse y compararse, parece que los procedimientos nominales *son* medición.

Los requisitos de la medición nominal son simples. A todos los miembros de un conjunto se les asigna el mismo valor numérico, y no se le asigna el mismo valor numérico a dos conjuntos. La medición nominal —al menos en una forma simple— fue expresada en la figura 26.1, donde los objetos del rango {0, 1} quedaron representados en las a, los objetos de U, las cinco personas, por medio de la regla: "si x es hombre, asignar 0; si x es mujer, asignar 1". Ésta es la manera en que se cuantifica la medición nominal cuando está involucrada únicamente una dicotomía. Cuando la partición contiene más de dos categorías, debe utilizarse algún otro método. La cuantificación de medición nominal básicamente equivale a contar objetos en las casillas de los subconjuntos o particiones.

considera bueno desde el punto de vista de la medición (escalación). La conversión de estas mediciones a puntuaciones estándar o *Z*, resulta en unidades que pueden considerarse cuantitativamente iguales. Los métodos de escalación que utilizan la curva normal para obtener mediciones en la escala de intervalo pueden, cuando mucho, considerarse aproximaciones con precisión desconocida. Comrey y Lee (1995) presentan un método de este tipo en el capítulo 5 de su libro.

Medición de razón (escalas)

El nivel más alto de medición es el de razón, y el ideal de medición de los científicos es la escala de razón. Una *escala de razón*, además de poseer las características de las escalas nominal, ordinal y de intervalo, posee un cero absoluto o natural con significado empírico. Si una medición es cero en una escala de razón, entonces existe una base para afirmar que un objeto no posee la característica medida. Puesto que existe un cero absoluto o natural, es posible realizar todas las operaciones aritméticas, incluyendo la multiplicación y la división. Los números de la escala indican las cantidades reales de la propiedad medida. Si existiera una escala de razón del *rendimiento*, entonces sería posible decir que un alumno con una puntuación de 8 en la escala posee un rendimiento dos veces mayor que un alumno con una puntuación de 4 en la misma escala.

Uno de los principales problemas en las ciencias sociales y del comportamiento es que la operación de suma no puede definirse (Comrey, 1950). Además, no existen sustitutos satisfactorios reales para el operador de suma en las ciencias sociales y del comportamiento que permita al investigador obtener una escala de medición de razón. Hubo algunos procedimientos de escalación que fueron complejos y parcialmente exitosos, pero, en general, los datos con los que trabajan los científicos sociales y del comportamiento no son siquiera aproximadamente cercanos a datos de una escala de razón.

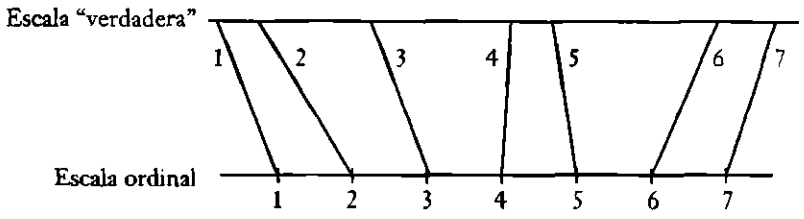
Comparación de escalas: consideraciones prácticas y estadísticos

Las características básicas de los cuatro tipos de medición y sus escalas acompañantes ya se han analizado. ¿Qué tipo de escalas se utilizan en la investigación educativa y del comportamiento? Se utilizan principalmente la nominal y la ordinal, aunque existe una alta posibilidad de que muchas escalas y pruebas utilizadas en la medición psicológica y educativa se aproximen a la medición de intervalo lo suficientemente para propósitos prácticos, como se verá más adelante.

Primero, considere la medición nominal. Cuando los objetos se dividen en dos, tres o más categorías con base en la pertenencia a un grupo —sexo, identificación étnica, casado-soltero, protestante-católico-judío, etcétera— la medición es nominal. Cuando las variables continuas se convierten en atributos, como cuando los objetos se dividen en alto-bajo y viejo-joven, se obtiene lo que puede llamarse medición cuasi-nominal: aunque sujetos de, por lo menos, un orden de rango, los valores son, en efecto, colapsados a 1 y 0.

Resulta instructivo estudiar las operaciones numéricas que son, en un sentido estricto, legítimas con cada tipo de medición. En la medición nominal se permite, por supuesto, el conteo del número de casos en cada categoría y subcategoría. Los estadísticos de frecuencia, como los porcentajes de χ^2 y ciertos coeficientes de correlación (coeficientes de contingencia) pueden utilizarse. Esto suena poco; pero en realidad es bastante. Un buen principio que debe recordarse es éste: si no es posible utilizar cualquier otro método, casi

FIGURA 26.4



siempre es posible realizar una partición cruzada con los participantes. Si se estudia la relación entre dos variables y no se tiene forma adecuada de medirlos de manera ordinal o de intervalo, quizá se pueda encontrar una forma de dividir los objetos de estudio en por lo menos dos grupos. Por ejemplo, al estudiar la relación entre la motivación de los miembros de un consejo de educación para convertirse en miembros del consejo y su religión, como lo hicieron Gross, Mason y McEachern (1958), se pide a jueces expertos que dividan la muestra de miembros del consejo en aquellos con "buena" motivación y aquellos con motivación "pobre". Después se puede hacer una partición cruzada de la religión respecto a la dicotomía de motivación, y así estudiar la relación.

Las puntuaciones de pruebas de inteligencia, aptitud y personalidad son, *hablando de forma básica y estricta*, ordinales. Éstas indican de forma más o menos precisa, no las *cantidades* de los rasgos de inteligencia, aptitud y personalidad de los individuos, sino más bien las *posiciones del orden de rango* de los individuos. Para verlo, es necesario darse cuenta de que las escalas ordinales no poseen las características deseables de igualdad de intervalos o ceros absolutos. Las puntuaciones de pruebas de inteligencia constituyen algunos ejemplos. Un individuo con una puntuación de cero en una medida de inteligencia no necesariamente carece de ella, ya que no existe un cero absoluto en la escala de una prueba de inteligencia. El cero es arbitrario y al no tener un cero absoluto la suma de *cantidades* de inteligencia no tiene ningún sentido, puesto que los puntos de cero arbitrarios conducen a sumas diferentes. Sumar a dos personas cuando cada una tiene una puntuación de inteligencia de 70 no es equivalente a una persona con un CI de 140. En una escala con un punto cero arbitrario se realiza la siguiente suma: 2 + 3 = 5. Entonces, la suma es 5 unidades escalares por arriba de cero. Pero si el punto cero arbitrario es impreciso y el punto del cero "real" está 4 puntos más abajo que la posición del cero arbitrario de la escala, entonces los anteriores 2 y 3 en realidad deberían ser 6 y 7, ¡y 6 + 7 = 13!

La falta de un cero real en las escalas ordinales no es tan seria como la falta de intervalos iguales. Aun sin un cero real, pueden añadirse *distancias* dentro de la escala, siempre y cuando tales distancias sean iguales (empíricamente). La situación podría parecerse a la indicada en la figura 26.4. La escala en la parte superior (escala "verdadera") indica los valores "verdaderos" de una variable. La escala de la parte inferior (escala ordinal) indica la escala de orden de rango utilizada por un investigador. En otras palabras, un investigador ha ordenado por rango a siete personas bastante bien; pero sus valores numéricos ordinales, que se ven con intervalos iguales, no son "verdaderos", aunque puedan ser representaciones bastante precisas de los hechos empíricos.

Estrictamente hablando, los estadísticos que pueden utilizarse con escalas ordinales incluyen las medidas de orden de rango, tales como el coeficiente de correlación de orden de rango, r, la W de Kendall y el análisis de varianza de orden de rango, las medianas y los percentiles. Si únicamente son legítimos dichos estadísticos (y otros similares), ¿cómo es

En el estado que guarda actualmente la medición, no se puede estar seguro de que los instrumentos de medición tengan intervalos iguales. Es importante plantear la pregunta: ¿qué tan serias son las distorsiones y errores introducidos al tratar las mediciones ordinales como si fueran mediciones de intervalo? Al tener cuidado en la construcción de instrumentos de medición, y especial cuidado en la interpretación de los resultados, las consecuencias evidentemente no son serias. Los métodos estadísticos más poderosos dependen menos de la escala de medición subyacente que de las propiedades de distribución de los datos.

El mejor procedimiento parecería ser tratar las mediciones ordinales como si fueran mediciones de intervalo; pero estando constantemente alertas a la posibilidad de desigualdades grandes en los intervalos. Debe aprenderse lo más posible acerca de las características de las herramientas de medición. A través de la apropiada refinación de los métodos de medición y de los procedimientos de escalación, es posible obtener datos que sean aproximadamente normales en su forma. Con datos de este tipo, se pueden utilizar métodos paramétricos de análisis estadístico más poderosos. El investigador debe estar consciente de que es incorrecto ignorar las propiedades escalares de los datos. Por ejemplo, sería inapropiado que un investigador interpretara un grupo con una media de 50 como el doble de un grupo que tuviera una media de 25. Mucha información útil se ha obtenido al tratar datos ordinales como de intervalo, lo que ha resultado en avances científicos en psicología, sociología y educación. En pocas palabras, es muy improbable que los investigadores sean conducidos por mal camino al seguir este consejo, si son cuidadosos al aplicarlo. Para encontrar una útil revisión de la literatura sobre el problema de las escalas de medición y estadística, revise Gardner (1975) o Michell (1990).

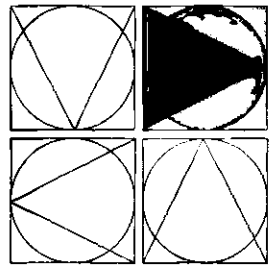
RESUMEN DE CAPÍTULO

1. La medición es un componente importante de investigación.
2. Sin medición o cuantificación de información, muchos métodos de análisis estadístico no podrían utilizarse.
3. Stevens define la medición como el proceso de asignación de números a objetos y eventos, de acuerdo con alguna regla.
4. Stevens define cuatro conjuntos de reglas: nominal, ordinal, de intervalo y de razón.
5. La mayor parte de los datos de las ciencias sociales y del comportamiento son ordinales. Sin embargo, a través de ciertos métodos y supuestos de escalación, pueden considerarse como datos de escala de intervalo.
6. Comrey afirma que una consideración importante es que los datos de las ciencias del comportamiento pueden considerarse de intervalo, si el proceso de medición genera datos que tengan una distribución normal.
7. La medición implica un isomorfismo entre los números y la realidad.
8. Continúa la discusión sobre cuál es la mejor forma de manejar datos de las ciencias sociales y del comportamiento.

SUGERENCIAS DE ESTUDIO

1. ¿Cuál es el primer paso en la medición?
2. De acuerdo a Stevens, ¿cuáles son las reglas que forman parte del proceso de medición?

3. Dé un ejemplo de la ciencia o de la vida diaria que ilustre la medición ordinal.
4. Un artículo interesante escrito hace muchos años por Prokasy (1962) es relevante aun para la discusión actual sobre el uso de métodos paramétricos para datos ordinales. Lea el artículo de Prokasy y, después, revise el capítulo 1 de Cliff (1996).
5. Lea el artículo de F. M. Lord sobre el tratamiento estadístico de datos de futbol americano (en Kirk, 1972). En él se describe, de manera humorística, cómo la gente percibe y utiliza los números. ¿Los números de una escala nominal pueden sumarse?



CAPÍTULO 27

CONFIABILIDAD

Después de asignar valores numéricos a los objetos o eventos de acuerdo con reglas, deben enfrentarse dos grandes problemas de medición: la confiabilidad y la validez. Ya se ha diseñado un sistema de medición y se han administrado los instrumentos de medición a un grupo de participantes. Ahora deben preguntarse y responderse las siguientes preguntas: ¿cuál es la confiabilidad del instrumento de medición? ¿Cuál es su validez?

Si no se conoce la confiabilidad ni la validez de los propios datos, es posible que haya poca fe en los resultados obtenidos y en las conclusiones obtenidas a partir de ellos. Éstas son dos propiedades psicométricas clave que deben ser satisfechas para responder a las muchas críticas hechas a los datos de las ciencias sociales y del comportamiento, así como a los métodos de medición. Los datos de las ciencias sociales y de educación, derivados de la conducta humana y de productos humanos están, como se vio en el capítulo 26, un poco alejados de las propiedades del interés científico; por lo tanto, su validez puede cuestionarse. La preocupación por la confiabilidad proviene de la necesidad de fiarse de la medición. Los datos provenientes de todos los instrumentos de medición en psicología y educación contienen errores de medición. Dependiendo del grado en que contengan errores, los datos que produzcan serán fiables o no.

Definiciones de confiabilidad

Sinónimos de confiabilidad son *estabilidad, fiabilidad, consistencia, reproductibilidad, predictibilidad* y *falta de distorsión*. Por ejemplo, las personas confiables son aquellas cuyo

comportamiento es consistente, predecible y fiable; lo que hacen mañana y la siguiente semana será consistente con lo que hacen hoy y con lo que hicieron la semana pasada; se dice que son estables. Por otro lado, las personas poco confiables son aquellas cuyo comportamiento es mucho más variable; son impredeciblemente variables. En algunas ocasiones hacen algo; y en otras, algo distinto; carecen de estabilidad. Se dice que son inconsistentes.

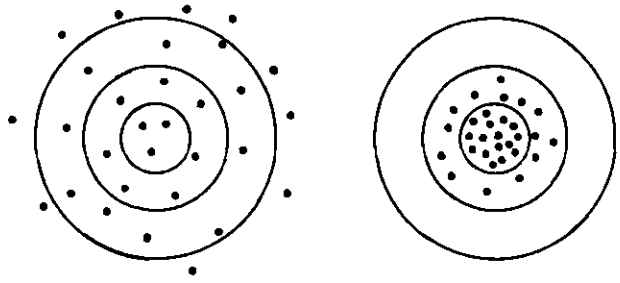
Lo mismo sucede con las mediciones en psicología y en educación: son más o menos variables de una ocasión a otra. O son estables o relativamente predecibles, o son inestables y relativamente impredecibles; son consistentes o no lo son. Si son confiables, entonces se puede depender de ellas; si no son confiables, no se puede depender de ellas.

La definición de confiabilidad se enfoca de tres maneras: un enfoque se sintetiza con la pregunta: si se mide el mismo conjunto de objetos una y otra vez, con el mismo instrumento de medición o uno comparable, ¿se obtendrán iguales o similares resultados? La pregunta implica una definición de confiabilidad en términos de *estabilidad, fiabilidad* y *predictibilidad.* Es la definición que se ofrece en discusiones elementales del tema.

Un segundo enfoque se sintetiza con la pregunta: ¿las medidas obtenidas a partir de un instrumento de medición son las medidas "verdaderas" de la propiedad que se mide? Ésta es una definición de *falta de distorsión*. Comparada con la primera definición, se aleja más del sentido común y de la intuición; sin embargo, es también más fundamental. Estos dos enfoques o definiciones se resumen en las palabras *estabilidad* y *falta de distorsión*. Sin embargo, como se verá más adelante la definición sobre la falta de distorsión implica la definición de estabilidad. La confiabilidad se refiere al grado en el que la medición concuerda consigo misma. En el capítulo 28 se tratará la validez. Con frecuencia los términos "confiabilidad" y "validez" se confunden, no obstante existe una clara distinción entre ellos. La confiabilidad no tiene nada que ver con la veracidad de la medición. Algunos autores se han referido a la confiabilidad como precisión (véase Magnusson, 1967; Tuckman, 1975). Esto es verdad, pero con frecuencia se confunde con el significado de precisión en términos de validez. La validez también tiene que ver con la precisión, pero de una manera diferente que la confiabilidad. La confiabilidad se relaciona con la precisión con la que un instrumento de medición mide aquello que se desea. La palabra clave aquí es "aquello". Si se tiene una prueba que se considera que mide habilidad matemática, no se sabe si la prueba mide, en realidad, habilidad matemática. Si la prueba es altamente confiable, solamente se sabe que está midiendo "algo" con precisión. El asegurarse de que la prueba de habilidad matemática en realidad mide habilidad matemática, implica involucrarse con aspectos de validez.

Existe un tercer enfoque en la definición de confiabilidad, el cual no sólo ayuda a lograr una mejor definición y a resolver tanto problemas teóricos como prácticos, sino que también implica otros enfoques y definiciones. Se puede investigar qué tanto *error de medición* existe en un instrumento de medición. Recuerde que existen dos tipos generales de varianza: sistemática y por el azar. La *varianza sistemática* se inclina hacia una dirección —las puntuaciones tienden a ser todas negativas o todas positivas, o todas altas o todas bajas—. En este caso el error es constante o está sesgado. La *varianza por el azar* o *del error* se autocompensa— las puntuaciones tienden a inclinarse ahora hacia este lado, ahora hacia este otro—. Los errores de medición son errores aleatorios; representan la suma de diversas causas. Entre dichas causas están los elementos comunes del azar o aleatorios —presentes en todas las medidas debido a causas desconocidas—, la fatiga temporal o momentánea, las condiciones fortuitas que en un momento en particular afectan al objeto medido o al instrumento de medición, las fluctuaciones en la memoria y en el estado de ánimo, y otros factores que son temporales y cambiantes. Dependiendo del grado en que los errores de medición estén presentes en un instrumento de medición, el instrumento

FIGURA 27.1



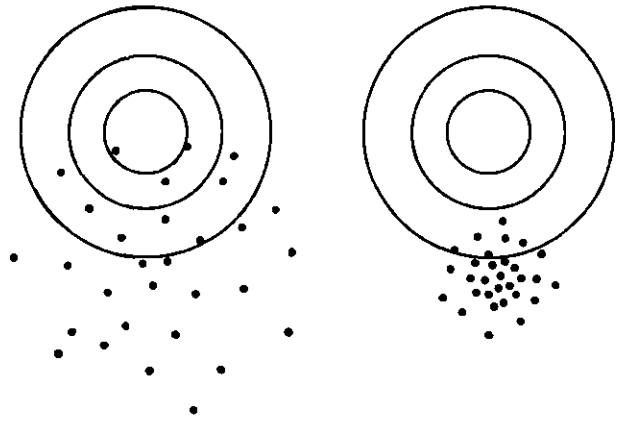
Rifle antiguo Rifle nuevo

será poco confiable. En otras palabras, la confiabilidad puede definirse como la ausencia relativa de errores de medición en un instrumento de medición.

La confiabilidad es la *falta de distorsión* o *precisión* de un instrumento de medición. Recuerde que una medida altamente confiable sólo indica que está midiendo algo con precisión o de forma consistente. Puede ocurrir que no esté midiendo lo que se cree que mide. Un ejemplo para ilustrar lo anterior es la báscula que tenemos en nuestros hogares. Suponga que esta báscula siempre sobrestima el peso de una persona por 5 kilogramos. Si alguien se coloca sobre esta báscula 50 veces durante el periodo de una hora, encontrará muy poca fluctuación del peso registrado en la báscula. La báscula es precisa en el sentido de que indica consistentemente el mismo peso. Sin embargo, es imprecisa en el sentido de que siempre da un peso equivocado por 5 kilogramos. La báscula sería considerada confiable, pero no válida.

Considere que un deportista desea comparar la precisión de dos armas. Una es una pieza antigua fabricada hace un siglo, pero que se encuentra aún en buenas condiciones. La otra es un arma moderna fabricada por un armero experto. Ambas piezas se encuentran fijas en bases de granito y son accionadas hacia un blanco por un pistolero experto. Cada

FIGURA 27.2



Rifle antiguo Rifle nuevo

arma se dispara igual número de veces. En la figura 27.1 se presenta el patrón hipotético de tiros a un blanco para cada una. El blanco de la izquierda representa el patrón de tiros producido por el arma antigua; observe que los tiros se encuentran considerablemente dispersos. Ahora considere que el patrón de tiros en el blanco de la derecha está más junto. Los tiros se encuentran agrupados de forma cercana alrededor del blanco.

Suponga que se asignan números a los círculos del blanco: 3 al centro, 2 al círculo siguiente, 1 al círculo externo y 0 a cualquier tiro que salga del blanco. Es obvio que si se calculan medidas de variabilidad, por ejemplo, una desviación estándar, de los dos patrones de tiro, el rifle antiguo tendría una medida de variabilidad mucho más grande que el rifle más nuevo. Estas medidas pueden considerarse índices de confiabilidad. La medida menor de variabilidad del rifle nuevo indica mucho menos error y, por lo tanto, mucho mayor precisión. El rifle nuevo es confiable; el rifle antiguo es menos confiable.

Ahora analice la figura 27.2. Aquí se tiene el mismo patrón de tiros de ambos rifles; aunque no están centrados en el blanco como en la figura 27.1. El rifle nuevo seguiría considerándose más confiable que el antiguo, pero debido a que ambos se salen del blanco, entonces la puntería no es precisa. Aquí los patrones de la precisión de los tiros de los rifles miden confiabilidad; mientras que la precisión de la puntería de los rifles mide validez. La figura 27.1 ilustra una manera burda de demostrar confiabilidad con validez; en cambio, la figura 27.2 demuestra confiabilidad con poca o ninguna validez. Es posible tener confiabilidad sin validez, pero no a la inversa. La confiabilidad por sí misma resulta poco útil para evaluar la mayoría de las mediciones. Como se indicó antes, una medición puede ser errónea consistentemente. No existe garantía de que el instrumento de medición sea bueno. No obstante, la ausencia de una confiabilidad alta sí indica que el instrumento de medición es pobre.

De forma similar, las mediciones en psicología y educación poseen mayores y menores confiabilidades. Se aplica un instrumento de medición, por ejemplo, una prueba de rendimiento aritmético, a un grupo de niños —generalmente sólo una vez—. La meta, por supuesto, es múltiple: se busca obtener la puntuación "verdadera" de cada niño. En la medida en que se fallen las puntuaciones "verdaderas", el instrumento de medición, la prueba, resulta poco confiable. Las puntuaciones aritméticas "verdaderas" y "reales" de cinco niños, por ejemplo, son 35, 31, 29, 22, 14. Otro investigador desconoce estas puntuaciones "verdaderas". Los resultados obtenidos son 37, 30, 26, 24, 15. Aunque en ningún caso se logró la puntuación "verdadera", todas poseen el mismo orden de rango. La confiabilidad y precisión del investigador son sorprendentemente altas.

Suponga que las cinco puntuaciones hubiesen sido 24, 37, 26, 15, 30. Éstas son las mismas cinco puntuaciones; aunque presentan un orden de rango muy diferente. En este caso, la prueba no sería confiable a causa de su falta de precisión. Para demostrar esto de

TABLA 27.1 *Puntuaciones y órdenes de rango "verdaderos", confiables y no confiables obtenidos de cinco niños*

(1) Puntuaciones "verdaderas"	(Rango)	(2) Puntuaciones de una prueba confiable	(Rango)	(3) Puntuaciones de una prueba no confiable	(Rango)
35	(1)	37	(1)	24	(4)
31	(2)	30	(2)	37	(1)
29	(3)	26	(3)	26	(3)
22	(4)	24	(4)	15	(5)
14	(5)	15	(5)	30	(2)

forma más compacta, los tres conjuntos de puntuaciones, con sus órdenes de rango, se han colocado unos junto a otros en la tabla 27.1. Las órdenes de rango de la primera y segunda columnas covarían de manera exacta. El coeficiente de correlación del orden de rango es 1.00. Aun cuando las puntuaciones de la prueba de la segunda columna no son exactas, se encuentran en el mismo orden de rango. Con base en esto, por medio del uso de un coeficiente de correlación del orden de rango, la prueba es confiable. Sin embargo, el coeficiente de correlación entre los rangos de la primera y tercera columnas es cero, de tal modo que la última prueba no es confiable por completo.

Teoría de la confiabilidad

El ejemplo presentado en la tabla 27.1 sintetiza lo que se debe saber acerca de la confiabilidad. El tratamiento que en este capítulo se da a la confiabilidad está basado en la teoría clásica de las pruebas. Existe un tratamiento mucho más avanzado de confiabilidad realizado por Cronbach, Gleser, Nanda y Rajaratnam (1972), llamado teoría de generalización. Aquí se tratará el modelo más tradicional de confiabilidad. Para hacerlo, es necesario formalizar los conceptos intuitivos y describir una teoría de la confiabilidad, la cual no sólo es elegante conceptualmente, sino que también es poderosa prácticamente. Resulta útil unificar las ideas sobre medición y proporciona un fundamento para comprender varias técnicas analíticas. La teoría también se relaciona de forma adecuada con el modelo de varianza enfatizado en análisis previos.

Cualquier conjunto de medidas posee una varianza total; es decir, después de aplicar un instrumento a un conjunto de objetos y de obtener un conjunto de números (puntuaciones), es posible calcular una media, una desviación estándar y una varianza. Aquí solamente se tratará la varianza, la cual, como se vio antes, es una varianza total obtenida, ya que incluye varianzas debidas a múltiples causas. En general, cualquier *varianza total obtenida* (o suma de cuadrados) incluye la varianza sistemática y del error.

Cada persona posee una puntuación obtenida, X_t. (La "*t*" significa "total".) Algunos autores se refieren a ella como la puntuación observada. Algunas ocasiones sólo se anota "*O*" o X_o. Ésta sería la medición que se hace de un objeto, persona, cosa o evento. La puntuación observada tiene dos componentes: un componente "verdadero" y un componente de error. Se supone que cada persona tiene una puntuación "verdadera", X_∞. (El símbolo "∞" de infinito se utiliza para representar lo "verdadero".) Un símbolo alternativo que el lector puede encontrar en la literatura es T o X_T. Dicha puntuación sería conocida sólo por un ser omnisciente, porque el sistema de medición es imperfecto. Note además lo que se estableció anteriormente. La puntuación verdadera puede incluir propiedades diferentes de la propiedad que se desea medir. El problema para medir esa propiedad es de validez. El otro componente es la puntuación de error, X_e o E; en este caso, error no significa un error que se haya cometido, sino que la puntuación de error es algún incremento o decremento que resulta de varios de los factores responsables de la incapacidad para medir la puntuación verdadera. Por ejemplo, un estudiante quizá tenga una puntuación observada menor que la puntuación "verdadera" debido a que esa persona estuvo enferma el día del examen. Por lo tanto, se puede afirmar que la diferencia entre la puntuación real y la observada es un error. Algunos errores son contables y otros no lo son.

La lógica conduce a una ecuación básica simple para la teoría:

$$X_t = X_\infty + X_e \tag{27.1}$$

o

$$X_0 = X_T + X_e$$

o

$$O = T + E$$

Esto establece, de forma sucinta, que cualquier puntuación observada está formada de dos componentes: un componente "verdadero" y un componente de error. La única parte de esta definición que representa un problema real es X_∞, que se concibe como la puntuación que un individuo obtendría si todas las condiciones internas y externas fueran "perfectas" y si el instrumento de medición fuera también "perfecto". De manera más realista se considera que es la media de un gran número de aplicaciones de la prueba a la misma persona. Simbólicamente, $X_\infty = (X_1 + X_2 + \ldots + X_n)/n$. Lord y Novick (1968) llaman a la puntuación "verdadera" el valor esperado de una puntuación observada, el cual puede interpretarse como la puntuación promedio que un individuo obtendría si toma un número infinito de mediciones independientes repetidas. Considérese lo siguiente: si una persona deseara conocer su estatura, ella puede medirse una vez. ¿Dará esto su estatura "verdadera"? Es poco probable, ya que el aparato de medición es falible. Por lo tanto, la persona haría bien en tomar múltiples mediciones de su estatura y, después, calcular la media de las estaturas. Esta media estaría más cerca de su estatura verdadera que cualquier medición hecha de forma aislada. Si el número de mediciones se acerca al infinito, la media se iría acercando cada vez más a la estatura verdadera.

Con un poco de álgebra simple, la ecuación 27.1 se extiende para producir una ecuación más útil en términos de varianza:

$$V_T = V_\infty + V_E \tag{27.2}$$

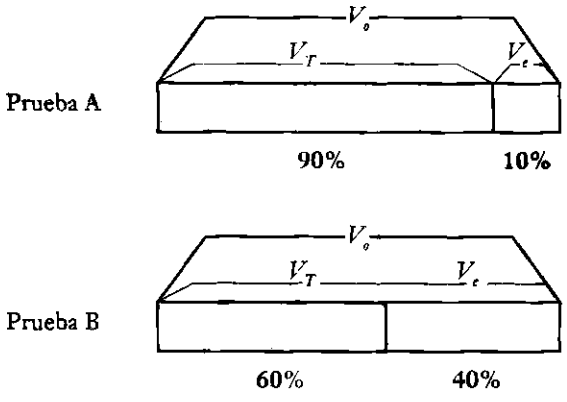
o

$$V_O = V_T + V_e$$

La ecuación 27.2 indica que la varianza total obtenida, de una prueba, se forma de dos componentes de varianza: un componente "verdadero" y un componente de "error". Si, por ejemplo, fuese posible aplicar el mismo instrumento al mismo grupo 4 367 929 veces, y después calcular las medias de las 4 367 929 puntuaciones de cada persona, se tendría un conjunto de mediciones "casi verdaderas" del grupo. En otras palabras, estas medias son las X_∞, del grupo. Entonces podría calcularse la varianza de las X_∞, produciendo V_∞. Este valor siempre debe ser menor que V_t o V_O, la varianza calculada a partir del conjunto de puntuaciones originales obtenido (las X_t u O), debido a que las puntuaciones originales contienen error. Sin embargo, las puntuaciones "verdaderas" o "casi verdaderas" no poseen error, ya que éste se ha eliminado por medio del proceso del cálculo de promedios. En otras palabras, si no hubiese errores de medición en las X_t u O, entonces $V_t = V_\infty$ o $V_O = V_T$. Pero siempre existen errores de medición, y se supone que si se conocieran las puntuaciones de error y se restaran de las puntuaciones obtenidas, entonces se obtendrían las puntuaciones "verdaderas".

Nunca se conocen las puntuaciones "verdaderas" ni tampoco se conocen realmente las puntuaciones de error. No obstante, es posible estimar la varianza del error y, al hacerlo, en efecto es posible sustituir la ecuación 27.2 y resolverla. Ésta es la esencia de la idea, aunque se han omitido ciertos supuestos y pasos de la discusión. Un diagrama muestra las ideas de forma más clara. Sean las varianzas totales de dos pruebas representadas por medio de dos barras. Una prueba es altamente confiable; la otra lo es sólo moderadamente, como se indica en la figura 27.3. Las pruebas A y B tienen la misma varianza total, pero el 90% de la prueba A es varianza "verdadera" y el 10% es varianza del error. Únicamente el 60% de la prueba B es varianza "verdadera" y el 40% restante varianza del error. Por lo tanto, la prueba A es mucho más confiable que la prueba B.

FIGURA 27.3



La confiabilidad se define, por decirlo de alguna manera, a través del error; a mayor error, menor confiabilidad; y a menor error, mayor confiabilidad. Hablando de forma práctica, lo anterior significa que si se estima la varianza del error de una medida, entonces también se puede estimar la confiabilidad de la medida, lo cual conduce a dos definiciones de confiabilidad equivalentes:

1. La confiabilidad es la proporción de la varianza "verdadera" respecto de la varianza total obtenida de los datos producidos por un instrumento de medición.
2. La confiabilidad es la proporción de la varianza del error respecto de la varianza total producida por un instrumento de medición, restado de 1.00; donde el índice 1.00 indica una confiabilidad perfecta.

Resulta más fácil escribir las definiciones en forma de ecuación:

$$r_{tt} = \frac{V_\infty}{V_t} = \frac{V_T}{V_O} \tag{27.3}$$

$$r_{tt} = 1 - \frac{V_e}{V_t} = 1 - \frac{V_e}{V_o} \tag{27.4}$$

donde r_{tt} es el coeficiente de confiabilidad y los otros símbolos fueron ya definidos antes. La ecuación 27.3 es teórica y no puede utilizarse para realizar cálculos. La ecuación 27.4 es tanto teórica como práctica; se utiliza tanto para conceptualizar la idea de confiabilidad como para estimar la confiabilidad de un instrumento. Una ecuación alternativa a (27.4) es:

$$r_{tt} = \frac{V_t - V_e}{V_t} = \frac{V_o - V_e}{V_o} \tag{27.5}$$

Esta ecuación alternativa de la confiabilidad será útil para ayudar a comprender lo que es la confiabilidad.

Dos ejemplos computacionales

Para mostrar la naturaleza de la confiabilidad, en la tabla 27.2 se muestran dos ejemplos. Uno, denominado I en la tabla, es un ejemplo de alta confiabilidad; el otro, denominado II, es un ejemplo de baja confiabilidad. Note con cuidado que se utilizan exactamente los mismos números en ambos casos. La única diferencia es que están ordenados de manera distinta. La situación en ambos casos es: a cinco individuos se les aplicó una prueba con cuatro reactivos. (Lo cual es poco realista, por supuesto, aunque ayudará a ilustrar varias cuestiones.) Los datos de los cinco individuos se encuentran en los renglones; las sumas de los individuos se muestran a la derecha de los renglones (Σ_t). Las sumas de los reactivos se presentan en la parte inferior de cada tabla (Σ_{tt}). Además, las sumas de los individuos en los reactivos impares (Σ_{impar}) y las sumas de los individuos de los reactivos pares (Σ_{par}) se presentan en la extrema derecha de cada subtabla. Los cálculos necesarios para el análisis de varianza de dos factores se muestran debajo de las tablas de datos.

Para volver estos ejemplos más realistas, imagine que los datos son puntuaciones en una escala de 6 puntos respecto a, por ejemplo, las actitudes hacia la escuela. Una puntuación elevada significa una actitud altamente favorable; una puntuación baja, una actitud poco favorable (o nada favorable). (Sin embargo, no hace ninguna diferencia cuáles son las puntuaciones. Inclusive pueden ser unos y ceros resultantes de marcar los reactivos de una prueba de rendimiento: correcto es igual a 1, e incorrecto es igual a 0.) En I, el individuo 1 tiene una actitud altamente favorable hacia la escuela; mientras que el individuo 5 tiene una actitud poco favorable hacia la escuela. Éstas ya están indicadas por las sumas de los individuos (o las medias): 21 y 5. Dichas sumas (Σ_t) son las puntuaciones comúnmente producidas por pruebas. Por ejemplo, si se quisiera conocer la media del grupo, se calcularía como (21 + 18 + 14 + 10 + 5)/5 = 13.60.

La varianza de estas sumas proporciona uno de los términos de las ecuaciones 27.4 y 27.5, pero no el otro: V_t pero no V_e. Utilizando el análisis de varianza es posible calcular tanto V_t como V_e. Los análisis de varianza de I y II indican cómo se hace esto. No es necesario ocuparse demasiado de estos cálculos, ya que son secundarios al tema principal.

El análisis de varianza produce las varianzas: entre reactivos, entre individuos y residual o del error. Las razones F de los reactivos no son significativas en I ni en II. (Observe que ambos cuadrados medios son 2.27. Obviamente deben ser iguales, dado que se calculan a partir de las sumas en la parte baja de las dos subtablas.) En realidad, tales varianzas no representan un interés central —únicamente se desea remover la varianza debida a los reactivos, de la varianza total—. El interés central reside en las varianzas individuales y en las varianzas del error, que se encuentran encerradas por un círculo en las subtablas. La varianza total de las ecuaciones 27.3, 27.4 y 27.5 es interesante, ya que es un índice de las diferencias entre individuos. Es una medida de las diferencias individuales. En lugar de escribir V_t, entonces se escribe V_{ind}, lo cual significa la varianza resultante de las diferencias individuales. Al utilizar (27.4) o (27.5) se obtienen coeficientes de confiabilidad de .92 para los datos de I, y de .45 para los datos de II. Los datos hipotéticos de I son confiables; los de II no lo son en la misma medida.

Con la ecuación 27.4:

$$r_{tt} = 1 - \frac{V_e}{V_{ind}} = 1 - \frac{.81}{10.08} = .92 \quad r_{tt} = 1 - \frac{2.60}{4.70} = .45$$

Con la ecuación 27.5:

$$r_{tt} = \frac{V_{ind} - V_e}{V_{ind}} = \frac{10.08 - 0.81}{10.08} = .92 \quad r_{tt} = \frac{4.70 - 2.60}{4.70} = .45$$

Impares-pares:

$$r_{tt} = .91 \quad r_{tt} = .32$$

Quizás la mejor forma para entender lo anterior sea regresar a la ecuación 27.3. Ahora se escribe $r_{tt} = V_\infty / V_{ind}$. Si se tuviera un camino directo para calcular V_∞, se podría calcular rápidamente r_{tt}, pero como se vio antes, no existe un camino directo. Sin embargo, existe una forma para estimarlo. Si se encuentra una forma para estimar V_e, la varianza de error, entonces el problema está resuelto debido a que V_e puede restarse de V_{ind} para producir un estimado de V_∞. En efecto, es posible ignorar V_∞ y restar la proporción V_e / V_{ind} de 1 y obtener r_{tt}. Ésta es una forma perfectamente aceptable para calcular r_{tt} y para conceptualizar la confiabilidad. La lógica de $V_{ind} - V_e$ tal vez sea más fructífera y se ligue mejor con la discusión previa sobre los componentes de la varianza.

En el capítulo 13 se estableció que cada problema estadístico tiene una cantidad total de varianza, y que cada fuente de varianza contribuye a esta varianza total. Ahora se tradu-

TABLA 27.2 *Demostración de confiabilidad y cálculo de los coeficientes de confiabilidad (ejemplos hipotéticos)*

I: $r_{tt} = .92$

Individuos	Reactivos *a*	*b*	*c*	*d*	Σ_t	$\Sigma_{impares}$	Σ_{pares}
1	6	6	5	4	21	11	10
2	4	6	5	3	18	9	9
3	4	4	4	2	14	8	6
4	3	1	4	2	10	7	3
5	1	2	1	1	5	2	3
Σ_{tr}	18	19	19	12	$\Sigma X_t = 68$ $(\Sigma X_t)^2 = 4\,624$ $\Sigma X_t^2 = 288$		

$$C = \frac{(68)^2}{20} = 231.20$$

Total = 288 − 231.20 = 56.80

$$\text{Entre reactivos} = \frac{1\,190}{5} - 231.20 = 6.80$$

$$\text{Entre individuos} = \frac{1\,086}{4} - 231.2 = 40.30$$

Fuente	*gl*	*sc*	*cm*	*F*
Reactivos	3	6.80	2.27	2.80 (n.s.)
Individuos	4	40.30	(10.08)	12.44 (.001)
Residual	**12**	**9.70**	**(0.81)**	
Total	19	56.80		

II: $r_{tt} = .45$

Individuos	Reactivos *a*	*b*	*c*	*d*	Σ_t	$\Sigma_{impares}$	Σ_{pares}
1	6	4	5	1	16	11	5
2	4	1	5	4	14	9	5
3	4	6	4	2	16	8	8
4	3	6	4	3	16	7	9
5	1	2	1	2	6	2	4
Σ_{tr}	18	19	19	12	$\Sigma X_t = 68$ $(\Sigma X_t)^2 = 4\,624$ $\Sigma X_t^2 = 288$		

$$C = \frac{(68)^2}{20} = 231.20$$

Total = 56.80

Entre reactivos = 6.80

$$\text{Entre individuos} = \frac{1\,000}{4} - 231.20 = 18.80$$

Fuente	*gl*	*sc*	*cm*	*F*
Reactivos	3	6.80	2.27	1 (n.s.)
Individuos	4	18.80	(4.70)	1.81 (n.s.)
Residual	**12**	**31.20**	**(2.60)**	
Total	19	56.80		

cirá el razonamiento del capítulo 13 al problema presente. En muestras aleatorias de la misma población, V_e y V_∞ deben ser iguales estadísticamente. Pero si V_e, la varianza entre grupos, es significativamente mayor que V_d, la varianza dentro de grupos (error), entonces existe algo en V_e más allá y por encima del azar. Esto es, V_e incluye la varianza de V_d y, además, un poco de varianza sistemática.

De forma similar, puede decirse que si V_{ind} es significativamente mayor que V_e, entonces existe algo en V_{ind} más allá y por encima de la varianza del error. Dicho exceso de varianza parecería que se debe a diferencias individuales en aquello que se esté midiendo. La medición apunta hacia las puntuaciones "verdaderas" de los individuos. Cuando se dice que la confiabilidad es la precisión de un instrumento de medición, se quiere indicar que un instrumento confiable de medición más o menos mide las puntuaciones "verdaderas" de individuos, siendo que el "más o menos" depende de la confiabilidad del instrumento. El hecho de que se midan las puntuaciones "verdaderas" puede inferirse únicamente a partir de las *diferencias* "verdaderas" entre individuos; aunque ninguna de ellas pueda, por supuesto, medirse de forma directa. Lo que se hace es inferir las diferencias "verdaderas" a partir de las diferencias empíricas y falibles medidas, las cuales están siempre, en cierta medida, corruptas por errores de medición.

Ahora, si existe alguna manera de eliminar de V_{ind} el efecto de los errores de medición, alguna manera de liberar a V_{ind} del error, entonces el problema se resuelve con facilidad. Tan sólo se resta V_e de V_{ind} para obtener un estimado de V_∞. Entonces la proporción de la varianza "pura" respecto de toda la varianza, "pura" e "impura", es el estimado de la confiabilidad del instrumento de medición. Para resumirlo simbólicamente:

$$r_{tt} = \frac{V_\infty}{V_{ind}} = \frac{V_{ind} - V_e}{V_{ind}} = 1 - \frac{V_e}{V_{ind}}$$

Los cálculos reales se presentan en la parte final de la tabla 27.2.

Regresando a los datos de la tabla 27.2, analice si es posible "observar" la confiabilidad de I y la no-confiabilidad de II. Observe primero las columnas donde están registrados los totales de los individuos (Σ_t). Note que las sumas de I tienen un mayor rango que las de II: 21 – 5 = 16 y 16 – 6 = 10. Dados los mismos individuos, a mayor confiabilidad de una medida, mayor será el rango de las sumas de los individuos. Piense en el extremo: un instrumento completamente no confiable produciría sumas parecidas a las sumas producidas por números aleatorios y, por supuesto, la confiabilidad de los números aleatorios es aproximadamente de cero. (La razón F no significativa para individuos, 1.81 en II, indica que r_{tt} = .45 no es estadísticamente significativo.)

Ahora examine los órdenes de rango de los valores bajo los reactivos *a*, *b*, *c* y *d*. En I los cuatro órdenes de rango son casi iguales. Aparentemente cada reactivo de la escala de actitud está midiendo la misma cuestión. Dependiendo del grado en que los reactivos individuales produzcan los mismos órdenes de rango de individuos, la prueba será confiable. Los reactivos permanecen unidos, por decirlo así; son consistentes internamente. Note también que los órdenes de rango de los reactivos de I son casi los mismos que los órdenes de rango de las sumas.

Los órdenes de rango de los valores de los reactivos de II son bastante diferentes. Los órdenes de rango de *a* y *c* concuerdan bastante; son iguales a los de I. Sin embargo, los órdenes de rango de *a* y *b*, *a* y *d*, *b* y *d*, *c* y *d*, no concuerdan muy bien. O los reactivos están midiendo cuestiones diferentes, o no están midiendo de forma muy consistente. Esta falta de congruencia de los órdenes de rango se refleja en los totales de los individuos. A pesar de que los órdenes de rango de los totales es similar a los órdenes de rango de los totales de

I, el rango o varianza es considerablemente menor, y existe una falta de dispersión entre las sumas (por ejemplo, los tres números 16).

Se concluye la consideración de estos dos ejemplos examinando ciertas cifras en la tabla 27.2, que no fueron consideradas anteriormente. En el lado derecho de I y II se presentan las sumas de los reactivos impares ($\Sigma_{impares}$) y las sumas de los reactivos pares (Σ_{pares}). Tan sólo se suman los valores de los reactivos impares a través de los renglones: $a + c$: 6 + 5 = 11, 4 + 5 = 9, 4 + 4 = 8, etcétera, en I. Después se suman también los valores de los reactivos pares en I: $b + d$: 6 + 4 = 10, 6 + 3 = 9, etcétera. Si hubiera más reactivos, por ejemplo, a, b, c, d, e, f, g, entonces se sumarían: $a + c + e + g$ para las sumas impares, y $b + d + f$ para las sumas pares. Para calcular el coeficiente de confiabilidad, se calcula la correlación producto-momento entre las sumas impares y las sumas pares, y después se corrige el coeficiente resultante con la fórmula de Spearman-Brown. Tanto las sumas de los reactivos impares como de los pares son, por supuesto, las sumas de únicamente la mitad de los reactivos de una prueba. Por ende, son menos confiables que las sumas de todos los reactivos. La fórmula de Spearman-Brown corrige el coeficiente impar-par (y otros coeficientes partidos) para el menor número de reactivos utilizados en el cálculo del coeficiente. (Se explicará más sobre esto en una sección posterior de este capítulo. También se pueden consultar varias pruebas buenas y libros de medición tales como el de Anastasi y Urbina, 1997; Brown, 1983; Friedenberg, 1995; o Sax, 1997.) Los r_n impar-par para I y II son .91 y .32, respectivamente; bastante cerca de los resultados del análisis de varianza de .92 y .45. (Con más participantes y más reactivos, los estimados generalmente son más cercanos.)

Esta simple operación quizá parezca desconcertante. Para observar que ésta es una variación del mismo tema sobre la varianza y el orden de rango, observe primero el orden de rango de las sumas de los dos ejemplos. Los órdenes de rango de Σ_{impar} y Σ_{par} son casi iguales en I, pero bastante diferentes en II. La lógica es la misma que antes. Evidentemente, los reactivos están midiendo la misma cuestión en I, pero en II los dos conjuntos de reactivos no son consistentes. Para reconstruir la discusión sobre la varianza, recuerde que al sumar la suma de los reactivos impares con la suma de los reactivos pares de cada persona, se obtiene la suma total o $\Sigma_{impares} + \Sigma_{pares} = \Sigma_t$.

Interpretación del coeficiente de confiabilidad

Si r, el coeficiente de correlación, se eleva al cuadrado, se convierte en un coeficiente de determinación. Éste brinda la proporción o porcentaje de la varianza compartida por dos variables. Si $r = .90$, entonces las dos variables comparten $(.90)^2 = 81\%$ de la varianza total de las dos variables en común. El coeficiente de confiabilidad es también un coeficiente de determinación. Teóricamente indica cuánta varianza, de la varianza total de una variable medida, es "verdadera". Si se tuvieran las puntuaciones "verdaderas" y se pudieran correlacionar con las puntuaciones de la variable medida, y se elevara al cuadrado el coeficiente de correlación resultante, entonces se obtendría el coeficiente de confiabilidad.

Una representación simbólica servirá para aclarar esto. Sea $r_{t\infty}$ el coeficiente de correlación entre las puntuaciones obtenidas y las puntuaciones "verdaderas", X_∞. El coeficiente de confiabilidad se define de la siguiente manera:

$$r_{tt} = (r_{t\infty})^2 \qquad \textbf{(27.6)}$$

Aunque no es posible calcular $r_{t\infty}$ de forma directa, es útil entender la lógica del coeficiente de confiabilidad en dichos términos teóricos. La correlación de la puntuación verdadera con la puntuación observada con frecuencia se conoce como el *índice de confiabilidad*.

Puesto que una puntuación verdadera es algo que existe pero que no puede medirse, es obvio que el índice de confiabilidad no puede calcularse directamente. Como resultado, el coeficiente de confiabilidad no puede obtenerse de manera directa, por lo menos a través de este método. No obstante, existen varias formas para calcular la confiabilidad de las mediciones. Magnusson (1967) se refiere a ellas como métodos prácticos para estimar la confiabilidad. El primero consiste en aplicar el mismo instrumento de medición al mismo grupo de personas, en dos ocasiones diferentes. El lapso de tiempo entre las dos ocasiones depende del tipo y del propósito de las mediciones. Por lo común, se elige un intervalo de tiempo entre ambas aplicaciones, para que haya suficiente disminución del recuerdo sobre las respuestas. La realización adecuada del procedimiento conduce a dos mediciones por persona, las cuales, dadas en pares, se utilizan en una fórmula para calcular la correlación. Dicha correlación entre las puntuaciones de la ocasión 1 y de la ocasión 2 se denomina *confiabilidad test-retest.* Sirve para medir la estabilidad a través del tiempo. Ésta no es una buena manera para calcular el coeficiente de confiabilidad si el abandono escolar es alto o si los organismos que se están midiendo pasarán por un cambio drástico en el desarrollo, entre el periodo 1 y el periodo 2. Si el instrumento de medición es una prueba de vocabulario, la confiabilidad test-retest puede no resultar fructífera si la prueba se aplica, en dos o más ocasiones, a niños que están expuestos a un ambiente educativo donde su vocabulario se desarrolla rápidamente. Otra interpretación teórica es considerar que cada X_∞ puede ser la media de un gran número de X_n derivadas de la aplicación de una prueba a un individuo un gran número de veces, si lo demás permanece igual. La idea que subyace a esto se explicó anteriormente. La primera aplicación de la prueba produce, por ejemplo, un cierto orden de rango de los individuos. Si la segunda, tercera o más mediciones tienden a producir aproximadamente el mismo orden de rango, entonces la prueba es confiable, lo cual representa una interpretación de estabilidad o test-retest de la confiabilidad.

Otro método que puede utilizarse para calcular el coeficiente de confiabilidad consiste en desarrollar dos *formas equivalentes* o *paralelas* del instrumento de medición. En términos de prueba, esto implicaría crear dos formas de la prueba. Las dos formas serían equivalentes, pero no idénticas. Estarían compuestas de reactivos similares, posiblemente del mismo banco de reactivos. Cada persona estaría sujeta a mediciones por medio de los dos instrumentos. Como resultado, cada persona tendría, entonces, dos puntuaciones y, nuevamente, los pares de puntuaciones serían utilizados en una fórmula de correlación para calcular la correlación. Tal correlación sería considerada como una forma paralela o equivalente de confiabilidad. Dicho método posee la ventaja de minimizar las deserciones escolares. Además, tampoco hay que preocuparse demasiado respecto a si las personas que se están midiendo recordarán las respuestas. Sin embargo, las formas paralelas tienen algunos problemas. Por un lado, se requiere que el investigador realice dos formas de la prueba, las que necesitarían tener medias y desviaciones estándar que sean equivalentes estadísticamente. También, el procedimiento deseable requeriría que las personas que se miden tengan que estar sujetas a mediciones durante un periodo más largo y por ende serían susceptibles a la fatiga y el aburrimiento. Si esto sucede, entonces se afectaría su desempeño en los últimos reactivos, lo que podría contribuir a disminuir el coeficiente de confiabilidad.

La tercera categoría para calcular el coeficiente de confiabilidad se denomina *consistencia interna.* Existen varios métodos para obtener la consistencia interna. Cada método depende de ciertos supuestos que pueden hacerse sobre las mediciones. El primero se llama *confiabilidad por mitades*; el segundo, *coeficiente alfa*, y el tercero, *fórmulas 20 y 21 de Kuder-Richardson* (KR-20, KR-21). Aunque en el siguiente análisis se utilizará el término *prueba* para designar al instrumento de medición, no necesariamente tiene que ser una prueba en sí. Como brevemente se mencionó y demostró antes, la confiabilidad por mita-

des implica dividir la prueba en dos mitades. El objetivo es obtener dos mitades iguales o equivalentes, lo cual se logra sumando todas las respuestas a los reactivos de la primera mitad, o sumando todas las respuestas a los reactivos de la segunda mitad. Si todos los reactivos son homogéneos, entonces las dos mitades serán iguales. Si la prueba inicia con los reactivos más fáciles y progresa hacia los más difíciles, entonces el método mencionado previamente no será efectivo en producir mitades iguales. El método recomendado aquí sería sumar todas las respuestas a los reactivos impares para crear un total, y luego sumar todas las respuestas a los reactivos pares para crear el otro total. En cualquiera de los casos anteriores, cada persona tendrá dos puntuaciones de mitad de suma. Estas puntuaciones se correlacionan utilizando la fórmula estándar. La correlación resultante se nombraría "confiabilidad por mitades". Como se demostró en Magnusson (1967), Allen y Yen (1979) y en el trabajo clásico de Gullikson (1950) con reactivos homogéneos, a mayor tamaño de la prueba (más reactivos), habrá mayor confiabilidad; a menor tamaño de la prueba (menos reactivos), habrá menor confiabilidad. Con el método de confiabilidad por mitades, ya no se está hablando acerca de una confiabilidad de la prueba completa: la confiabilidad por mitades subestimará la confiabilidad real, pues ahora se trata de la correlación de dos mitades de la prueba. Al utilizar la confiabilidad por mitades se necesita utilizar una de tres fórmulas para estimar la confiabilidad de la prueba completa, basado en valores de la mitad de ella.

Una de estas fórmulas es la fórmula profética de Spearman-Brown, la cual tiene otros usos además de la estrategia por mitades. Con el uso de esta fórmula, junto con el supuesto de que las mitades son iguales, puede calcularse un estimado de la confiabilidad de la prueba completa. La fórmula de Spearman-Brown es:

$$r_{tt} = \frac{nr_{tt}}{1 + (n - 1)r_{tt}}$$

Para la estrategia por mitades, n se establece igual a 2. La r_{tt} es la confiabilidad por mitades, y la r_{tt} es la confiabilidad estimada para la prueba completa.

Las otras dos fórmulas son distintas en apariencia, pero ambas tienen el mismo propósito. Antes de describirlas, es necesario reiterar que la fórmula de Spearman-Brown puede aplicarse a otras situaciones de confiabilidad (véase Anastasi y Urbina, 1997). También podría emplearse cuando el investigador esté relativamente seguro de que las dos mitades son iguales. Si existe cualquier duda respecto a la homogeneidad de las mitades, no debe utilizarse la fórmula Spearman-Brown, ya que sobrestimará la confiabilidad de la prueba completa. En su lugar, es preferible utilizar la fórmula de Rulon o la fórmula de Guttman (Magnusson, 1967). Ambas toman en cuenta las diferencias entre las mitades. Tanto la fórmula de Rulon como la fórmula de Guttman estiman la confiabilidad de la prueba completa sin el uso de la confiabilidad por mitades.

La fórmula de Rulon es

$$r_{tt} = 1 - \frac{V_d}{V_t} = 1 - \frac{V_{(a-b)}}{V_t}$$

y la fórmula de Guttman es

$$r_{tt} = 2\left[1 - \frac{(V_a + V_b)}{V_t}\right]$$

donde a representa el total de la primera mitad de puntuaciones; y b, el total de la segunda mitad de puntuaciones. V_d es la varianza de la diferencia de las puntuaciones ($d = a - b$), V_t es la varianza de las puntuaciones totales ($t = a + b$). V_a es la varianza del total de la primera mitad de puntuaciones; y V_b, la varianza del total de la otra mitad de puntuaciones.

Para sintetizar, los reactivos de la prueba se consideran homogéneos. Esta interpretación, en efecto, se reduce a la misma idea de otras interpretaciones: precisión. Tome cualquier muestra aleatoria de reactivos de la prueba y cualquier otra muestra aleatoria diferente de reactivos de la misma. Trate cada muestra como una subprueba separada. Entonces, cada individuo tendrá dos puntuaciones: una X_p para una submuestra, y otra X_t para la otra submuestra. Se correlacionan los dos conjuntos, y se continúa el proceso indefinidamente. La intercorrelación promedio de las submuestras (correlacionadas por medio de la fórmula Spearman-Brown) demuestra la consistencia interna de la prueba. Pero esto significa realmente que cada submuestra —si la prueba es confiable— tiene éxito en producir aproximadamente el mismo orden de rango de los individuos. Si no es así, entonces la prueba no es confiable.

La confiabilidad por mitades está basada en dos mitades que generalmente se consideran equivalentes o paralelas. Si este concepto se lleva más allá al considerar cada reactivo como una prueba paralela separada, es posible derivar algunas de las medidas de confiabilidad que se encuentran comúnmente en la literatura sobre investigación psicológica y educativa. En 1937, Kuder y Richardson desarrollaron esta idea, la cual resultó en dos de las fórmulas de confiabilidad más utilizadas para la consistencia interna: KR-20 y KR-21. Están numeradas de esta forma a causa de que la KR-20 fue la vigésima ecuación en su artículo, y la KR-21 fue la vigésimo primera ecuación. Ambas asumen que cada reactivo tiene la misma media y la misma varianza. Las fórmulas de Kuder-Richardson son aplicables a instrumentos de medición (por ejemplo, pruebas) con un sistema dicotómico o binario de calificación de respuesta. Un ejemplo de calificación dicotómica son los reactivos que se califican como correctos (1) o incorrectos (0). Las pruebas con respuestas de verdadero-falso también se consideran como un sistema dicotómico de calificación. Si se elige que p sea la proporción de receptores de la prueba que responden correctamente el reactivo i (o que se considera "verdadero"), entonces q_i es la proporción que responde incorrectamente el reactivo i (o que se considera "falso"). k es el número de reactivos en la prueba. Con esta información, la fórmula KR-20 se ve así:

$$r_{tt} = \frac{k}{k-1}\left(\frac{V_t - \sum p_i q_i}{V_t}\right)$$

Si se asume que cada reactivo tiene las mismas p_i y q_i, entonces $\sum p_i q_i$ puede reemplazarse por kp_iq_i. Al hacer esto se llega a KR-21.

$$r_{tt} = \frac{k}{k-1}\left(\frac{V_t - kp_iq_i}{V_t}\right)$$

la cual puede simplificarse aún más a:

$$r_{tt} = \frac{k}{k-1}\left(1 - \frac{Mk - M^2}{kV_t}\right)$$

donde k es el número de reactivos y M es la media del total de las puntuaciones. En esencia KR-21 es un caso especial de KR-20, donde p_iq_i (también conocido como dificultades o

respaldo de los reactivos) son iguales. Si un investigador desea obtener el estimado de confiabilidad más conservador, para un instrumento con reactivos que usan calificación binaria, entonces se recomienda esta fórmula. Observe que este coeficiente subestimará KR-20 si las dificultades o respaldo de los reactivos tienen un rango amplio.

A manera de recordatorio, las fórmulas KR-20 y KR-21 son aplicables cuando los reactivos de un instrumento de medición (por ejemplo, una prueba) tienen calificación binaria o la escala de respuestas es dicótoma. Si el formato de calificación o de respuesta no es binario, esta fórmula no puede utilizarse. En el periodo entre el desarrollo de Kuder-Richardson en 1937, y el desarrollo del coeficiente alfa de Cronbach en 1951, se desarrollaron muchas pruebas psicológicas con base en un sistema binario de respuesta. Con la creación de Cronbach (1951), los investigadores fueron capaces de evaluar la confiabilidad de consistencia interna de su instrumento, el cual tenía diferentes escalas de calificación y de respuesta. De hecho, a través de una prueba matemática es posible demostrar que las fórmulas de Kuder-Richardson son casos especiales del coeficiente alfa de Cronbach o alfa de Cronbach. De este rango de coeficientes de confiabilidad, el coeficiente alfa es el más general. Con éste ahora es posible que un investigador encuentre la confiabilidad de instrumentos que utilicen escalas de Likert. La fórmula del alfa de Cronbach es la siguiente:

$$r_{tt} = \alpha = \frac{k}{k-1}\left(1 - \frac{\sum V_i}{V_t}\right)$$

Un método alternativo para escribir el coeficiente alfa, utilizando la intercorrelación entre reactivos, es

$$r_{tt} = \frac{n\bar{r}_{tt}}{1 + (n-1)\bar{r}_{tt}}$$

donde $\bar{r}_{tt}$ es la media de las correlaciones inter-reactivos. Lo que esto significa, esencialmente, es que si se correlacionara cada reactivo con cada uno de los demás reactivos del instrumento, se encontrara la media de dichas correlaciones y después se insertara la media de las correlaciones inter-reactivo en la fórmula de Spearman-Brown, entonces se obtendría el coeficiente alfa o la fórmula de Kuder-Richardson.

Cabe señalar que el ejemplo computacional realizado anteriormente en este capítulo constituye un ejemplo donde se puede utilizar el análisis de varianza para determinar el coeficiente de confiabilidad, y debe ser equivalente al coeficiente alfa.

El error estándar de la media y el error estándar de medición

Dos aspectos importantes de la confiabilidad son la confiabilidad de las medias y la confiabilidad de las medidas individuales, los cuales se relacionan con el error estándar de la media y el error estándar de la medición. En estudios de investigación, generalmente el error estándar de la media y de estadísticos relacionados —como el error estándar de las diferencias entre medias y el error estándar de un coeficiente de correlación— es el más importante de ellos. Puesto que el error estándar de la media se discutió de manera considerable en un capítulo anterior, sólo es necesario decir aquí que la confiabilidad de estadísticos específicos es otro aspecto del problema general de confiabilidad. El error

TABLA 27.3 *Confiabilidad y error estándar de medición (ejemplo hipotético)*

	X_t	X_∞	X_e
	2	1	1
	1	2	−1
	3	3	0
	3	4	−1
	6	5	1
Σ:	15	15	0
M:	3	3	0
V:	2.8	2.0	.80

$$r_{tt} = 1 - \frac{V_e}{V_t} = 1 - \frac{V_e}{V_O} = 1 - \frac{.80}{2.80} = 0.71 \qquad r_{t\infty} = 0.845$$

$$r_{tt} = \frac{V_\infty}{V_t} = \frac{2.00}{2.80} = 0.71 \qquad r_{tt} = r^2_{t\infty} = (.845)^2 = 0.71$$

$$VE_{med} = V_t(1 - r_{tt}) = 2.80\ (1 - 0.71) = 0.81$$

$$EE_{med} = DE_t\sqrt{1 - r_{tt}} = \sqrt{VE_{med}}\sqrt{0.81} = 0.90$$

estándar de medición, o su cuadrado, la varianza estándar de medición, necesita definirse e identificarse, aunque sea de manera breve. Esto se hará mediante un ejemplo simple.

Un investigador mide las actitudes de cinco individuos y obtiene las puntuaciones presentadas bajo la columna llamada X_t, en la tabla 27.3. Suponga, además, que las puntuaciones "verdaderas" de actitud de los cinco individuos son aquellas presentadas bajo la columna llamada X_∞. (Sin embargo, recuerde que en la realidad nunca es posible conocer estas puntuaciones.) Puede notarse que el instrumento es confiable. A pesar de que sólo una de las puntuaciones obtenidas es exactamente igual a su puntuación acompañante "verdadera", las diferencias, entre las puntuaciones obtenidas que son diferentes y las puntuaciones "verdaderas", son pequeñas. Tales diferencias se presentan bajo la columna llamada "X_e": son "puntuaciones de error". Evidentemente el instrumento es bastante preciso. El cálculo de r_{tt} confirma dicha impresión: .71.

Una medida muy directa de la confiabilidad del instrumento puede obtenerse al calcular la varianza o la desviación estándar o las puntuaciones de error (X_e). La varianza de las puntuaciones de error y las varianzas de las puntuaciones X_t y X_∞ se calcularon y se incluyeron en la tabla 27.3. La varianza de las puntuaciones de error ahora se nombran, justificadamente, como *varianza estándar de medición*, la cual podría llamarse con mayor precisión "varianza estándar de los errores de medición". La raíz cuadrada de dicho estadístico se denomina *error estándar de medición*. La varianza estándar de medición se define de la siguiente manera:

$$VE_{med} = V_t\,(1 - r_{tt}) \tag{27.7}$$

En efecto, sólo es posible calcular tal estadístico, si se conoce el coeficiente de confiabilidad. Note que si existe alguna forma para estimar VE_{med}, entonces es posible calcular el coeficiente de confiabilidad. Esto requiere de mayor investigación.

Se inicia con la definición de confiabilidad dada anteriormente: $r_{tt} = V_\infty / V_t = 1 - V_e / V_t$. Una ligera manipulación algebraica produce la varianza estándar de medición:

$$r_{tt} = 1 - \frac{V_e}{V_t}$$

$$r_{tt} V_t = V_t - V_e$$

$$V_e = V_t - r_{tt} V_t$$

$$V_e = V_t(1 - r_{tt})$$

La parte derecha de la ecuación es igual a la parte derecha de la ecuación 27.7. Por lo tanto, $V_e = VE_{med}$, o la varianza de error utilizada anteriormente en el análisis de varianza es la varianza estándar de medición. La varianza estándar de medición y el error estándar de medición del ejemplo se calcularon en la tabla 27.3, y son .81 y .90, respectivamente. Como muestran los libros de texto sobre medición (por ejemplo, Anastasi y Urbina, 1997), sirven para interpretar puntuaciones individuales de pruebas. Dicha interpretación no será discutida aquí; tales estadísticos se han incluido sólo para demostrar la conexión entre la teoría original y las formas para determinar la confiabilidad.

Otro cálculo de la tabla 27.3 requiere de una explicación. Si se correlacionan las puntuaciones X_t y X_∞, se obtiene un coeficiente de correlación de .845. Ahora se obtiene este coeficiente $r_{t\infty}$ de forma directa, y se eleva al cuadrado para obtener el coeficiente de confiabilidad (ecuación 27.6). Este último es, por supuesto, igual al anterior: .71.

Incremento de la confiabilidad

El principio que subyace al incremento de la confiabilidad es el llamado anteriormente principio *maxmincon*, en una forma ligeramente diferente: "Maximizar la varianza de las diferencias individuales y minimizar la varianza del error." La ecuación 27.4 indica con claridad tal principio. A continuación se describe el procedimiento general.

Primero, se escriben sin ambigüedades los reactivos de los instrumentos de medición psicológica o educativa. Un evento ambiguo llega a interpretarse en más de una forma. Un reactivo ambiguo permite que la varianza del error se introduzca silenciosamente, debido a que los individuos pueden interpretar el reactivo de forma diferente. Dichas interpretaciones tienden a ser aleatorias y, por lo tanto, incrementan la varianza del error y disminuyen la confiabilidad.

Segundo, si un instrumento no es lo suficientemente confiable, deben añadirse más reactivos del mismo tipo y calidad, por lo común, aunque no necesariamente, incrementará la confiabilidad en una cantidad predecible. El añadir más reactivos incrementa la posibilidad de que la X_t de cualquier individuo esté cerca de su X_∞. Ello es una cuestión del muestreo de la propiedad del espacio o del reactivo. Con pocos reactivos, puede surgir un error grande por el azar. Con más reactivos puede no ser tan grande. La probabilidad de que se balancee por otro error aleatorio en sentido inverso es mayor cuando hay más reactivos. En síntesis, una mayor cantidad de reactivos incrementa la probabilidad de una medición precisa. (Recuerde que cada X_t es la suma de los valores de los reactivos, para cada individuo.)

En tercer lugar, la especificación de instrucciones claras y estándar tiende a reducir los errores de medición. Siempre se debe tener mucho cuidado al escribir las instrucciones para expresarlas con claridad, ya que las instrucciones ambiguas incrementan la varianza del error. Además, los instrumentos de medición deben aplicarse siempre bajo condiciones estándar, bien controladas y similares. Si las situaciones de aplicación difieren, de nuevo puede introducirse varianza del error. En los campos de la psicología y educación,

una prueba que tiene uniformidad de aplicación y calificación se denomina *prueba estandarizada*. Por lo tanto, las pruebas estandarizadas son aquellas que se han sujetado al rigor de la reducción de la varianza del error.

¿Entonces cómo saber si se han escrito reactivos ambiguos o claros? ¿Cómo saber si los reactivos añadidos para intentar incrementar la confiabilidad son del mismo tipo y calidad? Existe un conjunto de procedimientos estadísticos llamados *análisis de reactivos*, que ayudan a responder tales preguntas. El análisis de reactivos se utiliza para incrementar tanto la confiabilidad como la validez de una prueba, lo cual se logra al evaluar cada reactivo de forma separada para determinar si el reactivo es bueno o pobre. Si el reactivo mide o no lo que se desea que mida es cuestión de validez. La validez se analiza en el capítulo 28. En pruebas donde las respuestas se evalúan como correctas e incorrectas (como las pruebas cognitivas), los reactivos se evalúan en términos de su nivel de dificultad. En pruebas donde no hay respuestas correctas o incorrectas (como las que se encuentran en pruebas afectivas), se utilizaría el índice de acuerdos en lugar de la dificultad. El índice de dificultad es una razón simple del número de personas que responden correctamente el reactivo y el número total de personas que toman la prueba. El índice de acuerdos se calcula como la razón del número de personas que selecciona una respuesta, entre el número total de personas que responden la prueba. Por lo tanto, en esencia, el índice de dificultad y el índice de acuerdos son similares en su cálculo.

$$\text{Dificultad del reactivo} = \frac{\textit{número de personas que responden correctamente un reactivo}}{\textit{número total de personas que toma la prueba}}$$

$$\text{Índice de acuerdos} = \frac{\textit{número de personas que selecciona una respuesta}}{\textit{número total de personas que toma la prueba}}$$

Para el índice de dificultad, a mayor valor, más fácil será el reactivo. Lo anterior indica que más personas respondieron correctamente el reactivo. Reactivos con índices de 0.0 o 1.00 contribuyen muy poco a la prueba, en términos de la información que brindan acerca de las diferencias entre las personas. Cuando cada estudiante responde correctamente casi todos los reactivos en una prueba fácil de matemáticas, esto revela muy poco acerca de la diferencia de las personas en habilidades matemáticas. Por otro lado, una prueba que consista de reactivos demasiado difíciles tampoco revela qué tanto difieren los individuos. No importa cuáles sean sus habilidades, todos los individuos responderán de forma incorrecta esos reactivos. Por regla general, la mayoría de los creadores de pruebas concuerdan en que los mejores reactivos, en términos de dificultad y de acuerdo, son aquellos con valores entre .5 y .7. Algunos recomiendan combinar reactivos de diferentes niveles de dificultad, pero que tengan un índice general entre .5 y .7.

Después de la dificultad y del acuerdo, el siguiente índice para el análisis de reactivos es el *índice de discriminación de reactivos*. Dicho estadístico es el que indicará al investigador (en pruebas cognitivas) qué tan efectivamente el reactivo fue capaz de discriminar entre puntuaciones altas y puntuaciones bajas. Se considera un buen reactivo a aquel que es contestado correctamente por las personas con alta puntuación, y contestado erróneamente por aquellos con baja puntuación. Cuando así sucede, el reactivo tiene la discriminación máxima. El índice de discriminación de reactivos funciona mejor para pruebas cognitivas, las cuales son pruebas que tienen respuestas correctas e incorrectas. En pruebas de tipo afectivo (por ejemplo, de personalidad), donde no hay respuestas correctas e

incorrectas, se utiliza la correlación de la puntuación del reactivo con la puntuación total, aunque ésta también puede utilizarse con pruebas cognitivas.

Con el índice de discriminación de los reactivos, el investigador primero determina el grupo con puntuación más alta y el grupo con puntuación más baja. Para hacerlo se utilizan las puntuaciones totales. Es recomendable que los dos grupos sean iguales en términos del número de personas; éste varía dependiendo del número de personas que tomó la prueba. Después se cuenta el número de personas, dentro de cada grupo, que respondieron correctamente el reactivo. Se calcula una puntuación de diferencia entre el número de personas en el grupo de alta puntuación, que respondieron correctamente el reactivo, y el número de personas del grupo de baja puntuación que respondieron correctamente el mismo reactivo. El índice de discriminación del reactivo es la razón de la diferencia y el número de personas en el grupo de alta puntuación. Se podría haber utilizado como denominador de este cálculo el número de personas del grupo de baja puntuación; pero el número debe ser el mismo:

$$\text{Índice de discriminación del reactivo } i = \frac{P_A - P_B}{\text{\# de personas en el grupo de alta puntuación}}$$

donde P_A es el número de personas en el grupo de alta puntuación que respondieron correctamente el reactivo, y P_B es el número de personas del grupo de baja puntuación que respondieron correctamente el mismo reactivo.

Valores de 0.0, 1.0 y –1.0 son raros. Si el índice es negativo, el reactivo posee discriminación invertida. Esto indicaría al investigador que algo anda definitivamente mal con este reactivo. Se espera que los reactivos tengan valores positivos; a mayor valor, mayor discriminación.

En el caso de la correlación del reactivo con la puntuación total, el investigador, en esencia, correlacionaría la puntuación de cada reactivo o respuesta con la puntuación total. La idea aquí es que si el reactivo es parte de un todo —un todo que mide algo que se desea— debe tener una alto valor de correlación con el total. Recuerde, puesto que se espera que los reactivos sean homogéneos, la correlación de cada reactivo con la puntuación total debe ser alta. Un reactivo que tiene una baja correlación con el total se interpreta como un reactivo que está midiendo algo que difiere de aquello que los demás reactivos están midiendo. El reactivo no es homogéneo con los demás reactivos. Con las computadoras de alta velocidad y la disponibilidad de programas estadísticos, un investigador obtiene dichas correlaciones muy fácilmente. Friedenberg (1995) ofrece una presentación muy buena sobre la manera de calcular tales índices.

El análisis de reactivos con el empleo de estos métodos más tradicionales funciona relativamente bien. Sin embargo, existe un nuevo desarrollo caracterizado por mejorías claras respecto a los métodos tradicionales. Este "nuevo elemento" en el análisis de reactivos se denomina *Teoría de Respuesta al Item* o *TRI*. La TRI involucra mucho más matemáticas que el método tradicional. Su meta principal consiste en clasificar la dificultad o acuerdo de los reactivos. A causa de su complejidad matemática, es mejor realizarlo por medio de programas computacionales. Una compañía llamada Assessment Systems Corporation distribuye varios de los programas a través de Lawrence Erlbaum Associates. Este método esencialmente implica el uso de la *curva característica del reactivo* (ítem) con la teoría del *rasgo latente*. En la teoría del rasgo latente se asume que el desempeño de la prueba puede ser explicado por la posición de quien toma la prueba, sobre una característica hipotética e inobservable (por ejemplo, un rasgo). No se implica que el rasgo cause el comportamiento ni que dicho rasgo exista física o fisiológicamente. Los rasgos latentes son meros constructos estadísticos creados a partir de datos empíricos. La medición básica utilizada en la TRI es

una probabilidad. Es la probabilidad de que una persona con una habilidad específica o rasgo latente responda correctamente un reactivo, con un nivel específico de dificultad. Con reactivos que no se califican como correctos e incorrectos, la TRI aun puede calcular la probabilidad de que una persona con cierta característica dé una respuesta específica, basada en los acuerdos de tal reactivo.

La curva característica del reactivo es una gráfica de la relación entre la puntuación que obtiene en la prueba la persona que la toma y el desempeño en un reactivo en particular. La puntuación de la prueba, por supuesto, mide qué cantidad del atributo o rasgo tiene el individuo. El desempeño en un reactivo en particular por lo común se expresa en forma de probabilidad o proporción. Los mejores reactivos tenderán a exhibir un patrón donde aquellos con altas puntuaciones tiendan a responder correctamente el reactivo; mientras que aquellos con puntuaciones bajas tiendan a responder incorrectamente el mismo reactivo. A mayor pendiente de la curva, de las puntuaciones bajas hacia las puntuaciones altas (pendiente positiva), mayor será el poder discriminativo de ese reactivo. Los reactivos con discriminación negativa tienen una pendiente negativa y tienen un problema que requiere mayor análisis. La curva característica del reactivo también puede ofrecer una medida de la dificultad del reactivo. Al tomar el nivel .50 de probabilidad o proporción y encontrar la puntuación total correspondiente de la prueba para ese nivel, esta puntuación total puede utilizarse como medida de la dificultad. La puntuación total de la prueba correspondería al punto donde el 50% de quienes tomaron la prueba respondieron correctamente el reactivo. Esto difiere ligeramente del índice de dificultad del reactivo que se analizó antes; pero es tan útil como él. Por medio del uso del ajuste matemático y estadístico de la curva, un investigador obtiene índices de discriminación y dificultad de las curvas características de los reactivos. El ajuste de la curva no lineal utilizado en estos procedimientos va más allá del alcance de este libro. Se refiere al lector a estupendas obras que tratan el tema: Allen y Yen (1979), Baker (1992), Crocker y Algina (1986) y Wright y Stone (1979).

El valor de la confiabilidad

Para ser interpretable, una prueba debe ser confiable. A menos que se pueda depender de los resultados de la medición de las propias variables, no es posible determinar, con alguna confianza, las relaciones entre las variables. Puesto que la medición no confiable es medición sobrecargada de error, la determinación de relaciones se convierte en una tarea difícil y poco convincente. ¿Es bajo un coeficiente de correlación obtenido entre dos variables, debido a que una o ambas medidas no sean confiables? ¿Una razón *F* del análisis de varianza es no significativa debido a que la relación hipotetizada no existe, o debido a que la medida de la variable dependiente no es confiable?

La confiabilidad, aunque no es el aspecto más importante de la medición, es bastante importante. En cierto sentido, esto es como el problema del dinero: su ausencia constituye el verdadero problema. Una confiabilidad alta no es garantía de buenos resultados científicos; pero no puede haber buenos resultados científicos sin confiabilidad. En resumen, la confiabilidad es una condición necesaria, pero no suficiente, del valor de los resultados de la investigación y su interpretación.

En este punto es necesario plantear la pregunta: ¿qué tan alto se requiere que sea el coeficiente de confiabilidad? No existe una respuesta rápida y rigurosa a esta pregunta. Por alguna razón, diversos investigadores han establecido .70 como el límite entre confiabilidades aceptables y no aceptables; sin embargo, no existe ninguna evidencia para apoyar esta regla arbitraria. De hecho, la mayoría de los autores de los libros de texto (sobre medición) no establecen dicho valor. Anastasi y Urbina (1997), por ejemplo, no

mencionan tal regla. Nunnally (1978) afirma que un nivel satisfactorio de confiabilidad depende de cómo se utilice la medida. En algunos casos un valor de confiabilidad de .50 o .60 es aceptable; mientras que en otras un valor de .90 es apenas aceptable. Un valor bajo de confiabilidad puede ser aceptable si el instrumento de medición posee una validez alta. Gronlund (1985) señala que la mayoría de las pruebas realizadas por maestros poseen confiabilidades de entre .60 y .85, y aun así son útiles en decisiones instruccionales. Gronlund también brinda consideraciones que deben tenerse al decidir si un valor de confiabilidad es aceptable. Todas las consideraciones se centran en qué tipo de decisión se toma al utilizar la prueba o el instrumento de medición. Si la decisión tomada por medio de la prueba es importante, final, irreversible, inconfirmable, concierne a individuos o tiene consecuencias duraderas, entonces es necesario un alto nivel de confiabilidad. Si la decisión tiene poca importancia, tomada en una etapa temprana, reversible, confirmable por medio de otros datos, concierne a grupos o tiene efectos temporales, entonces es aceptable un valor bajo de confiabilidad.

RESUMEN DEL CAPÍTULO

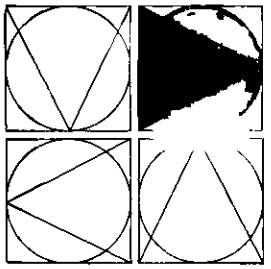
1. Este capítulo examina principalmente la teoría clásica de la confiabilidad. También contempla algunos de los desarrollos "más novedosos" en esta área.
2. La confiabilidad se define como la consistencia o estabilidad del instrumento de medición.
3. La teoría clásica de las pruebas creó la ecuación: $X_t = X_\infty + X_e$, donde X_t es la puntuación observada, X_∞ es la puntuación verdadera y X_e es la puntuación de error.
4. La confiabilidad y la validez se confunden a menudo debido a que ambas tratan con la precisión de las mediciones. No obstante, la confiabilidad está poco relacionada con el hecho de si el instrumento realmente mide lo que se desea. Su aspecto de precisión se refiere a la medición de la puntuación "verdadera".
5. Una medición puede ser confiable e inválida al mismo tiempo. El instrumento de medición puede medir algo de forma imprecisa todo el tiempo.
6. El índice de confiabilidad es de interés; es la correlación entre las puntuaciones verdaderas y las puntuaciones observadas. Sin embargo, las puntuaciones verdaderas no son observables.
7. El coeficiente de confiabilidad es el cuadrado del índice de confiabilidad.
8. Métodos prácticos para obtener el coeficiente de confiabilidad son:

   test-retest, formas paralelas, consistencia interna

9. La consistencia interna puede obtenerse a través de uno de los siguientes métodos: por mitades, por las fórmulas 20 y 21 de Kuder-Richardson, y por el coeficiente alfa de Cronbach.
10. El error estándar de medición indica qué cantidad de error hay en el coeficiente de confiabilidad.
11. Para incrementar la confiabilidad se pueden escribir mejores reactivos, añadir más reactivos similares y estandarizar la administración y la calificación del instrumento de medición y las respuestas.
12. El análisis de reactivos brinda información sobre qué tan buenos o qué tan pobres son los reactivos dentro del instrumento de medición.
13. Qué tan alto debe ser el coeficiente de confiabilidad, para ser aceptable, depende del tipo de decisión a tomar y de las condiciones bajo las cuales se determinó el coeficiente.

SUGERENCIAS DE ESTUDIO

1. ¿En qué difiere la teoría de la generalización de la teoría clásica de las pruebas?
2. De las siguientes, ¿cuál considera usted que es más útil para los investigadores:? ***a)*** validez, o ***b)*** confiabilidad. Justifique su elección.
3. Describa algunos de los problemas con ***a)*** la confiabilidad test-retest y ***b)*** las formas paralelas de confiabilidad. Señale un ejemplo donde usted usaría y no usaría cada una de éstas.
4. Dadas las siguientes situaciones enlistadas abajo, ¿cuál coeficiente de confiabilidad sería el más adecuado para cada una?
   a) Una prueba de mecanografía aplicada a un grupo de alumnos en un curso sobre el uso del procesador de palabras
   b) Una lista de problemas psicológicos utilizada por terapeutas
   c) Una prueba cognitiva de rendimiento
   d) Una prueba de ortografía con palabras de cuatro letras
   e) El número de actos "agresivos" de un chimpancé macho en un zoológico, durante el mismo periodo diario de 10 minutos
   f) Después de que un grupo de estudiantes completó una prueba, ésta se dividió en dos partes y se calcularon las puntuaciones separadas para cada estudiante: la correlación de las dos puntuaciones fue de 0.79
5. ¿Cuántos componentes diferentes puede encontrar, que fueran parte del término de error, en la ecuación de la teoría clásica de las pruebas: $X_t = X_\infty + X_e$?
6. Ofrezca una explicación referente a por qué una puntuación o medida "verdadera" nunca puede alcanzarse.
7. La confiabilidad por mitades de una prueba es de .70. ¿Cuál es la confiabilidad estimada de la prueba completa?
8. Si una confiabilidad test-retest de una prueba con 50 reactivos es de .65, ¿cuál sería la confiabilidad estimada si se añadieran 50 reactivos similares a la prueba?



CAPÍTULO 28

VALIDEZ

El tema de la validez es complejo, controvertido y especialmente importante en la investigación del comportamiento. Aquí, quizás más que en cualquier otra parte, se cuestiona la naturaleza de la realidad. Sin embargo, no es posible estudiar la validez sin investigar, tarde o temprano, el significado de las variables. Sin embargo, no es posible estudiar la validez sin tarde o temprano investigar sobre la naturaleza y el significado de las propias variables.

Cuando se miden ciertas propiedades físicas y atributos relativamente simples de personas, la validez no representa un gran problema. Más bien existe, con frecuencia, congruencia cercana y directa entre la naturaleza del objeto que se mide y el instrumento de medición. Por ejemplo, la longitud de un objeto puede medirse colocando palos marcados con un sistema numérico estándar (pies o metros) sobre el objeto. El peso es más indirecto, pero no difícil: un objeto ubicado en un contenedor desplaza al contenedor hacia aba-

jo. El movimiento del contenedor hacia abajo se registra sobre un índice calibrado (libras u onzas). Por lo tanto, con ciertos atributos físicos existe poca duda de aquello que se está midiendo.

Por otro lado, suponga que un científico educativo desea estudiar la relación entre inteligencia y rendimiento escolar o la relación entre autoritarismo y estilo de enseñanza. Ahora no existen reglas que utilizar, ni escalas con que medir el grado de autoritarismo, ni atributos físicos o de comportamiento claros que indiquen, sin lugar a dudas, el estilo de enseñanza. En tales casos es necesario inventar formas indirectas para medir propiedades psicológicas y educativas. Estas formas son, en ocasiones, tan indirectas que la validez de la medición y sus productos se vuelven dudosos.

Tipos de validez

La definición más común de validez se sintetiza en la pregunta: ¿estamos midiendo lo que creemos que estamos midiendo? El énfasis en esta pregunta está en lo que se mide. Por ejemplo, un maestro ha construido una prueba para medir la *comprensión* de los procedimientos científicos y ha incluido en la prueba sólo reactivos *factuales* sobre procedimientos científicos. La prueba no es válida ya que aunque quizás mida de manera confiable el *conocimiento factual* de los alumnos sobre los procedimientos científicos, no mide su *comprensión* de dichos procedimientos. En otras palabras, quizás mida bastante bien aquello que mide; pero no mide lo que el maestro en realidad intentaba medir.

Aunque la definición más común de validez fue expresada antes, debe enfatizarse de inmediato que no existe una validez única. Una prueba o escala es válida de acuerdo con el propósito científico o práctico de quien la utiliza. Los educadores pueden estar interesados en la *naturaleza* del rendimiento en matemáticas de los alumnos de preparatoria. Entonces ellos estarían interesados en *lo que* mide una prueba de rendimiento o aptitud matemática. Por ejemplo, ellos podrían querer conocer los factores involucrados en el desempeño de una prueba de matemáticas y sus contribuciones relativas a este desempeño. Por otro lado, podrían estar interesados en conocer a los alumnos que probablemente tendrán éxito y a aquellos que probablemente no lo tendrán en las matemáticas de preparatoria. Quizás tengan poco interés en *lo que* mide una prueba de aptitud matemática, y estén interesados ante todo en una *predicción* exitosa. En estos dos usos de las pruebas están implicados diferentes tipos de validez. Ahora se examinará un desarrollo extremadamente importante en la teoría de las pruebas: el análisis y el estudio de los diferentes tipos de validez. Aunque existan varios tipos, el investigador debe diseñar el estudio de validación sólo con un tipo de validez en mente. Algunos investigadores calculan todos los coeficientes de validez sólo para descubrir que cada uno adquiere un valor diferente.

La clasificación más importante de los tipos de validez es la que creó un comité conjunto de la Asociación Psicológica Americana, la Asociación Americana de Investigación Educativa y el Consejo Nacional de Mediciones utilizadas en Educación. Se incluyen tres tipos de validez: de *contenido*, *relacionada con el criterio* y de *constructo*. Cada una de éstas se examinará de forma breve, aunque se pondrá un mayor énfasis en la validez de constructo, ya que tal vez sea la forma más importante de validez, desde el punto de vista de la investigación científica.

Validez de contenido y validación de contenido

Una profesora universitaria de psicología ha impartido un curso para estudiantes del último año, donde enfatizó la comprensión de los principios del desarrollo humano. Ella prepara

una prueba de tipo objetivo. Al querer conocer su validez, examina críticamente la relevancia de cada uno de los reactivos de la prueba, para entender los principios del desarrollo humano. Además les pide a dos colegas que evalúen el contenido de la prueba. Naturalmente, les informa a sus colegas lo que está tratando de medir. Ella está investigando la *validez de contenido* de la prueba.

La validez de contenido es la *representatividad* o la *adecuación de muestreo* del contenido —la sustancia, la materia, el tema— de un instrumento de medición. La *validación de contenido* está guiada por la pregunta: ¿la sustancia o contenido de esta medida es representativa del contenido del universo de contenido de la propiedad que se mide? Cualquier propiedad psicológica o educativa posee un universo teórico de contenido, que consiste en todas las posibles cosas que se dicen u observan acerca de la propiedad. Los miembros de este universo, *U*, pueden denominarse "reactivos". La propiedad puede ser el "rendimiento aritmético", por dar un ejemplo relativamente simple. *U* posee un número infinito de miembros: todos los reactivos posibles utilizando números, operaciones aritméticas y conceptos. Una prueba con alta validez de contenido sería teóricamente una muestra representativa de *U*. Si fuera posible elegir aleatoriamente reactivos de *U* en número suficiente, entonces cualquiera de estas muestras de reactivos supuestamente formaría una prueba con una alta validez de contenido. Si *U* comprende los subconjuntos *A*, *B* y *C*, que son operaciones aritméticas, conceptos aritméticos y manipulaciones numéricas, respectivamente, entonces cualquier muestra de *U* lo suficientemente grande representaría a *A*, *B* y *C* de forma casi igual. La validez de contenido de la prueba sería satisfactoria.

Por desgracia, la mayoría de las veces no es posible elegir muestras aleatorias de reactivos de un universo de contenido; dichos universos sólo existen en teoría. Es verdad que es posible y deseable armar grandes grupos de reactivos, especialmente en el área de rendimiento, y obtener muestras aleatorias a partir de dichos grupos, con propósitos de prueba. Pero la validez de contenido de dichos grupos está siempre en duda, no importa qué tan abundantes y qué tan "buenos" sean los reactivos.

Si no es posible satisfacer la definición de validez de contenido, ¿cómo puede lograrse un nivel razonable de validez de contenido? La validación de contenido consiste esencialmente de juicio. Solo o con otros, el investigador juzga la representatividad de los reactivos. Se puede plantear la pregunta: ¿este reactivo mide la propiedad *M*? Expresado de manera más completa, se podría plantear la pregunta: ¿este reactivo es representativo del universo de contenido de *M*? Si *U* tiene subconjuntos, tales como los que se indicaron antes, entonces se deben plantear preguntas adicionales; por ejemplo: ¿este reactivo es miembro del subconjunto M_1 o del subconjunto M_2?

Algunos universos de contenido son más obvios y más fáciles de juzgar que otros; el contenido de muchas pruebas de rendimiento, por ejemplo, parecería obvio. Se dice que puede suponerse la validez de contenido de tales pruebas. Mientras que esta afirmación parece razonable, y mientras el contenido de la mayoría de las pruebas de rendimiento está "autovalidado" en el sentido de que, hasta cierto grado, el individuo que escribe la prueba define la propiedad que se está midiendo (por ejemplo, un maestro que escribe una prueba de ortografía o aritmética para la clase), es peligroso asumir la adecuación de la validez de contenido sin realizar esfuerzos sistemáticos para verificar el supuesto. Por ejemplo, un investigador educativo que comprueba hipótesis acerca de las relaciones entre el rendimiento en estudios sociales y otras variables, puede suponer la validez de contenido de una prueba de estudios sociales. Sin embargo, la teoría a partir de la cual se derivaron las hipótesis quizá requiera *comprensión* y *aplicación* de ideas de estudios sociales; mientras que la prueba utilizada puede tener un contenido casi puramente factual. La prueba carece de validez de contenido en su propósito. De hecho, el investigador no está comprobando en realidad las hipótesis establecidas.

Entonces, la validación de contenido es básicamente de juicio. Los reactivos de una prueba deben estudiarse y se debe ponderar la representatividad supuesta de cada reactivo en el universo, lo cual quiere decir que cada reactivo debe juzgarse respecto a su supuesta relevancia respecto a la propiedad que se mide; no es una tarea fácil. Por lo común, otros jueces "competentes" deben juzgar el contenido de los reactivos. De ser posible, el universo de contenido debe estar claramente definido; es decir, se les deben facilitar a los jueces instrucciones específicas para realizar juicios, así como las especificaciones sobre lo que están juzgando. Después, es posible utilizar algún método para agrupar los juicios independientes. Una excelente guía para la validez de contenido de pruebas de rendimiento es Bloom (1956), quien representa un intento exhaustivo por determinar y discutir objetivos educativos en relación con la medición. El trabajo de Bloom se denominó "taxonomía de Bloom".

Existe otro tipo de validez que es muy similar a la validez de contenido. Ésta se llama *validez aparente* o *de facie*, la cual no es una validez en el sentido técnico; se refiere a aquello que la prueba aparenta medir. Individuos entrenados o sin entrenamiento observarían la prueba y decidirían si ésta mide lo que se supone que debe medir. No se calcula la cuantificación del juicio ni tampoco un índice del acuerdo entre jueces. La validez de contenido es cuantificable a través del empleo de índices de concordancia de las evaluaciones de los jueces. Uno de dichos índices es la *Kappa* de Cohen (Cohen, 1960).

Validez relacionada con el criterio y validación

Como su burdo y desafortunado nombre lo indica, la *validez relacionada con el criterio* se estudia al comparar las puntuaciones de una prueba o escala con una o más variables externas, o criterios, que se sabe o se considera que miden el atributo que se estudia. Un tipo de validez relacionada con el criterio es la llamada *validez predictiva*. El otro tipo es la *validez concurrente*, que difiere de la predictiva en la dimensión del tiempo. La validez predictiva involucra el uso de desempeños del criterio futuros; mientras que la validez concurrente mide el criterio casi al mismo tiempo. En este sentido, la prueba sirve para evaluar el estatus presente del individuo.

La validez concurrente con frecuencia se utiliza para validar una prueba nueva. Para cada examinado se toman por lo menos dos medidas concurrentes. Una de ellas sería la prueba nueva y la otra sería una prueba o medida existente. La validez concurrente se calcularía al correlacionar los dos conjuntos de calificaciones. En el área de las pruebas de inteligencia, las pruebas nuevas e inclusive las revisiones de pruebas antiguas, se utiliza generalmente la prueba de Stanford-Binet o la prueba de Wechsler como criterio concurrente.

Cuando se predice el éxito o fracaso de los estudiantes a partir de sus medidas de aptitud académica, se está considerando la validez predictiva relacionada con el criterio. ¿Qué tan bien predice la prueba (o pruebas) el promedio final o el de la licenciatura? Aquí el enfoque no es tanto lo que la prueba mide, sino su habilidad predictiva. De hecho, en la validación relacionada con el criterio, la cual es con frecuencia investigación práctica y aplicada, el interés básico está más centrado en el criterio, es decir, en los resultados prácticos, que en los predictores. (En la investigación básica esto no es así.) A mayor correlación entre una medida o medidas de aptitud académica y el criterio, por ejemplo la calificación promedio, mejor será la validez. Breve y nuevamente, se enfatiza el criterio y su predicción. Thorndike (1996) ofrece un análisis sobre lo que constituye un buen criterio.

El término *predicción* está generalmente asociado con el futuro. Esto es desafortunado ya que, en la ciencia, predicción no necesariamente significa pronóstico. Se "predice" una

variable dependiente a partir de una variable independiente. Se "predice" la existencia o no-existencia de una relación; ¡incluso se "predice" algo que sucedió en el pasado! Este amplio significado de predicción es el que se utiliza aquí. En cualquier caso, la validez relacionada con el criterio está caracterizada por la predicción sobre un criterio *externo* y por la verificación de un instrumento de medición, ya sea ahora o en el futuro, contra un resultado o medida. En cierto sentido todas las pruebas son predictivas, pues "predicen" cierto tipo de resultado, una situación presente o futura. Las pruebas de aptitud predicen el rendimiento futuro; las pruebas de rendimiento, el rendimiento y competencia presentes y futuras, y las pruebas de inteligencia, la habilidad presente y futura para aprender y resolver problemas. Aun cuando se mide el autoconcepto, se predice que si la puntuación del autoconcepto es tal, entonces el individuo será de tal o cual manera ahora y en el futuro.

La mayor dificultad de la validación relacionada con el criterio es el criterio mismo. Obtener un criterio puede ser incluso difícil. ¿Qué criterio puede utilizarse para validar una medida de eficacia de un profesor? ¿Quién debe juzgar la eficacia de un profesor? ¿Qué criterio puede utilizarse para probar la validez predictiva de una prueba de aptitud musical?

Aspectos de decisión de la validez

Como se indicó antes, la validez relacionada con el criterio está asociada generalmente con resultados y problemas prácticos. El interés no se centra tanto en lo que está detrás del desempeño en la prueba, sino en su utilidad para resolver problemas prácticos y tomar decisiones. Se utilizan cientos de pruebas con los propósitos predictivos de evaluar y seleccionar candidatos potencialmente exitosos en educación, negocios y otras ocupaciones. ¿Ayuda materialmente una prueba o un conjunto de pruebas para decidir sobre la asignación de individuos a empleos, clases, escuelas y otros aspectos similares? Cualquier decisión implica una elección entre tratamientos, asignaciones o programas. Cronbach (1971) señala que para tomar una decisión, se predice el éxito de la persona bajo cada tratamiento y luego se utiliza alguna regla para traducir la predicción en una tarea o recomendación. Una prueba con alta validez relacionada con el criterio ayuda a los investigadores a tomar decisiones exitosas al asignar personas a tratamientos, considerando *tratamientos* en un sentido amplio. Un comité o jefe de admisiones decide si admite o no en la universidad a un solicitante, con base en una prueba de aptitud académica. En efecto, tal uso de las pruebas es bastante importante, y la validez predictiva de las pruebas también tiene gran importancia. Se recomienda el lector al ensayo de Cronbach para una buena exposición de los aspectos de toma de decisión de pruebas y validez.

Taylor y Russell (1939) realizaron una gran contribución en esta área, pues demostraron que las pruebas con poca validez aun pueden utilizarse de manera efectiva con propósito de decisiones. Desarrollaron la tabla Taylor-Russell, que utiliza tres piezas de información: el coeficiente de validez, la tasa de selección y la tasa base. La tasa de selección se refiere al número de personas (solicitantes) que se elegirán del número total de personas. Si hubiera sólo 10 plazas y 100 solicitantes, la tasa de selección sería .10 o 10%. La tasa base es la proporción de personas en la población con ciertas características. Estos datos por lo general se reportan en la prensa. La tasa base de mujeres es, por ejemplo, .52 o 52% de la población de Estados Unidos. Sin utilizar una prueba, si se reúnen aleatoriamente 100 personas en un cuarto, 52 de ellas serían mujeres. Cada uno de los tres componentes puede variar y el hacerlo tiene un efecto sobre la precisión de la selección. Es decir, ayuda a tomar una mejor decisión. Anastasi y Urbina (1997) ofrecen una buena explicación sobre la forma en que funciona este método. El lector interesado necesitará

consultar el artículo original de Taylor y Russell para ver el rango completo de tablas. En esencia, es posible realizar una mejor predicción utilizando una prueba con poca validez si la tasa de selección es pequeña. Desde 1939 este método ha sufrido algunas modificaciones y adiciones, entre las que se incluyen las de Abrahams, Alf y Wolfe (1971); Pritchard y Kazar (1979) y Thomas, Owen y Gunst (1977).

Predictores y criterios múltiples

Se utilizan tanto los predictores múltiples como los criterios múltiples. Más adelante, cuando se estudie la regresión múltiple, se enfocarán los predictores múltiples y la manera de manejarlos estadísticamente. Los criterios múltiples pueden manejarse de forma separada o juntos, aunque esto último no es fácil. En la investigación práctica por lo común debe tomarse una decisión. Si existe más de un criterio, ¿cómo se pueden combinar mejor para tomar una decisión? Por supuesto, debe considerarse la importancia relativa de los criterios. ¿Se desea un administrador con alta habilidad en solución de problemas, con alta habilidad en relaciones públicas o con ambas? ¿Cuál es más importante para un trabajo en particular? Es altamente probable que se haga común el uso tanto de los predictores múltiples como de los criterios múltiples, conforme se comprendan mejor los métodos multivariados y se utilice rutinariamente la computadora en la investigación predictiva.

Validez de constructo y validación de constructo

La validez de constructo es uno de los avances científicos más significativos de la teoría y de la práctica de la medición moderna. Representa un avance significativo ya que liga conceptos y prácticas psicométricos con conceptos teóricos. El trabajo clásico en el área es el de Cronbach y Meehl (1955). Cuando los expertos en medición investigan la validez de constructo de una prueba, casi siempre desean saber qué propiedad o propiedades psicológicas o de otro tipo pueden "explicar" la varianza de las pruebas. Buscan conocer el "significado" de las pruebas. Si se trata de una prueba de inteligencia, ellos desean saber qué factores subyacen al desempeño en la prueba. Plantean la pregunta: ¿qué factores o constructos explican la varianza del desempeño en la prueba? ¿Esta prueba mide habilidad verbal y habilidad de razonamiento abstracto? ¿"Mide" también la pertenencia a una clase social? Ellos preguntan, por ejemplo, qué proporción de la varianza total de la prueba es explicada por cada uno de los constructos como habilidad verbal, habilidad de razonamiento abstracto y pertenencia a una clase social. En síntesis, buscan explicar las diferencias individuales en las puntuaciones de la prueba. Su interés por lo general está centrado en las propiedades que se miden, más que en las pruebas utilizadas para lograr la medición.

Los investigadores por lo común inician con los constructos o variables que tienen relación. Suponga que un investigador ha descubierto una correlación positiva entre dos medidas: una de tradicionalismo educativo y la otra sobre la percepción de las características asociadas con un "buen" profesor. Los individuos con puntuaciones altas en la medida de tradicionalismo ven al "buen" profesor como eficiente, moral, minucioso, industrioso, concienzudo y confiable. Los individuos con puntuaciones bajas en la medida de tradicionalismo quizá vean al "buen" profesor de una forma diferente. El investigador ahora desea saber *por qué* existe dicha relación, es decir, lo que está detrás de ella. Para lograr esto, debe estudiarse el significado de los constructos incluidos en la relación: "percepción del 'buen' maestro" y "tradicionalismo". La manera de estudiar estos significados implica un problema de validez de constructo. Este ejemplo fue tomado de Kerlinger y Pedhazur (1968).

Se puede ver que la validación de constructo y la investigación científica empírica están íntimamente relacionadas. No es simplemente cuestión de validación de una prueba.

Debe intentarse validar la teoría que está detrás de la prueba. Cronbach (1990) indica que existen tres partes en la validación de constructo: sugerir qué constructos posiblemente explican el desempeño en la prueba, derivar hipótesis a partir de la teoría que incluye al constructo y comprobar empíricamente las hipótesis. Tal planteamiento es una precisión del modelo científico general analizado en capítulos anteriores.

El aspecto más importante sobre la validez de constructo que la separan de otros tipos de validez es su preocupación por la teoría, los constructos teóricos y la investigación científica empírica, incluyendo la comprobación de relaciones hipotetizadas. La validación de constructo en medición contrasta en forma notable con modelos que definen la validez de una medida, principalmente por su éxito al predecir el criterio. Por ejemplo, un aplicador de pruebas puramente empírico podría decir que una prueba es válida si distingue de manera eficiente entre individuos con altos o bajos niveles de cierto rasgo. El *porqué* de que la prueba sea exitosa al separar los subconjuntos de un grupo no tiene gran importancia. Es suficiente con que lo haga.

Convergencia y discriminación

Observe que la comprobación de hipótesis alternativas es particularmente importante en la validación de constructo, ya que se requiere tanto de la convergencia como de la discriminación. *Convergencia* significa que la evidencia de diferentes fuentes, reunida de diferentes maneras, indica un significado similar o igual al del constructo. Diferentes métodos de medición deben convergir en el constructo. La evidencia producida al aplicar el instrumento de medición a diferentes grupos en diferentes lugares debe producir significados similares o, si no es así, entonces debe explicar las diferencias. Por ejemplo, una medida del autoconcepto de niños debe ser capaz de ofrecer interpretaciones similares en distintas partes del país. Si no es capaz de ofrecer dichas interpretaciones en cierta localidad, entonces la teoría debe ser capaz de explicar por qué —de hecho debe predecir tal diferencia—.

Discriminación significa que es posible diferenciar empíricamente el constructo de otros constructos que puedan ser similares, y que se puede señalar lo que *no está relacionado* con el constructo. En otras palabras, se señala qué otras variables están correlacionadas con el constructo y de qué manera lo están. Sin embargo, también se indica cuáles variables no deben estar correlacionadas con el constructo. Por ejemplo, se señala que una escala para medir el *conservadurismo* debe correlacionarse sustancialmente, y de hecho lo hace, con medidas de *autoritarismo* y *rigidez* —la teoría predice esto— pero no se correlaciona con medidas de *aceptación social* (véase Kerlinger, 1970). A continuación se ejemplificarán estas ideas.

Un ejemplo hipotético de validación de constructo

Suponga que un investigador está interesado en los determinantes de la creatividad y la relación de la creatividad con el rendimiento escolar. El investigador nota que las personas más sociables, quienes muestran afecto hacia otros, también parecen ser menos creativos que aquellos que son menos sociables y afectuosos. El objetivo consiste en probar la relación implicada de una manera controlada. Una de las primeras tareas es obtener o construir una medida de la característica sociable-afectuoso. El investigador, conjeturando que esta combinación de rasgos quizá sea un reflejo de un interés más profundo en el amor por los demás, lo llama *amorismo*. Se asume que existen diferencias individuales respecto al amorismo, es decir, algunas personas lo poseen en gran cantidad, otras en cantidad moderada y otras muy poco.

El primer paso es construir un instrumento para medir el amorismo. La literatura ofrece poca ayuda, puesto que los psicólogos científicos han estudiado muy poco la naturaleza fundamental del amor. No obstante, se ha medido la sociabilidad. El investigador debe construir un nuevo instrumento, basando su contenido en conceptos intuitivos y racionales sobre lo que es el amorismo. La confiabilidad de la prueba, que fue probada con grupos grandes, oscila entre .75 y .85.

La pregunta ahora es si la prueba es o no válida. El investigador correlaciona el instrumento y lo llama escala *A*, con las medidas independientes de sociabilidad. Las correlaciones son moderadamente altas, pero se necesita mayor evidencia para afirmar que la prueba posee validez de constructo. Se deducen ciertas relaciones que deben existir o no entre el amorismo y otras variables. Si el amorismo es la tendencia general de amar a los demás, entonces debe correlacionarse con características tales como ser cooperativo y amistoso. Las personas con alto amorismo enfrentarán los problemas de una forma orientada al *yo*; en contraste con las personas con bajo amorismo, quienes enfrentarán los problemas de una forma orientada a la tarea.

Con base en este razonamiento, el investigador aplica la escala *A* y una escala para medir subjetividad a un grupo de estudiantes del primer año de preparatoria. Para medir el nivel de cooperación se realiza una observación del comportamiento del mismo grupo de estudiantes en el salón de clase. Las correlaciones entre las tres medidas son positivas y altas. Observe que no se esperaría una correlación alta entre las medidas. Si las correlaciones fueran demasiado altas, entonces se dudaría con respecto a la validez de la escala *A*; quizás estaría midiendo subjetividad o nivel de cooperación, pero no amorismo.

Debido a que conoce las desventajas de la medición psicológica, el investigador no está satisfecho. Estas correlaciones positivas tal vez se deban a un factor común a las tres pruebas, pero irrelevante para el amorismo; por ejemplo, la tendencia a dar respuestas "correctas" o deseables. (Sin embargo, esto podría descartarse a causa de que la medida de observación del cooperativismo se correlaciona positivamente con el amorismo y la subjetividad.) Por lo tanto, con un nuevo grupo de participantes, el investigador aplica las escalas de amorismo y subjetividad, evalúa la conducta de cooperativismo de los participantes y, además, aplica una prueba de creatividad que demostró ser confiable en otra investigación.

El investigador establece la relación entre amorismo y creatividad en la forma de una hipótesis: la relación entre la escala *A* y la medida de creatividad será negativa y significativa. Las correlaciones entre amorismo y cooperativismo, y entre amorismo y subjetividad serán positivas y significativas. También se formulan hipótesis de "verificación": la correlación entre cooperativismo y creatividad no será significativa, será cercana a cero; pero la correlación entre subjetividad y creatividad será positiva y significativa. Esta última relación se predice con base en hallazgos previos de investigación. Los seis coeficientes de correlación se presentan en la matriz de correlación de la tabla 28.1. Las cuatro medidas se denominan de la siguiente forma: A, amorismo; B, cooperativismo; C, subjetividad, y D, creatividad.

La evidencia de la validez de constructo de la escala *A* es buena. Todas las *r* resultaron tal como se predijo; de especial importancia son las *r* entre D (creatividad) y las otras variables. Note que hay tres tipos diferentes de predicción: positiva, negativa y cero; las tres resultaron tal como se predijo. Lo anterior ilustra lo que se llamaría *predicción diferencial* o *validez diferencial* —o discriminación—. No es suficiente predecir, por ejemplo, que la medida que se supone refleja la propiedad estudiada esté correlacionada en forma positiva con una variable teóricamente relevante. Se debería, deduciendo a partir de la teoría, predecir más de una de dichas relaciones positivas. Además, deberían predecirse relaciones de cero entre la variable principal y las variables "irrelevantes" con la teoría. En el

TABLA 28.1 *Intercorrelaciones de cuatro medidas hipotéticas (N = 90)*[a]

	B	C	D
A	.50	.60	–.30
B		.40	.05
C			.50

[a] A = Amorismo; B = Cooperativismo; C = Subjetividad; D = Creatividad. Los coeficientes de correlación de .25 o mayores son significativos al nivel .01.

ejemplo anterior, aunque se esperaba que el cooperativismo se correlacionara con el amorismo, no hubo una razón teórica para esperar que se correlacionara en lo absoluto con la creatividad.

Un ejemplo de diferente tipo es el investigador que introduce deliberadamente una medida que invalidaría otras relaciones positivas, si dicha variable se correlaciona con la variable cuya validez se estudia. Un gran problema de las escalas de personalidad y de actitud es el fenómeno que involucra el deseo de ser aceptado socialmente, que se mencionó antes. La correlación entre la variable estudiada y una variable teóricamente relacionada tal vez se deba a que ambos instrumentos estén midiendo el deseo de aceptación social más que las variables para las que fueron diseñados. Dicha tendencia se verifica, en parte, si se incluye una medida del deseo de aceptación social junto con otras medidas.

A pesar de que todas las evidencias conduzcan al investigador a creer que la escala *A* posee validez de constructo, aún pueden existir dudas. Por lo tanto, se realiza un estudio donde los alumnos con alto y bajo nivel de amorismo deben resolver problemas. Se predice que los alumnos con bajo nivel de amorismo resolverán los problemas con más éxito que aquellos con alto amorismo. Si los datos apoyan la predicción, esto representa mayor evidencia de la validez de constructo de la medida de amorismo. Esto es, por supuesto, un hallazgo significativo en sí mismo. No obstante, probablemente un procedimiento como éste sea más apropiado para medidas de rendimiento y de actitud. Por ejemplo, es posible manipular las comunicaciones para cambiar actitudes. Si las puntuaciones de actitud cambian de acuerdo con la predicción teórica, entonces ello sería evidencia de la validez de constructo de la medida de actitud, ya que las puntuaciones quizá no cambiarían de acuerdo con la predicción si la medida no estuviera midiendo el constructo.

El método multirrasgo–multimétodo

Una contribución significativa e influyente de Campbell y Fiske (1959) en la comprobación de la validez es el empleo de las ideas de convergencia y discriminación y de matrices de correlación, para aportar evidencia sobre la validez. Para explicar el método se usarán algunos datos de un estudio sobre actitudes sociales de Kerlinger (1967, 1984). Se ha encontrado que existen dos dimensiones básicas de las actitudes sociales, que corresponden a descripciones filosóficas, sociológicas y políticas del liberalismo y conservadurismo. Se aplicaron dos tipos de escalas diferentes a estudiantes de educación de posgrado y a grupos fuera de las universidades en Nueva York, Texas y Carolina del Norte. Un instrumento, la Escala de Actitudes Sociales, contenía *afirmaciones* usuales de actitud: 13 reactivos liberales y 13 conservadores. El segundo instrumento, Referentes-I o REF-I, utilizaba *referentes* de actitud (palabras y frases cortas: *propiedad privada*, *religión* y *derechos civiles*, por ejemplo) como reactivos, de los cuales 25 eran referentes liberales y 25 eran referentes conservadores.

Las muestras, las escalas y parte de los resultados se describen en Kerlinger (1972). Los datos reportados en la tabla 28.2 fueron obtenidos de una muestra de Texas, $N = 227$ estudiantes de posgrado.

Entonces, se tienen dos tipos de instrumentos de actitud completamente diferentes: uno con reactivos de referencia y el otro con reactivos afirmativos, o método 1 y método 2, respectivamente. Las dos dimensiones básicas medidas fueron el liberalismo (L) y el conservadurismo (C). ¿Miden liberalismo y conservadurismo las subescalas L y C de las dos escalas? Parte de la evidencia se muestra en la tabla 28.2, la cual presenta la correlación entre las cuatro subescalas de los dos instrumentos, así como los coeficientes de confiabilidad de la subescala, calculados a partir de las respuestas a las dos escalas.

En un análisis multirrasgo-multimétodo se utiliza más de un atributo y más de un método en el proceso de validación. Los resultados de correlacionar variables dentro y entre métodos pueden presentarse en la llamada matriz multirrasgo-multimétodo. La matriz presentada en la tabla 28.2 es la forma más simple posible de realizar un análisis de este tipo: dos variables y dos métodos. Por lo común se desearía utilizar más variables.

La parte más importante de la matriz es la diagonal que contiene las correlaciones entre los métodos; en la tabla 28.2 este resultado se ubica en la sección inferior izquierda de la tabla. Los valores diagonales deben ser sustanciales, pues reflejan las magnitudes de las correlaciones entre las mismas variables, medidas de forma distinta. Estos valores, expresados en itálicas en la tabla (.53 y .54) son bastante altos.

En este ejemplo, la teoría exige correlaciones cercanas a cero o correlaciones bajas negativas entre L y C (véase Kerlinger, 1967 para mayor profundidad sobre esto). La correlación entre L_1 y C_1 es –.07 y entre L_2 y C_2 es –.09, lo cual coincide con la teoría. La correlación cruzada entre L y C, es decir, la correlación entre L del método 1 y C del método 2, o entre L_1 y C_2, es –.37, mayor de lo que la teoría predice (se adoptó un límite superior de –.30). Entonces, con excepción de la correlación cruzada de –.37 entre L_1 y C_2, se sostiene la validez de constructo de la escala de actitudes sociales. Por supuesto que se desearía mayor evidencia que los resultados obtenidos con una muestra, y que también se desearía una explicación respecto a la alta correlación negativa de método cruzado entre L_1 y C_2. No obstante, el ejemplo ilustra las ideas básicas del método múltirrasgo-multimétodo para la validez.

Campbell y Fiske (1959) utilizaron terminología específica para describir cada correlación en la tabla. Las correlaciones *monométodo-monorrasgo* son las confiabilidades. Éstas se encuentran en la diagonal principal de la matriz; en la tabla 28.2 son los valores .85, .88, .81 y .82 encerrados en paréntesis. Las correlaciones *heterométodo-monorrasgo* representan

TABLA 28.2 *Correlaciones entre dimensiones de actitudes sociales a través de dos métodos de medición, modelo multirrasgo-multimétodo, muestra de Texas ($N = 227$)*[a]

		Método 1 (Referentes)		*Método 2 (Afirmaciones)*	
		L_1	C_1	L_2	C_2
Método 1	L_1	(.85)			
(Referentes)	C_1	–.07	(.88)		
Método 2	L_2	*.53*	–.15	(.81)	
(Afirmaciones)	C_2	–.37	*.54*	–.09	(.82)

[a] Método 1: referentes; método 2: afirmaciones; L = liberalismo; C = conservadurismo. Las cifras en paréntesis sobre la diagonal son índices de confiabilidad de la consistencia interna; las cifras en itálicas (.53 y .54) son correlaciones del cruce de los métodos L-L y C-C (validez).

la validez que se analizó anteriormente, que son los valores .53 y .54 escritos en itálicas en la tabla 28.2. Existen otros dos tipos de correlación: la *monométodo-heterorrasgo* (los valores −.07 y −.09), y la *heterométodo-heterorrasgo* (que fueron −.37 y −.15). Campbell y Fiske afirman que para obtener evidencia completa de la validez de constructo, las correlaciones deben seguir un patrón establecido. Si no se logran cubrir los requisitos se debilitan los aspectos de la validez. Algunos artículos han intentado resolver este problema al relajar algunos de los requisitos. Tales artículos afirman haber logrado un grado de éxito parcial.

El modelo del método multirrasgo-multimétodo constituye un ideal. Si es posible debe realizarse. En realidad la investigación y la medición de constructos importantes como el conservadurismo, la agresividad, la calidez del profesor, la necesidad de rendimiento, la honestidad, etcétera, finalmente lo requieren. Sin embargo, en muchas situaciones de investigación es difícil o aun imposible aplicar dos o más medidas de dos o más variables con muestras relativamente grandes. Aunque siempre deben realizarse esfuerzos para estudiar la validez, la investigación no debe abandonarse sólo porque no es posible aplicar el método completo.

Ejemplos de investigación de la validación concurrente

Wood (1994) ofrece un buen ejemplo de cómo validar una prueba que utiliza datos médicos y psicológicos. Aquí el criterio es una medición física real. Wood desarrolló un instrumento llamado instrumento de evaluación de la eficiencia del autoexamen de mama (Breast Self-Examination Proficiency Rating Instrument, BSEPRI), el cual mide qué tanto conocimiento tiene quien toma la prueba, respecto al autoexamen de mama. Las participantes en el estudio eran estudiantes de enfermería. A la mitad de ellas se les dieron instrucciones sobre el autoexamen y a la otra mitad no. Una prueba *t* demostró que quienes recibieron instrucciones obtuvieron puntuaciones significativamente mayores que quienes no las recibieron. Wood obtuvo la validez concurrente al correlacionar las puntuaciones de palpación del instrumento con la habilidad de los estudiantes para detectar protuberancias en un modelo de silicón.

Iverson, Guirguis y Green (1998) examinaron la validez concurrente en una forma breve de la escala Wechsler de inteligencia para adultos-revisada (WAIS-R). Esta forma breve consistía de siete escalas. Fue desarrollada para evaluar pacientes con diagnóstico de un trastorno del espectro de la esquizofrenia. Las puntuaciones del CI calculadas por medio de esta forma breve tienen una alta correlación con las puntuaciones del CI de la escala completa. Los CI verbales, los CI de ejecución y los CI de la escala completa, calculados con la forma breve, estaban altamente correlacionados con los CI de la escala completa. Las correlaciones (coeficientes de validez) oscilaron entre .95 y .98. En general, la forma breve de siete subescalas mostró validez concurrente adecuada y sirve para evaluar el funcionamiento intelectual de personas con trastornos psicóticos. Iverson y colaboradores correlacionaron la prueba nueva (forma breve) con la prueba establecida (escala completa) para obtener una medida de validez concurrente. Comrey (1993) utilizó un procedimiento similar para crear la forma breve de las escalas de personalidad de Comrey (Comrey Personality Scales, CPS). Con el uso de datos ya existentes Comrey extrajo los "mejores" reactivos de cada escala (que se analizan más adelante) y calculó dos puntuaciones totales: una de la forma breve y otra de la forma original. La correlación de las dos puntuaciones produjo un valor de validez concurrente.

Ejemplos de investigación de validación de constructo

En cierto sentido, cualquier tipo de validación es validación de constructo. Loevinger (1957) argumenta que la validez de constructo, desde un punto de vista científico, constituye

el total de la validez. En el otro extremo, Bechtoldt (1959) argumenta que la validez de constructo no tiene lugar en la psicología. Horst (1966) dice que es muy difícil aplicar las ideas de Cronbach y Meehl dentro de la teoría lógica y práctica de la psicometría. Sin embargo, cuando se prueban hipótesis y cuando se estudian relaciones empíricamente, se involucra la validez de constructo. Debido a su importancia, ahora se examinarán dos ejemplos de investigación sobre la validación de constructo.

Una medida de antisemitismo

En un intento inusual por validar su medida sobre antisemitismo, Glock y Stark (1966) utilizaron las respuestas a dos frases incompletas respecto a los judíos: "Es una pena que los judíos..." y "No puedo entender por qué los judíos..." Quienes calificaron las frases consideraron lo que cada sujeto había escrito y caracterizaron las respuestas como imágenes negativas, neutrales o positivas sobre los judíos. Entonces, cada sujeto fue considerado individualmente como poseedor de una de tres imágenes diferentes sobre los judíos. Cuando las respuestas al índice de creencias antisemitas (Index of Anti-Semitic Beliefs), la medida que se estaba validando, se dividieron en sin-antisemitismo, antisemitismo medio, antisemitismo medio alto y antisemitismo alto, los porcentajes de respuestas negativas a las dos frases incompletas fueron, respectivamente: 28, 41, 61, 75. Esto representa una buena evidencia de validez, ya que los individuos categorizados desde sin-antisemitismo hasta antisemitismo alto por medio de la medida a ser validada, el índice de creencias antisemitas, respondieron a una medida completamente diferente de antisemitismo, las dos con frases incompletas, de manera congruente con su categorización por medio del índice.

Una medida de personalidad

En un capítulo posterior se discutirá una importante herramienta analítica llamada *análisis factorial*. No obstante, es necesario mencionar este método para la comprensión de la validación de constructo. En años recientes, el análisis factorial parece ser el método de elección para muchas personas involucradas con la validez de constructo. El análisis factorial es esencialmente un método para encontrar aquellas variables que tienen algo en común. Si algunos reactivos de una prueba de personalidad están diseñados para medir extroversión, entonces, en un análisis factorial, dichos reactivos deben cargarse mucho hacia un factor y poco hacia los otros.

A mediados de los años cincuenta, el profesor Andrew L. Comrey, de la Universidad de California en Los Ángeles, realizó la tarea de examinar todas las pruebas de personalidad publicadas reconocidas. Su objetivo inicial era tratar de determinar cuál era la medida correcta (válida) de personalidad. Para esto, el doctor Comrey utilizó un análisis factorial. Contrariamente a sus expectativas iniciales, surgió una nueva prueba de personalidad de carácter único. La prueba de personalidad de Comrey, ahora conocida como las escalas de personalidad de Comrey (Comrey Personality Scales) (CPS), fue de las primeras pruebas desarrolladas por medio del uso del análisis factorial. En 1970, después de un proceso de 15 años de investigación y construcción de la prueba, se publicaron las escalas de personalidad de Comrey (véase Comrey y Lee, 1992 para encontrar un resumen y el procedimiento). El constructo de Comrey de personalidad consta de ocho dimensiones principales:

Confianza contra defensividad
Disciplina contra falta de compulsión
Conformismo social contra rebeldía
Actividad contra falta de energía

Estabilidad emocional contra neuroticismo
Extroversión contra introversión
Masculinidad contra feminidad (renombrados dureza mental contra sensibilidad)
Empatía contra egocentrismo

Desde 1970, Comrey ha publicado diversos artículos que apoyan la validez de sus escalas de personalidad. Esto se hizo al aplicar las CPS, o una forma traducida de las CPS, a diferentes grupos de personas. Después de obtener los datos, cada grupo de éstos fue analizado factorialmente. En cada caso surgieron los mismos ocho factores. Aunque esto no afirma que existan exclusivamente ocho factores de personalidad, los datos lo sustentan. En una investigación reciente realizada por Brief, Comrey y Collins (1994), las CPS fueron traducidas al ruso y aplicadas a 287 participantes hombres y 170 participantes mujeres. Los datos apoyaron seis de las ocho subescalas. Las únicas subescalas que no recibieron suficiente apoyo fueron la de Empatía contra Egocentrismo y la de Actividad contra Falta de Energía.

En un artículo breve, Comrey, Wong y Backer (1978) presentan un procedimiento simple para validar la escala de Conformidad Social contra Rebeldía. En un estudio, Comrey y colaboradores reclutaron a dos grupos de participantes: asiáticos y no-asiáticos. La percepción tradicional de los asiáticos es que son más conformistas socialmente que los no-asiáticos. Existe alguna evidencia que apoya esta afirmación, tal como una fuerte influencia paterna, fuertes valores tradicionales, etcétera. [El estudio de Scattone y Saetermoe (1997) es uno de los que ha demostrado lo anterior.] Por lo tanto, en el estudio de Comrey y colaboradores, la idea establecida respecto a la diferencia entre asiáticos y no-asiáticos sobre conformismo social fue utilizada como el criterio o "medida externa". Todos los participantes respondieron las escalas de personalidad de Comrey, aunque sólo la subescala de Conformismo Social contra Rebeldía era de interés para dicho estudio. Con el uso de una prueba *t*, estos investigadores demostraron una diferencia estadísticamente significativa entre asiáticos y no-asiáticos en la escala de Conformismo Social contra Rebeldía. El estudio podría utilizarse como ejemplo para ilustrar la validez discriminante.

El segundo estudio de este artículo demostró la validez convergente. Se espera que la Conformidad Social esté relacionada con la afiliación y filosofía políticas. Generalmente se piensa que los conservadores son más conformistas socialmente que los liberales, a quienes se considera más rebeldes. En este estudio algunas personas completaron las escalas de personalidad de Comrey y respondieron preguntas respecto a su afiliación política. Comrey y colaboradores encontraron una correlación estadísticamente significativa entre la afiliación política y las puntuaciones en la escala de Conformismo Social contra Rebeldía, lo cual proporcionó información adicional respecto a la validez de esa escala. A pesar de que este artículo es breve, está bien presentado. El estudiante aprenderá mucho con la lectura del artículo.

Medición de la democracia

¿Qué quiere decir *democracia*? El término se utiliza constantemente. ¿Pero qué se quiere decir cuando se usa? Aún más difícil, ¿cómo se mide? Bollen (1980) definió y midió "democracia", la utilizó como variable y demostró la validez de constructo de su índice de democracia política (Index of Political Democracy). Él examinó con sumo cuidado sus usos y definiciones previas, explicó la teoría subyacente al constructo y extrajo de medidas anteriores facetas importantes de la democracia política para construir su medida. Ésta contiene dos grandes aspectos —libertad política y soberanía popular— los cuales pueden llamarse variables latentes. Cada aspecto tiene tres facetas: *libertad de prensa*, *libertad de oposición de grupo* y *sanción gubernamental* (ausencia de) por libertades políticas; y *elecciones*

justas, selección ejecutiva y selección legislativa para la soberanía popular. Estos seis "indicadores" se utilizan para medir la democracia política de los países. Cada indicador está definido operacionalmente y se utiliza una escala de 4 puntos para aplicarlos a cualquier nación. La soberanía popular, por ejemplo, se mide al evaluar en qué grado la élite de un país representa al pueblo: derecho del voto, igual peso de los votos y proceso electoral justo. Los seis indicadores se combinan en un índice o puntuación única (véase Bollen, 1979, para una descripción detallada del índice y su puntuación). Note que "indicador" o "indicador social" es un término importante en la investigación social contemporánea. Por desgracia existe poco acuerdo respecto a cuáles son exactamente los indicadores. Se han definido de varias formas como índices de condiciones sociales, estadísticos e incluso como variables. En el artículo de Bollen se consideran variables. Para un análisis sobre las definiciones véase a Jaeger (1978).

A través del análisis factorial y otros procedimientos, Bollen encontró evidencia empírica para apoyar la confiabilidad y la validez de constructo del índice. Él demostró, por ejemplo, que los seis indicadores son manifestaciones de una variable latente subyacente, que es la "democracia política". También demostró que el índice está altamente correlacionado con otras medidas de democracia. Finalmente, se calcularon valores del índice para un gran número de países. Estos valores parecen coincidir con el grado de democracia (en una escala de 0 a 100) de los países; por ejemplo, Estados Unidos, 92.4; Canadá, 99.5; Cuba, 5.2; República de Estados Árabes, 38.7; Suecia, 99.9; Unión Soviética, 18.2; Israel, 96.8. Evidentemente Bollen logró medir con éxito un constructo en extremo complejo y difícil.

Otros métodos de validación de constructo

Además del método multirrasgo-multimétodo y de los métodos utilizados en los estudios anteriores, existen otros métodos para la validación de constructo. Cualquiera que aplique pruebas está familiarizado con la técnica de correlación de los reactivos con las puntuaciones totales. Al usar la técnica se supone que la puntuación total es válida. El reactivo es válido de acuerdo con el grado en que éste mida lo mismo que la puntuación total (véase capítulo 27 o Friedenberg para el estudio del análisis de reactivos).

Para estudiar la validez de constructo de cualquier medida, siempre es útil correlacionar la medida con otras medidas. El ejemplo sobre el amorismo analizado antes ilustró el método y las ideas que están detrás de él. Sin embargo, ¿no sería más valioso correlacionar una medida con un gran número de otras medidas? ¿Existe una mejor manera de aprender sobre un constructo que conocer sus correlatos? El análisis factorial constituye un método refinado para hacer esto, pues indica, en efecto, qué medidas miden la misma cosa y en qué grado miden aquello que miden.

El análisis factorial es un método poderoso e indispensable de la validación de constructo. Bollen (1980) lo utilizó en la validación del índice de democracia política y Comrey lo empleó para desarrollar una prueba completa de personalidad. Aunque ya fue descrito brevemente antes y se estudiará en detalle en un capítulo posterior, su gran importancia para la validación de medidas hace obligatorio describirlo aquí. Se trata de un método para reducir un gran número de medidas a un número más pequeño, llamadas *factores*, al descubrir cuáles "van juntas" (por ejemplo, cuáles miden la misma cosa) y las relaciones entre los grupos de medidas que van juntas. Por ejemplo, se pueden aplicar 20 pruebas a un grupo de individuos, suponiendo que cada una mide algo diferente. Sin embargo, quizá se encuentre que estas 20 pruebas son lo suficientemente redundantes como para ser explicadas con sólo cinco medidas o factores.

Una definición de validez en términos de varianza: la relación de la varianza entre la confiabilidad y la validez

El tratamiento de varianza de la validez presentado aquí es una extensión del tratamiento de confiabilidad presentado en el capítulo 27. Ambos tratamientos siguen la presentación de Guilford de la validez.

En el capítulo anterior la confiabilidad se definió como

$$r_{tt} = \frac{V_\infty}{V_t} \tag{28.1}$$

que es la proporción de la varianza "verdadera" entre la varianza total. Es teórica y empíricamente útil definir la validez de forma similar:

$$Val = \frac{V_\omega}{V_t} \tag{28.2}$$

donde *Val* es la validez, V_ω la varianza del factor común y V_t la varianza total de la medida. Por lo tanto, la validez se considera como la proporción de la varianza total de una medida, que es varianza del factor común.

Por desgracia, todavía no es posible presentar el significado completo de dicha definición, ya que se requiere de la comprensión de la llamada teoría factorial y ésta no se estudiará sino hasta después en el presente libro. A pesar de esta dificultad debe intentarse una explicación de la validez en términos de varianza para lograr una visión completa del tema. Además, la expresión matemática de la validez y la confiabilidad unificará y aclarará ambos temas. De hecho, la confiabilidad y la validez se considerarán como partes de un todo unificado.

La varianza del factor común es la varianza de una medida que es compartida por otras medidas. En otras palabras, la varianza del factor común es la varianza que dos o más pruebas tienen en común.

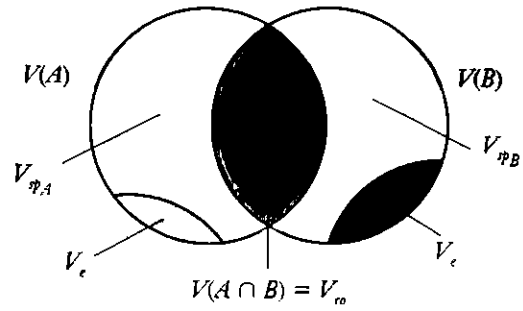
En contraste con la varianza del factor común de una medida está su *varianza específica*, V_σ, la varianza sistemática de una medida que no es compartida por cualquier otra medida. Si una prueba mide habilidades que miden otras pruebas, entonces se tiene varianza de factor común; si también mide habilidades que ninguna otra prueba mide, entonces se tiene varianza específica. La figura 28.1 expresa tales ideas y también añade el concepto de la varianza del error. Los círculos A y B representan las varianzas de las pruebas A y B. La intersección de A y B, $A \cap B$, es la relación de los dos conjuntos. De forma similar, $V(A \cap B)$ es la varianza del factor común. También se indican las varianzas específicas y las varianzas del error de ambas pruebas.

Entonces, desde este punto de vista y siguiendo el razonamiento sobre la varianza bosquejado en el capítulo anterior, cualquier varianza total de una medida posee varios componentes: *varianza del factor común*, *varianza específica* y *varianza del error*, lo cual se expresa en la ecuación:

$$V_t = V_\omega + V_\sigma + V_e \tag{28.3}$$

Para tener la capacidad de hablar de proporciones de la varianza total, se dividen los términos de la ecuación 28.3 entre la varianza total:

FIGURA 28.1



$$\frac{V_t}{V_t} = \frac{V_{co}}{V_t} + \frac{V_{es}}{V_t} + \frac{V_e}{V_t} \tag{28.4}$$

¿Cómo es que las ecuaciones 28.1 y 28.2 embonan aquí? El primer término a la derecha del signo de igual, V_{co}/V_t es el miembro derecho de la ecuación (28.2). Por lo tanto, la validez puede ser considerada como esa parte de la varianza total de una medida que no es varianza específica ni varianza del error, lo cual en forma algebraica se observa así:

$$\frac{V_{co}}{V_t} = \frac{V_t}{V_t} - \frac{V_{es}}{V_t} - \frac{V_e}{V_t} \tag{28.5}$$

Por medio de la definición dada en el capítulo anterior, la confiabilidad puede definirse como:

$$r_{tt} = 1 - \frac{V_e}{V_t} \tag{28.6}$$

Lo que puede escribirse como:

$$r_{tt} = \frac{V_t}{V_t} - \frac{V_e}{V_t} \tag{28.7}$$

Sin embargo, la parte derecha de la ecuación es parte del término de la derecha de la ecuación (28.5). Si se modifica la ecuación (28.5) ligeramente, se obtiene:

$$\frac{V_{co}}{V_t} = \frac{V_t}{V_t} - \frac{V_e}{V_t} - \frac{V_{sp}}{V_t} \tag{28.8}$$

Esto debe significar, entonces, que la validez y la confiabilidad son relaciones de varianza cercanas. La confiabilidad es igual a los primeros dos miembros de la derecha de (28.8). Por lo tanto, al incorporar (28.1) resulta:

$$r_{tt} = \frac{V_t}{V_t} - \frac{V_e}{V_t} = \frac{V_\infty}{V_t} \tag{28.9}$$

Si se sustituye en (28.8), se obtiene:

$$\frac{V_\omega}{V_t} = \frac{V_\infty}{V_t} - \frac{V_\sigma}{V_t} \tag{28.10}$$

De esta forma se observa que la proporción de la varianza total de una medida es igual a la proporción de la varianza total que es varianza "verdadera", menos la proporción que es varianza específica. O bien, la validez de una medida es esa porción de la varianza total de la medida, que comparte varianza con otras medidas. Teóricamente la varianza válida no incluye varianza debida al error, ni tampoco incluye varianza que sea específica únicamente a esta medida.

Todo esto puede resumirse de dos maneras. Primero, se suma en una ecuación o dos. Suponga que se tiene un método para determinar la varianza (o varianzas) del factor común de una prueba. (Posteriormente se verá que el análisis factorial es dicho método.) Para simplificar, considere que hay dos fuentes de varianza del factor común en una prueba —y ninguna otra—. Llame a estos factores *A* y *B*, que pueden ser habilidad verbal y habilidad aritmética, o tal vez actitudes liberales y actitudes conservadoras. Si se añade la varianza de *A* a la varianza de *B*, se obtiene la varianza del factor común de la prueba, la cual se expresa por medio de las ecuaciones:

$$V_\omega = V_A + V_B \tag{28.11}$$

$$\frac{V_\omega}{V_t} = \frac{V_A}{V_t} + \frac{V_B}{V_t} \tag{28.12}$$

Entonces, utilizando (28.2) y sustituyendo en (28.12), se obtiene:

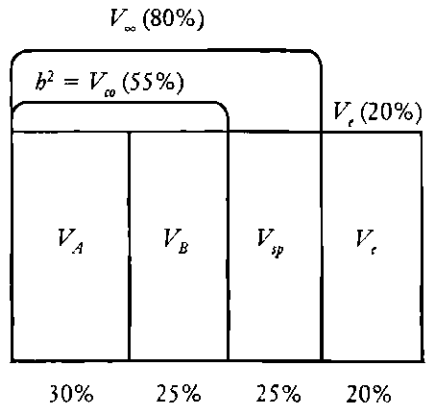
$$Val = \frac{V_A}{V_t} + \frac{V_B}{V_t} \tag{28.13}$$

La varianza total de una prueba, como se indicó antes, incluye la varianza del factor común, la varianza específica para la prueba y no para otra prueba (por lo menos en lo que se refiere a la presente información) y la varianza del error. Las ecuaciones 28.3 y 28.4 así lo expresan. Al sustituir en (28.4) la igualdad de (28.12) se obtiene:

$$\frac{V_t}{V_t} = \underbrace{\overbrace{\frac{V_A}{V_t} + \frac{V_B}{V_t}}^{h_2} + \frac{V_{es}}{V_t}}_{r_{tt}} + \frac{V_e}{V_t} \tag{28.14}$$

Los primeros dos términos del lado derecho de (28.14) están asociados con la validez de la medida, y los primeros tres términos de la derecha están asociados con la confiabilidad de la medida. Estas relaciones ya se han indicado. La varianza del factor común o el componente de validez de la medida se denomina h^2 (*aspectos comunes*), un símbolo que por lo común se utiliza para indicar la varianza del factor común de una prueba. Como siempre, la confiabilidad se denomina r_{tt}.

FIGURA 28.2



Comentar todas las implicaciones de esta formulación de validez y confiabilidad desviaría demasiado el tema en este momento. Todo lo que se necesita ahora es intentar aclarar la formulación con un diagrama y un breve análisis.

La figura 28.2 representa un intento por expresar la ecuación 28.14 en forma de diagrama. La figura indica la contribución de las distintas varianzas a la varianza total (considerada igual al 100%). Cuatro varianzas, tres varianzas sistemáticas y una varianza del error conforman la varianza total en dicho modelo teórico. Naturalmente, los resultados prácticos nunca son tan claros. Sin embargo, es notable lo bien que el modelo funciona. Pensar en términos de varianza también es valioso para conceptualizar y analizar los resultados de medición.

Se indica la contribución de cada fuente de varianza. De la varianza total, el 80% es varianza confiable; de la varianza confiable, el factor *A* contribuye con un 30% y el factor *B* contribuye con un 25% y otro 25% es específico de esta prueba. El restante 20% de la varianza total es varianza del error. La prueba se considera bastante confiable, puesto que una proporción importante de la varianza total es confiable o varianza "verdadera". La interpretación de la validez resulta más difícil. Si sólo hubiera un factor, por ejemplo *A*, y contribuyera con el 55% de la varianza total, entonces se podría decir que una proporción considerable de la varianza total sería varianza válida. Se sabría que buena parte de la medición confiable sería la medición de la propiedad conocida como *A*. Ésta sería una afirmación sobre la validez de constructo. Hablando prácticamente, los individuos medidos con la prueba serían ordenados por rangos respecto a *A*, con una confiabilidad adecuada.

No obstante, con el ejemplo hipotético anterior la situación es más compleja. La prueba mide dos factores, *A* y *B*. Podría haber tres conjuntos de órdenes de rango, uno resultante de *A*, uno de *B* y uno *específico*. Mientras que la confiabilidad repetida podría ser alta, si se pensara que se está midiendo únicamente *A*, al grado en que se pensara, la prueba no sería válida. Sin embargo, se podría tener una puntuación para cada individuo, una en *A* y una en *B*. En tal caso la prueba sería válida. Note que aunque se pensara que la prueba está midiendo únicamente *A*, las predicciones con un criterio podrían tener éxito, especialmente si el criterio tuviera mucho de *A* y de *B* en sí mismo. La prueba podría tener validez predictiva aun cuando su validez de constructo fuera cuestionable.

De hecho, los modernos desarrollos en medición indican que tales puntuaciones múltiples han empezado a formar parte, cada vez más, de un procedimiento aceptado.

Relación estadística entre confiabilidad y validez

Aunque aparecen en capítulos diferentes, los temas sobre la confiabilidad y la validez no están separados —ambos tratan con el nivel de excelencia de un instrumento de medición—. En capítulos anteriores se ha visto que es posible tener una medida confiable que no sea válida. Sin embargo, un instrumento de medición sin confiabilidad estaría destinado automáticamente al grupo de los instrumentos "pobres". También se ha mencionado brevemente que si se tiene una medida válida, entonces también se tiene una confiable. En el capítulo 27 se explicó lo que le sucede al coeficiente de confiabilidad cuando se incrementa el tamaño de la prueba. ¿Qué sucede con la validez al incrementarse el tamaño de la prueba? ¿Se ve igualmente afectada que la confiabilidad por el incremento del tamaño? La respuesta contundente es "no". El trabajo clásico de Gullekson (1950) presenta fórmulas para demostrar la relación. Si se añaden suficientes reactivos a la prueba para duplicar el coeficiente de confiabilidad, el coeficiente de validez sólo se incrementa un 41%. Las fórmulas proféticas de la validez por lo general incluyen al coeficiente de confiabilidad de cierta manera y forma. Por ejemplo, existe una fórmula para predecir el coeficiente de validez máximo, con base en el coeficiente de confiabilidad. Con el uso de dicha fórmula es posible obtener un coeficiente de validez más alto que el de confiabilidad. No obstante, en la práctica resulta muy difícil obtener un coeficiente de validez que sea más alto que el de confiabilidad. El razonamiento aquí es que se esperaría que una prueba que se correlaciona consigo misma debería ser mayor que la misma prueba correlacionada con una medida o criterio externo.

Si fuera posible eliminar los errores de medición de la prueba y del criterio, entonces se tendría esencialmente una correlación entre las puntuaciones verdaderas de ambas medidas. Se ha estudiado que los errores de medición tienden a reducir los valores del coeficiente. Es posible, en un sentido hipotético, encontrar cuál podría ser el coeficiente de validez, si se pudiera eliminar el error de medición (i) en el criterio y en la prueba, (ii) sólo en el criterio y (iii) sólo en la prueba. Dichas correcciones son denominadas *correcciones por atenuación*. Si se permite que r_{xy} sea la correlación entre el criterio x y la prueba y, la fórmula para corregir ambas por atenuación es:

$$xy\ corregido\ \ r_{xy} = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}}$$

La fórmula para determinar cuál sería la validez si se tuviera un *criterio perfecto* es:

$$r_{\infty y} = \frac{r_{xy}}{\sqrt{r_{xx}}}$$

La fórmula para determinar el coeficiente de validez si se tuviera una *prueba perfecta* es:

$$r_{x \infty} = \frac{r_{xy}}{\sqrt{r_{yy}}}$$

Estas fórmulas no deben utilizarse para tomar decisiones sobre individuos; aunque son útiles para determinar si vale la pena hacer una prueba o un criterio más confiable. Tales fórmulas muestran lo que le sucedería a la validez conforme se hicieran cambios en la confiabilidad.

La validez y confiabilidad de instrumentos de medición psicológicos y educativos

Las mediciones pobres llegan a invalidar cualquier investigación científica. La mayor parte de las críticas a la medición psicológica y educativa, hechas tanto por profesionales como por otras personas, se centra en la validez. Así es como debe ser. Lograr confiabilidad es, en gran parte, un aspecto técnico. Sin embargo, la validez es mucho más que técnica; se centra dentro de la esencia de la propia ciencia. También se centra en la filosofía. La validez de constructo, en particular, tiene un gran sentido filosófico, debido a que se relaciona con la naturaleza de la "realidad" y con la naturaleza de las propiedades que se miden.

A pesar de las dificultades para lograr mediciones psicológicas, sociológicas y educativas válidas y confiables, se ha progresado mucho en este siglo. Existe una creciente comprensión de que todos los instrumentos de medición deben ser examinados crítica y empíricamente, respecto a su confiabilidad y validez. Terminaron los días de tolerancia a la medición inadecuada. Las demandas impuestas por profesionales, las herramientas teóricas y estadísticas disponibles y aquellas que se van desarrollando rápidamente, así como la creciente sofisticación de los estudiantes de posgrado en psicología, sociología y educación, han establecido nuevos estándares más altos que deben ser estimulantes saludables para la imaginación, tanto de los que trabajan en investigación como de quienes desarrollan la medición científica.

RESUMEN DE CAPÍTULO

1. La validez trata con la precisión. ¿El instrumento mide lo que se supone que debe medir?
2. Existen tres tipos de validez
   - de contenido
   - relacionada con el criterio
   - de constructo
3. La validez de contenido se refiere a la adecuación de la representatividad o muestreo del contenido de la prueba.
4. La validez aparente es similar a la validez de contenido, pero no es cuantitativa e incluye una mera inspección visual de la prueba, por parte de revisores sofisticados o no-sofisticados.
5. Existen dos métodos bajo la validez relacionada con el criterio: concurrente y predictiva.
6. La característica distintiva entre la validez concurrente y la predictiva es la relación temporal entre el instrumento y el criterio.
7. Un instrumento con alta validez relacionada con el criterio ayuda a los usuarios de pruebas a tomar mejores decisiones en términos de ubicación, clasificación, selección y evaluación.
8. La validez de constructo busca explicar las diferencias individuales en puntuaciones de pruebas. Trata con conceptos abstractos que pueden contener dos o más dimensiones.
9. La validez de constructo requiere tanto de convergencia como de discriminación.
10. La convergencia establece que los instrumentos que pretendan medir la misma cosa deben estar altamente correlacionados.

11. La discriminación se demuestra cuando instrumentos que se supone miden cosas diferentes tienen una baja correlación.
12. Un método utilizado para demostrar tanto la convergencia como la discriminación es la matriz multirrasgo-multimétodo de Campbell y Fiske (1959).
13. La relación entre la validez y la confiabilidad es susceptible de demostrarse matemáticamente.
14. El conocimiento respecto a la interpretación de las mediciones es importante para los estudios de investigación.
15. Dos temas menos tradicionales respecto a la interpretación y la validez son: la comprobación en referencia al criterio y la comprobación en referencia a la información (o medición con probabilidad admisible).

SUGERENCIAS DE ESTUDIO

1. La literatura sobre la medición es vasta. Las siguientes referencias se eligieron por su excelencia particular o por su relevancia para temas importantes sobre medición. Sin embargo, algunos de los análisis son técnicos y difíciles. El estudiante encontrará análisis elementales sobre confiabilidad y validez en la mayor parte de los libros sobre medición.

Allen, M. J. y Yen, W. M. (1979). *Introduction to measurement theory.* Belmont, California: Brooks/Cole.

Cronbach, L. J. y Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302. [Una muy importante contribución a la medición moderna y a la investigación del comportamiento.]

Cureton, E. (1969). Measurement theory, en R. Ebel, V. Noll y R. Bauer (eds.), *Encyclopedia of educational research* (4a. ed.), 785-804. Nueva York: Macmillan. [Un panorama firme y general de la medición, con énfasis en la medición educativa.]

Horst, P. (1966). *Psychological measurement and prediction.* Belmont, California: Wadsworth.

Tryon, R. (1957). Reliability and behavior domain validity: A reformulation and historical critique. *Psychological Bulletin, 54,* 229-249. [Éste es un excelente e importante artículo sobre confiabilidad. Contiene un buen ejemplo trabajado.]

Los siguientes artículos sobre antologías de la medición constituyen fuentes valiosas de los clásicos en el campo. Especialmente los volúmenes de Mehrens y Ebel y de Jackson y Messick.

Anastasi, A. (ed.). (1966). *Testing problems in perspective.* Washington, DC: American Council on Education.

Barnette, W. L. (ed.). (1976). *Reading in psychological tests and measurement* (3a. ed.). Baltimore, MD: Williams y Wilkins.

Chase, C. y Ludlow G. (eds.). (1966). *Readings in educational and psychological measurement.* Boston: Houghton Mifflin.

Jackson, D. y Messick, S. (eds.). (1967). *Problems in human assessment.* Nueva York, McGraw-Hill.

Mehrens, W. y Ebel, R. (eds.). (1967). *Principles of educational and psychological measurement.* Skokie, Illinois: Rand McNally.

2. Un método importante para la validez de estudios es la validez cruzada. Los estudiantes avanzados pueden beneficiarse del ensayo de Mosier en el libro de Chase y Ludlow mencionado anteriormente. Se puede encontrar un breve resumen del ensayo de Mosier en Guilford (1954, p. 406).
3. Los estudiantes más avanzados también querrán saber algo sobre las fijaciones de respuesta —una amenaza para la validez, particularmente para la validez de reactivos e instrumentos de personalidad, actitud y valores—. Las *fijaciones de respuesta* son tendencias a responder los reactivos de ciertas maneras —alto, bajo, aprobar, desaprobar, en extremo, etcétera, independientemente del contenido de los reactivos—. Las puntuaciones resultantes están, por lo tanto, sistemáticamente sesgadas. La literatura es extensa y no puede citarse aquí. Sin embargo, una excelente exposición se encuentra en Nunnally (1978), capítulo 16, especialmente pp. 655 y sig. Los defensores de los efectos de las fijaciones de respuesta en los instrumentos de medición son muy duros en sus afirmaciones. Rorer (1965) ha atacado enfáticamente el tema de las fijaciones de respuesta.

   La posición tomada en este libro es que las fijaciones de respuesta realmente suceden y que en algunas ocasiones tienen efectos considerables, pero que las fuertes declaraciones de los partidarios son exageradas. La mayor parte de la varianza en las medidas bien construidas parece deberse a las variables medidas, y relativamente muy poco a las fijaciones de respuesta. Los investigadores deben estar conscientes de las fijaciones de respuesta y sus posibles efectos negativos sobre los instrumentos de medición, pero no deben tener miedo de utilizar los instrumentos. Si se tomara demasiado en serio a las escuelas de pensamiento sobre las fijaciones de respuesta y sobre lo que se ha llamado el efecto del experimentador (en educación es el efecto Pigmalión) explicado antes, se tendría que abandonar la investigación del comportamiento con excepción, quizás, de la investigación que se realiza con las llamadas medidas no invasivas.
4. Imagine que usted aplicó una prueba con seis reactivos a seis personas. Las puntuaciones de cada reactivo de cada persona se presentan abajo. Suponga que también aplicó otra prueba con seis reactivos a otras seis personas. Las puntuaciones también se incluyen abajo. Las puntuaciones de la primera prueba, I, se presentan a la izquierda; las puntuaciones de la segunda prueba, II, se presentan a la derecha.

I							II						
	Reactivos							*Reactivos*					
Personas	*a*	*b*	*c*	*d*	*e*	*f*	Personas	*a*	*b*	*c*	*d*	*e*	*f*
1	6	6	7	5	6	5	1	6	4	5	6	6	3
2	6	4	5	5	4	5	2	6	2	7	4	4	4
3	5	4	7	6	4	3	3	5	6	5	3	4	2
4	3	2	5	3	4	4	4	3	4	4	5	4	5
5	2	3	4	4	3	2	5	2	1	7	1	3	5
6	2	1	3	1	0	2	6	2	3	3	5	0	2

Las puntuaciones en II son las mismas que en I, excepto que el orden de las puntuaciones de los reactivos (*b*), (*c*), (*d*) y (*f*) se ha cambiado.

   a) Realice un análisis de varianza de dos factores con cada uno de los conjuntos de puntuaciones. Compare e interprete las razones *F*. Ponga especial atención a la razón *F* para *Personas* (individuos).

b) Calcule $r_{tt} = (V_{ind} - V_e)/ V_{ind}$ para I y II. Interprete las dos r_{tt}. ¿Por qué son tan diferentes?

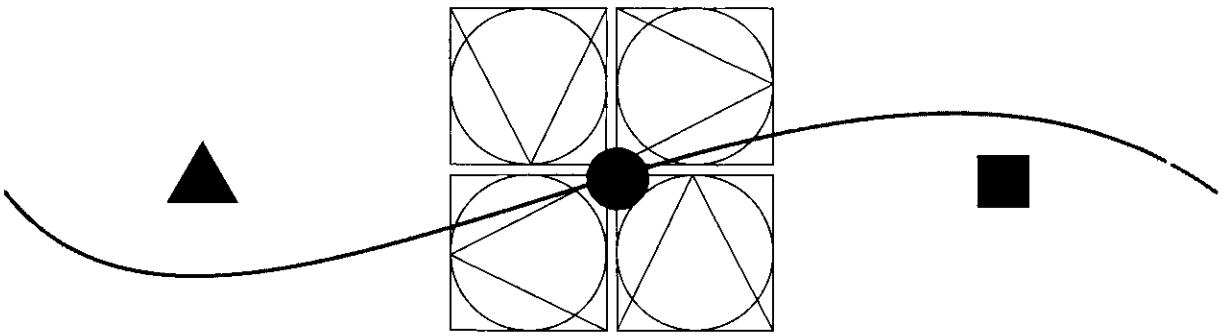
c) Sume los reactivos impares a través de los renglones; sume los reactivos pares. Compare los órdenes de rango y los rangos de los totales impares, de los totales pares y de los totales de los seis reactivos. Los coeficientes de correlación entre los reactivos impares y pares, corregidos, son .98 y .30. Explique por qué son tan diferentes. ¿Qué significan?

d) Suponga que había 100 personas y 60 reactivos. ¿Habría cambiado esto los procedimientos y el razonamiento subyacente? ¿Habría afectado, el efecto de cambiar el orden de, por ejemplo, cinco a diez reactivos, a las r_{tt} tanto como en estos ejemplos? Si no fuese así, ¿por qué no?

[Respuestas: ***a***) I: $F_{reactivos} = 3.79$ (.05); $F_{personas} = 20.44$ (.001). II: $F_{reactivos} = 1.03$ (n.s); $F_{personas} = 1.91$ (n.s). ***b***) I: $r_{tt} = .95$; II: $r_{tt} = .48$.]

Parte Nueve

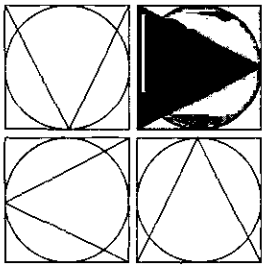
Métodos de observación y de recolección de datos



Capítulo 29
Entrevistas e inventarios de entrevistas

Capítulo 30
Pruebas y escalas objetivas

Capítulo 31
Observaciones del comportamiento y sociometría



CAPÍTULO 29

ENTREVISTAS E INVENTARIOS DE ENTREVISTAS

La entrevista es quizás la técnica de uso más frecuente para obtener información de la gente. Ha sido y aún es utilizada en todo tipo de situaciones prácticas: el abogado obtiene información de un cliente; el médico conoce sobre el paciente; el oficial de admisiones o el profesor determina la adecuación de estudiantes a determinadas escuelas, departamentos y programas. Sin embargo, sólo hasta hace poco se ha utilizado la entrevista de forma sistemática para propósitos científicos, tanto en el laboratorio como en el campo.

Los métodos de recolección de datos se clasifican de acuerdo a qué tan directos son. Si se desea saber algo sobre las personas, se les puede preguntar directamente. Ellos ofrecen o no una respuesta. Por otro lado, es posible preguntar de forma indirecta. Se puede utilizar un estímulo ambiguo como una fotografía borrosa, una mancha de tinta o una pregunta vaga, y después preguntar respecto a las impresiones de los estímulos, bajo el supuesto de que los entrevistados darán la información requerida sin saber que lo están haciendo. Esta técnica es bastante indirecta. La mayor parte de los métodos de recolección de datos utilizados en la investigación psicológica y sociológica son relativamente directos o moderadamente indirectos. En pocas ocasiones se utilizan medios muy indirectos.

Las entrevistas y los inventarios (cuestionarios) por lo general son bastante directos, lo cual representa tanto una fortaleza como una debilidad. Tienen fortaleza porque gran

cantidad de la información requerida en la investigación social científica se obtiene de los entrevistados por medio de preguntas directas. Aunque las preguntas deben manejarse con sumo cuidado, los entrevistados pueden, y generalmente lo hacen, dar mucha información de forma directa. No obstante, existe información de naturaleza más difícil que los entrevistados quizá no estén dispuestos a dar fácil y directamente —por ejemplo, información sobre sus ingresos, relaciones sexuales y algunas actitudes hacia la religión o hacia los grupos minoritarios—. En tales casos, las preguntas directas llegan a generar datos que no son válidos. Sin embargo, si se manejan en forma apropiada, aun el material personal o polémico puede obtenerse exitosamente por medio de entrevistas e inventarios.

La entrevista es probablemente uno de los métodos más antiguos y más utilizados para conseguir información. Posee importantes cualidades que las pruebas y escalas objetivas y las observaciones del comportamiento no tienen. Una entrevista puede proporcionar una gran cantidad de información si se utiliza con un inventario bien realizado. Es flexible y se adapta a situaciones individuales, y puede usarse con frecuencia cuando ningún otro método es posible o adecuado. Estas cualidades la hacen especialmente adecuada para la investigación con niños. Los métodos y consideraciones sobre las entrevistas con niños pueden encontrarse en Aldridge y Wood (1998) y en Poole y Lamb (1998). Minkes, Robinson y Weston (1994) ofrecen explicaciones sobre la forma de entrevistar a niños con discapacidades. Ellis (1989) describe cómo conducir una entrevista con niños superdotados canadienses. Si un entrevistador sabe que el entrevistado, especialmente un niño, no entiende una pregunta, puede, dentro de ciertos límites, repetir o replantear la pregunta. Las preguntas sobre deseos, aspiraciones y ansiedades pueden plantearse de tal manera que produzcan información precisa. De mayor importancia, quizás, es el hecho de que la entrevista permite explorar el contexto y las razones de las respuestas a las preguntas. McReynolds (1989) sintetiza el estado de los instrumentos de medición clínica, de los cuales uno es el *inventario de entrevista.*

La mayor desventaja de la entrevista y de su inventario acompañante es de índole práctica. Las entrevistas toman mucho tiempo. El obtener información de un individuo llega a tomar tanto como una hora e incluso dos horas. Una gran inversión de tiempo implica esfuerzo y dinero. Andrews (1974) determina los requisitos de un estudio de investigación bien conducido, que utilice la entrevista, en términos del reclutamiento, entrenamiento, selección y supervisión. Uno de los componentes importantes de la entrevista es la supervisión. Andrews menciona por lo menos nueve responsabilidades que un supervisor debe tener. Por lo tanto, siempre que una técnica más económica responda a los propósitos de investigación, no deben utilizarse las entrevistas. Emory (1976) cita investigación realizada respecto a las características del entrevistador. Tales estudios hallaron evidencia de que características triviales podían influir en los resultados de la entrevista. Por ejemplo, Emory cita el hecho de que las mujeres son mejores entrevistadores que los hombres, y que los hombres casados son mejores que las mujeres solteras. También cita un estudio realizado por el Centro Nacional de Investigación de Opinión (National Opinion Research Center, NORC) que relaciona ciertas características con la calidad de la entrevista. El entrenar entrevistadores para que produzcan el mismo nivel de calidad requiere de tiempo, recursos e inclusive quizás de experiencia previa.

Las entrevistas e inventarios como herramientas de la ciencia

Las entrevistas e inventarios han sido utilizados, en su mayor parte, simplemente para reunir los llamados hechos. El uso más importante de la entrevista debe ser el estudio de

relaciones y la prueba de hipótesis. En otras palabras, la entrevista es un instrumento de medición psicológica y sociológica. Quizás más preciso, los productos de entrevistas —las respuestas de los entrevistados a preguntas cuidadosamente elaboradas— pueden traducirse en medidas de variables. Por lo tanto, las entrevistas y los inventarios de entrevista están sujetos a los mismos criterios de confiabilidad, validez y objetividad que otros instrumentos de medición.

Una entrevista sirve para tres propósitos principales:

1. Como un dispositivo exploratorio para ayudar a identificar variables y relaciones, para sugerir hipótesis y para guiar otras fases de la investigación.
2. Ser el principal instrumento de la investigación. En dicho caso, en el inventario de entrevista se incluyen preguntas diseñadas para medir las variables de la investigación. Estas preguntas se consideran después como reactivos en un instrumento de medición, más que como meros dispositivos para reunir información.
3. Puede complementar otros métodos: hacer un seguimiento de resultados inesperados, validar otros métodos y profundizar en las motivaciones de entrevistados y en las razones por las que responden como lo hacen.

Al utilizar entrevistas como herramientas de investigación científica, debe plantearse la pregunta: ¿los datos del problema de investigación pueden obtenerse de una manera mejor y más fácil? Lograr confiabilidad, por ejemplo, no constituye un problema pequeño. Los entrevistadores deben tener un entrenamiento; las preguntas deben probarse con anterioridad y revisarse para eliminar ambigüedades y redacción inadecuada. ¿Vale la pena el esfuerzo? Tampoco la validez es un problema pequeño. Deben realizarse esfuerzos especiales para eliminar los prejuicios del entrevistador; las preguntas deben probarse para encontrar sesgos desconocidos. El problema de investigación particular y la naturaleza de la información buscada debe, en el análisis final, determinar si se utilizará o no la entrevista. Cannell y Kahn (1968, capítulo 15) y Warwick y Lininger (1975, capítulo 7) proporcionan una guía detallada respecto a si una entrevista debe o no ser empleada.

La entrevista

La *entrevista* es una situación interpersonal cara a cara donde una persona (el entrevistador) le plantea a otra persona (el entrevistado) preguntas diseñadas para obtener respuestas pertinentes al problema de investigación. Existen dos tipos generales de entrevista: la *estructurada* y la *no estructurada*, o *estandarizada* y *no estandarizada* (véase Cannell y Kahn, 1968). En la entrevista estandarizada las preguntas, su secuencia y su redacción son fijas. Se permite cierta libertad al entrevistador al plantear las preguntas, pero ésta es relativamente poca. El *Manual del Entrevistador* (1976) producido por el Instituto de Investigación Social (Institute for Social Research) de la Universidad de Michigan establece que la perspectiva de la entrevista está evolucionando de la perspectiva tradicional. La entrevista se considera como una interacción, una relación de roles activos entre entrevistador y entrevistado, donde el entrevistador es incluso un maestro. Cannell y Kahn (1968) y Dohrenwend y Richardson (1963) ofrecen información adicional sobre este tema. El nivel de libertad se especifica de antemano. Las entrevistas estandarizadas utilizan inventarios de entrevista que se han preparado cuidadosamente para obtener información concerniente al problema de investigación.

Las entrevistas no estandarizadas son más flexibles y abiertas. A pesar de que los propósitos de investigación determinan las preguntas planteadas, su contenido, secuencia y

redacción están en manos del entrevistador. Por lo común no utilizan ningún inventario. En otras palabras, la entrevista no estandarizada y no estructurada constituye una situación abierta, en contraste con la entrevista estandarizada y estructurada, que es una situación cerrada. Ello no significa que una entrevista no estandarizada sea casual; debe ser planeada tan cuidadosamente como la estandarizada. Green y Tull (1988) afirman que las entrevistas no estructuradas obtienen información que las entrevistas estructuradas no ofrecen. Con el modelo informal de la entrevista no estructurada el investigador obtiene ideas respecto a las motivaciones del entrevistado. Las entrevistas no estructuradas algunas veces se denominan *entrevistas profundas*. Son especialmente útiles para realizar estudios exploratorios. El interés central aquí lo representa la entrevista estandarizada. Sin embargo, se reconoce que muchos problemas de investigación pueden requerir, y muchas veces es así, un tipo de entrevista en que el entrevistador tenga permitido utilizar preguntas alternativas que se ajusten a entrevistados particulares y a cuestiones particulares. El procedimiento real de la conducción de una entrevista no se analiza en este libro. El lector encontrará una guía en la sección de sugerencias de estudio del presente capítulo.

El inventario de entrevista

La conducta de entrevistar es, en sí misma, un arte, pero la planeación y la redacción de un inventario de entrevista lo es todavía más. Es poco común que un novato produzca un buen inventario, al menos sin considerable estudio y práctica previos. Existen varias razones para ello; las principales probablemente sean el significado múltiple y la ambigüedad de las palabras, la ausencia de un enfoque directo y constante en los problemas e hipótesis de estudio, la falta de apreciación del inventario como un instrumento de medición y la ausencia de antecedentes y experiencia necesarios.

Tipos de información y reactivos de los inventarios

La mayor parte de los inventarios incluyen tres tipos de información: información de la carátula (identificación), información de tipo censal (o sociológica) e información del problema. Con excepción de la identificación, estos tipos de información se analizaron en un capítulo anterior. No obstante, debe mencionarse la importancia de identificar cada inventario de manera precisa y completa. El investigador cuidadoso debe aprender a identificar con letras, números u otros símbolos cada inventario y cada escala. Además, debe registrarse de forma sistemática la identificación de la información de cada individuo. Existen dos tipos de reactivos de inventario que se usan comúnmente: *de alternativa fija* (o cerrados) y *de preguntas abiertas*. Se utiliza también un tercer tipo de reactivo, de alternativas fijas: los reactivos *de escala*.

Reactivos de alternativa fija

Como su nombre lo indica, los reactivos de alternativa fija ofrecen al entrevistado una opción entre dos o más alternativas. A estos reactivos también se les llama preguntas *cerradas* o *de encuesta*. El tipo más común de reactivo de alternativa fija es dicotómico: plantea preguntas que pueden responderse como *sí* o *no*, *de acuerdo* o *en desacuerdo* y otro tipo de respuestas de dos opciones. Con frecuencia se añade una tercera posibilidad: *no sé* o *indeciso*.

Un ejemplo de un reactivo de alternativa fija sería:

¿Considera usted que el gobierno de Estados Unidos ya encontró una cura para el SIDA, pero que la está ocultando?

Sí ..... [ ]
No ..... [ ]
No sé ..... [ ]

A pesar de que los reactivos de alternativa fija poseen las claras ventajas de lograr una mayor uniformidad de la medición y, por lo tanto, mayor confiabilidad, de forzar al entrevistado a responder en una forma que se ajuste a las categorías previamente establecidas y de ser fáciles de codificar, también tienen ciertas desventajas. La mayor de ellas es su superficialidad: sin sondeo, generalmente no van más allá de la superficie de la respuesta. También pueden irritar al entrevistado que no encuentre ninguna alternativa adecuada para él. Peor aún, es posible que fuercen respuestas. Un entrevistado puede elegir una alternativa para ocultar ignorancia o elegir alternativas que no representan con precisión los hechos u opiniones. Tales dificultades no implican que los reactivos de alternativa fija sean malos o inútiles. Por el contrario, se utilizan con buenos resultados si se redactan de manera juiciosa, si se utilizan con un sondeo y si se mezclan con reactivos abiertos. Un *sondeo* es un dispositivo utilizado para encontrar información de los entrevistados sobre un tema, sus marcos de referencia o, más común, para aclarar y establecer las razones de las respuestas dadas. El sondeo incrementa el poder de "obtención de respuesta" de las preguntas, sin cambiar su contenido. Ejemplos de sondeo son: "Dígame más sobre esto." "¿Cómo es eso?" "¿Puede explicarlo?" (véase Warwick y Lininger, 1975, pp. 210-215).

Reactivos abiertos

Los reactivos abiertos o de final abierto representan un desarrollo extremadamente importante en la técnica de la entrevista. Las preguntas *abiertas* son aquellas que brindan un marco de referencia para las respuestas de los entrevistados, pero poniendo un mínimo de restricción a las respuestas y a su expresión. Aunque su contenido está determinado por el problema de investigación, no imponen ninguna otra restricción sobre el contenido ni forma de las respuestas del entrevistado. Más tarde se darán ejemplos.

Las preguntas abiertas tienen importantes ventajas, aunque también tienen desventajas. Sin embargo, si se redactan y utilizan apropiadamente se minimizan las desventajas. Las preguntas abiertas son flexibles; tienen la posibilidad de profundizar; le permiten al entrevistador aclarar malos entendidos (a través del sondeo), establecer la falta de conocimiento de un entrevistado, detectar ambigüedades, promover la cooperación y lograr *rapport* y mejores estimados de las verdaderas intenciones, creencias y actitudes de los entrevistados. Su empleo tiene también otra ventaja: en ocasiones las respuestas a preguntas abiertas *sugieren* posibilidades de relaciones e hipótesis. Los entrevistados algunas veces darán respuestas inesperadas que tal vez indiquen la existencia de relaciones no anticipadas originalmente.

Un tipo especial de pregunta abierta es la pregunta de *embudo*. En realidad se trata de un conjunto de preguntas dirigidas a obtener información sobre un solo tema importante o sobre un solo conjunto de temas relacionados. El embudo inicia con una pregunta general y se va reduciendo progresivamente al punto o puntos específicos importantes. Warwick y Lininger (1975) señalan que el mérito del embudo es que permite la libre respuesta en las primeras preguntas y después se reduce a preguntas y respuestas específicas; y que también facilita el descubrimiento de los marcos de referencia del entrevistado. Otra forma de pregunta de embudo inicia con una pregunta abierta general y continúa con reactivos específicos cerrados. La mejor forma de reconocer las buenas preguntas abiertas y de embudo es por medio del estudio de ejemplos.

Para obtener información sobre prácticas de crianza de niños, Sears, Maccoby y Levin (1957) utilizaron varias buenas preguntas abiertas y de embudo. Una de ellas, con comentarios del autor entre corchetes, es:

Ejemplo

> Por supuesto que todos los bebés lloran. [Note que el entrevistador tranquiliza al padre respecto al llanto de su hijo.] Algunas madres consideran que si se levanta a un bebé cada vez que llora, se le malcría. Otros piensan que no se debe dejar llorar demasiado tiempo a un bebé. [El marco de referencia se ha expresado claramente. También se ha tranquilizado a la madre pues no importa cómo maneje el llanto de su bebé.] ¿Qué piensa usted respecto a esto?
>
> (a) ¿Qué hizo con X respecto a esto?
>
> (b) ¿Qué hizo a la medianoche?

Este conjunto de preguntas de embudo no sólo evalúa actitudes, sino que también sondea prácticas específicas.

Reactivos de escala

Un tercer tipo de reactivo de inventario es el reactivo de escala. Una *escala* es un conjunto de reactivos verbales, a cada uno de los cuales un individuo responde expresando grados de acuerdo o desacuerdo, o algún otro modo de respuesta. Los reactivos de escala tienen alternativas fijas y colocan al individuo entrevistado en algún punto de la escala. (Serán analizados con mayor profundidad en el capítulo 30.) El uso de reactivos de escala en inventarios de entrevista es un desarrollo que promete mucho, debido a que se combinan los beneficios de las escalas con los de las entrevistas. Por ejemplo, se puede incluir una escala para medir actitudes hacia la educación en un inventario de entrevista sobre el mismo tema. Las puntuaciones de la escala se obtienen de esta forma para cada entrevistado y se verifican contra datos de preguntas abiertas. Es posible medir la *tolerancia hacia la inconformidad*, como lo hizo Stouffer (1955), al tener una escala para medir esta variable incluida en el inventario de entrevista.

Criterios para la redacción de preguntas

Los criterios o preceptos para la redacción de preguntas se han desarrollado a través de la experiencia y la investigación. Algunos de los más importantes se presentan más adelante en forma de preguntas. Se anexaron breves comentarios a las preguntas. Al enfrentarse con la necesidad real de redactar un inventario, el estudiante deberá consultar tratados más extensos, ya que el siguiente análisis, en congruencia con la postura del resto del capítulo, pretende ser sólo una introducción al tema. Si se desea una guía práctica véase Emory (1976, capítulo 8), quien proporciona un buen resumen y puntos clave para elaborar el inventario; así como Noelle-Neuman (1970) y Warwick y Lininger (1975). Emory enfatiza la forma de probar el instrumento antes de usarlo realmente, cómo secuenciar los reactivos o preguntas y qué hacer bajo ciertas situaciones. Otras recomendaciones son Atkinson (1971), Beed y Stimson (1985) y Mishler (1986).

1. *¿Está relacionada la pregunta con el problema y los objetivos de investigación?* Con excepción de preguntas de información factual y sociológica, todos los reactivos del inventario deben estar en función del problema de investigación. Esto significa que el

propósito de cada pregunta es generar información que sirva para probar las hipótesis de la investigación.

2. *¿Es apropiado el tipo de pregunta?* Alguna información puede obtenerse mejor con preguntas abiertas —razones para comportamientos, intenciones y actitudes—. Por otro lado, otro tipo de información puede obtenerse de forma más expedita por medio de preguntas cerradas. Si todo lo que se requiere de un entrevistado es la opción preferida de dos o más alternativas, y estas alternativas pueden especificarse con claridad, sería inútil utilizar una pregunta abierta (véase Dohrenwend y Richardson, 1963; Schuman y Presser, 1979; Warwick y Lininger, 1975).
3. *¿El reactivo es claro y sin ambigüedades?* Un reactivo o afirmación ambigua es aquella que permite o invita a interpretaciones alternativas, de las cuales resultan respuestas diferentes. Las preguntas denominadas de doble sentido, por ejemplo, son ambiguas debido a que proporcionan dos o más marcos de referencia en lugar de uno solo. Los entrevistados, aun cuando no se confundan con la complejidad y alternativas ofrecidas por la siguiente pregunta, difícilmente responderían utilizando un marco común de referencia y comprendiendo lo que se pide. "¿Cómo le va a usted y a su familia este año? ¿La pregunta se refiere a finanzas, felicidad marital, estado de salud o a qué?

   Se ha realizado un gran trabajo respecto a la redacción de reactivos. Si se siguen ciertos preceptos, esto ayuda al redactor a evitar ambigüedades. En primer lugar, deben evitarse las preguntas que contengan más de una idea a la que el entrevistado pueda reaccionar. Un reactivo como "¿considera usted que las metas educativas de la preparatoria moderna y los métodos de enseñanza utilizados para lograr estos objetivos son educativamente adecuados?" es una pregunta ambigua, debido a que al entrevistado se le cuestiona acerca de los objetivos educativos y de los métodos de enseñanza en la misma pregunta. En segundo lugar, se deben evitar términos y expresiones ambiguas. Es posible plantear la siguiente pregunta a un entrevistado: ¿piensa usted que los maestros de su escuela reciben un trato justo?" Se trata de un reactivo ambiguo debido a que "trato justo" puede referirse a diferentes tipos de trato. El término *justo* también puede significar "justicia", "equidad", "no demasiado bueno", "imparcial" y "objetivo". La pregunta requiere de un contexto claro, es decir, de un marco de referencia explícito. (Sin embargo, algunas preguntas ambiguas se utilizan deliberadamente para producir distintos marcos de referencia.)
4. *¿La pregunta es de tipo conducente?* Las preguntas conducentes sugieren respuestas y como tales, amenazan la validez. Si se le pregunta a alguien "¿ha leído usted sobre la situación de la escuela local?" se puede obtener un número desproporcionado de respuestas "sí", ya que la pregunta implicaría que es malo no haber leído sobre la situación de la escuela local.
5. *¿La pregunta demanda conocimiento e información que el entrevistado no posee?* Para contrarrestar la invalidez de una respuesta debida a falta de información, resulta sensato utilizar preguntas de filtro de información. Antes de preguntarle a una persona lo que piensa acerca de la UNESCO, primero debe preguntársele si sabe lo que significa y es la UNESCO. Existe otro método: se le explica al entrevistado brevemente lo que es la UNESCO y luego se le pregunta lo que piensa de ella.
6. *¿La pregunta demanda material personal o delicado que el entrevistado pueda negarse a proporcionar?* Se requiere de técnicas especiales para obtener información de naturaleza personal, delicada o polémica. Pregunte sobre los ingresos u otros asuntos personales más tarde en la entrevista, después de haber establecido el *rapport*. Si se pregunta respecto a algo que es desaprobado socialmente, primero debe mostrarse que algunas personas piensan en un sentido y que otras piensan en otro sentido. En

efecto, no debe provocarse que el entrevistado se desapruebe a sí mismo. Se le debe asegurar que todas las respuestas serán confidenciales.

7. *¿La pregunta está cargada de deseo de aceptación social?* La gente tiende a dar respuestas que son socialmente deseables, respuestas que indican o implican la aprobación de actos o cosas que son generalmente consideradas como "buenas". Se le puede preguntar a una persona sobre sus sentimientos hacia los niños. Se supone que todos deben amar a los niños. A menos que se sea cuidadoso, se obtendrá una respuesta estereotipada sobre los niños y el amor. También, si se le pregunta a una persona si vota, hay que tener cuidado, ya que se supone que todos deben votar. Si se le pregunta a un entrevistado sobre sus reacciones ante los grupos minoritarios, de nuevo se corre el riesgo de obtener respuestas inválidas. La mayor parte de las personas educadas, sin importar cuáles sean sus "verdaderas" actitudes, están conscientes de la desaprobación de los prejuicios. Entonces, una buena pregunta es aquella donde los entrevistados no son conducidos a expresar meros sentimientos socialmente deseables. Al mismo tiempo, tampoco se debe preguntar a los entrevistados de forma que se enfrenten con la necesidad de dar respuestas socialmente indeseables.

El valor de las entrevistas y de los inventarios de entrevistas

Cuando la entrevista se acompaña de un inventario adecuado de valor comprobado, constituye una herramienta de investigación potente e indispensable que produce datos que ninguna otra herramienta de investigación ofrece. Es adaptable, capaz de utilizarse con todo tipo de entrevistados en muchos tipos de investigaciones y única por su adecuación para hacer exploraciones profundas. ¿Pero equilibran sus fortalezas a sus debilidades? ¿Cuál es su valor en la investigación del comportamiento, en comparación con otros métodos de recolección de datos?

La herramienta más natural con la cual comparar a la entrevista es el llamado cuestionario. Como se señaló antes, "cuestionario" es un término que se emplea casi para cualquier clase de instrumento que incluye preguntas o reactivos a los que responde un individuo. Aunque el término se utiliza de manera intercambiable con "inventario", parece estar más asociado con instrumentos autoadministrados que poseen reactivos cerrados o de alternativa fija.

El *instrumento autoadministrado* posee ciertas ventajas. Siendo todos o la mayoría de sus reactivos de tipo cerrado, se alcanza mayor uniformidad de estímulo y, por lo tanto, mayor confiabilidad. A este respecto, tiene las ventajas de las escalas y pruebas escritas de tipo objetivo, si se elaboran y se prueban previamente de manera adecuada. Una segunda ventaja es que, si son anónimos y confidenciales, se alienta la honestidad y la franqueza. Este tipo de instrumento también se aplica a muchas personas de manera relativamente fácil. Una ventaja un tanto dudosa es que puede enviarse por correo a los entrevistados. Además, también es económica: su costo por lo general es una fracción del de las entrevistas.

Las desventajas de los instrumentos autoadministrados (cuando se envían por correo) parecen sobrepasar sus ventajas. La principal desventaja es un bajo porcentaje de recuperación. Una segunda desventaja es que quizá no sea tan uniforme como parece. La experiencia ha demostrado que con frecuencia la misma pregunta tiene diferente significado para distintas personas. Como se explicó antes, esto puede manejarse en la entrevista. Sin embargo, no es posible hacer algo para resolver dicha situación cuando se autoadministra el instrumento. En tercer lugar, si sólo se utilizan reactivos cerrados, el instrumento muestra las mismas debilidades de los reactivos cerrados analizadas con anterioridad. Por otro lado, si se utilizan reactivos abiertos, es posible que el entrevistado se niegue a escribir las

respuestas, lo cual reduce la muestra de respuestas adecuadas. Muchas personas no son capaces de expresarse adecuadamente a través de la escritura, y a muchos que sí pueden hacerlo, les disgusta.

Debido a estas desventajas, probablemente la entrevista sea superior al cuestionario autoadministrado. (Esta objeción, por supuesto, no incluye las escalas de personalidad y de actitud que están cuidadosamente elaboradas.) El mejor instrumento disponible para estudiar el comportamiento, las intenciones futuras, los sentimientos, las actitudes y las razones del comportamiento de las personas parece ser la entrevista estructurada, en combinación con un inventario de entrevista que incluya reactivos cerrados, abiertos y de escala. Por supuesto, la entrevista estructurada debe elaborarse y construirse con cuidado, así como aplicarse únicamente por entrevistadores hábiles. Sus principales desventajas son el costo en tiempo, energía y dinero, y el alto nivel de habilidad necesarios para su elaboración. Una vez que se superan sus desventajas, la entrevista estructurada se vuelve una poderosa herramienta.

El grupo focal y la entrevista de grupo: otro método de entrevista

Quizás este tema pertenezca a un capítulo previo cuando se expusieron los métodos cualitativos. Algunos investigadores equiparan al método del grupo focal con la investigación cualitativa (Calder, 1977). Algunos se han referido a este método como *entrevistas de grupo* (Wells, 1974). Basch (1987) reporta que este método fue expuesto por Bogardus en 1926, pero que sólo se utilizó ocasionalmente a partir de entonces hasta los años ochenta. Quienes utilizaban primordialmente el método del grupo focal, hasta hace poco, eran los investigadores de mercado y de negocios. Basch (1987) considera que el método del grupo focal es prometedor en áreas diferentes de la mercadotecnia. Él considera que podría ser una técnica de investigación para mejorar la investigación, práctica y teoría de la educación para la salud. El método proporciona una visión profunda de la gente. Sudman, Bradburn y Schwarz (1996) creen que la metodología del grupo focal sirve para determinar la manera en que los entrevistados producen y procesan información.

La técnica del grupo focal implica entrevistar a dos o más personas al mismo tiempo. El tamaño del grupo focal debe ser lo suficientemente grande para generar diversos puntos de vista, pero lo suficientemente pequeño para ser manejable. Krueger (1994) recomienda de siete a diez personas por grupo focal, lo cual permitirá a cada persona tener la oportunidad de participar en la discusión. Existe un moderador que conduce la discusión de forma abierta y libre. Este moderador o facilitador requiere estar bien entrenado. Es función del moderador hacer que la discusión no se aleje demasiado del tema de interés. El tema puede ser cualquiera. Las respuestas de los entrevistados no son solicitadas de forma activa. No se dan sugerencias directas. En investigación de mercado o de consumo el tema se referiría a un producto o servicio. En psicología el interés sería, por ejemplo, el lenguaje utilizado por hombres homosexuales afroamericanos (Mays, Cochran, Bellinger, Smith, Henley *et al.*, 1992). En el área de salud, un grupo focal se utilizaría para determinar los temores acerca de los cinturones de seguridad o las bolsas de aire. Una de las metas consiste en examinar las actitudes y el comportamiento de la gente. La otra meta es descubrir lo que cada participante piensa sobre el tema que se discute. Las opiniones y descripciones surgen de los entrevistados. El investigador espera ser capaz de descubrir, a través de las discusiones, los discernimientos importantes que después sirvan para resolver problemas. Calder (1977) afirma que el método del grupo focal es útil para descubrir infor-

mación que se utilice para diseñar un estudio cuantitativo de investigación. Algunos han utilizado el grupo focal como un medio para desarrollar cuestionarios. La investigación de grupo focal también ayuda a los investigadores a desarrollar constructos que empleen en estudios futuros. Calder lo llama "conocimiento precientífico".

Una de las ventajas de los grupos focales es su costo, pues cuesta muy poco organizarlos. Los mayores costos residirían en conseguir y pagar al moderador. Además, los participantes podrían recibir un pago simbólico por su tiempo. El grupo focal también se realiza de forma rápida. Se dispone de las ideas de los entrevistados rápidamente y se realiza una videofilmación de las sesiones para analizarlas después con mayor profundidad. Es muy bueno para generar hipótesis para posteriores investigaciones. En la investigación de mercado, el grupo focal permite al cliente (fabricante) que encargó el estudio, ser un participante activo en la participación grupal. Así, dicha persona es capaz de obtener la información de primera mano. Lo anterior es posible debido a que los grupos se organizan de un tamaño que sea manejable. La interacción entre los entrevistados puede generar intercambios estimulantes que resulten en información útil, que no se obtiene con otros métodos de investigación. Además, como se mencionó antes, los grupos focales son muy flexibles. Un moderador experto va dirigiendo, y aun permitiendo que ideas prometedoras fluyan.

Sin embargo, el grupo focal no es muy recomendable para producir información concreta. Una decisión no debe basarse únicamente en la información reunida con dicho método. Además, ha sido criticado por los investigadores cuantitativos como "no científico" e indigno de confianza. Las preguntas no son estandarizadas y pueden variar de un grupo a otro. Con el uso de grupos muy pequeños, los datos de los grupos focales sufren en su posibilidad de generalización. A diferencia de la investigación de encuesta estructurada, el grupo focal no implica mucho esfuerzo por asegurarse de que el grupo sea representativo. Como sucede en la dinámica de cualquier grupo, siempre habrá unos cuantos individuos que dominen la conversación. Entonces, el moderador necesita contar con suficiente experiencia para minimizar la situación sin cortar el flujo de comunicación. La entrevista de grupo focal requiere de mucha paciencia y habilidad. Berger (1991) ofrece algunas sugerencias útiles para el moderador. También bosqueja lo que debe contener un reporte sobre un grupo focal. Algunos participantes ven al grupo focal como una oportunidad para ventilar sus emociones. Por lo tanto, los temas sensibles no deben explorarse por medio de grupos focales. En la sección de sugerencias de estudio se presentan algunas muy buenas referencias sobre los grupos focales. Éstos constituyen un método cualitativo de investigación, y como tales son capaces de ofrecer información rica que no pueden explotar los métodos cuantitativos. Son muy adecuados para conocer lo que desean los clientes o lo que la gente piensa acerca de ciertas políticas y reglas. Los grupos enfocados han probado su eficacia en el estudio de organizaciones.

Algunos ejemplos de investigación de grupos focales

Audience Studies Incorporated (ASI) es una compañía de investigación de mercados que ha operado a las afueras de Hollywood, California, durante muchos años. Se invita a los consumidores a una demostración, en la que se presenta un programa y comerciales de televisión. Por su participación se sortean premios tales como champú, crema dental, analgésicos, etcétera. Durante la demostración del programa y de los comerciales, los participantes utilizan un dispositivo electrónico de calificación para comunicar sus puntos de vista acerca de lo que están viendo. Las respuestas se graban. Los participantes también completan cuestionarios después de cada comercial o programa. A partir de este grupo se elige a varias personas para participar en grupos focales. Los fabricantes de productos generalmente comisionan a ASI para que conduzca estos grupos focales para obtener ideas sobre cómo funciona su producto en relación con la competencia. En varias ocasiones

participan representantes del fabricante en los grupos focales, para obtener información de manera directa. Por ejemplo, si un fabricante está pensando en desarrollar un nuevo producto, la información obtenida de los grupos focales brinda ideas sobre lo que debe llevar el producto en términos de manufactura y mercadotecnia.

Mays *et al.* (1992) utilizó un grupo focal que incluía hombres homosexuales afroamericanos. El tema de discusión era la conducta sexual y el VIH. Con este método, Mays *et al.* fueron capaces de recopilar el argot que emplean los varones afroamericanos homosexuales. Estos resultados son útiles para comparar a los hombres homosexuales afroamericanos con homosexuales americanos blancos, y también para construir cuestionarios diseñados para descubrir la conducta sexual de varones homosexuales afroamericanos. El conocimiento obtenido a partir de dicho estudio también serviría para educar a consejeros y profesionales de la salud que tratan con homosexuales afroamericanos. Mays y sus colaboradores (1992, p. 432) afirman lo siguiente:

> Al utilizar la terminología presentada aquí, en la conducción de investigación relacionada con el VIH, es importante recordar que los procesos lingüísticos y cognitivos están inmersos en un contexto. Al evaluar la conducta sexual de hombres homosexuales negros, el planteamiento de preguntas que incluyen su argot también debe originarse desde el marco de referencia de su experiencia.

Sussman, Burton, Dent, Stacy y Flay (1991) publicaron que se debe tener precaución con el uso de los grupos focales, pues consideran que tales grupos pueden inducir ciertos *efectos grupales* que sesguen las respuestas. Su estudio exploró el extenso procedimiento de los grupos focales, que incluye un cuestionario previo de grupo. El cuestionario tiene material que se cubrirá durante la sesión del grupo y puede afectar a los miembros del grupo al comprometerlos con una posición antes de que comience la discusión grupal. Estos investigadores consideran que la gente depende de las respuestas de otras personas, y de esta manera convergen en una norma colectiva. Es decir, algunos entrevistados tendrán juicios más extremos después de las discusiones grupales.

Uno de los efectos de las normas colectivas es el *efecto de polarización del grupo*. El involucrarse en un grupo puede sesgar a los participantes a responder de maneras más extremas. Específicamente Sussman *et al.* (1991) buscaron una polarización de actitudes (un efecto de sesgo por influencia del grupo). La discusión en el grupo focal se dirigió hacia cómo reclutar adolescentes que consuman tabaco para una clínica contra el tabaquismo. Se utilizaron 31 grupos focales; a cada uno se les administraron cuestionarios de pretest y de postest. Los datos obtenidos apoyaron la existencia de un efecto de polarización del grupo. Después de participar en un grupo focal, los entrevistados manifestaron una evaluación más alta de las estrategias de reclutamiento autogeneradas. También reportaron que si fuesen fumadores, dichas estrategias los inducirían a unirse al programa. El estudio demostró que los grupos focales podrían no generan nuevas estrategias. Sin embargo, sí parecen ser efectivos en inducir en los participantes una actitud más favorable hacia las soluciones autogeneradas de problemas.

RESUMEN DEL CAPÍTULO

1. La entrevista constituye el método más antiguo y universal para extraer grandes cantidades de información de la gente.
2. Los métodos de recolección de datos utilizados en una entrevista pueden ser clasificados de acuerdo a qué tan directos son en sus preguntas y planteamientos.

3. Las entrevistas requieren de mucho tiempo. Por lo tanto, la recolección de datos es costosa en términos de tiempo, esfuerzo y dinero.
4. Las entrevistas requieren de entrevistadores entrenados y de un cuestionario bien desarrollado.
5. La entrevista se utiliza para tres propósitos principales: como un dispositivo exploratorio para generar ideas e hipótesis, como el instrumento principal utilizado en un estudio y como complemento para otros métodos y/o como instrumento de seguimiento.
6. La entrevista es una situación interpersonal cara a cara. Existen dos tipos: estructurada o no estructurada.
7. Un tipo de entrevista no estructurada es la entrevista de grupo o grupos focales.
8. Se buscan tres tipos de información en los inventarios de entrevista: de identificación, de tipo censal (sociológica) y del problema.
9. Los tipos de reactivos utilizados en un inventario de entrevista son: reactivos de alternativa fija, reactivos abiertos y reactivos de escala.
10. Existen siete criterios para la redacción de reactivos y preguntas en el inventario.
11. El método del grupo focal es una entrevista no estructurada que utiliza un número pequeño de participantes. Es de bajo costo y de rápida realización.
12. La investigación de grupos focales es de tipo cualitativo.
13. Los grupos focales tienen un problema de generalización.

SUGERENCIAS DE ESTUDIO

1. A continuación se incluyen varias referencias valiosas sobre la entrevista y unas cuantas sobre el inventario de entrevista.

Obras clásicas

Cannell, C. y Kahn, R. (1968). Interviewing, en G. Lindzey y E. Aronson (eds.), *The handbook of social psychology*, vol. II (2a. ed.). Reading, MA: Addison-Wesley, 526-595.

Survey Research Center. (1976). *Interviewer's manual* (ed. rev.). Ann Arbor, Michigan: Institute for Social Research, University of Michigan. [Una excelente guía sobre los aspectos prácticos de la entrevista.]

Warwick, D. y Lininger, C. (1975). *The sample survey: Theory and practice.* Nueva York: McGraw-Hill.

Trabajos más recientes

Beed, T. W. y Stimson, R. J. (1985). *Survey Interviewing: Theory and techniques.* Nueva York: Routledge, Chapman, and Hall.

Bowden, J. C. (1995). *An investigator's guide to interviewing and interrogation.* Orlando, Florida: Bowden.

Knale, S. (1996). *Interviews: An introduction to qualitative research interviewing.* Thousand Oaks, California: Sage.

Lukas, S. (1993). *Where to start and what to ask: The assessment handbook.* Nueva York: Norton.

Mollica, R. F. y Caspi-Yavin, Y. (1991). Measuring torture and torture-related symptoms. *Psychological Assessment, 3*, 581-587. [Explica por qué los instrumentos y técnicas actuales de entrevista son inadecuados cuando se usan para entrevistar personas que han sido torturadas.]

Myers, J. (1996). *Interviewing young children about body touch and handling.* Chicago, Illinois: University of Chicago Press.

2. Por fortuna existen buenos inventarios de entrevista en abundancia. El lector debe estudiar uno o dos de ellos con cuidado. Los inventarios sugeridos a continuación están bien construidos y son muy interesantes. Note que los inventarios publicados casi siempre incluyen explicaciones metodológicas extensas. El estudiante aprenderá mucho sobre la construcción de escalas de entrevista con el estudio de este material.

Campbell, A., Converse, P. y Rodgers, W. (1976). *The quality of American life.* Nueva York: Russell Sage Foundation, app. B. [Un inventario grande con muchos reactivos de escala y cuidadosas instrucciones para el entrevistador. También es un estudio muy importante.]

Free, L. y Cantril, H. (1967). *The political beliefs of Americans.* New Brunswick, Nueva Jersey: Rutgers University Press, app. B. [Presenta buenas preguntas, sondeos y reactivos de alternativa fija.]

Glock, C. y Stark, R. (1966). *Christian beliefs and anti-Semitism.* Nueva York: Harper & Row. [Al final del libro se presenta el inventario completo, principalmente con reactivos de alternativa fija.]

3. Los ejemplos dados en el punto 2 son todos de investigación de encuesta, el campo de investigación donde el arte y la técnica de la entrevista se desarrolló y utilizó en primera instancia. Sin embargo, las entrevistas son y han sido utilizadas en lo que puede denominarse estudios "normales", es decir, estudios cuyo único o principal interés es encontrar relaciones entre variables. El estudio de Burt (1980) sobre las actitudes hacia la violación es un buen ejemplo; a continuación se mencionan otros ejemplos.

Estudios "normales"

Beckman, L. J. y Mays, V. M. (1985). Educating community gatekeepers about alcohol abuse in women: Changing attitudes, knowledge and referral practices. *Journal of Drug Education, 15,* 289-309. [Se utilizó una entrevista telefónica para evaluar el efecto de dos talleres sobre el conocimiento, actitudes y prácticas de referencia hacia las mujeres que sufren de abuso de alcohol.]

Campbell, A. y Schuman, H. (1968). *Racial Attitudes in fifteen cities.* Ann Arbor, Michigan: Institute for Social Research, University of Michigan. [Una combinación de "encuesta" y preguntas actitudinales enfocadas a la comprensión de las actitudes raciales y su cambio.]

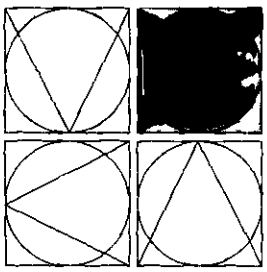
Doob, A. y MacDonald, G. (1979). Television viewing and fear of victimization: Is the relationship causal? *Journal of Personality and Social Psychology, 37,* 170-179. [Se realizó una entrevista de puerta en puerta para determinar si la gente que ve más la televisión tiene más temores que la gente que la ve menos. Se realizó una comparación entre aquellos que viven en un área de baja criminalidad y quienes viven en un área de alta criminalidad.]

Gersch, I. S. y Nolan, A. (1994). Exclusion: What the children think. *Educational Psychology in Practice, 10,* 35-45. [Se diseñó, aplicó y analizó un inventario de entrevista para medir las actitudes y experiencias de los niños hacia la escuela. Dicho instrumento se utilizó para evaluar a estudiantes que fueron excluidos.]

Jones, S. L. (1996). The association between objective and subjective care-giver burden. *Archives of Psychiatric Nursing, 10*, 77-84. [En tres momentos de recolección de datos se utilizaron entrevistas telefónicas para encontrar asociaciones entre las cargas objetivas y subjetivas de cuidadores.]

4. Los siguientes son libros y artículos que tratan con la teoría y práctica de grupos focales en ciencias sociales y del comportamiento. Los grupos focales con frecuencia sirven como generadores de ideas y para comprobar hipótesis en un ambiente informal. Consiga uno de ellos y lea los capítulos sobre metodología.

Berger, A. A. (1991). *Media research techniques.* Newbury Park, California: Sage.
Greenbaum, T. L. (1993). *The handbook for focus group research.* Nueva York: Lexington Books.
Morgan, D. L. (1988). *Focus groups as qualitative research.* Newbury Park, California: Sage.
Templeton, J. F. (1994). *The focus group: A strategic guide to organizing, conducting, and analyzing the focus group interview.* Chicago, Illinois: Probus Publishing.
Vaughn, S. (1996). *Focus group interviews in education and psychology.* Thousand Oaks, California: Sage.



CAPÍTULO 30

PRUEBAS Y ESCALAS OBJETIVAS

El método de observación y recolección de datos más utilizado en las ciencias del comportamiento es el de las pruebas y escalas. La gran cantidad de tiempo que pasan los investigadores en la construcción o búsqueda de medidas de variables es tiempo bien invertido, ya que la adecuada medición de las variables de investigación constituye uno de los aspectos nodales del trabajo científico sobre el comportamiento. En general, se ha puesto muy poca atención a la medición de las variables de estudios de investigación. ¿Qué utilidad pueden tener los intrigantes e importantes problemas de investigación, el sofisticado diseño de investigación y el intrincado análisis estadístico, si las variables de los estudios de investigación se miden pobremente? Por fortuna se ha progresado mucho en la comprensión de la teoría de la medición psicológica y educativa, así como en el mejoramiento de la práctica de la medición. En este capítulo se examina algo de la tecnología que subyace a los procedimientos objetivos de medición.

Objetividad y métodos objetivos de observación

La objetividad, una característica central y esencial de la metodología científica, es fácil de definir, aunque evidentemente difícil de entender. También es polémica. La objetividad es el acuerdo entre jueces expertos respecto a lo que se observa. Los métodos objetivos de observación son aquellos donde, cualquiera que siga las reglas prescritas, asignará los mismos valores numéricos a objetos y conjuntos de objetos como lo haría cualquier otra persona. Un procedimiento objetivo es aquel en el que el acuerdo entre los observadores se encuentra en su nivel máximo. En términos de varianza, la varianza de los observadores se encuentra en su nivel mínimo, lo cual quiere decir que la varianza de juicio, es decir, la varianza debida a las diferencias entre jueces en la asignación de valores numéricos a objetos, se aproxima a cero. Es posible encontrar una extensa discusión sobre la objetividad en Kerlinger (1979, pp. 9-13 y 262-264). La importancia de la comprensión de la objetividad en la ciencia no puede sobrestimarse. Es especialmente importante comprender que la objetividad científica es metodológica y tiene poco o nada que ver con la objetividad como una supuesta característica de los científicos. El hecho de que un científico como persona sea o no sea objetivo no es el punto importante. Lo importante es que la objetividad científica es inherente a los procedimientos metodológicos, caracterizados por el acuerdo entre jueces expertos —y nada más—.

Todos los métodos de observación son inferenciales: se realizan inferencias respecto a las propiedades de los miembros de conjuntos con base en los valores numéricos asignados a tales miembros, por medio de entrevistas, pruebas, escalas y observaciones directas del comportamiento. Los métodos difieren respecto a si son directos o indirectos, en el grado en que las inferencias son realizadas a partir de las observaciones en bruto. Las inferencias realizadas por medio de métodos objetivos de observación, por lo general, son muy largas, a pesar de que aparentan ser directas. La mayoría de dichos métodos permiten un alto grado de acuerdo entre los observadores, ya que los participantes registran marcas en papel y tales marcas están restringidas a dos o más opciones de alternativas dadas por el observador. A partir de las marcas en el papel, el observador infiere las características de los individuos y conjuntos de individuos que hacen las marcas. En un tipo de métodos objetivos, el observador (o juez) hace las marcas en papel, observa al objeto o los objetos de medición y elige entre las alternativas dadas. En tal caso, también se realizan inferencias sobre las propiedades del objeto o los objetos observados, a partir de las marcas en papel. La principal diferencia reside en quién hace las marcas.

Debe reconocerse que todos los métodos de observación poseen cierta objetividad. No existe una clara dicotomía, en otras palabras, entre los llamados métodos objetivos y otros métodos de observación. Más bien existe una diferencia en el grado de objetividad. Nuevamente, si se piensa en los grados de objetividad como grados de acuerdo entre observadores, desaparece la ambigüedad y confusión que con frecuencia se asocia con el problema.

Entonces existe el acuerdo de que lo que aquí se llama métodos objetivos de observación y medición no poseen el monopolio de objetividad ni de inferencia, sino que son más objetivos y no menos inferenciales que cualquier otro método de observación y medición. Los métodos que se expondrán en el presente capítulo por ningún motivo abarcan todos los métodos posibles, pues el tema es grande y muy variado. Se consideran únicamente como medidas de variables, vistas y evaluadas de la misma forma que todas las demás medidas de variables.

Pruebas y escalas: definiciones

Una *prueba* es un procedimiento sistemático en el que se presenta a los individuos un conjunto de estímulos construidos a los cuales responden. Las respuestas permiten que quien realiza la prueba asigne a los examinados valores numéricos o conjuntos de valores numéricos, a partir de los cuales se pueden realizar inferencias sobre si el examinado posee aquello que se supone que la prueba está midiendo. Esta definición dice un poco más que la que afirma que la prueba es un instrumento de medición.

Una *escala* es un conjunto de símbolos o valores numéricos, construida de tal manera que los símbolos o valores numéricos puedan ser asignados por una regla a los individuos (o a sus comportamientos) a quienes se aplica la escala, y donde la asignación indica si el individuo posee lo que se supone que mide la escala. Al igual que una prueba, una escala es un instrumento de medición. De hecho, con excepción del significado excesivo asociado con la prueba, se observa que prueba y escala se definen de manera similar. Sin embargo, estrictamente hablando, la escala se utiliza de dos maneras: para indicar un instrumento de medición y para indicar los valores numéricos sistematizados del instrumento de medición. Se utiliza en ambos sentidos sin demasiada preocupación por las distinciones. No obstante, se debe recordar que: *las pruebas son escalas, pero las escalas no necesariamente son pruebas*. Es así porque las escalas, por lo general, no tienen el significado de competencia ni de éxito o fracaso que tienen las pruebas.

Tipos de medidas objetivas

La mayor parte de los cientos, quizás miles, de pruebas y escalas pueden dividirse en las siguientes clases: pruebas de inteligencia y aptitud, pruebas de rendimiento, medidas de personalidad, escalas de actitud y valores, y medidas objetivas diversas. A continuación se discutirá cada uno de estos tipos de medida, desde un punto de vista de investigación.

Pruebas de inteligencia y aptitud

En la investigación psicológica y educativa con frecuencia se necesita de una medida de inteligencia o aptitud, ya sea como variable independiente o como variable dependiente. Para evaluar los efectos de programas educativos de uno u otro tipo sobre el rendimiento académico, por ejemplo, generalmente es necesario controlar la inteligencia, de tal manera que las diferencias encontradas entre los grupos de tratamiento experimental no puedan atribuirse a diferencias en la inteligencia, más que a los tratamientos. Existe un gran número de buenas pruebas de inteligencia que utilizan los investigadores, quizás sea el caso de una de las llamadas pruebas de recopilación (ómnibus). (Una prueba de *recopilación* es aquella que tiene diferentes tipos de reactivos: verbales, numéricos, espaciales y otros, en un instrumento.) Dichas pruebas por lo general son altamente verbales y se correlacionan de forma sustancial con el rendimiento académico. Los manuales de Buros (1998) son guías útiles para dichas pruebas. Anastasi (1988) ofrece una lista clasificada de pruebas representativas en su libro, y divide cada prueba en clasificaciones separadas, tales como de inteligencia, de personalidad, etcétera.

Una *aptitud* es la habilidad potencial para el logro. Las pruebas de aptitud se utilizan principalmente para orientación y consejería. También se emplean en investigación, particularmente como variables control. Una *variable control* es aquella cuyo efecto sobre una variable dependiente quizá requiera anularse. Por ejemplo, al estudiar el efecto de un programa correctivo de lectura sobre el rendimiento en lectura, puede ser necesario atribuir la aptitud verbal a las posibles diferencias de grupo en habilidad verbal. De forma

similar, tal vez sea necesario controlar otras variables de influencia potencial: habilidades numéricas y espaciales, por ejemplo. Las pruebas de aptitud resultan útiles en tales casos.

Pruebas de rendimiento

Las *pruebas de rendimiento* miden la eficiencia, maestría y comprensión presentes, en áreas generales y específicas del conocimiento. En su mayoría, constituyen mediciones de la eficacia de la instrucción y el aprendizaje y son, por supuesto, enormemente importantes en educación y en la investigación educativa. De hecho, como se mencionó antes, con frecuencia el rendimiento es la variable dependiente en investigaciones que incluyen métodos de instrucción.

Las pruebas de rendimiento a menudo se clasifican de varias formas. Para los propósitos de este libro, se dividen, primero, en pruebas estandarizadas y en pruebas de construcción especial. Las *pruebas estandarizadas* son grupos de pruebas ya publicadas, que se basan en un contenido educativo general común a un gran número de sistemas educativos. Son los productos de un alto grado de competencia y habilidad profesional en la redacción de pruebas y, como tales, por lo general son bastante confiables y casi siempre válidas. Están dotadas con minuciosas tablas de normas (promedios) que pueden utilizarse con propósitos comparativos. Las *pruebas de construcción especial* son generalmente pruebas realizadas *ex profeso* por maestros para medir logros más específicos y limitados. Por supuesto, también pueden ser elaboradas por investigadores educativos para medir áreas limitadas de rendimiento.

En segundo lugar, las pruebas estandarizadas de rendimiento pueden, a su vez, clasificarse en pruebas generales y especiales. Las *pruebas generales* son baterías de pruebas que miden las áreas más importantes del rendimiento académico: uso del lenguaje, vocabulario, lectura, aritmética y estudios sociales. Las pruebas de rendimiento especial, como su nombre lo indica, son pruebas de materias individuales tales como historia, ciencia e inglés.

Los investigadores rara vez eligen las pruebas de rendimiento debido a que los sistemas escolares son quienes las seleccionan. Sin embargo, cuando a los investigadores se les da la oportunidad de elegir, deben evaluar cuidadosamente el tipo de prueba de rendimiento que requiere su problema de investigación. Suponga que la variable de investigación en un estudio es "el rendimiento en la comprensión de conceptos". Muchas, quizás la mayoría de las pruebas utilizadas en las escuelas, no son adecuadas para medir dicha variable. En tales casos, los investigadores pueden elegir una prueba especialmente diseñada para medir la comprensión de conceptos, o ellos mismos diseñar la prueba. La construcción de una prueba de rendimiento es un trabajo enorme, aunque no es posible comentar aquí los detalles. Se refiere al estudiante a pruebas especializadas. Por desgracia, existen pocos libros sobre la construcción de pruebas y escalas para propósitos de investigación. La mayoría de los libros y otros textos sobre medición se enfocan, casi siempre, en la construcción y el uso de instrumentos con propósitos de aplicación. Sin embargo, los investigadores que necesiten construir medidas de rendimiento de cualquier tipo encontrarán una excelente guía en los trabajos de Adkins (1974); Cangelosi (1990); Gronlund (1988); Haladyna (1997); Hopkins (1989), y Osterlind (1989). Los investigadores que necesiten construir escalas de actitud encontrarán en el libro de Edwards (1957) un documento invaluable. Dawes (1972) cubre algunos métodos que Edwards no cubrió.

Medidas de personalidad

La medición de rasgos de personalidad constituye el problema más complejo de la medición psicológica. La razón es simple: la personalidad humana es extremadamente compleja. Para propósitos de medición, la personalidad puede considerarse como la organización particular de los rasgos del individuo. Un *rasgo* es una característica de un individuo, revelada

a través de comportamientos recurrentes en situaciones diferentes. Se dice que un individuo es compulsivo o que posee el rasgo de compulsividad debido a que se observa que esa persona es notoriamente pulcra tanto en su vestimenta como en su lenguaje, siempre es puntual, desea que todo esté muy ordenado, y le disgustan y evita las irregularidades.

El principal problema en la medición de la personalidad es la validez. Medir la validez de los rasgos de personalidad requiere que se conozca qué son tales rasgos, la forma en que interactúan y cambian, y cómo se relacionan entre sí, lo cual constituye un requisito muy grande e incluso severo. La cuestión no es, como les gusta señalar a los críticos ingenuos, que la personalidad no pueda medirse debido a que es demasiado esquiva, demasiado compleja, demasiado existencial, o que los esfuerzos de medición no hayan tenido éxito, sino más bien que se ha logrado cierto grado de éxito en una tarea tan difícil. No obstante, el problema de la validez es considerable.

Existen dos modelos generales para la construcción y validación de medidas de personalidad: el *método a priori* y el *método teórico* o *de constructo*. En el *método a priori*, los reactivos se construyen para reflejar la dimensión de personalidad que se mide. Puesto que el introvertido con frecuencia es una persona que se aísla, quizá se redacten reactivos sobre la preferencia por estar solo. Por ejemplo, esto podría incluir reactivos que indiquen la preferencia por evitar fiestas para medir la introversión. Como la persona ansiosa estará nerviosa y desorganizada cuando se encuentre bajo estrés, se podrían redactar reactivos que sugieran estas condiciones, para medir la ansiedad. Entonces, en el método *a priori* el redactor de la escala reúne o redacta reactivos que midan, de forma ostensible, rasgos de personalidad.

Éste es esencialmente el modelo de los primeros redactores de pruebas de personalidad. Aunque no existe nada inherentemente malo con el método —de hecho, tendrá que utilizarse, especialmente en las primeras etapas de la construcción de pruebas y escalas— los resultados llegan a ser confusos. Los reactivos no siempre miden lo que se cree que están midiendo. En algunas ocasiones incluso se descubre que un reactivo que se pensaba que medía, por ejemplo, responsabilidad social, en realidad mide una tendencia a estar de acuerdo con afirmaciones socialmente deseables. Por tal razón, el método a priori, utilizado de forma única, resulta insuficiente. Por lo anterior, las pruebas de personalidad generalmente carecen de validez de contenido.

El método de validación que se utiliza a menudo con las escalas de personalidad *a priori* es el *método de grupo conocido*. Para validar una escala de responsabilidad social, es posible buscar un grupo de individuos con un alto nivel conocido de responsabilidad social, y otro con un bajo nivel conocido de responsabilidad social. Si la escala logra diferenciar con éxito a ambos grupos, entonces se considera que tiene validez.

La medida de personalidad *a priori* y otras medidas continuarán empleándose en la investigación del comportamiento. Sin embargo, debe evitarse su uso ingenuo y a ciegas. Debe verificarse su validez de constructo y su validez relacionada con el criterio, especialmente por medio del análisis factorial y otras formas empíricas. Las medidas de personalidad, al igual que otras medidas, se han utilizado con frecuencia tan sólo porque quienes las usan piensan que miden aquello que afirman que están midiendo.

El *método teórico* o *de constructo* de la construcción de medidas de personalidad enfatiza las relaciones de la variable medida con otras variables, las relaciones surgidas de la teoría que subyace a la investigación. Aunque la construcción de escalas debe ser siempre, en cierto grado, *a priori*, mientras mayor sea el número de medidas de personalidad sujetas a las pruebas de validez de constructo, mayor será la confianza en su fidelidad. No es suficiente aceptar simplemente la validez de una escala de personalidad ni aun aceptar su validez debido a que ha logrado diferenciar con éxito, por ejemplo, a artistas de científicos, a maestros de no maestros, a personas normales de personas neuróticas. A final de cuentas,

debe establecerse su validez de constructo, es decir, su uso exitoso en la predicción de una gran variedad de relaciones teóricas.

Escalas de actitud

Las actitudes, aunque se tratan de forma separada aquí y en la mayor parte de los análisis de libros de texto, en realidad son parte integral de la personalidad. Los teóricos modernos también consideran la inteligencia y la aptitud como partes de la personalidad. Sin embargo, la medición de la personalidad trata principalmente de rasgos. Un *rasgo*, como se mencionó en la sección anterior, es una característica relativamente perdurable del individuo, a responder de cierta manera en todas las situaciones. Si alguien es dominante, exhibirá comportamiento dominante en la mayoría de las situaciones. Si alguien es ansioso, presentará una conducta ansiosa en la mayor parte de sus actividades. Por otro lado, una *actitud* es una predisposición organizada a pensar, sentir, percibir y comportarse hacia un referente u objeto cognitivo. Se trata de una estructura perdurable de creencias que predispone al individuo a comportarse de manera selectiva hacia los referentes de actitud.

Un *referente* es una categoría, una clase o un conjunto de fenómenos: objetos físicos, eventos, conductas e inclusive constructos (véase Brown, 1958). La gente tiene actitudes hacia muchas cosas diferentes: grupos étnicos, instituciones, religión, aspectos y prácticas educativas, la Suprema Corte, derechos civiles, propiedad privada, etcétera. Dicho en otras palabras, se tienen actitudes hacia algo "de afuera". Un rasgo tiene una referencia subjetiva; una actitud tiene una referencia objetiva. Alguien que tiene una actitud hostil hacia los extranjeros quizá sea hostil únicamente con los extranjeros; pero alguien que tiene un rasgo de hostilidad es hostil hacia todas las personas (al menos potencialmente).

Existen tres tipos principales de escalas de actitud: escalas de puntuación sumada, escalas de intervalos aparentemente iguales y escalas acumulativas (o Guttman). Una escala de puntuación sumada (un tipo de las cuales se denomina escala tipo Likert) es un conjunto de reactivos de actitud, todos los cuales son considerados con un "valor de actitud" aproximadamente igual, y donde cada uno de los participantes responde con grados de acuerdo o desacuerdo (intensidad). Las puntuaciones de los reactivos de dicha escala se suman, o se suman y promedian, para producir una puntuación de actitud del individuo. Como en todas las escalas de actitud, el propósito de la escala de puntuación sumada es ubicar a un individuo en algún punto de un continuo del nivel de acuerdo de la actitud en cuestión.

Es importante hacer notar dos o tres características de las escalas de puntuación sumada, pues muchas escalas comparten estas características. Primero, U, el universo de reactivos, se considera un conjunto de reactivos con igual "valor de actitud", como se indicó en la definición anterior. Ello significa que no existe una escala de reactivos como tal; un reactivo es igual a cualquier otro reactivo respecto a su valor de actitud. Se "clasifica" a los individuos que responden los reactivos. La clasificación surge de las sumas (o promedios) de las respuestas de los individuos. Cualquier subconjunto de U es teóricamente igual que cualquier otro subconjunto de U: se ordenaría a un conjunto de individuos por rango de la misma manera si se utiliza U_2 o U_1.

En segundo lugar, las escalas de puntuación sumada permiten la expresión de la intensidad de la actitud. Los participantes pueden estar simplemente de acuerdo o estar fuertemente de acuerdo. Esto conlleva ventajas, así como desventajas. La principal ventaja es que resulta una mayor varianza. Cuando existen cinco o siete categorías posibles de respuesta, resulta obvio que la varianza de la respuesta debe ser mayor que con sólo dos o tres categorías (por ejemplo, de acuerdo, en desacuerdo y sin opinión). Por desgracia, la varianza de las escalas de puntuación sumada con frecuencia parece contener varianza debida a la fijeza de respuesta. Los individuos tienen tendencias diferentes a usar ciertos tipos

de respuestas: respuestas extremas, respuestas neutrales, respuestas en acuerdo y respuestas en desacuerdo. Dicha varianza de respuesta confunde la varianza de la actitud (y el rasgo de personalidad). Las diferencias individuales producidas por las escalas de actitud de puntuaciones sumadas (y medidas de rasgos calificadas de manera similar) han mostrado deberse, en parte, a la fijeza de respuesta y a otras extrañas fuentes de varianza similares. La literatura sobre la fijeza de respuesta es muy amplia y no puede citarse en detalle. La exposición de Nunnally (1978) está bien equilibrada. Mientras que la fijeza de respuesta puede considerarse como una ligera amenaza para la validez de la medición, su importancia ha sido sobrestimada y la evidencia disponible no justifica las fuertes afirmaciones negativas hechas por los partidarios de la fijeza de respuesta. En otras palabras, mientras se debe estar consciente de las posibilidades y amenazas, no hay que paralizarse por el incremento del riesgo (véase Rorer, 1965, para un análisis más profundo).

A continuación se presentan dos reactivos de puntuación sumada de una escala construida por Burt (1980, p. 222) para el estudio de actitudes hacia la violación. Tales reactivos se desarrollaron para medir el estereotipo del rol sexual. Se utilizó una escala de 7 puntos que va desde un fuerte acuerdo (7) hasta un fuerte desacuerdo (1). Los valores entre paréntesis (y los valores intermedios) se asignan a las respuestas indicadas.

Hay algo malo con la mujer que no desea casarse y formar una familia.
Una mujer debe ser virgen cuando se casa.

Las escalas de intervalos aparentemente iguales de Thurstone se construyen a partir de principios diferentes. Mientras que el producto final, un conjunto de reactivos de actitud, puede utilizarse para el mismo propósito de asignación de puntuaciones individuales de actitud, las escalas de intervalos aparentemente iguales también logran el importante propósito de escalar los reactivos de actitud. A cada reactivo se le asigna un valor de escala que indica la fortaleza de actitud de una respuesta de acuerdo para el reactivo. El universo de reactivos se considera un conjunto ordenado; es decir, los reactivos difieren en su valor de escala. El procedimiento de clasificación encuentra estos valores de escala. Además, los reactivos de la escala final a utilizarse son tan selectos que los intervalos entre ellos son iguales, lo cual representa una importante y deseable característica psicométrica.

Los siguientes reactivos de intervalos aparentemente iguales, con los valores de escala de los reactivos, provienen de la escala de Actitud hacia la Iglesia, de Thurstone y Chave (1929, pp. 61-63, 78):

Considero que la Iglesia es la mayor institución en Estados Unidos hoy. (Valor de escala: 0.2)
Creo en la religión, pero en raras ocasiones voy a la iglesia. (Valor de escala: 5.4)
Pienso que la Iglesia es un obstáculo para la religión, ya que aun depende de lo sobrenatural, la superstición y el mito. (Valor de escala: 9.6)

En la escala de Thurstone y Chave, a menor valor de escala del reactivo, más positiva sería la actitud hacia la Iglesia. El primer y tercer reactivos son respectivamente el más bajo y el más alto en la escala. El segundo reactivo, por supuesto, tiene un valor intermedio. La escala total contenía 45 reactivos con valores de escala a través de todo el continuo. Sin embargo, por lo general, las escalas de intervalos aparentemente iguales contienen bastante menos reactivos.

El tercer tipo de escala, la *escala acumulativa* o *Guttman*, consta de un conjunto relativamente pequeño de reactivos homogéneos que son unidimensionales (o que supuestamente lo son). Una escala unidimensional mide una variable y sólo una. La escala obtiene su nombre de la relación acumulativa entre los reactivos y las puntuaciones totales de los

individuos. Por ejemplo, a cuatro niños se les plantean tres preguntas aritméticas: (a) 28/7 = ?, (b) 8 × 4 = ? y (c) 12 + 9 = ? El niño 1, quien resuelve (a) correctamente, tiene muchas posibilidades de resolver (b) y (c) también correctamente. El niño 2, quien resuelve (a) de forma incorrecta pero (b) de forma correcta, también tiene muchas posibilidades de resolver (c) correctamente. El niño 3, quien resuelve correctamente sólo (c), tiene muy pocas probabilidades de resolver (a) y (b) de forma correcta. La situación se sintetiza de la siguiente manera (la tabla incluye la puntuación del cuarto niño, quien no resuelve ninguna pregunta de forma correcta):

	(a)	(b)	(c)	Puntuación total
Niño 1	1	1	1	3
Niño 2	0	1	1	2
Niño 3	0	0	1	1
Niño 4	0	0	0	0

(1 = Correcto; 0 = Incorrecto)

Note la relación entre el patrón de respuestas a los reactivos y las puntuaciones totales. Si se conoce la puntuación total de un niño, se puede predecir su patrón, si la escala es acumulativa, solamente si el conocimiento de las respuestas correctas de los reactivos más difíciles predice las respuestas de los reactivos más fáciles. Observe también que se escalan ambos reactivos y las personas.

De manera similar, es posible plantear a la gente varias preguntas acerca de un objeto actitudinal. Si bajo análisis los patrones de respuesta se ordenan entre sí de la forma indicada arriba (o por lo menos de manera muy similar), entonces se dice que las preguntas o reactivos son unidimensionales. De esta manera, las personas pueden ser ordenadas de acuerdo con sus respuestas en la escala (véase Edwards, 1957, capítulo 7, para una mayor discusión sobre las escalas acumulativas unidimensionales).

Resulta obvio que estos tres métodos de construcción de escalas de actitud son muy diferentes. Considere que métodos similares o iguales pueden utilizarse con otros tipos de escalas de personalidad u otras escalas. La escala de puntuación sumada se concentra en los participantes y sus ubicaciones dentro de la escala. La escala de intervalos aparentemente iguales se concentra en los reactivos y sus ubicaciones dentro de la escala. De manera interesante, ambos tipos de escalas producen casi los mismos resultados en lo que se refiere a la confiabilidad y a la ubicación de los individuos en órdenes de rango actitudinales. Las escalas acumulativas se concentran en la posibilidad de escalar de los conjuntos de reactivos y en la posición de los individuos en la escala.

De los tres tipos de escalas, la de puntuaciones sumadas parece ser el más útil para la investigación del comportamiento, pues es más fácil de desarrollar y, como se indicó antes, produce casi los mismos resultados que la escala de intervalos aparentemente iguales, de construcción más laboriosa. Utilizadas con cuidado y con el conocimiento de sus debilidades, las escalas de puntuaciones sumadas se adaptan a muchas de las necesidades de los investigadores del comportamiento. Las escalas acumulativas parecen ser de menor utilidad y menos aplicables de manera general. Si se utiliza un objeto cognitivo de corte claro, una escala acumulativa breve y bien construida genera medidas confiables de un número de variables psicológicas: tolerancia, conformismo, identificación grupal, aceptación de la autoridad, permisividad, etcétera. También debe señalarse que el método puede mejorarse y alterarse de varias formas. Dawe (1972) y Edwards (1957) describen la forma de construir y evaluar escalas acumulativas, así como escalas de puntuación sumada y escalas de intervalos aparentemente iguales.

Escalas de valores

Los *valores* son preferencias de tipo cultural hacia objetos, ideas, gente, instituciones y comportamientos (Kluckhohn, 1951, pp. 388-433). Mientras que las actitudes son organizaciones de creencias sobre cosas "de afuera", es decir, predisposiciones a comportarse hacia los objetos o referentes de actitudes, los valores expresan preferencias por formas de conducta y estados de existencia (Rokeach, 1968). Términos como *igualdad, religión, libre empresa, derechos civiles* y *obediencia* expresan valores. Dicho de manera sencilla, los valores expresan lo "bueno", lo "malo", los "debería" y los "debiera" del comportamiento humano. Los valores colocan las ideas, las cosas y los comportamientos en un continuo de aprobación-desaprobación. Implican opciones entre cursos de acción y de pensamiento.

Con el propósito de brindar al lector una idea de los valores, se presentan tres reactivos. Se puede pedir a los individuos que expresen su aprobación o desaprobación sobre el primero y segundo reactivos, quizás en forma de puntuación sumada, y que elijan una de las tres alternativas del tercer reactivo.

> Por el propio bien y por el bien de la sociedad, una persona debe ser controlada por la tradición y la autoridad.
> Ahora más que nunca debemos fortalecer la familia, el estabilizador natural de la sociedad.
> ¿Cuál de los siguientes aspectos es el más importante para desarrollar una vida plena: la educación, el logro o la amistad?

Por desgracia, los valores han recibido escasa atención científica, aun cuando éstos y las actitudes conforman gran parte de la producción verbal de las personas y son, probablemente, influyentes determinantes del comportamiento. Por lo tanto, la medición de valores padece desatención. Sin embargo, los valores sociales y educativos tal vez se conviertan en el centro de mucho más trabajo teórico y empírico en el futuro, ya que los científicos sociales se han vuelto cada vez más conscientes de que los valores constituyen influencias importantes en el comportamiento individual y de grupo (véase Dukes, 1955; Haddock y Zanna, 1998; Hendrick, Hendrick y Dicke, 1998: Hogan, 1973; Lubinski, Schmidt y Benbow, 1996; Pittel y Mendelsohn, 1966; Robinson, 1996). Una fuente de escalas de valores es Levitin (1969). Un ensayo muy sugestivo y valioso que apareció hace 45 años es el trabajo realizado por Kluckhohn (1951). Otro ensayo sobre la medición de valores que aún es importante es el de Thurstone (1959).

Tipos de escalas y reactivos objetivos

Dos tipos generales de reactivos que se usan con frecuencia son aquellos en que las respuestas son independientes y aquellos en que no son independientes. *Independencia* aquí significa que la respuesta de una persona a un reactivo no está relacionada con su respuesta a otro reactivo. Todos los reactivos de verdadero-falso, *sí-no, de acuerdo-en desacuerdo* y de tipo Likert pertenecen al tipo independiente. El sujeto responde cada reactivo libremente, con un rango de dos o más respuestas posibles, de las cuales puede elegir sólo una. Por otro lado, los reactivos no independientes obligan al sujeto a elegir un reactivo o alternativa que excluye la elección de otros reactivos o alternativas. Tales formas de escalas y reactivos se denominan de elección forzada. El sujeto se enfrenta con dos o más reactivos o subreactivos y se le pide que elija uno o más de ellos de acuerdo con algún criterio, e incluso criterios.

Dos ejemplos simples mostrarán la diferencia entre reactivos independientes y no independientes. Primero, es posible dar al sujeto un conjunto de instrucciones que permitan

independencia de respuesta; en segundo lugar, se le puede indicar un conjunto de instrucciones contrastante, con opciones más limitadas (no independientes):

Ejemplos

> Indique junto a cada una de las siguientes afirmaciones qué tanto las aprueba, utilizando una escala del 1 al 5, donde el 1 significa "No lo apruebo en lo absoluto" y el 5 significa "Lo apruebo muchísimo".
>
> A continuación se dan 40 pares de afirmaciones. De cada par escoja la que apruebe más. Márquela con una palomita (✓).

Las ventajas de los reactivos independientes son la economía y aplicabilidad de la mayoría de los análisis estadísticos a sus respuestas. Además, cuando se responde cada reactivo se obtiene un máximo de información, donde cada reactivo contribuye a la varianza. También se requiere de menos tiempo para aplicar las escalas independientes, aunque quizá sufran del sesgo provocado por la fijeza de respuesta. Los individuos pueden dar respuestas similares o iguales para cada reactivo: pueden apoyarlos todos con entusiasmo o con indiferencia dependiendo de su predilección particular de respuesta. La varianza importante de una variable llega, entonces, a confundirse por la fijeza de respuesta.

El tipo de escala de elección forzada evita, por lo menos en cierto grado, el sesgo de respuesta. Sin embargo, al mismo tiempo, sufre por la falta de independencia, y por su costo excesivo y alta complejidad. No obstante, existen algunos investigadores, como Comrey (1970), que han construido una escala del sesgo de respuesta dentro de la prueba de personalidad. Las escalas de elección forzada en ocasiones también agotan la tolerancia y la paciencia del sujeto, lo cual se traduce en menor cooperación. Aun así, muchos expertos consideran que los instrumentos de elección forzada son prometedores para la medición educativa y psicológica. Mientras que otros expertos se muestran escépticos.

Entonces, las escalas y los reactivos se dividen en tres tipos: *de acuerdo-en desacuerdo* (o aprobación-desaprobación o *verdadero-falso*, y similares), de orden de rango y de elección forzada. Cada uno de ellos se analiza brevemente. En la literatura pueden encontrarse explicaciones más detalladas (véase Edwards, 1957; Guilford, 1954).

Reactivos de acuerdo-desacuerdo

Existen tres formas generales de reactivos de acuerdo y desacuerdo:

1. Aquellos que permiten una de dos respuestas posibles.
2. Aquellos que permiten una de tres o más respuestas posibles.
3. Aquellos que permiten más de una elección de tres o más respuestas posibles.

Las dos primeras formas proporcionan alternativas como "de acuerdo-desacuerdo"; "sí-no"; "lo apruebo-sin opinión-lo desapruebo"; "lo apruebo mucho–lo apruebo-lo desapruebo mucho"; "1, 2, 3, 4, 5". Los participantes eligen una de las respuestas proporcionadas para reportar sus reacciones a los reactivos. Al hacerlo, dan reportes sobre sí mismos o indican sus reacciones a los reactivos. La mayor parte de las escalas de personalidad y de actitud utilizan dichos reactivos. Si una persona está construyendo un instrumento que utilice este método, es muy importante la manera en que se redacte la escala. Por ejemplo, suponga que se elaboró la siguiente escala de Likert de 5 puntos:

1 = En desacuerdo
2 = En ligero desacuerdo
3 = Neutral

4 = En ligero acuerdo
5 = De acuerdo

El problema aquí se centra en la manera en que el lector interpreta el significado de cada punto de la escala. Un sujeto puede elegir un "2" y otro puede elegir un "4". Ambos pudieron haber hecho la misma interpretación. Si alguien está en ligero acuerdo, entonces esa persona quizá también esté en ligero desacuerdo. De ahí surge la confusión.

El tercer tipo de escala de este tipo presenta un número de reactivos: se dan instrucciones a los participantes para indicar aquellos reactivos que los describen, reactivos con los que están de acuerdo o simplemente reactivos que ellos eligen. El listado de adjetivos es un buen ejemplo. Se le presenta al sujeto una lista de adjetivos, donde algunos indican rasgos deseables como analítico, generoso y considerado; y donde otros indican rasgos indeseables como cruel, egoísta y vulgar. Se les pide que marquen aquellos adjetivos que los caracterizan. (Por supuesto, este tipo de instrumento también sirve para caracterizar a otras personas.) Quizás una forma mejor sería una lista con todos los adjetivos positivos de escalas de valores conocidas, donde se les pide a los participantes que seleccionen un número específico de sus propias características personales. La escala de intervalos aparentemente iguales y su sistema de respuesta donde se marcan los reactivos de actitud con los que se está de acuerdo es, por supuesto, la misma idea. La idea es útil, especialmente con el desarrollo de escalas factoriales, de métodos de escalación y el creciente uso de los métodos de elección.

La calificación de los reactivos de acuerdo-en desacuerdo llega a ser problemática debido a que no todos los reactivos, o los componentes, reciben respuestas. (Con una escala de puntuación sumada o una escala de puntuación ordinaria, los participantes generalmente responden a todos los reactivos.) Sin embargo, en general, se pueden utilizar sistemas simples de asignación de valores numéricos a las diversas opciones. Por ejemplo, de acuerdo-en desacuerdo puede ser 1 y 0; sí-no puede ser 1, 0, –1 o, evitando los signos negativos: 2, 1, 0. A las respuestas de los reactivos de puntuación sumada descritos anteriormente simplemente se les asigna del 1 al 5 o del 1 al 7.

El principal aspecto que los investigadores deben tener en mente es que el sistema de puntuación debe producir datos interpretables, congruentes con el sistema de puntuación. Si se utilizan puntuaciones de 1, 0, –1, los datos deben ser capaces de proveer una interpretación escalada; es decir, 1 es "alto" o "mucho", –1 es "bajo" o "poco" y 0 está en medio. Un sistema de 1, 0 puede significar alto y bajo o simplemente presencia o ausencia de un atributo. Tal sistema puede ser útil y poderoso, como se vio anteriormente cuando se estudiaron variables como sexo, raza, clase social, etcétera. En síntesis, los datos producidos por sistemas de puntuación deben tener significados claramente interpretables en cierto sentido cuantitativo. Se refiere al lector al análisis de Ghiselli (1964, pp. 44-49) sobre el significado de las puntuaciones. No obstante, algunos expertos han criticado el uso de 0-1 o de los sistemas binarios de puntuación. Durante el desarrollo de sus escalas de personalidad, Comrey descubrió que los reactivos que utilizan un esquema de respuesta binaria están sujetos a problemas y distorsiones que no necesariamente se obtendrían si la escala tuviera 3 puntos o más. Los resultados del estudio de Comrey sobre las escalas se resumen en Comrey (1978) y Comrey y Lee (1992).

Se han desarrollado varios sistemas para ponderar reactivos; aunque la evidencia indica que las puntuaciones ponderadas y no ponderadas dan en gran parte los mismos resultados. Los estudiantes parecen encontrar esto difícil de creer. (Note que se habla acerca de la ponderación de las respuestas a los reactivos.) Aunque el asunto no está completamente establecido, existe fuerte evidencia de que en pruebas y medidas con el suficiente número de reactivos —como 20 o más— la ponderación diferencial de reactivos no genera mucha diferencia en los resultados finales. Tampoco la ponderación diferencial de respuestas pro-

duce mucha diferencia (véase Guilford, 1954; Nunnally, 1978). Tampoco se produce ninguna diferencia, en términos de varianza, si se transforma las ponderaciones de las puntuaciones de manera lineal. Se puede hacer que los participantes utilicen un sistema, +1, 0, –1 y, por supuesto, utilizar las puntuaciones en un análisis. Sin embargo, se puede añadir una constante de 1 a cada puntuación, produciendo 2, 1, 0. Las puntuaciones transformadas son más fáciles de trabajar, ya que no tienen signos negativos.

Reactivos y escalas de orden de rango

El segundo grupo de tipos de escalas y reactivos es ordinal o de orden de rangos, que es una forma simple y muy útil de escala o reactivo. Una escala completa puede ordenarse por rangos; es decir, se le pide a los participantes que ordenen todos los reactivos de acuerdo a algún criterio específico. Por ejemplo, si se desea comparar los valores educativos de administradores, maestros y padres, se les presenta un número de reactivos que presuntamente midan valores educativos a los miembros de cada grupo con las instrucciones de que los ordenen de acuerdo con sus preferencias.

En su estudio sobre actitudes hacia la liberación femenina, Taleporos (1977) desarrolló una escala de orden de rango de problemas sociales. Se les pidió a los participantes que ordenaran los siguientes problemas sociales: *adicción a las drogas*, *contaminación ambiental*, *discriminación racial*, *discriminación sexual*, *crimen violento* y *asistencia social*. Taleporos esperaba que los dos grupos que estaba estudiando ordenaran los temas sociales de manera similar, con excepción del tema de la *discriminación sexual*. Se sostuvo su hipótesis. Su estudio representó un uso productivo de la escalación de orden de rangos.

Las escalas de orden de rangos tienen tres ventajas analíticamente convenientes:

1. Las escalas de los individuos pueden interrelacionar y analizar fácilmente. Los órdenes de rangos compuestos de los grupos de individuos también se correlacionan fácilmente.
2. Los valores de escala de un conjunto de estímulos pueden calcularse utilizando uno de los métodos de orden de rango de escalación (véase Guilford, 1954).
3. Las escalas escapan parcialmente de la fijeza de respuesta y de la tendencia a mostrarse de acuerdo con los reactivos socialmente deseables.

Reactivos y escalas de elección forzada

La esencia de un método de elección forzada es que el sujeto debe elegir entre alternativas que en apariencia se perciben casi igualmente favorables (o desfavorables). Estrictamente hablando, el método no es nuevo. Las escalas de comparaciones de pares y de orden de rangos son métodos de elección forzada. Lo que es diferente acerca del método de elección forzada, como tal, es que se determinan los valores de discriminación y preferencia de los reactivos, y se aparean aquellos reactivos que son aproximadamente iguales en ambos. Así se controlan en cierta medida, la fijeza de respuesta y la "deseabilidad del reactivo". (La *deseabilidad del reactivo* significa que un reactivo puede ser elegido sobre otro simplemente porque expresa una cuestión deseable reconocida comúnmente. Si a un hombre se le pregunta si es descuidado o eficiente, tiende a decir que es eficiente, aunque sea descuidado.)

El método de las comparaciones apareadas (o comparaciones de pares) posee un largo y respetable pasado psicométrico. Sin embargo, se ha utilizado principalmente con el propósito de determinar valores de escala (Guilford, 1954). Aquí las comparaciones de pares se consideran como un método de medición. La esencia del método es que conjuntos de pares de estímulos, o reactivos de diferentes valores en un solo continuo o en dos continuos o factores diferentes, se presentan al sujeto con las instrucciones de que elija un miembro de cada par con base en algún criterio establecido. El criterio podría ser: el que

mejor caracterice al sujeto o el que el sujeto prefiera. Los reactivos de los pares pueden ser palabras solas, enunciados e inclusive párrafos. Por ejemplo, Edwards, en su inventario de preferencia personal (Personal Preference Schedule), apareó de manera efectiva afirmaciones que expresan distintas necesidades. Un reactivo que mide la necesidad de autonomía, por ejemplo, está apareado con otro reactivo que mide la necesidad de cambio. Se le pide al sujeto que elija uno de esos reactivos. Se supone que la persona elegirá el reactivo que se ajuste a sus necesidades. Una característica única de la escala es que los valores del deseo de aceptación social de los miembros apareados se determinaron de forma empírica y los pares se relacionaron de acuerdo con esto. El instrumento produce perfiles de puntuaciones de necesidad para cada individuo.

De alguna manera los dos tipos de técnicas de comparaciones de pares, 1) la determinación de valores de escala de estímulos y 2) la medición directa de variables, constituyen los métodos psicométricos más satisfactorios. Son simples y económicos debido a que solamente existen dos alternativas. Además, se puede obtener una gran cantidad de información con una cantidad limitada de material. Por ejemplo, si un investigador tiene únicamente 10 reactivos, cinco de la variable *A* y cinco de la variable *B*, se puede construir una escala de 5×5 o 25 reactivos, ya que cada reactivo *A* puede aparearse sistemáticamente con cada reactivo *B*. (La puntuación es simple: asignar un "1" a *A* o a *B* en cada reactivo, dependiendo de la alternativa que el sujeto elija.) Más importante aún, los reactivos de comparación de pares obligan a los participantes a elegir. Aunque esto puede molestar a algunos participantes, especialmente si consideran que ningún reactivo representa lo que elegirían (es decir, elegir entre cobarde y débil para categorizarse a sí mismo), es en realidad una actividad humana acostumbrada. Debemos elegir cada día de nuestras vidas. Inclusive se puede argumentar que los reactivos de acuerdo-en desacuerdo son artificiales y que los reactivos de elección son "naturales". En un estudio sobre el concepto del interés social de Adler (valorar cosas diferentes al yo), Crandall (1980) utilizó comparaciones de pares para desarrollar su escala de interés social (Social Interest Scale). Los jueces calificaron 90 rasgos respecto a su relevancia para el interés social. Se utilizaron 48 pares, donde un miembro de cada par tenía relevancia para el interés social y el otro miembro no la tenía. Entonces, después de la realización de un análisis de reactivos para determinar cuáles eran los reactivos más discriminantes, se desarrolló una escala de 15 reactivos. Por desgracia, Crandall no reporta la forma de la escala. Sin embargo, la idea es buena: utilizó la fortaleza de las comparaciones de pares para encontrar buenos reactivos para una escala final.

Los reactivos de elección forzada con más de dos partes pueden asumir un número de formas con tres, cuatro o cinco partes, las cuales son homogéneas o heterogéneas respecto a lo favorable y a lo no favorable. Se analiza e ilustra sólo uno de estos tipos para demostrar los principios que subyacen a dichos reactivos. Por medio de un análisis factorial, un procedimiento conocido como la *técnica de los incidentes críticos* o algún otro método, se reúnen y seleccionan los reactivos. Por lo común se descubre que algunos reactivos discriminan entre grupos de criterio y que otros no lo hacen. Ambos tipos de reactivos —llámense discriminantes e irrelevantes— se incluyen en cada conjunto de reactivos. Además, se determinan los valores de preferencia para cada reactivo.

Un reactivo típico de elección forzada es una *tétrada*. Una forma útil de tétrada consiste en dos pares de reactivos, un par con un alto valor de preferencia y el otro par con un bajo valor de preferencia, donde un miembro de cada par es discriminativo (válido) y el otro miembro del par es irrelevante (no válido). Un esquema de dicho reactivo de elección forzada es:

alta preferencia-discriminante
alta preferencia-irrelevante

baja preferencia-discriminante
baja preferencia-irrelevante

Se dirige al sujeto para que elija el reactivo de la tétrada que más prefiera, o que constituya la mejor descripción de sí mismo (o de alguien más), etcétera. También se dirige a esta persona para que seleccione el reactivo menos preferido o menos descriptivo de sí mismo.

La idea básica detrás de este reactivo más bien complejo es, como se indicó antes, que se controla la fijeza de respuesta y el deseo de aceptación social. El sujeto no puede decir, al menos teóricamente, cuáles son los reactivos discriminantes y cuáles los irrelevantes; tampoco se pueden elegir los reactivos con base en los valores de preferencia. Así, se contrarresta la tendencia a evaluarse a sí mismo (o a otros) demasiado alto o demasiado bajo y, por lo tanto, la validez presuntamente se incrementa (Guilford, 1954).

Un reactivo de elección forzada de un tipo hasta cierto punto diferente, construido por el primer autor de este libro con fines ilustrativos del uso de reactivos de investigación real, es:

consciente
agradable
respondiente
sensible

Uno de los reactivos (sensible) es un reactivo *A* y otro (consciente) es un reactivo *B*. (*A* y *B* se refieren a factores adjetivados.) Los otros reactivos son presuntamente irrelevantes. Se puede pedir a los participantes que elijan uno o dos reactivos que son muy importantes que un maestro posea.

Los métodos de elección forzada parecen ser muy promisorios. Aun así, existen dificultades técnicas y psicológicas, entre las cuales la más importante parece ser la falta de independencia de los reactivos, la quizás demasiado compleja naturaleza de algunos reactivos y la resistencia de los participantes ante las opciones difíciles. Se refiere al lector a los estudios de Guilford (1954) o de Bock y Jones (1968) sobre el tema: éstos son de autoridad, objetivas y breves; y también a las revisiones de Scott (1968) y Zavala (1965). (Para la revisión de referencias más recientes sobre reactivos y escalas de elección forzada véase Borg, 1988; Bownas y Bernardin, 1991; Closs, 1978; Deaton, Glasnapp y Poggio, 1980; Hyman y Sharp, 1983; May y Forsyth, 1980; Presser y Schuman, 1980; Ray, 1990; y Stanley, Wandzilak, Ansorge y Potter, 1987.)

Medidas ipsativas y normativas

Una distinción que se ha vuelto importante y que generalmente es mal entendida, en investigación y medición, es aquella que existe entre las medidas normativas e ipsativas. Las *medidas normativas* son el tipo común de medidas obtenidas con pruebas y escalas; pueden variar de manera independiente, es decir, se ven relativamente poco afectadas por otras medidas y, para su interpretación, se refieren a la media de las medidas de un grupo, siendo que los conjuntos de medidas de individuos poseen medias y desviaciones estándar diferentes. Las *medidas ipsativas*, por otro lado, se ven afectadas de manera sistemática por otras medidas y, para su interpretación, se refieren a la misma media, siendo que cada conjunto de medidas del individuo posee la misma media y desviación estándar. Para terminar con esta más bien opaca palabrería, sólo piense en un conjunto de rangos, del 1 al 5, donde el 1 indica "el primero", "el más alto" o "el más"; y el 5 indica "el último", "el más bajo" y "el menos", con el 2, 3 y 4 señalando posiciones intermedias. Sin importar quién utilice estos rangos, la suma y la media de los rangos es siempre la misma, 15 y 3, y la desviación estándar es siempre la misma, 1.414. Los rangos, entonces, son medidas ipsativas.

Si los valores 1, 2, 3, 4 y 5 estuvieran disponibles para calificar, por ejemplo, cinco objetos, y fueran cuatro personas las que calificaran los cinco objetos, se obtendría algo como lo siguiente:

	Personas			
	1	2	2	3
	2	2	1	2
Objetos	3	4	5	3
	4	3	5	3
	5	5	4	2
Sumas:	15	16	17	13
Medias:	3.0	3.2	3.4	2.6

Note que las sumas y las medias (y las desviaciones estándar también) son diferentes. Éstas son medidas normativas. Teóricamente con las medidas normativas no existen restricciones en el valor que el individuo *A* le puede asignar al objeto *C*—con excepción, por supuesto, de los números del 1 al 5—.

No obstante, con las medidas ipsativas, el procedimiento —en este caso de orden de rangos— ha creado restricciones sistemáticas. Cada individuo debe utilizar 1, 2, 3, 4 y 5 tan sólo una vez, y todos deben ser utilizados, lo cual indica que cuando cinco objetos están siendo ordenados por rango y se asigna uno, por ejemplo, rango 1, sólo quedan cuatro rangos por asignar. Después de que se asigna el 2 al siguiente objeto, sólo quedan tres, etcétera, hasta el último objeto, al que debe asignársele el 5. Un razonamiento similar se aplica a otro tipo de procedimientos y medidas ipsativos: comparaciones de pares, tétradas y pentadas de elección forzada o metodología Q.

La limitación importante en los procedimientos ipsativos es que, estrictamente hablando, no se pueden aplicar los estadísticos usuales, puesto que éstos dependen de los supuestos que los procedimientos ipsativos violan sistemáticamente. Además, el procedimiento ipsativo genera correlaciones negativas espurias entre reactivos. En un instrumento de comparaciones de pares, por ejemplo, la selección de un miembro de un par automáticamente excluye la selección del otro miembro. Ello significa una falta de independencia y una correlación negativa entre reactivos, en función del procedimiento instrumental. Sin embargo, la mayor parte de las pruebas estadísticas se basan en el supuesto de independencia de los elementos que entran en las fórmulas estadísticas. Además, el análisis de correlaciones, tal como el análisis factorial, puede distorsionarse seriamente por las correlaciones negativas. Por desgracia, estas limitaciones no se han comprendido o se han subestimado por los investigadores que, por ejemplo, han tratado los datos ipsativos de forma normativa (Hicks, 1970). Se invita al lector a demostrar el comportamiento de las escalas ipsativas estableciendo una pequeña matriz de números ipsativos, generados de forma hipotética por medio de las respuestas en una escala de comparaciones de pares. Utilice números 1 y 0 y calcule las r entre reactivos para los individuos.

Elección y construcción de medidas objetivas

Una de las tareas más difíciles para el investigador del comportamiento, cuando se enfrenta con la necesidad de medir variables, es encontrar el camino a través de un gran número de medidas ya existentes. Si existe una buena medida para una variable en particular, parece

tener poco sentido construir una medida nueva. De cualquier manera, la pregunta es: ¿existe una buena medida? La respuesta a esta pregunta quizá requiera de una gran búsqueda y estudio. El investigador debe saber, primero, qué tipo de variable se va a medir. Se ha tratado de ofrecer una guía dentro de la estructura recién proporcionada. Se debe saber claramente si la variable es una aptitud, rendimiento, personalidad, actitud o algún otro tipo de variable. El segundo paso es consultar uno o dos libros de texto que analicen medidas y pruebas psicológicas. Después, se deben consultar las bien conocidas guías de Buros. Aunque Buros ofrece una excelente guía sobre pruebas publicadas, muchas buenas medidas no se han publicado comercialmente. Por lo tanto, debe buscarse en la literatura de aparición periódica. A pesar de que muchas escalas no están disponibles de manera comercial, se pueden reproducir (con permiso) y utilizarse con propósitos de investigación. Otras fuentes valiosas son Andrulis (1977); Comrey, Backer y Glaser (1973); Fischer y Corcoran (1994); Goldman, Saunders y Busch (1996); Keyser y Sweetland (1987) y Taulbee (1983).

Fuentes valiosas de información sobre pruebas y escalas son las revistas *Psychological Bulletin*, *Journal of Psychoeducational Assessment*, *Applied Psychological Measurement*, *Educational and Psychological Measurement*, *Journal of Educational Measurement*, *Psychological Assessment* y *Journal of Experimental Education*.

Tal vez un investigador encuentre que no existe una medida que mida el atributo deseado. O, si existe una medida, quizá sea insatisfactoria para los propósitos. Por consiguiente, el investigador debe construir una nueva medida o instrumento, o abandonar la variable. La construcción de pruebas y escalas objetivas constituye una tarea larga y ardua. No existen atajos. Un instrumento pobremente construido llega a ocasionar más daño que beneficio, ya que puede conducir al investigador a conclusiones erróneas. Entonces, el investigador que debe construir un nuevo instrumento tiene que seguir ciertos procedimientos conocidos y guiarse por criterios psicométricos aceptados.

Se ha logrado un enorme progreso en la medición objetiva de la inteligencia, las aptitudes, el rendimiento, la personalidad y las actitudes. Sin embargo, las opiniones están divididas, en ocasiones de forma muy marcada, sobre el valor de la medición objetiva. El avance más impresionante se ha logrado en la medición objetiva de la inteligencia, las aptitudes y el rendimiento. Los avances en la medición de la personalidad y las actitudes no han sido tan impresionantes. El problema es, por supuesto, la validez, especialmente la validez de las medidas de personalidad.

Dos o tres desarrollos recientes son muy alentadores. Uno es la creciente comprensión de la complejidad que implica la medición de cualquier variable de personalidad y de actitud. El segundo lo constituyen los avances técnicos para llevarla a cabo. Otro desarrollo muy relacionado consiste en el empleo del análisis factorial como ayuda en la identificación de variables y como guía en la construcción de medidas. Un tercer desarrollo (que se estudió en un capítulo anterior) es el creciente conocimiento, comprensión y maestría del problema de validez en sí mismo, y en especial la comprensión de que la validez y la teoría psicológica están interrelacionados.

RESUMEN DEL CAPÍTULO

1. La prueba o escala es el método más utilizado en las ciencias del comportamiento para la recolección de datos.
2. Una meta consiste en desarrollar y utilizar pruebas que sean objetivas. Sin embargo, la objetividad no es fácil de comprender.
3. La objetividad científica no depende de las características del científico.

4. La objetividad científica incluye el acuerdo entre jueces expertos. Los métodos de observación y la recolección de datos poseen diferentes grados de objetividad.
5. Una prueba es un procedimiento sistemático para determinar el comportamiento de individuos.
6. Una escala es un conjunto de símbolos o valores numéricos construidos de tal manera que tales símbolos o valores numéricos puedan asignarse a individuos utilizando alguna regla.
7. Las pruebas de aptitud miden el potencial de logro de la persona. Se usan principalmente para orientación y consejería.
8. Las pruebas de rendimiento miden la eficiencia, maestría y comprensión presentes, de áreas generales y específicas del conocimiento. Las pruebas elaboradas por maestros son consideradas como pruebas de rendimiento.
9. La medición de los rasgos de personalidad constituye el problema más complejo de la medición psicológica. La personalidad es muy compleja respecto a problemas de validez.
10. Existen dos métodos para la construcción y validación de medidas de personalidad: el método *a priori* y el método *de constructo*.
11. Las escalas de actitud miden la predisposición de un individuo a pensar, sentir, percibir y comportarse hacia otra persona, idea u objeto.
12. Existen tres tipos de escalas de actitud: la escala de puntuaciones sumadas, las escalas de intervalos aparentemente iguales y las escalas acumulativas o Guttman.
13. La escala de puntuaciones sumadas es la escala utilizada con mayor frecuencia en las ciencias del comportamiento.
14. Las escalas de valores miden la preferencia expresada por una persona hacia formas de conducta. Incluyen religión y libre empresa.
15. Existen dos tipos de escalas objetivas: independientes y no independientes.
16. En las escalas objetivas independientes la respuesta de una persona a un reactivo no se relaciona con su respuesta a otro reactivo. En los reactivos no independientes una respuesta a un reactivo podría conducir al sujeto a preguntas más profundas.
17. Las escalas y los reactivos pueden dividirse en tres tipos: *de acuerdo-en desacuerdo*, de orden de rangos y de elección forzada.
18. Las medidas normativas no se ven afectadas por otras medidas. Sin embargo, las medidas ipsativas sí son afectadas por otras medidas.
19. El investigador debe dedicar tiempo para determinar si ya existe una prueba para el estudio. Se cuenta con un número de fuentes publicadas y no publicadas de pruebas. Sólo se debe crear una prueba nueva si no existe una para los propósitos del investigador.

SUGERENCIAS DE ESTUDIO

1. Las siguientes referencias ayudarán a los estudiantes a encontrar su camino en el largo y difícil, pero importante, camino de las pruebas y escalas objetivas, especialmente en educación.

Adkins, D. (1974). *Test construction: Development and interpretation of achievement tests* (2a. ed.) Columbus, Ohio: Charles E. Merrill. [Un libro invaluable para estudiantes e investigadores.]

Bloom, B. (ed.). (1976). *Taxonomy of educational objectives. The classification of educational goals: Handbook 1, cognitive domain.* Nueva York: David McKay. [Este libro básí-

co e inusual intenta establecer un fundamento para la medición cognitiva, por medio de la clasificación de los objetivos educativos y por la presentación de muchos preceptos y ejemplos. Las páginas 201-207, que bosquejan el libro, son útiles para quienes construyen pruebas y para los investigadores educativos.]

Impara, J. C. y Plake, B. S. (eds.). (1998). *Buros 13th mental measurements yearbook.* Lincoln, Nebraska: Buros Institute. [Descripciones y revisiones de pruebas publicadas y medidas de todos tipos. Véase también ediciones anteriores.]

Mehrens, W. y Ebel, R. (eds.). (1967). *Principles of educational and psychological measurement.* Chicago: Rand McNally. [Una valiosa colección de muchas de las contribuciones clásicas a la medición y a la teoría y práctica de las pruebas.]

2. Para comprender la lógica y la elaboración de instrumentos psicológicos de medición, resulta útil estudiar las explicaciones relativamente completas sobre cómo se desarrollan. Las siguientes referencias escritas describen el desarrollo de interesantes e importantes reactivos e instrumentos de medición.

Allport, G., Vernon, P. y Lindzey, G. (1951). *Study of values. Manual of directions* (ed. rev.). Boston: Houghton Mifflin.

Comrey, A. L. (1961). Factored homogeneous item dimensions in personality research. *Educational and Psychological Measurement, 21,* 417-431.

Comrey, A. L. y Lee, H. B. (1992). *A first course in factor analysis* (2a. ed.). Hillsdale, Nueva Jersey: Lawrence Erlbaum.

Edwards, A. (1953). *Personal preference schedule, manual.* Nueva York: Psychological Corp. [Mide necesidades en formato de elección forzada (comparaciones de pares).]

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140.* [Monografía original de Likert donde describe su técnica; se trata de una importante guía en la medición de actitudes.]

Thurstone, L. y Chave, E. (1929). *The measurement of attitude.* Chicago: University of Chicago Press. [Este clásico describe la construcción de la escala de intervalos aparentemente iguales, para medir actitudes hacia la Iglesia.]

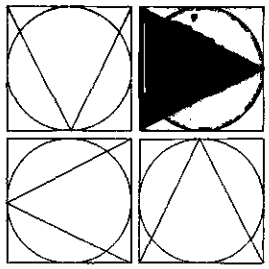
Woodmansee, J. y Cook, S. (1967). Dimensions of verbal racial attitudes: Their identification and measurement. *Journal of Personality and Social Psychology, 7.* 240-250. [Probablemente la mejor medida para actitudes hacia los negros. El inventario aparece en el volumen de Robinson, Rusk y Head, citado en las sugerencias de estudio 3.]

3. Las siguientes son tres útiles antologías de escalas de actitud, valores y otras escalas. Su utilidad reside no sólo en las muchas escalas que contienen, sino también en las críticas perspicaces que se enfocan en la confiabilidad, la validez y otras características de las escalas.

Robinson, J., Rusk, J. y Head, K. (1968). *Measures of political attitudes.* Ann Arbor: Institute for Social Research, University of Michigan.

Robinson, J. y Shaver, P. (1969). *Measures of social psychological attitudes.* Ann Arbor: Institute for Social Research, University of Michigan.

Shaw, M. y Wright, J. (1967). *Scales for the measurement of attitudes.* Nueva York: McGraw-Hill.



CAPÍTULO 31

OBSERVACIONES DEL COMPORTAMIENTO Y SOCIOMETRÍA

Todos observan los actos de otros. Se observa a otras personas y se les escucha hablar. Se infiere lo que los otros quieren decir cuando expresan algo; y se infieren las características, motivaciones, sentimientos e intenciones de otros con base en estas observaciones. Se dice

"ella es una juez perspicaz de la gente", queriendo decir que sus observaciones sobre el comportamiento son agudas y que se considera que sus inferencias sobre lo que está detrás del comportamiento son válidas. Sin embargo, este tipo de observaciones diarias de la mayoría de la gente resultan insatisfactorias para la ciencia. Los científicos sociales también deben observar el comportamiento humano; pero no se satisfacen con observaciones no controladas. Ellos buscan observaciones confiables y objetivas a partir de las cuales puedan realizar inferencias válidas. Tratan la observación del comportamiento como parte de un procedimiento de medición: asignan valores numéricos a objetos de acuerdo con reglas, en este caso de acuerdo con actos o secuencias de actos de comportamiento humano.

A pesar de que esto puede parecer simple y directo, evidentemente no lo es: existe mucha controversia y debate respecto a la observación y a los métodos de observación. Los críticos del punto de vista de que las observaciones del comportamiento deben ser controladas rigurosamente —punto de vista adoptado en este capítulo y en el resto del libro— afirman que es demasiado estrecho y artificial. Los críticos dicen que en lugar de eso, las observaciones deben ser naturales: los observadores deben estar inmersos en situaciones realistas y naturales presentes, y deben observar el comportamiento tal y como ocurre de manera natural, por así decirlo. Sin embargo, como se verá, la observación del comportamiento es extremadamente compleja y difícil.

Existen básicamente dos formas de observación: se puede observar a la gente hacer y decir cosas, y se puede preguntar a la gente sobre sus propias acciones y sobre el comportamiento de otros. Las principales maneras para obtener información son experimentando algo de forma directa o pidiéndole a alguien que informe lo que sucedió. En el presente capítulo se incluyen principalmente los eventos que se ven y se escuchan y la observación del comportamiento, así como la solución de problemas científicos que surgen a partir de dichas observaciones. También se examina de manera breve un método para evaluar las interacciones e interrelaciones de miembros de grupos: la sociometría, la cual es una forma especial y valiosa de observación. Los miembros de grupos se observan entre sí y registran las reacciones entre ellos, de tal manera que los investigadores evalúen el estado sociométrico de los grupos.

Problemas en la observación del comportamiento

El observador

El principal problema de la observación del comportamiento reside en el observador. Una de las dificultades de las entrevistas es que el entrevistador forma parte del instrumento de medición. Dicho problema casi no existe en las pruebas y escalas objetivas. En la observación del comportamiento, el observador es tanto una fortaleza, como una debilidad cruciales. Esto se debe a que el observador debe asimilar la información derivada de las observaciones, para luego realizar inferencias acerca de los constructos. El observador mira cierto comportamiento —por ejemplo, un niño que golpea a otro niño— y debe procesar, de alguna manera, tal observación y hacer una inferencia de que el comportamiento es una manifestación del constructo "agresión" o "comportamiento agresivo", o incluso "hostilidad". La fortaleza y debilidad del procedimiento es el poder de inferencia del observador. Si no fuera por la inferencia, una máquina observadora sería mejor que un observador humano. La fortaleza radica en que el observador puede relacionar el comportamiento observado con el constructo o variables de un estudio al unir el comportamiento con el constructo. Una de las dificultades recurrentes de la medición consiste en lograr cerrar el abismo que existe entre el comportamiento y el constructo.

La debilidad básica del observador consiste en que se pueden realizar inferencias incorrectas a partir de las observaciones. Considere dos casos extremos. Suponga, por un lado, que un observador, que se muestra muy hostil ante la educación escolar religiosa, observa las clases en una escuela religiosa. Queda claro que los prejuicios de esta persona pueden invalidar las observaciones. El observador tal vez califique a un maestro adaptable como inflexible debido a la existencia de un prejuicio o a la percepción de que la enseñanza en escuelas religiosas es inflexible. O tal vez ese mismo observador juzgue el comportamiento realmente estimulante de un maestro de una escuela religiosa como insulso. Por otro lado, suponga que un observador pueda ser completamente objetivo y no sabe nada sobre la educación pública o religiosa. En cierto sentido cualquier observación que realice no estará sesgada; pero será inadecuada. La observación del comportamiento humano requiere de un conocimiento competente sobre dicho comportamiento y aun del significado del comportamiento.

Sin embargo, existe otro problema: el observador puede afectar los objetos de observación en tanto que forma parte de la situación de observación. Sin embargo, en realidad y por fortuna, éste no constituye un problema severo. De hecho, representa un problema para el novato, quien parece creer que la gente actúa de forma diferente, inclusive artificial, cuando se le observa. Parece ser que los observadores ejercen muy poco efecto en las situaciones que observan. Se percibe que los individuos y los grupos se adaptan más bien rápidamente a la presencia de un observador y que actúan como normalmente lo harían. Esto no quiere decir que el observador no pueda ejercer un efecto. Quiere decir que si el observador es cuidadoso para no interferir y para evitar que las personas observadas sientan que se están haciendo juicios, entonces el observador, como estímulo influyente, es prácticamente anulado. Babbie (1995) afirma que no existe una protección completa para el efecto del observador. Sin embargo, el conocimiento y la sensibilidad ante este problema ofrecen una protección parcial.

Validez y confiabilidad

En la superficie, nada parece más natural cuando se observa el comportamiento, que creer que se está midiendo lo que se dice que se está midiendo. Sin embargo, cuando se da una carga interpretativa al observador, la validez puede verse afectada (así como la confiabilidad). A mayor carga de interpretación, mayor será el problema de la validez. No obstante, ello no significa que no deba darse una carga interpretativa al observador.

Un aspecto simple de la validez de las medidas de observación es su poder predictivo. ¿Predicen criterios relevantes de forma confiable? El problema, como siempre, reside en los criterios. Las medidas independientes de las mismas variables son infrecuentes. ¿Es posible afirmar que la medida de observación del comportamiento de un maestro es válida debido a que se correlaciona positivamente con las calificaciones de los superiores? Se podría tener una medida independiente sobre necesidades orientadas hacia uno mismo, ¿pero sería esta medida un criterio adecuado para la observación de dichas necesidades?

Una clave importante para el estudio de la validez de medidas de observación del comportamiento parece ser la validez de constructo. Si las variables que se miden a través de un método de observación están inmersas dentro de un marco teórico, entonces debe existir cierta relación. ¿Verdaderamente existen? Suponga que una investigación incluye la teoría de la autoeficacia de Bandura (1982), y que se ha construido un sistema de observación cuyo propósito es medir la competencia en el desempeño. En efecto, la teoría indica que la autoeficacia percibida, o la autopercepción de competencia, afecta la competencia del desempeño real de una persona: a mayor autoeficacia, mayor será la competen-

cia en el desempeño. Si se encuentra que la autopercepción de la competencia y las medidas de la competencia real observada al efectuar cierta tarea prescrita son positivas y altas, entonces se apoya la hipótesis derivada de la teoría. Sin embargo, esto también constituye evidencia de la validez de constructo del sistema de observación.

La confiabilidad de los sistemas de observación es un asunto simple, aunque por ningún motivo es fácil. Con frecuencia se define como el acuerdo entre observadores. A partir de este punto de vista, los registros de películas, videocintas y audiocintas ayudan a lograr una confiabilidad muy alta. No obstante, el acuerdo entre observadores tiene defectos potenciales. Por ejemplo, la magnitud de un índice de acuerdo en parte se debe al acuerdo por azar y, por lo tanto, requiere corrección. Quizás el camino más fácil a seguir sea utilizar diferentes métodos para evaluar la confiabilidad, tal como se haría con cualquier medida utilizada en investigación del comportamiento: acuerdo de los observadores, confiabilidad repetida y el método de análisis de varianza. La evaluación de la confiabilidad y el acuerdo entre observadores son problemas especialmente difíciles de la observación directa, ya que los estadísticos comunes dependen del supuesto de que las medidas son independientes —aunque con frecuencia no son independientes—. La mayor parte del trabajo realizado en esta área proviene o se basa en el trabajo de Cohen y Fleiss sobre el coeficiente kappa (Fleiss, 1986; Fleiss y Cohen, 1973). El uso de la teoría de la generalizabilidad es prometedora en la medición de la confiabilidad para datos nominales (Li y Lautenschlager, 1997). Algunos desarrollos provienen de las ciencias de la salud, donde la observación y el acuerdo desempeñan un papel importante en el análisis y en las decisiones (véase Dunn, 1989, 1992). Medley y Mitzel (1963) ofrecen una profunda, pero técnicamente compleja, exposición sobre la confiabilidad de valoraciones en un marco del análisis de varianza. Hollenbeck (1978) y Rowley (1976) discuten acerca de la confiabilidad de las observaciones cuando las medidas son nominales. Revisiones más recientes sobre la confiabilidad de las observaciones y la confiabilidad intercalificadora se encuentran en Dewey (1989), McDermott (1988), Perreault y Leigh (1989), Schouten (1986), Topf (1986), Zegers (1991) y Zwick (1988). El artículo de Perreault y Leigh trata el tema desde el punto de vista de la investigación de mercado. Topf revisa el uso de las medidas de confiabilidad nominales en investigación de enfermería, y McDermott analiza su aplicación en la psicología escolar. Chan (1987), Oud y Sattler (1984), y Powers (1985) han desarrollado programas computacionales para ayudar a los investigadores a calcular estadísticos del acuerdo entre observadores.

Entonces, es necesario definir con precisión y sin ambigüedades lo que se va a observar. Si se está midiendo la *curiosidad*, se debe decir al observador qué es un comportamiento curioso. Si se mide la *cooperatividad*, se debe explicar de alguna manera al observador de qué forma se distingue el comportamiento cooperativo de otros tipos de comportamiento. Esto quiere decir que se debe proporcionar al observador alguna forma de definición operacional de la variable que se está midiendo; la variable debe definirse en términos de comportamiento.

Categorías

La tarea fundamental del observador consiste en asignar categorías a los comportamientos. Recuerde, del trabajo previo sobre particiones, que las categorías deben ser exhaustivas y mutuamente excluyentes. Para satisfacer la condición de exhaustividad primero debe definirse U, el universo de comportamientos que se van a observar. En algunos sistemas de observación esto no es difícil de lograr. McGee y Snyder (1975), para comprobar la hipótesis de que la gente que sala sus alimentos antes de probarlos, percibe que el control del

comportamiento está dentro del individuo (control disposicional), más que en el ambiente (control situacional), simplemente observaron la conducta de salar la comida de los participantes en restaurantes. En otros sistemas de observación resulta más difícil. Muchos, o la mayoría de los sistemas de observación citados en la enorme antología de instrumentos de observación de Simon y Boyer (1970), *Mirrors for Behavior*, son complejos y difíciles de utilizar. ¡Este trabajo consiste de 14 volúmenes de instrumentos de observación del comportamiento! La mayoría de tales instrumentos, 67 de 79, son utilizados para observaciones educativas. Los lectores que pretendan utilizar observación de comportamiento en su investigación deben consultar dicha información, en especial el volumen 1, que contiene un análisis general en las páginas 1-24.

En concordancia con el énfasis de este libro —de que el propósito de la mayoría de la observación es medir variables— se cita un sistema de observación en el salón de clases del interesante y creativo trabajo de Kounin y sus colegas (Kounin y Doyle, 1975; Kounin y Gump, 1974). El sistema reportado es más complejo que el sistema de observación de salar y probar los alimentos, pero mucho menos complejo que muchos sistemas de observación en el salón de clases. La variable observada fue la de compromiso con la tarea, que se observó al videofilmar 596 lecciones y luego observarlas en video para obtener medidas de compromiso. Las medidas se clasificaron como alto compromiso con la tarea y bajo compromiso con la tarea. Los autores también midieron la continuidad en las lecciones creando categorías que reflejaban mayor o menor continuidad en las mismas. Cuando se observaba el comportamiento de niños, utilizaron las categorías para registrar los comportamientos pertinentes observados.

Unidades de comportamiento

Decidir qué unidades utilizar en la medición del comportamiento humano continúa siendo un problema sin resolver. Aquí con frecuencia se enfrenta un conflicto entre las demandas de la confiabilidad y la validez. Teóricamente es posible lograr un alto grado de confiabilidad utilizando unidades pequeñas y fáciles de observar y registrar. Puede intentarse definir el comportamiento de manera operacional al hacer una lista de un gran número de actos de comportamiento, pudiendo lograr así, por lo general, un alto grado de precisión y confiabilidad. Sin embargo, al hacer esto también puede reducirse tanto el comportamiento que ya no guarde demasiada semejanza con el comportamiento que se intentaba observar, con lo cual se pierde validez.

Por otra parte, pueden utilizarse amplias definiciones "naturales" y quizás lograr un alto grado de validez. Se podría instruir a los observadores para que observen la *cooperatividad* y definir el *comportamiento de cooperación* como "aceptar los métodos, sugerencias e ideas de otras personas; trabajar armónicamente con otros para lograr metas" o alguna definición más bien general. Si los observadores han tenido experiencia de grupo y comprenden los procesos grupales, entonces podría esperarse que pudieran evaluar de manera válida el comportamiento como cooperativo o no cooperativo, al utilizar dicha definición. Una definición tan general e incluso tan vaga como ésta le permite al observador capturar, si le es posible, toda la gama del comportamiento cooperativo. No obstante, su gran ambigüedad permite que se hagan diferentes interpretaciones, disminuyendo probablemente su confiabilidad.

Algunos investigadores que siguen un modelo fuertemente operacional insisten en que se realicen definiciones sumamente específicas de las variables observadas. Ellos enlistarán diversos comportamientos específicos que el observador debe anotar; ningún otro se observa ni se registra. Modelos extremos como éste pueden producir una confiabi-

lidad alta, pero también pueden perder parte del aspecto esencial de las variables observadas. Suponga que se hace una lista de 10 tipos específicos de comportamientos para *cooperatividad* y que el universo de posibles comportamientos consta de 40 o 50 tipos. En efecto, se perderán aspectos importantes de la cooperatividad. Aunque aquello que se mide puede medirse de forma confiable, quizá resulte bastante trivial o irrelevante, en parte, para la variable.

Cooperatividad

Éste es el problema molar-molecular de cualquier procedimiento de medición en las ciencias sociales. El *modelo molar* toma "todos" de comportamiento más grandes como unidades de observación. Unidades completas de interacción pueden especificarse como blancos de observación. El comportamiento verbal puede separarse en intercambios completos entre dos o más individuos, o en párrafos u oraciones completas. En contraste, el *modelo molecular* toma segmentos más pequeños de comportamiento como unidades de observación. Cada intercambio completo o parcial puede registrarse. Las unidades de comportamiento verbal quizá sean palabras o frases cortas. Los observadores molares empiezan con una variable general definida de forma amplia, como se dijo antes, y observan y registran una variedad de comportamientos bajo la única rúbrica. La interpretación que den al significado del comportamiento que observan depende de la experiencia y conocimiento que tengan. Por otra parte, los observadores moleculares tratan de sacar su propia experiencia, conocimiento e interpretación de la escena de observación; registran lo que ven y nada más.

Inferencia del observador

Los sistemas de observación difieren en otra importante dimensión: la *cantidad de inferencia* requerida del observador. Los sistemas moleculares requieren relativamente de poca inferencia. El observador simplemente anota si un individuo hace o dice algo. Por ejemplo, un sistema puede requerir que el observador anote cada unidad de interacción, que puede estar definida como cualquier intercambio verbal entre dos individuos. Si ocurre un intercambio, se anota; si no ocurre, no se anota. Otra categoría podría ser "golpea a otro niño" . Cada vez que un niño golpee a otro, ello se anota. No se hacen inferencias en este sistema —si, por supuesto, es acaso posible escapar a las inferencias (por ejemplo, "golpea")—. Se registra el comportamiento puro lo más posible.

Son escasos los sistemas de observadores con un nivel tan bajo de inferencia por parte del observador. La mayor parte de los sistemas requieren de cierto nivel de inferencia. Un investigador puede estar realizando investigación sobre el comportamiento del consejo de educación, y decide que un análisis con poca inferencia se ajusta al problema, y utiliza reactivos de observación como "sugiere un curso de acción", "interrumpe a otro miembro del consejo", "plantea una pregunta", "da una orden al superintendente", y otras similares. Puesto que dichos reactivos son ambiguos comparativamente, la confiabilidad de la observación necesita ser alta.

Los sistemas que requieren que el observador utilice altos niveles de inferencia son más comunes y probablemente más útiles en la mayor parte de la investigación. Los sistemas de observación de alta inferencia proveen al observador categorías denominadas, las cuales requieren de mayor o menor interpretación del comportamiento observado. Por ejemplo, suponga que se mide la *dominancia*, que se define como los intentos realizados por un individuo para mostrar superioridad intelectual (o de otro tipo) sobre otros indivi-

duos, con poco reconocimiento de las metas de grupo y de las contribuciones de otros. Esto, por supuesto, requerirá de un mayor nivel de inferencia del observador, y los observadores tendrán que entrenarse para que exista acuerdo sobre lo que constituyen comportamientos dominantes. Sin dicho entrenamiento y acuerdo —y probablemente sin experiencia en procesos grupales— la confiabilidad puede verse amenazada. Weick (1968) presenta una sofisticada exposición sobre la inferencia en la observación, y también analiza los sesgos en la observación y sugiere soluciones metodológicas para minimizar los efectos del sesgo. Señalamientos similares son pertinentes cuando se intentan medir muchas variables psicológicas y sociológicas: cooperación, competencia, agresividad, democracia, aptitud verbal, rendimiento y clase social, por ejemplo. Para revisar discusiones más recientes sobre observación e inferencia en estas áreas se recomienda leer Alexander, Newell, Robbins y Turner (1995); Borich y Klinzing (1984); Chavez (1984); Hartmann y Wood (1990); Jaffe (1997); Nurius y Gibson (1990), y Timberlake y Silva (1994). Los artículos de Borich y Klinzing y de Chavez se aplican a las observaciones en salón de clases. Hartmann y Wood tratan los sistemas de observación del comportamiento utilizados en modificación conductual. Nurius y Gibson tratan la observación e inferencia clínica encontrada en el trabajo social. En estrecha relación están los artículos de Jaffe y el de Alexander *et al.*, que tratan las observaciones clínicas. Timberlake y Silva tratan la observación e inferencia obtenida al observar la conducta de animales.

No es posible realizar generalizaciones uniformes sobre las virtudes relativas de sistemas con diferentes niveles de inferencia. Tal vez el mejor consejo para el neófito sea buscar un nivel medio de inferencia. Las categorías demasiado vagas, con muy poca especificación sobre qué se va a observar, ponen una carga excesiva sobre el observador. Es muy fácil que diferentes observadores den distintas interpretaciones al mismo comportamiento. Las categorías demasiado específicas, aunque reducen la ambigüedad y la incertidumbre, pueden tender a ser demasiado rígidas e inflexibles, e incluso triviales. Lo mejor es que el lector estudie varios sistemas exitosos, poniendo especial atención a las categorías de comportamiento y a las definiciones (instrucciones) ligadas a las categorías para guía del observador.

Generalización y aplicabilidad

Los sistemas de observación difieren considerablemente en su *generalización*, o en el grado de *aplicabilidad* a las situaciones de investigación distintas de aquellas para las que fueron diseñadas originalmente. Algunos sistemas son bastante generales: están diseñados para utilizarse con muchos problemas de investigación diferentes. El reconocido análisis de grupo de interacción de Bales (1951) es uno de dichos sistemas generales. Es un sistema de baja inferencia en el que todo el comportamiento verbal y no verbal, presuntamente en cualquier grupo, puede clasificarse dentro de una de 12 categorías: "muestra solidaridad", "está de acuerdo", "pide opinión", etcétera. Las 12 categorías están agrupadas dentro de tres grandes conjuntos: social emocional positivo, social emocional negativo y de tarea neutral.

Sin embargo, algunos sistemas fueron construidos para situaciones particulares de investigación, para medir variables particulares. El ejemplo anterior sobre salar los alimentos, es bastante específico y difícilmente se aplica a otras situaciones. El sistema de Kounin y Doyle (1975), que aunque fue construido específicamente para la investigación de Kounin, se aplica en muchas situaciones del salón de clases. De hecho, la mayor parte de los sistemas elaborados para problemas de investigación específicos se utilizan a menudo con ciertas modificaciones, en otros problemas de investigación.

Resulta necesario enfatizar que los "pequeños" sistemas de observación sirven para medir variables específicas. Suponga, por ejemplo, que la atención de los alumnos de escuela primaria sea una variable clave en una teoría sobre el rendimiento escolar. La atención (como rasgo o hábito), por sí mismo, ejerce poco efecto sobre el rendimiento: considere que la correlación es cero. Es una variable clave debido a que interactúa con cierto método de enseñanza y tiene un efecto pronunciado indirecto sobre el rendimiento. Asumiendo que esto es así, se debe medir la atención. Parece claro que se tendrá que observar el comportamiento del alumno, mientras se esté utilizando el método en cuestión y un método de "control". En tal caso, se necesita encontrar o diseñar un sistema de observación que se enfoque en la atención. Para evaluar la influencia del ambiente del salón de clases, por ejemplo, Keeves (1972) concluyó que era necesario medir la atención al observar a los estudiantes a quienes les pedía que pusieran atención a una tarea asignada por el maestro. Se asignaron puntuaciones que indicaban atención o la falta de ella. Este "pequeño" sistema de observación era confiable y aparentemente válido. Es probable que sistemas con objetivos específicos como éste incrementen su empleo en la investigación del comportamiento, especialmente en educación.

Muestreo del comportamiento

El muestreo, la última característica de las observaciones, estrictamente hablando no es una característica. Es una forma para obtener observaciones. Antes de usar un sistema de observación en investigación, debe decidirse cuándo y cómo se aplicará el sistema. Si se va a observar el comportamiento de los maestros en el salón de clase, ¿cómo se muestrearán los comportamientos? ¿Se observarán todos los comportamientos específicos en un periodo de clase? ¿O se harán muestreos de forma sistemática y aleatoria de comportamientos específicos? En otras palabras, debe diseñarse y utilizarse un plan de muestreo de algún tipo.

Existen dos aspectos del muestreo del comportamiento: el muestreo de eventos y el muestreo de tiempo. El *muestreo de eventos* es la selección de la observación de las ocurrencias integrales de comportamiento o eventos de cierta clase. Ejemplos de eventos integrales son los berrinches, las peleas y disputas, los juegos, los intercambios verbales sobre temas específicos, las interacciones entre alumnos y maestros en el salón de clases, etcétera. El investigador que estudia eventos debe saber cuándo ocurrirán dichos eventos y debe estar presente cuando sucedan, como con eventos del salón de clases; o esperar hasta que sucedan, como en las peleas.

El muestreo de eventos posee tres virtudes: 1) Los eventos son situaciones naturales de la vida y, por lo tanto, tienen una validez inherente que las muestras de tiempo por lo común no poseen. 2) Un evento integral posee una continuidad de comportamiento que los actos de comportamiento fragmentados de las muestras de tiempo no poseen. Si se observa una situación de solución de problemas desde el inicio hasta el final, entonces se está contemplando una unidad completa y natural de comportamiento individual y grupal. Al hacerlo, se logra una unidad completa y realista más grande del comportamiento individual y social. Como se estudió en un capítulo anterior, cuando se expusieron los experimentos de campo y los estudios de campo, las situaciones naturales impactan y se acercan a la realidad psicológica y social de una manera que los experimentadores normalmente no logran. 3) La tercera virtud del muestreo de eventos implica una característica importante de muchos eventos de comportamiento: en algunas ocasiones son inusuales y poco frecuentes. Por ejemplo, se puede estar interesado en las decisiones tomadas en reuniones administrativas y legislativas; o tal vez interesarse en el último paso de la solución de

problemas. Los métodos disciplinarios de los maestros constituyen una variable. Dichos eventos y muchos otros son relativamente poco frecuentes. Como tales, pueden perderse fácilmente por el muestreo de tiempo; por lo tanto, requieren de un muestreo de eventos. Sin embargo, si se toma el punto de vista más activo de observación promulgado por Weick (1968), es posible arreglar las situaciones para asegurarse de la ocurrencia más frecuente de eventos que suceden en pocas ocasiones.

El *muestreo de tiempo* es la selección de unidades de comportamiento para observación en diferentes momentos del tiempo. Pueden seleccionarse de formas sistemáticas o aleatorias para obtener muestras del comportamiento. Un buen ejemplo es el comportamiento del maestro. Suponga que se estudian las relaciones entre ciertas variables como el estado de alerta, la justicia y la iniciativa del maestro, por una parte; y la iniciativa y cooperación del alumno, por la otra. Se pueden seleccionar muestras aleatorias de maestros y después tomar muestras de tiempo de sus actos de comportamiento. Tales muestras de tiempo pueden ser sistemáticas: tres observaciones de 5 minutos en momentos específicos durante cada una de, por ejemplo, cinco horas de clase, siendo las horas de clase el primero, tercero y quinto periodos de un día, y el segundo y cuarto periodos del día siguiente. O pueden ser al azar: cinco periodos de 5 minutos de observación seleccionados aleatoriamente de un universo especificado de periodos de 5 minutos. De hecho existen muchas maneras de establecer y seleccionar muestras de tiempo. Como siempre, la forma en que se eligen dichas muestras, su duración y su número debe estar determinada por el problema de investigación. En un fascinante estudio sobre el liderazgo y el poder de la influencia grupal en niños pequeños, Merei (1949) señala que el muestreo de tiempo sólo mostraría líderes dando órdenes y al grupo obedeciendo; mientras que observaciones prolongadas mostrarían los mecanismos internos del hecho de dar órdenes y obedecer.

Las muestras de tiempo tienen la importante ventaja de incrementar la probabilidad de obtener muestras representativas de comportamiento. Sin embargo, ello es así sólo con los comportamientos que ocurren con mucha frecuencia. Los comportamientos poco frecuentes tienen una alta probabilidad de escapar de la red de muestreo, a menos que se seleccionen muestras enormes. El comportamiento creativo, compasivo y hostil, por ejemplo, quizá sea muy poco frecuente. Aun así, el muestreo de tiempo constituye una contribución positiva al estudio científico del comportamiento humano.

Como se explicó antes, las muestras de tiempo carecen de continuidad, de un contexto adecuado y, aun, de naturalidad. Lo anterior es cierto particularmente cuando se utilizan pequeñas unidades de tiempo y de comportamiento. Sin embargo, no hay razón para que el muestreo de eventos y el muestreo de tiempo no puedan combinarse algunas veces. Si se estudian recitaciones en el salón de clases, se selecciona una muestra aleatoria de los periodos de clase de un maestro en diferentes momentos, y se observan todas las recitaciones durante los periodos que se muestrearon, en su totalidad.

Algunas referencias muy buenas sobre el muestreo de eventos y el muestreo de tiempo se encuentran en Arrington (1943), Martin y Bateson (1993), Wright (1960) y Zeren y Makosky (1986). El artículo de Zeren y Makosky es sobresaliente, pues describe un ejercicio de salón de clases para enseñar a los estudiantes a realizar observaciones sistemáticas de comportamiento humano espontáneo. Además se describen tres técnicas de observación (muestreo de tiempo, muestreo de eventos y calificación de rasgos) y se compara su uso en comportamientos simulados presentados en televisión. La actividad en la clase incluyó una conferencia sobre métodos de observación, un ejercicio en donde se utilizaba uno de los tres métodos y una discusión en la clase. El artículo expone cómo enseñar el método científico para reunir datos de observación, la importancia de elaborar definiciones operacionales precisas para el acuerdo intercalificadores y el cálculo de coeficientes de confiabilidad.

Escalas de calificación

Hasta este punto, se ha hablado únicamente acerca de la observación de *comportamiento real*. Los observadores observan y escuchan directamente a los objetos en cuestión. Se sientan en el salón de clases y observan las interacciones maestro-alumno y alumno-alumno. O quizá vean y escuchen a un grupo de niños que resuelve un problema, frente a un espejo de doble vista.

Sin embargo, existe otra clase de observación del comportamiento que necesita mencionarse. Este tipo de observación se denominará *comportamiento recordado* o *comportamiento percibido*. Está clasificado convenientemente bajo el tema de las escalas de calificación. Para medir el comportamiento recordado o percibido, por lo común se les presenta a los observadores un sistema de observación en forma de escala de algún tipo, y se les pide que evalúen una o más características de un objeto, cuando el objeto no esté presente. Para hacerlo, ellos deben evaluar basándose en observaciones pasadas o en percepciones sobre cómo es el objeto observado y sobre cómo se comportará. Una forma conveniente para medir tanto el comportamiento real como el comportamiento percibido o recordado son las escalas de calificación.

Una *escala de calificación* es un instrumento de medición que requiere que un calificador u observador asigne al objeto calificado categorías o continuos que poseen valores numéricos asignados a ellos. Las escalas de calificación son quizás los instrumentos de medición más comunes, probablemente debido a que en apariencia son fáciles de construir y, lo más importante, son fáciles y rápidas de utilizar. Por desgracia, su aparente facilidad de construcción es engañosa, y la facilidad de uso conlleva un precio alto: la falta de validez debida a un número de fuentes de sesgo que entran en las medidas de calificación. No obstante, con conocimiento, habilidad y cuidado, las calificaciones resultan valiosas.

Para revisar un excelente estudio de las escalas de calificación, véase Guilford (1954), Nunnally (1978), Nunnally y Bernstein (1993) y Torgeson (1958). Si se desea revisar una presentación relativamente poco técnica sobre las escalas de calificación se recomienda leer a Selltiz, Jahoda, Deutsch y Cook (1961). Aunque las escalas de calificación ya se mencionaron con anterioridad en este libro, no se analizaron de manera sistemática. Al leer lo que sigue, el estudiante debe tener en mente que las escalas de calificación son en realidad escalas objetivas y, como tales, deberían haberse incluido en el capítulo 30. Su exposición se reservó para este capítulo a causa de que la exposición del capítulo 30 está enfocada principalmente en medidas donde responde el sujeto a quien se está midiendo. Las escalas de calificación, por otro lado, son medidas de individuos y sus reacciones, características y comportamientos, realizadas por observadores. Entonces, el contraste está en la forma en que el sujeto se observa a sí mismo y cómo lo perciben los demás. Las escalas de calificación también sirven para medir objetos, productos y estímulos psicológicos, tales como la escritura manual, los conceptos, los ensayos, los protocolos de entrevista y los materiales de pruebas proyectivas.

Tipos de escalas de calificación

Existen cuatro o cinco tipos de escalas de calificación, dos de los cuales se analizaron en el capítulo 30. Se trata de los listados y los instrumentos de elección forzada. Ahora se consideran sólo tres tipos y sus características: la escala de calificación de categorías, la escala de calificación numérica y la escala de calificación gráfica. Son bastante similares y difieren sólo en algunos detalles.

La *escala de calificación de categorías* presenta a los observadores o jueces varias categorías, de donde ellos eligen la que mejor caracteriza el comportamiento o características del objeto que se califica. Suponga que se califica el comportamiento de una maestra en el salón de clases. Una de las características que se califica es, por ejemplo, el estado de alerta. Un reactivo de categoría podría ser similar al que se mostró en el primer ejemplo. Una forma diferente utiliza descripciones condensadas; un reactivo de este tipo sería como el que se presenta en el segundo ejemplo

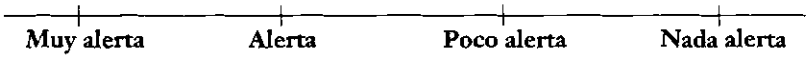
Ejemplos

¿Qué tan alerta es ella? (Marque una)
Muy alerta
Alerta
Poco alerta
Nada alerta

¿Emplea recursos? (Marque una)
Siempre emplea recursos; nunca le faltan ideas
Sus recursos son buenos
Algunas veces carece de ideas
No emplea recursos; rara vez tiene ideas

Las *escalas de calificación numérica* son, quizás, las más fáciles de construir y de utilizar. Además, también producen números que pueden usarse directamente en análisis estadísticos. Por otro lado, como los números representan intervalos iguales en la mente del observador, pueden alcanzar la medición intervalar (véase Guilford, 1954, p. 264). Cualquiera de las anteriores escalas de categorías puede convertirse fácil y rápidamente en escalas de calificación numérica, simplemente añadiendo números antes de cada una de las categorías. Los números 3, 2, 1, 0 o 4, 3, 2, 1 pueden agregarse al reactivo del *estado de alerta* mencionado anteriormente. Un método de calificación numérica conveniente consiste en el empleo del mismo sistema numérico, por ejemplo, 4, 3, 2, 1, 0 con cada reactivo. Éste es, por supuesto, el sistema utilizado en las escalas de actitud de puntuación sumada. Sin embargo, en las escalas de calificación, probablemente sea mejor dar tanto la descripción verbal como los valores numéricos.

En las *escalas de calificación gráfica* se combinan líneas o barras con frases descriptivas. El reactivo del estado de alerta, que se expuso antes, podría aparecer de la siguiente manera en forma gráfica:



Tales escalas incluyen muchas variedades: líneas verticales segmentadas, líneas continuas, líneas sin marcas, líneas divididas en intervalos iguales marcados (como la anterior) y otras. Probablemente se trata de las mejores formas de las escalas de calificación y las más utilizadas. Fijan un continuo en la mente del observador; sugieren intervalos iguales, y son claras y fáciles de comprender y de usar. Guilford (1954, p. 268) las sobrestima un poco cuando afirma: "son muchas las virtudes de las escalas de calificación gráfica, y sus fallas son pocas", pero su señalamiento se toma a bien.

Debilidades de las escalas de calificación

Las escalas de calificaciones tienen dos serias debilidades, una es extrínseca y la otra intrínseca. El defecto extrínseco consiste en que son aparentemente tan fáciles de construir

y de usar, que se utilizan de forma indiscriminada, a menudo sin conocimiento de sus defectos intrínsecos. No se hará una pausa para mencionar los errores que pueden escabullirse en la construcción y empleo inadecuados de las escalas de calificación. En lugar de eso, se alerta al lector en contra de su uso para cualquiera y todas las necesidades de medición. Primero debe plantearse la pregunta: ¿existe una mejor forma para medir mis variables? Si es así, es necesario utilizarla; si no, entonces se deben estudiar las características de las buenas escalas de calificación, trabajar con esmerado cuidado y poner los resultados de las calificaciones bajo prueba empírica y análisis estadístico adecuado.

El defecto intrínseco de las escalas de calificación es su tendencia al error constante o por sesgo, lo cual no es nuevo para el lector, por supuesto, pues dicho problema se abordó cuando se consideró la fijeza de respuesta. Sin embargo, con las calificaciones es especialmente amenazante para la validez. El error constante de calificación toma varias formas, de las cuales la más penetrante es el famoso *efecto de halo*. Se trata de la tendencia a valorar un objeto en la dirección constante de una impresión general del objeto. Casos diarios de halo son, por ejemplo, creer que un hombre es virtuoso porque nos agrada, y/o manifestar grandes elogios a los presidentes republicanos y condenar a los demócratas.

El halo se manifiesta con frecuencia en la medición, especialmente con las calificaciones. Los profesores evalúan más alta la calidad de las preguntas de una prueba de ensayo de lo que deberían, porque les simpatiza el examinado. O tal vez califiquen más alto (o más bajo) de lo que deberían la segunda, tercera y cuarta preguntas, debido a que la primera pregunta estuvo bien contestada (o mal contestada). La evaluación que hace el maestro del rendimiento de los niños, en la cual influye la docilidad o falta de docilidad de los niños, constituye otro caso de halo. Al evaluar a los individuos con escalas de calificación, existe la tendencia de que la calificación de una característica influya en las calificaciones de otras características. Es difícil evitar el halo. Parece ser particularmente fuerte en rasgos que no están claramente definidos, que no son fáciles de observar y que son moralmente importantes (véase Guilford, 1954, p. 279).

Dos fuentes importantes de error constante son el error por severidad y el error por flexibilidad. El *error por severidad* es la tendencia general de calificar demasiado bajo a todos los individuos, en todas las características. Es el del duro crítico: "nadie obtiene un 10 en mis cursos". El *error por flexibilidad* es la tendencia general opuesta de calificar demasiado alto. Éste es el caso del buen amigo que estima a todos, y la estimación se refleja en las calificaciones.

Una fuente exasperante de invalidez de las calificaciones es el *error de tendencia central*, que es la tendencia general a evitar todos los juicios extremos y a calificar justo en el centro de una escala de calificación. Se manifiesta particularmente cuando los calificadores no están familiarizados con los objetos que se están calificando.

Existen otros tipos de error de menor importancia que no se consideran aquí. Es más importante saber cómo enfrentar los tipos de error mencionados antes. Se trata de un tema complejo que no puede estudiarse aquí. Se refiere al lector a Guilford (1954, pp. 280-288, 383, 395-397), donde se discuten a detalle muchas estrategias para lidiar con el error. Los errores sistemáticos pueden tratarse en cierto grado por medio de las medias estadísticas. Guilford ha creado un método ingenioso que utiliza el análisis de varianza. La idea básica es que las varianzas debidas a los participantes, los jueces y las características se extraen de la varianza total de calificaciones; entonces se corrigen las calificaciones. Un método más fácil, cuando se califica una sola característica de los individuos, es el análisis de varianza de dos factores (grupos correlacionados). La confiabilidad también puede calcularse con facilidad. El uso del análisis de varianza para estimar la confiabilidad, como se aprendió antes, fue una contribución de Hoyt (1941). Ebel (1951) aplicó el análisis de varianza a la confiabilidad de las calificaciones.

Las escalas de calificación deben usarse en la investigación del comportamiento. Su uso no garantizado, expedito e inexperto ha sido justamente condenado. Pero esto no debe significar una condena general. Tienen virtudes que las hacen valiosas herramientas de investigación científica: requieren de menos tiempo que otros métodos, por lo común son interesantes y fácil de usar por los observadores, poseen un rango muy amplio de aplicación, y pueden utilizarse con un gran número de características. Hay que añadir el hecho de que es posible usarlas en conjunto con otros métodos, es decir, servirse de ellas como instrumentos de ayuda para las observaciones del comportamiento, y utilizarlas en conjunto con otros instrumentos objetivos, con entrevistas y aun con medidas proyectivas.

Ejemplos de sistemas de observación

Otros sistemas de observación del comportamiento (no mencionados antes) se resumen a continuación, para ayudar al estudiante a tener una idea de la variedad de sistemas posibles y de las maneras en que se construyen y utilizan dichos sistemas. Además, el lector comprenderá mejor cuándo es apropiada una observación del comportamiento.

Muestreo de tiempo del comportamiento de juego de niños con problemas auditivos

El comportamiento de juego se considera un componente importante del desarrollo normal del niño. No obstante, los niños con problemas auditivos tienen déficits de comunicación que interfieren con el desarrollo normal del juego. Se ha descubierto que los niños con problemas auditivos se involucran en juegos menos complejos y menos intercambio social, que los niños que no tienen estos problemas. En su estudio sobre el comportamiento de juego de niños con problemas auditivos, Esposito y Koorland (1989) utilizaron una técnica de muestreo de tiempo momentáneo para registrar el comportamiento de dos niños en diferentes lugares de juego (uno de 3.5 años y otro de 5 años). La meta era utilizar los datos para comparar el comportamiento de niños con problemas auditivos cuando se relacionan y cuando no se relacionan con niños sin problemas auditivos. Esto incluyó la observación y el registro del comportamiento de cada niño durante intervalos de 10 segundos en dos sesiones de 10 minutos por día, durante cuatro días a la semana, por dos semanas, durante el juego libre en interiores. Un ambiente estaba integrado y el otro no. Se codificó el comportamiento de juego libre de cada niño, de acuerdo con las categorías de juego definidas por Higginbotham, Baker y Neill (1980). Existen ocho categorías principales de juego que se clasifican, a su vez, en *juego social*, *juego cognitivo* y *sin juego*. Los investigadores encontraron diferencias en el comportamiento de juego entre los dos tipos de ambientes. Si la interacción de los compañeros durante el juego contribuye al desarrollo normal del niño, entonces sus resultados sugieren que los ambientes integrados son más adecuados para los niños con problemas auditivos.

Observación y evaluación de la enseñanza universitaria

En uno de los relativamente pocos —y mejores— estudios sobre los maestros y la enseñanza universitaria, Isaacson, McKeachie, Milholland y Lin (1964), después de mucho trabajo preliminar en los reactivos y sus dimensiones o factores, les pidieron a los estudiantes universitarios que calificaran y evaluaran a sus maestros, con base en el recuerdo de sus

observaciones e impresiones. Se ha publicado un número de estudios similares desde la aparición de dicho estudio; sin embargo, continúa siendo uno de los mejores. El sistema de observación presentado no se diseñó deliberadamente para medir variables, sino para ayudar a la evaluación del desempeño de los maestros. No obstante, sus dos dimensiones básicas pueden, por supuesto, utilizarse como variables de investigación. Un aspecto notable de los estudios que evalúan maestros universitarios es que los investigadores no parecen estar conscientes de que el propósito de dichos sistemas de observación debe ser la mejora de la instrucción (o que utilicen sus dimensiones como variables de investigación), y no propósitos de tipo administrativo (véase Kerlinger, 1971).

Isaacson *et al.* utilizaron una escala de calificación con 46 reactivos e instruyeron a los estudiantes a responder de acuerdo con la frecuencia de la ocurrencia de ciertos actos de comportamiento, y no de acuerdo con el hecho de si los comportamientos eran deseables o indeseables. Su interés básico se centraba en las dimensiones o variables subyacentes a los reactivos. Ellos encontraron seis de tales dimensiones (factores). La primera dimensión estaba relacionada con las habilidades generales de enseñanza.

Aunque los seis factores son importantes debido a que parecen mostrar diversos aspectos de la enseñanza (por ejemplo, la estructura, que es la organización que hace el instructor del curso y sus actividades; y el *rapport*, que es el aspecto más interactivo de la enseñanza y la amistad), el enfoque aquí será en el primer factor. A continuación se presentan tres de los reactivos:

> Él transmitía su material en una forma interesante. Él estimulaba la curiosidad intelectual de sus estudiantes. Él daba explicaciones claras y sus explicaciones iban al grano (p. 347).

Sin embargo, el reactivo más efectivo era incluso más general:

¿Cómo calificarías a tu instructor respecto a su habilidad general (completa) de enseñanza?

a) Un instructor sobresaliente y estimulante
b) Un instructor muy bueno
c) Un buen instructor
d) Un instructor adecuado, pero no estimulante
e) Un instructor pobre e inadecuado

Mientras que es posible cuestionar el hecho de denominar a este estudio y a otros parecidos como estudios de observación, sí existe cierta observación, aunque es bastante diferente al tener que recordarse y ser indirecta, global y altamente inferencial y, finalmente, mucho menos sistemática que la observación real. Se pide a los estudiantes que recuerden y califiquen comportamientos a los cuales pueda no haber puesto especial atención. No obstante, el estudio de Isaacson *et al.*, aunado a otros estudios, demostró que esta forma de observación puede utilizarse en la confiabilidad de la evaluación del instructor y del curso.

Evaluación de la observación del comportamiento

No cabe duda de que la observación objetiva del comportamiento humano ha avanzado más allá de la etapa rudimentaria. Los avances, al igual que otros avances metodológicos y de medición logrados en los pasados 10 a 20 años, han resultado sorprendentes. El incremento del dominio y sofisticación psicométricos y estadísticos se manifiestan en la observación y evaluación del comportamiento real y recordado. La investigación científica

social puede beneficiarse —y efectivamente lo hará— con estos avances. Muchos problemas de investigación educativa, por ejemplo, demandan enérgicamente observaciones del comportamiento: los niños en salones de clases que interactúan entre sí y con sus maestros, administradores y maestros que discuten problemas escolares en juntas de personal, consejos de educación que trabajan sobre decisiones políticas, etcétera. Tanto la investigación básica como la aplicada, en especial la investigación que involucra procesos y decisiones grupales, se benefician de la observación directa. Además puede utilizarse en estudios de campo, experimentos de campo y experimentos de laboratorio. He aquí un modelo metodológico que es esencialmente igual en las situaciones de campo y de laboratorio.

La dificultad del uso de sistemas de escala completa sin duda ha desmotivado el uso de la observación en la investigación del comportamiento. No obstante, las observaciones deben usarse cuando las variables de estudios de investigación sean de naturaleza interactiva e interpersonal, y cuando se desee estudiar las relaciones entre el comportamiento real, como las técnicas de manejo de clase o las interacciones grupales, y otros comportamientos o variables atributivas. Aunque es importante preguntar acerca del comportamiento, no hay un sustituto para observar, de forma tan directa como sea posible, lo que en realidad hace la gente cuando se enfrenta con diferentes circunstancias y diferentes personas. Además, quizá no sea necesario usar los sistemas de observación más grandes en la mayor parte de la investigación del comportamiento. Como se señaló anteriormente, es posible diseñar sistemas pequeños para propósitos específicos de investigación. El sistema limitado de Keeves (1972) mostró ser sumamente apropiado para sus propósitos. En cualquier caso, la investigación científica del comportamiento requiere de observaciones directas e indirectas del comportamiento, y los medios técnicos para la realización de dichas observaciones se están volviendo cada vez más adecuados y disponibles. En el siglo XXI se debe lograr una comprensión y mejoría considerables de los métodos de observación, así como el incremento de su uso significativo.

Sociometría

Constantemente se evalúa a la gente con quien uno trabaja, con quien va a la escuela y con quien vive en casa. Se juzga su adecuación para el trabajo, para el juego y para vivir con nosotros. Además los juicios se basan en nuestras observaciones sobre su comportamiento en distintas situaciones. Se dice que juzgamos con base en nuestra "experiencia". La forma de medición que ahora se considera sociometría está basada en muchas de esas observaciones informales. Nuevamente, el método se basa en observaciones recordadas y en los juicios inevitables que se hacen de la gente, después de observarla.

Sociometría y elección sociométrica

Sociometría es un término amplio que indica un número de métodos de reunión y análisis de datos sobre los patrones de elección, comunicación e interacción de los individuos de grupos. Se podría decir que la sociometría es el estudio y medición de la elección social. También se ha considerado como un medio para estudiar las atracciones y repulsiones de los miembros de un grupo.

Se le pide a una persona que elija a una o más personas, de acuerdo con uno o más criterios establecidos por el investigador: ¿con quién le gustaría trabajar? ¿Con quién le gustaría jugar? Entonces, la persona elige una, dos, tres o más opciones de entre los miembros de su propio grupo (generalmente) o de otros grupos. ¿Qué podría ser más simple y

natural? El método funciona bien tanto para los miembros de jardín de niños como para científicos atómicos.

La elección sociométrica debe comprenderse ampliamente: no significa tan sólo "elección personas", también puede significar "elección de líneas de comunicación", "elección de líneas de influencia" o "elección de grupos minoritarios". Las elecciones dependen de las instrucciones y preguntas dadas a los individuos. A continuación se presenta una muestra de una lista de preguntas e instrucciones sociométricas:

Ejemplo

¿Con quién le gustaría trabajar (jugar, sentarse, etcétera)?
¿Quiénes son los miembros de este grupo (grupo de edad, clase, club, por ejemplo) que a usted le agradan más (menos)?
¿Quiénes son los tres mejores (peores) alumnos de su clase?
¿A quién elegiría para representarlo en un comité para mejorar el bienestar de la facultad?
¿Quiénes son los cuatro individuos que tienen el mayor prestigio en su organización (clase, compañía, equipo)?
¿Cuáles son los dos grupos de gente más aceptables (menos aceptables) para usted como vecinos (amigos, socios de negocios, socios profesionales)?

En efecto, existen muchas posibilidades. Algunas de ellas se analizan en Lindzey y Byrne (1968). Además, tales posibilidades llegan a multiplicarse simplemente al preguntar: ¿quién piensa usted que lo elegiría para...? Y ¿quién piensa usted que el grupo elegiría para...? También se pide a los participantes que ordenen por rangos a otros utilizando criterios sociométricos, siempre y cuando no sean demasiados a ordenar; o se pueden utilizar escalas de calificación.

Si se solicita a los miembros de un grupo u organización que se califiquen entre sí utilizando uno o más criterios, las instrucciones sociométricas quedarían de manera similar a la siguiente:

Ejemplo

Aquí hay una lista de los miembros de su grupo. Califique a cada uno de acuerdo con el hecho de si a usted le gustaría trabajar con él en un comité encargado de establecer un conjunto de estatutos. Utilice los números 4, 3, 2, 1 y 0, donde 4 signifique que le gustaría mucho trabajar con él, 0 que no desearía trabajar con él en lo absoluto, y los otros números representan grados intermedios de cuánto le agradaría trabajar con él.

Evidentemente es posible utilizar otros métodos de medición. La principal diferencia radica en que la sociometría siempre implica ideas como la de elección, interacción, y comunicación sociales, y las influencias que están detrás de ellas.

Métodos de análisis sociométrico

Existen tres formas básicas de análisis sociométrico: matrices sociométricas, sociogramas o gráficas dirigidas e índices sociométricos. De todos los métodos de análisis sociométrico, las *matrices sociométricas,* que vamos a definir, quizás contengan las posibilidades e implicaciones más importantes para el investigador del comportamiento. Los *sociogramas* son diagramas o tablas de las elecciones realizadas en los grupos. Los sociogramas o gráficas dirigidas se tratarán muy poco, ya que se utilizan más para propósitos prácticos que de investigación, y su análisis matemático es difícil y requiere de mayor espacio del que se le

puede brindar aquí. Al lector que requiera mayor detalle y explicación se le recomienda consultar a Fienberg y Wasserman (1981) y Ove (1981). Los *índices sociométricos* son números sencillos calculados a partir de dos o más números producidos por datos sociométricos. Indican características sociométricas de individuos y grupos.

Matrices sociométricas

Se aprendió antes que una *matriz* es una tabla o arreglo rectangular de números o de otros símbolos. Para aquellos que no estén familiarizados con las matrices, se recomienda la lectura de Lindzey y Byrne (1968, pp. 470-473), donde encontrará un buen repaso sobre el análisis de matrices. En Kemeny, Snell y Thompson (1966, pp. 217-250, 384-406) se presentan explicaciones de operaciones elementales de matrices y matrices sociométricas. Una buena exposición elemental de matrices puede encontrarse en Davis (1973). Una revisión antigua, pero aun valiosa, de métodos matemáticos y estadísticos para el análisis de la estructura y comunicación de grupo es la que presentan Glanzer y Glaser (1959).

La sociometría casi siempre está relacionada con matrices cuadradas o de $n \times n$, donde n es igual al número de personas en un grupo. Los renglones de la matriz se denominan i, las columnas se denominan j. Por supuesto, i y j pueden representar cualquier número y cualquier persona en el grupo. Si se escribe a_{ij}, significa un dato en el renglón i y en la columna j de la matriz, o, de forma más simple, cualquier dato en la matriz. Es conveniente escribir *matrices sociométricas*, que son matrices de números que expresan todas las elecciones de los miembros de cualquier grupo.

Suponga que un grupo de cinco miembros respondió a la pregunta sociométrica "¿con quién le gustaría trabajar en tal o cual proyecto durante los próximos dos meses? Elija dos individuos". Las respuestas a la pregunta sociométrica son, evidentemente, *elecciones*. Si un miembro del grupo elige a otro miembro, la elección se representa por un 1. Si un miembro del grupo no elige a otro, la falta de elección está representada por un 0. (Si se hubiera incluido el rechazo, se podría utilizar −1.) La matriz sociométrica de elecciones, C, de esta situación grupal hipotética se presenta en la tabla 31.1.

Es posible analizar la matriz de varias formas. Pero primero es necesario asegurarse de saber cómo leer la matriz. Probablemente resulte más fácil leerla de izquierda a derecha, de i a j. El miembro i elige (o no elige) al miembro j. Por ejemplo, a elige b y e; c elige d y e. Algunas veces es conveniente expresarse en voz pasiva, "b fue elegido por a, d y e" o "c no fue elegido por ninguno".

TABLA 31.1 *Matriz sociométrica de elección: grupo de cinco miembros, pregunta de dos elecciones*[a]

		j				
		a	***b***	***c***	***d***	***e***
	a	0	1	0	0	1
	b	1	0	0	0	1
i	***c***	0	0	0	1	1
	d	0	1	0	0	1
	e	1	1	0	0	0
	Σ	2	3	0	1	4

[a] El individuo i elige al individuo j. Es decir, la tabla se lee por renglones: b elige a y e. También puede leerse por columnas: b es elegido por a, d y e. Las sumas de la sección inferior indican el número de elecciones que recibe cada individuo.

El análisis de una matriz por lo común se inicia observando quién elige a quién. Con una matriz simple esto es fácil. Existen tres tipos de elecciones: simple o de un factor, mutua o de dos factores y sin elección. Se analizarán primero las elecciones simples. (Esto se explicó en el párrafo anterior.) Una elección *simple* de un factor es cuando *i* elige *j*, pero *j* no elige *i*. En la tabla 31.1, *c* elige *d*, pero *d* no eligió *c*. Se escribe: $i \rightarrow j$, o $c \rightarrow d$. Una elección *mutua* es donde *i* elige *j* y *j* también elige *i*. En la tabla, *a* elige *b* y *b* elige *a*. Se escribe $i \leftrightarrow j$ o $a \leftrightarrow b$. Se podrían contar las elecciones mutuas en la tabla 31.1: $a \leftrightarrow b$, $a \leftrightarrow e$, $b \leftrightarrow e$.

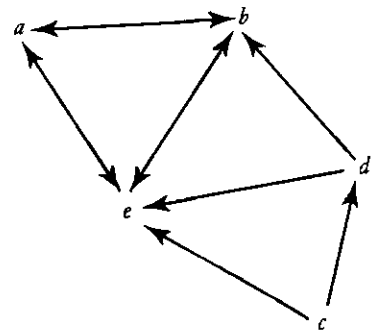
El grado en que cualquier miembro es elegido se percibe fácilmente al sumar las columnas de la matriz. En efecto, *e* es "popular": esta persona fue elegida por los demás miembros del grupo; *a* y *b* fueron elegidos dos y tres veces, respectivamente. De hecho *c* no es popular en lo absoluto: nadie eligió a esta persona; *d* tampoco es popular, ya que fue elegida sólo una vez. Si se les permite a los individuos un número ilimitado de elecciones, es decir, si se les instruye para elegir cualquier número de individuos, entonces las sumas de renglones cobran significado. Se les puede decir a los participantes que elijan una, dos, tres o más personas. Parece ser que tres es un número común de elecciones. El número permitido debe determinarse con base en los propósitos de investigación. Estas sumas se denominarían, por ejemplo, índices de *gregarismo*.

Existen otros métodos de análisis de matriz que son potencialmente útiles para los investigadores. Por ejemplo, por medio de operaciones de matrices relativamente simples se pueden determinar puntos y cadenas de influencia en grupos pequeños y grandes. Sin embargo, dichos aspectos van más allá del alcance de esta obra.

Sociogramas o gráficas dirigidas

Los análisis más simples son similares a aquellos que se acaban de exponer. Pero con una matriz más grande que la de la tabla 31.1 es casi imposible asimilar las complejidades de las relaciones de elección. Para esto, los *sociogramas* son útiles, siempre y cuando el grupo no sea demasiado grande. Ahora se cambia el nombre de "sociograma" por el de "gráfica dirigida", que es un término matemático más general que se aplica a cualquier situación donde *i* y *j* estén en alguna relación *R*. En lugar de decir "*i* elige *j*", es muy posible decir "*i* influye en *j*", "*i* se comunica con *j*", "*i* es amigo de *j*", o "*i* domina *j*". En lenguaje simbólico, se escribe: *iRj*. De manera específica, de los ejemplos expuestos arriba, se escribe: *iCj* (*i* elige *j*), *iIj* (*i* influye en *j*), *iCj* (*i* se comunica con *j*), *iFj* (*i* es amigo de *j*), *iDj* (*i* domina *j*). Cualquiera de estas interpretaciones puede describirse por una matriz como la de la tabla 31.1 y por medio de una gráfica dirigida. En la figura 31.1 se presenta una gráfica dirigida.

FIGURA 31.1



Con una simple ojeada se percibe que *e* es el centro de elección. A esta persona se le llamaría *líder*; o se diría que se trata de una persona agradable o competente. Aún más importante, note que *a*, *b* y *c* se eligieron entre sí. Conforman un eslabón. Un *eslabón* se define como tres o más individuos que se eligen entre sí. Festinger, Schachter y Back (1950) presentan un valioso método para identificar eslabones dentro de grupos. Al buscar más flechas dobles no se encuentra ninguna otra. Ahora se buscan individuos que no tengan flechas apuntando hacia ellos: *c* es uno de dichos individuos. Se afirma que *c* no es elegido, o que es rechazado. Si se desea más información sobre los eslabones y su identificación, consulte a Glanzer y Glaser (1959, pp. 326-327), quienes esbozan de forma sucinta métodos de la multiplicación de matrices binarias (1, 0), cuya aplicación ofrece ideas útiles respecto a la estructura grupal.

Observe que las gráficas dirigidas y las matrices dicen lo mismo. El número de elecciones que *a* recibe se observa añadiendo los números 1 en la columna *a* de la matriz. Se obtiene la misma información añadiendo el número de puntas de flecha que señalan hacia *a* en la gráfica. Para grupos de tamaño pequeño y mediano, y para propósitos descriptivos, las gráficas constituyen medios excelentes para sintetizar relaciones grupales. No son tan adecuadas para grupos más grandes (mayores de 20 miembros) ni para propósitos más analíticos, ya que se tornan difíciles de construir y de interpretar. Además, diferentes individuos pueden obtener gráficas diferentes utilizando los mismos datos. Las matrices son generales y, si se manejan apropiadamente, no son tan difíciles de interpretar. Con los mismos datos, diferentes individuos deben escribir exactamente las mismas matrices.

Índices sociométricos

En sociometría son posibles muchos índices. A continuación se muestran tres de ellos. El lector encontrará otros en la literatura. La presente exposición se basa, en su mayoría, en Proctor y Loomis (1951).

Un índice simple pero útil es:

$$EE_j = \frac{\sum c_j}{n-1} \tag{31.1}$$

donde EE_j = el estado de elección de la persona j; $\sum c_j$ = la suma de elecciones en la columna j; y n = el número de individuos en el grupo (se utiliza $n - 1$ porque no se puede contar al propio individuo). Para el sujeto E de la tabla 31.1, $CS_E = 4/4 = 1.00$ y para el sujeto $EE_a = 2/4 = .50$. *EE* revela qué tanto o qué tan poco es elegido un individuo; en pocas palabras, se trata del *estatus de elección*. Desde luego, es posible tener un índice de rechazo de elección. Simplemente se pone el número de los 0 en cualquier columna en el numerador de la ecuación 31.1.

Las medidas sociométricas grupales quizás sean más interesantes. Una medida de la cohesión de un grupo es:

$$Co = \frac{\sum (i \leftrightarrow j)}{\left[\frac{n(n-1)}{2}\right]} \tag{31.2}$$

La cohesión de grupo está representada por Co y $\sum(i \leftrightarrow j)$ igual a la suma de elecciones mutuas (o pares mutuos). Este útil índice es la proporción de las elecciones mutuas del número total de pares posibles. En un grupo de cinco miembros, el número total de pares posibles son cinco cosas, tomando dos a la vez:

$$\left\lfloor \frac{5}{2} \right\rfloor = \frac{5(5-1)}{2}$$

Si en una situación de elección ilimitada hubiera dos elecciones mutuas, entonces $C_o = 2/10 = .20$, un grado más bien bajo de cohesión. En el caso de elección limitada, la fórmula es:

$$Co = \frac{\Sigma(i \leftrightarrow j)}{\left[\frac{dn}{2}\right]} \tag{31.3}$$

donde d es igual al número de elecciones permitidas a cada individuo. Para C de la tabla 31.1 $Co = 3/(2 \times 5/2) = 3/5 = .60$, un grado sustancial de cohesión.

Usos de la sociometría en investigación

Debido a que los datos de la sociometría parecen ser tan diferentes a otros tipos de datos, los estudiantes tal vez encuentren difícil considerar la medición sociométrica como medición. No cabe duda de que los datos sociométricos son diferentes; pero son el resultado de observación y *son medidas*. Puesto que son medidas, también tienen los mismos problemas básicos de la medición, como la validez y la confiabilidad. Lindzey y Byrne (1968) analizan dichos aspectos de medición. Las mediciones sociométricas son útiles, por ejemplo, para clasificar individuos y grupos. En el estudio clásico del Bennington College, Newcomb (1943) midió el prestigio individual al pedir a los estudiantes que nombraran cinco estudiantes a los que ellos elegirían como los más valiosos para representar al Bennington College en una importante reunión de estudiantes, de todo tipo de universidades estadounidenses. El autor después agrupó a los estudiantes de acuerdo con la frecuencia de elección y relacionó esta medida de *prestigio sociométrico* con el conservadurismo político y económico. Al leer los ejemplos de tal sección, el estudiante debe comprender con claridad que la sociometría es un método de observación y recolección de datos que, como cualquier otro método de observación, obtiene medidas de variables.

Prejuicio en las escuelas

Rooney-Rebeck y Jason (1986) investigaron los efectos de la tutoría de compañeros de grupo cooperativo sobre las relaciones interétnicas de niños estadounidenses negros, blancos e hispanos de primero y tercer grados. Realizaron observaciones directas de las interacciones sociales en el patio de juegos, antes y después de un programa de intervención de ocho semanas. Los índices sociométricos se calcularon para medir las asociaciones interétnicas. Rooney-Rebeck y Jason encontraron un incremento en las interacciones interétnicas y en las elecciones sociométricas en los niños de primer grado, quienes también mostraron un incremento en sus calificaciones en aritmética y lectura. Sin embargo, no se encontraron cambios significativos entre los niños de tercer grado respecto a sus asociaciones interétnicas ni a su desempeño académico. A partir de tales resultados, parece ser que una estructura cooperativa de tutoría de compañeros en el salón de clases tiene beneficios para niños de primer grado, pero no necesariamente para niños de tercer grado. Los investigadores sugieren que ello puede deberse a una experiencia limitada en competencia académica y en prejuicios étnicos manifiestos, en los niños de primer grado.

Sociometría y estereotipos

Gross, Green, Storck y Vanyur (1980) utilizaron una combinación de calificaciones sociométricas y de estereotipos para estudiar las actitudes de las personas. Participantes de uno y otro sexo observaron una filmación, ya sea de un estímulo de una mujer homosexual o de un estímulo de un hombre homosexual. Los participantes se dividieron en tres grupos. A uno de los grupos de participantes se le informó, antes de ver la filmación, que la persona era homosexual. Al segundo grupo se le informó lo mismo, pero después de haber visto la filmación, y al tercer grupo no se le informó nada. Los resultados revelaron que las calificaciones de los rasgos eran más estereotípicas y las calificaciones sociométricas menos favorables para la persona del estímulo en ambas condiciones de revelación: inmediata o retardada. Aquellas personas identificadas como homosexuales fueron juzgadas de manera más estereotipada por los participantes del mismo sexo. Los hombres, por lo general, calificaron con mayor dureza a la persona en la situación retardada que en las condiciones de revelación inmediata.

Sociometría y estatus social

Se han realizado diversos estudios utilizando la sociometría para medir el estatus social. A continuación se presenta uno de ellos, el cual involucra el estatus social entre niños en edad escolar.

Inderbitzen, Walters y Bukowski (1997) estudiaron la relación entre grupos de estatus sociométrico y la ansiedad social en las relaciones entre compañeros adolescentes. Los participantes consistieron en 973 estudiantes del sexto grado de primaria al tercer grado de secundaria. El número de niños y niñas era casi igual. Los participantes completaron la *escala de ansiedad social para adolescentes* (Social Anxiety Scale for Adolescents) y una tarea de nominación sociométrica, la cual incluía descripciones comportamentales como el que más me gusta, el que menos me gusta, el que comienza las peleas la mayoría de las veces, el del mejor sentido del humor, el líder de la clase, el más fácil de influenciar y el más cooperador. Las nominaciones sociométricas se utilizaron después para clasificar a los estudiantes en grupos de estatus sociométrico estándar, tales como popular, promedio, rechazado, no tomado en cuenta y polémico; así como en subgrupos de rechazo: agresivo rechazado o sumiso rechazado. Los resultados indicaron que los estudiantes clasificados como rechazados y no tomados en cuenta reportaron mayor ansiedad social que aquellos clasificados como promedio, populares o polémicos. Además los estudiantes sumisos rechazados reportaron significativamente mayor ansiedad social que los estudiantes agresivos rechazados o promedio. A través del uso de la sociometría es posible descubrir la existencia de problemas entre compañeros adolescentes.

Raza, creencia y elección sociométrica

Graham y Cohen (1997) estudiaron la relación entre raza y sexo en las relaciones entre niños compañeros. Las relaciones se midieron a través de calificaciones sociométricas y amistades observadas. Dicho estudio observó a cada estudiante en una sola escuela primaria, incluyendo del 1o. al 6o. grados. La escuela fue elegida a causa de que tenía casi igual número de niños estadounidenses negros y blancos en cada clase. Además, los niños fueron divididos en dos grupos de edad: del 1o. al 3er. grados y del 4o. al 6o. grados. Independientemente de la edad, raza o sexo y de las medidas de relación, los niños favorecieron a los compañeros del mismo sexo más que a los compañeros de la misma raza, lo cual indica que los niños prefirieron las interacciones sociales con otros niños, sin importar la raza, y que las niñas prefirieron interactuar con otras niñas, sin importar la raza. Aunque los niños negros mayores tuvieron más amigos mutuos de la misma raza que de otra raza, los niños negros se mostraron más aceptantes de los niños blancos, que a la inversa. A

pesar de que hubo algunas preferencias por la misma raza, las evaluaciones entre razas fueron, por lo general, bastante positivas en ambas medidas de las relaciones entre compañeros.

La sociometría es un método simple, económico y natural de observación y de recolección de datos. Siempre que se involucren actos humanos, tales como la elección, la influencia, la dominación y la comunicación —especialmente en situaciones de grupo— pueden utilizarse métodos sociométricos, ya que poseen considerable flexibilidad. Si se les define de forma general, se adaptan a una amplia variedad de investigaciones tanto en el laboratorio como en el campo. Sus posibilidades de cuantificación y de análisis, aunque no siempre percibidas en la literatura, son recompensantes. La habilidad para utilizar la simple asignación de los 1 y los 0 es particularmente afortunada, debido a que pueden aplicarse métodos matemáticos poderosos a los datos, con resultados interpretables y significativos únicos. Los métodos de matrices constituyen el ejemplo sobresaliente, pues con ellos se descubren eslabones en grupos, canales de comunicación e influencia, patrones de cohesión, niveles de conexión, jerarquización, etcétera.

Como se mencionó antes, los métodos sociométricos, al igual que otros métodos, no carecen de defectos. Longshore (1982), por ejemplo, recomienda el empleo de otros métodos no intrusivos en lugar de la sociometría, cuando se estudian problemas delicados como la disgregación. Señala que los científicos sociales no han logrado hallazgos consistentes sobre la disgregación, a causa de que los investigadores se han preocupado más por los resultados de la disgregación, que por el amplio rango de condiciones bajo las cuales ocurre dicho fenómeno. Longshore considera que la evaluación de resultados a corto plazo debe evaluarse a través de medidas como la observación no intrusiva de los grupos de juego o las clases, en lugar de la sociometría.

RESUMEN DEL CAPÍTULO

1. Existe mucha controversia y debate respecto a la observación y a los métodos de observación.
2. Dos modelos básicos de observación son los siguientes:
   - Se observan los comportamientos abiertos de la gente (por ejemplo, lo que hacen y lo que dicen)
   - Se pregunta a la gente sobre sus propias acciones y sobre el comportamiento de otros
3. El observador puede representar un problema importante en el estudio. El observador tal vez haga inferencias incorrectas acerca del comportamiento observado.
4. El observador llega a afectar a los objetos de observación al formar parte del sistema observacional.
5. Las mediciones observadas están sujetas a requisitos de validez y confiabilidad. Al requerir que el observador interprete la observación, puede reducirse la validez.
6. La confiabilidad para las observaciones toma la forma de acuerdo entre jueces u observadores. El comportamiento que se observará a través de la observación directa debe establecerse claramente, con buenas definiciones operacionales.
7. Una tarea fundamental del observador consiste en categorizar los datos observados. Se crean categorías y, conforme se observan ciertos comportamientos, se hace una marca o nota en esa categoría.
8. Las unidades de comportamiento a veces son vagas o muy generales. Algunos investigadores utilizan definiciones operacionales muy estrictas.

9. Toda observación requiere de cierto nivel de interpretación por parte del observador.
10. Los diferentes sistemas de observación varían en su grado de generalización. Algunos son muy generales, mientras que otros son bastante específicos.
11. Los comportamientos pueden muestrearse con las técnicas del muestreo de eventos o del muestreo de tiempo. Cada una posee ventajas y desventajas.
12. Un tipo de observación implica presentar a los observadores con un sistema de observación en forma de escala de calificación. Se les pide que evalúen un objeto en términos de una o más características.
13. Existen cinco tipos de escalas de calificación:
    - listas de chequeo
    - instrumentos de elección forzada
    - escala de calificación de categorías
    - escala de calificación numérica
    - escala de calificación gráfica
14. Las escalas de calificación presentan dos serias debilidades:
    a) Puesto que son fáciles de construir y de utilizar, pueden crearse sin el conocimiento de sus defectos intrínsecos.
    b) Constantemente tienden al error o al sesgo.
15. La sociometría es un término amplio que indica el número de métodos de recolección y análisis de datos sobre patrones de elección, comunicación e interacción de individuos en grupos.
16. Existen tres formas básicas de análisis sociométrico: matrices sociométricas, sociogramas o gráficas dirigidas, e índices sociométricos.

SUGERENCIAS DE ESTUDIO

1. El lector debe estudiar en detalle uno o dos sistemas de observación del comportamiento. Para los alumnos de educación, el sistema de Medley y Mitzel (1963) producirá altos beneficios. Otros lectores desearán estudiar uno o dos sistemas distintos a éste, ya que el artículo tiene autoridad y claridad, y contiene muchos ejemplos. Las dos mejores referencias generales son las de Heyns y Lippitt (1954), y la de Weick (1968) en la primera y segunda ediciones del *Handbook of Social Psychology*. Una antología de 79 sistemas de observación se ha publicado por Simon y Boyer (1970) en cooperación con Research for Better Schools, Inc., un laboratorio regional de educación. El investigador que tenga la intención de realizar observaciones debe consultar esta amplia colección de sistemas. El estudiante de educación encontrará excelentes resúmenes y presentaciones sobre sistemas de observación educativa en Dunkin y Biddle (1974). Los siguientes artículos son valiosos. Boice señala la falta de entrenamiento para realizar observaciones del comportamiento y ofrece sugerencias para alcanzar dicho entrenamiento. Herbert y Attridge proporcionan criterios para los sistemas de observación. También señalan que el conocimiento de dichos sistemas es limitado.

Boice, R. (1983). Observational Skills. *Psychological Bulletin, 93*, 3-29.
Herbert, J. y Attridge, C. (1975). A guide for developers and users of observation systems and manuals. *American Educational Research Journal, 12*, 1-20.

2. Un investigador que estudia los patrones de influencia de los consejos de educación obtuvo la siguiente matriz, de un consejo de educación. (Note que es similar a una situación de elección ilimitada, ya que cada individuo puede influir en todos o en ninguno de los miembros del grupo.) La matriz se lee: *i* influye en *j*.

		j				
		a	***b***	***c***	***d***	***e***
	a	0	0	1	1	0
	b	0	0	0	0	1
i	***c***	1	0	0	1	0
	d	1	0	1	0	0
	e	0	1	0	0	0

   a) ¿Qué conclusiones se obtienen del estudio de esta matriz? ¿ El consejo está dividido? ¿Existe la posibilidad de que haya un conflicto?
   b) Dibuje una gráfica de la situación de influencia e interprétela.
   c) ¿Existe algún eslabón en el consejo? (Defina eslabón como se hizo en el texto.) Si es así, ¿quiénes son sus miembros?
   d) ¿Cuáles miembros poseen el menor número de canales de influencia? ¿Son ellos, entonces, mucho menos influyentes que los otros miembros, siempre y cuando lo demás se mantenga igual?
   [Respuestas: ***c***) Sí: **a, c, d**; ***d***) **b** y **e**.]

3. En el ejercicio 2 de las sugerencias de estudio, calcule la cohesión de grupo utilizando la ecuación 32.2.
   [Respuesta: $C_0 = .40$.]

4. Lea alguno de los siguientes artículos, los cuales aplican la sociometría. Ponga especial atención a la forma en que se realizó.

   Ray, G. E., Cohen, R., Secrist, M. E. y Duncan, M. K. (1997). Relating aggressive and victimization behaviors to children's sociometric status and friendships. *Journal of Social and Personal Relationships, 14*, 95-108. [El estudio se enfoca en la relación entre nominaciones de compañeros estudiantes de 9 a 12 años de edad, sobre comportamiento agresivo y de víctima, y el estatus sociométrico: popular, promedio, rechazado y el número de amigos mutuos.]

   Schwendinger, H. y Schwendinger, J. R. (1997). Charting subcultures at a frontier of knowledge. *British Journal of Sociology, 48*, 71-94. [El artículo describe un programa de investigación para estudiar las subculturas adolescentes por medio del uso de gráficas de grandes redes subculturales. Tales gráficas y redes se generan por métodos de tipo social y sociométricos para lograr entender fenómenos como la adolescencia y la delincuencia.]

5. Consulte uno de los siguientes estudios sobre muestreo de eventos y muestreo de tiempo. Tome nota sobre la forma en que los autores ejecutan cada uno de los procedimientos.

   Bass, R. F. y Aserlind, L. (1984). Interval and time-sample data collection procedures. Methodological issues. *Advances in Learning and Behavioral Disabilities, 3*, 1-39. [Muestreo de tiempo.]

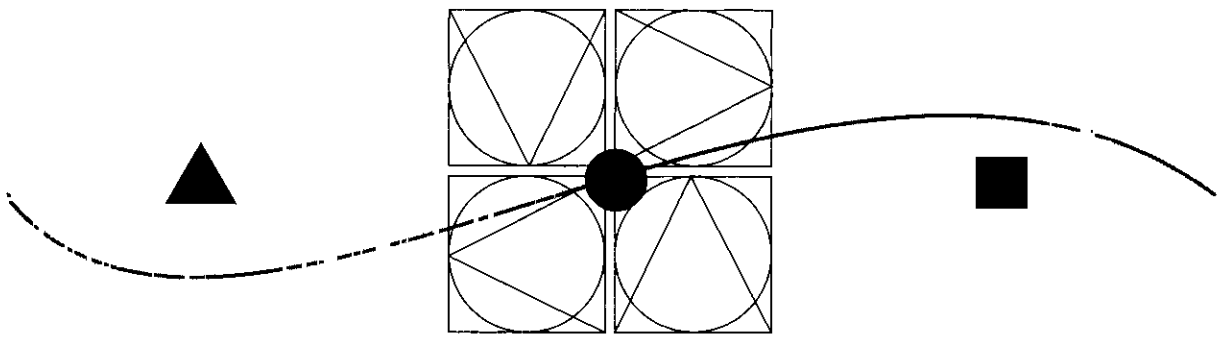
   Brown, K. W. y Moskowitz, D. S. (1998). Dynamic stability of behavior: The rhythms of our interpersonal lives. *Journal of Personality, 66*, 105-134. [Muestreo de eventos.]

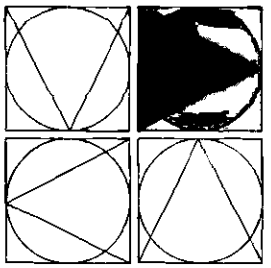
Child, G. H. (1997). A concurrent validity study of teacher's rating for nominated "problem" children. *British Journal of Educational Psychology, 67*, 457-474. [Muestreo de tiempo.]

Peregrine, P. N., Drews, D. R., North, M. y Slupe, A. (1993). Sampling techniques and sampling error in naturalistic observation: An empirical evaluation with implications for cross-cultural research. *Cross-cultural Research: Journal of Comparative Social Science, 27*, 232-246. [Compara el muestreo de eventos, de tiempo y por racimos.]

PARTE DIEZ

MÉTODOS MULTIVARIADOS





CAPÍTULO 32

ANÁLISIS DE REGRESIÓN MÚLTIPLE: FUNDAMENTOS

El *análisis de regresión múltiple* es un método para estudiar los efectos y magnitudes de los efectos de más de una variable independiente sobre una variable dependiente, utilizando los principios de correlación y regresión. Se pasará inmediatamente a la investigación y se dejará la explicación para más adelante.

Tres ejemplos de investigación

¿De qué manera se relacionan la contaminación del aire y el nivel socioeconómico con la mortalidad por padecimientos respiratorios? Lave y Seskin (1970), en su estudio de datos

de personas inglesas y estadounidenses, utilizaron el análisis de regresión múltiple para responder esa pregunta. En sus estudios realizados en barrios ingleses evaluaron los presuntos efectos de la contaminación del aire y el nivel socioeconómico, como variables independientes, sobre las tasas de mortalidad por cáncer pulmonar, bronquitis y neumonía, como variables dependientes.

El efecto general de las dos variables independientes sobre la variable dependiente se expresa por el cuadrado del coeficiente de correlación múltiple, o R^2. La interpretación de este coeficiente es similar a la de r^2, que se estudió con anterioridad. Recuerde que al elevar al cuadrado un coeficiente de correlación se produce un estimado de la cantidad de varianza compartida por dos variables. Dicho concepto es muy utilizado en el análisis de regresión.

Es la proporción de la varianza de la variable dependiente, en este caso la mortalidad, explicada por las dos variables independientes. Las R^2 entre la mortalidad debida a la bronquitis, por una parte, y la contaminación del aire y el nivel socioeconómico, por la otra, oscilaron entre .30 y .78 en diferentes muestras en Inglaterra y Gales, indicando así relaciones sustanciales. Las R^2 para las variables dependientes, mortalidades por cáncer pulmonar y por neumonía, fueron similares. La regresión múltiple también le permite al investigador aprender algo sobre las influencias relativas de las variables independientes. En la mayoría de las muestras, la contaminación del aire fue más importante que el nivel socioeconómico. Como análisis de "control", Lave y Seskin (1970) estudiaron otros tipos de cáncer que se suponía que no se verían afectados por la contaminación del aire. Las R^2 fueron consistentemente más bajas, tal como se esperaba. Extensiones de la investigación hacia áreas metropolitanas de Estados Unidos produjeron resultados similares. Los estudiantes deben tener en cuenta la explicación previa sobre la dificultad para interpretar resultados no experimentales. No obstante, Lave y Seskin construyeron un caso fuerte, a pesar de que algunas de sus interpretaciones resultaron cuestionables. Hasta este punto, sería adecuado que los lectores regresaran a la sección "Relaciones multivariadas y regresión" en el capítulo 5.

¿De qué manera se relacionan la moderación en la dieta, la ingesta energética y la actividad física con el peso corporal? Klesges, Isbell y Klesges (1992) utilizaron un análisis de regresión múltiple para intentar responder esa pregunta. El estudio es interesante a causa del número de problemas de salud asociados con la obesidad. Estos investigadores recolectaron los datos de 287 adultos, durante un año. La variable dependiente era el cambio de peso desde la línea base hasta el seguimiento un año después. Las variables independientes eran el peso en la línea base, el índice de la masa corporal, la puntuación de moderación, la edad, la ingesta energética total, el porcentaje de ingesta de grasa, el porcentaje de carbohidratos y los niveles de actividad física. Las mediciones de la estatura y del peso de tales participantes se utilizaron para estimar la masa corporal. Cada semana se tomaban las medidas del consumo de la dieta; la actividad física se midió a través de un cuestionario de actividad física con 16 reactivos que representaban actividades físicas. La restricción se midió con una escala de moderación. Se realizaron dos análisis separados de regresión; uno para los hombres y otro para las mujeres.

La R^2 entre la variable dependiente, el cambio en el peso, y una combinación lineal de las variables dependientes fue de .13 para los hombres y de .21 para las mujeres, lo cual indica que las variables independientes estudiadas por Klesges *et al.* explicaron sólo el 13 por ciento de la variabilidad observada en el cambio de peso de los hombres, y sólo el 21 por ciento en el caso de las mujeres. La cifra de los hombres no fue estadísticamente significativa al nivel $\alpha = .05$. No obstante, la ecuación de regresión para las mujeres fue significativa al nivel $\alpha = .01$. Sin embargo, a través del empleo de la regresión, estos investigadores fueron capaces de determinar que para los hombres el peso corporal y la

masa corporal iniciales representaron las dos variables más importantes para explicar el cambio en el peso.

En un estudio sobre la predicción del promedio de calificaciones de preparatoria, Holtzman y Brown (1968) utilizaron dos medidas de las variables independientes: hábitos y actitudes de estudio (*hábitos*) y la aptitud escolar (*aptitud*). Las correlaciones entre el promedio de preparatoria y los hábitos y la aptitud en primero de secundaria ($N = 1\,684$) fueron .55 y .61. La correlación entre hábitos y aptitud fue de .32. ¿Qué cantidad más de varianza se explicaría al añadir la medida de aptitud escolar a la medida de hábitos de estudio? Si se combinan hábitos y aptitud de manera óptima para predecir el promedio, se obtiene una correlación de .72. Entonces, la respuesta a la pregunta es $.72^2 - .55^2 = .52 - .30 = .22$, o 22 por ciento más de la varianza del promedio se explica al añadir la aptitud y los hábitos.

Se trata de ejemplos de análisis de regresión múltiple. La idea básica es la misma que en la correlación simple, excepto que se utilizan k variables independientes, cuando k es mayor que 1, para predecir la variable dependiente. En el análisis de regresión simple una variable, X, se utiliza para predecir otra variable, Y. En el análisis de regresión múltiple, las variables $X_1, X_2, \ldots, X_k$, sirven para predecir Y. El método y los cálculos se realizan de tal manera que den la "mejor" predicción posible, dadas las correlaciones entre todas las variables. En otras palabras, en lugar de decir "Si X entonces Y", se dice "Si $X_1, X_2, \ldots, X_k$, entonces Y", y los resultados de los cálculos indican qué tan "buena" es la predicción, y aproximadamente qué cantidad de la varianza de Y está explicada por la "mejor" combinación lineal de las variables independientes.

Análisis de regresión simple

Se afirma que se estudia la regresión de las puntuaciones de Y sobre las puntuaciones de X. Se busca estudiar de qué forma las puntuaciones de Y "regresan hacia", cómo "dependen de", las puntuaciones de X. Galton (véase Cowles, 1989), quien fue el primero en mencionar la noción de correlación, obtuvo la idea a partir del concepto de "regresión hacia la mediocridad", un fenómeno observado en estudios sobre la herencia. (El símbolo r empleado para el coeficiente de correlación, originalmente significó "regresión".) Los hombres altos tenderán a tener hijos más bajos; y los hombres bajos, hijos más altos. Entonces, las estaturas de los hijos tienden a "regresar a" o "a volver a" la media poblacional. Estadísticamente, si se desea predecir Y a partir de X, y la correlación entre X y Y es cero, entonces la mejor predicción es la media. Es decir, para cualquier X, por ejemplo X_7, sólo es posible predecir la media de Y. Sin embargo, a mayor correlación, habrá una mejor predicción. Si $r = 1.00$, entonces la predicción es perfecta. En la medida en que la correlación se aleje de 1.00, en esa misma medida las predicciones de X a Y serán menos perfectas. Si se grafican los valores de X y Y cuando $r = 1.00$, todos se encontrarán en una línea recta. A mayor correlación, más cerca estarán los valores graficados a la línea de regresión (véase capítulo 5).

Para ilustrar y explicar el concepto de regresión estadística se utilizan dos ejemplos ficticios con números simples. Los números usados en los ejemplos son iguales, excepto que se ordenan de manera diferente. Los ejemplos se toman del capítulo 15 donde, al considerar el análisis de varianza, se estudiaron los efectos de la prueba F de la correlación entre grupos experimentales. Los ejemplos se presentan en la tabla 32.1. En el ejemplo de la izquierda, denominado A, la correlación entre los valores de X y de Y es .90; mientras que en el ejemplo de la derecha, denominado B, la correlación es 0. En la tabla también se presentan ciertos cálculos necesarios para realizar el análisis de regresión: las sumas y las

TABLA 32.1 *Análisis de regresión de dos conjuntos de puntuaciones*

A. $r = .90$					B. $r = .00$				
Y	X	XY	Y'	d	Y	X	XY	Y'	d
1	2	2	1.2	−.2	1	5	5	3	−2
2	4	8	3.0	−1.0	2	2	4	3	−1
3	3	9	2.1	.9	3	4	12	3	0
4	5	20	3.9	.1	4	6	24	3	1
5	6	30	4.8	.2	5	3	15	3	2
Σ:	15	20	69	0	15	20	60		0
M:	3	4	$\Sigma d^2 = 1.90$		3	4	$\Sigma d^2 = 10.00$		
Σ^2:	55	90			55	90			

$$\Sigma y^2 = 55 - \frac{(15)^2}{5} = 10 \qquad \Sigma y^2 = 55 - \frac{(15)^2}{5} = 10$$

$$\Sigma x^2 = 90 - \frac{(20)^2}{5} = 10 \qquad \Sigma x^2 = 90 - \frac{(20)^2}{5} = 10$$

$$\Sigma xy = 69 - \frac{(15)(20)}{5} = 9 \qquad \Sigma xy = 60 - \frac{(15)(20)}{5} = 0$$

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{9}{10} = .90 \qquad b = \frac{0}{10} = 0$$

$$a = \bar{Y} - b\bar{X} = 3 - (.90)(4) = -.60 \qquad a = 3 - (0)(4) = 3$$

$$Y' = a + bX = -.60 + .90X \qquad Y' = 3 + (0)X$$

medias, las sumas de cuadrados de la desviación de X y Y ($\Sigma x^2 = \Sigma X^2 - (\Sigma X)^2/n$), los productos cruzados de desviación ($\Sigma xy = \Sigma XY - (\Sigma X)(\Sigma Y)/n$), y ciertos valores de regresión que se explicarán en breve.

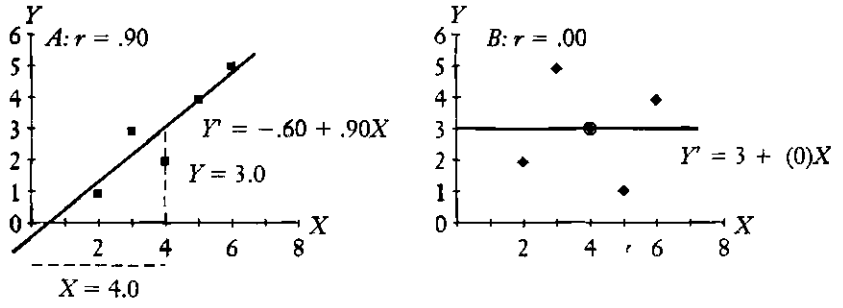
En primer lugar, observe la diferencia entre las puntuaciones en los conjuntos A y B. Difieren únicamente en el orden de las puntuaciones de la segunda columna o columna de las X. Los dos órdenes diferentes producen correlaciones muy diferentes entre las puntuaciones de X y Y. En el conjunto A, $r = .90$, y en el conjunto B, $r = .00$. En segundo lugar, observe los estadísticos en la parte inferior de la tabla. Σx^2 y Σy^2 son iguales tanto en A como en B, pero Σxy es 9 en A y 0 en B. Concentrándose en las puntuaciones del conjunto A, la ecuación básica de la regresión lineal simple es:

$$Y' = a + bX \qquad \textbf{(32.1)}$$

donde X = las puntuaciones de la variable independiente, a = la constante de la intersección, b = el coeficiente de regresión y Y' = las puntuaciones predichas de la variable dependiente. Una ecuación de regresión es una fórmula de predicción: los valores de Y se predicen a partir de los valores de X. La correlación entre los valores observados de X y Y, en efecto, determina cómo "funciona" la ecuación de predicción. La constante de la intersección, a, y el coeficiente de regresión, b, se explicarán en breve.

Los dos conjuntos de valores X y Y de la tabla 32.1 se grafican en la figura 32.1. Para cada gráfica se dibujaron líneas que "corren a través" de los puntos graficados. Si existiera una forma de colocar dichas líneas de tal manera que estuvieran simultáneamente lo más

FIGURA 32.1



cerca posible de todos los puntos, entonces las líneas deberían expresar la regresión de Y sobre X. La línea en la gráfica de la izquierda, donde $r = .90$, corre cerca de los puntos XY graficados. No obstante, en la gráfica de la derecha, donde $r = .00$, no es posible trazar la línea cerca de todos los puntos. En efecto, los puntos están ubicados aleatoriamente, puesto que $r = .00$.

Las correlaciones entre X y Y, $r = .90$ y $r = .00$, determinan las pendientes de las líneas de regresión (cuando las desviaciones estándar de X y Y son iguales, como sucede en este caso). La *pendiente* indica el cambio que se espera en Y con el cambio de una unidad de X. En el ejemplo de $r = .90$, cuando hay un cambio de 1 en X, se predice un cambio de .90 en Y, lo cual se expresa de manera trigonométrica como la longitud de la línea opuesta al ángulo formado por la línea de regresión, dividida entre la longitud de la línea adyacente al ángulo. En la figura 32.1, si se traza una línea perpendicular a la línea de regresión —el punto donde las medias de X y de Y intersectan, por ejemplo— hasta una línea trazada horizontalmente, a partir del punto donde la línea de regresión se intersecta con el eje Y, o en $Y = -.60$, entonces $3.6/4.0 = .90$. Un cambio de 1 en X significa un cambio de .90 en Y. Se han utilizado puntuaciones en bruto en la mayor parte de los ejemplos del presente capítulo, debido a que se ajustan mejor a los propósitos del texto. Sin embargo, un examen profundo de la regresión requiere de una explicación que utilice puntuaciones de desviación y puntuaciones estándar. El énfasis aquí, así como en cualquier parte del libro, está en los usos de investigación de los métodos y técnicas, y no en la estadística como tal. Por lo tanto, el estudiante debe complementar su estudio con buenas explicaciones básicas sobre la regresión simple y múltiple. Se recomienda remitirse a las referencias de la sección de "Sugerencias de estudio", al final del capítulo 33.

La gráfica de los valores de X y Y del ejemplo B (parte derecha de la figura 32.1) es bastante diferente. En el ejemplo A, es posible, fácil y visualmente, trazar una línea a través de los puntos y lograr una aproximación bastante precisa hacia la línea de regresión. Pero en el ejemplo B esto es difícilmente posible. La línea tan sólo se puede trazar por medio del uso de otras guías, las cuales se explicarán en breve. Otro aspecto que es importante destacar es el esparcimiento o dispersión de los puntos graficados alrededor de las dos líneas de regresión. En el ejemplo A, se presentan muy cerca de la línea. Si $r = 1.00$, entonces todos estarían sobre la línea. Por otra parte, cuando $r = .00$, se dispersan ampliamente alrededor de la línea. ***A menor correlación, habrá mayor dispersión.***

Para calcular los estadísticos de regresión de los dos ejemplos, se deben calcular las sumas de cuadrados y los productos cruzados de desviación, lo cual se efectuó en la parte inferior de la tabla 32.1. La fórmula para calcular la *pendiente* o *coeficiente de regresión*, b, es:

$$b = \frac{\Sigma xy}{\Sigma x^2} \tag{32.2}$$

Las dos b son .90 y .00. La *constante de intersección*, a, se calcula por medio de la fórmula:

$$a = \bar{Y} - b\bar{X} \tag{32.3}$$

Las a para los dos ejemplos son −.60 y 3; para el ejemplo A, $a = 3 - (.90)(4) = -.60$. La constante de intersección es el punto donde la línea de regresión se intersecta con el eje Y. Para trazar la línea de regresión, se coloca una regla entre la constante de intersección en el eje Y y el punto donde se unen la media de Y y la media de X. (En la figura 32.1 dichos puntos se indican con pequeños cuadrados en el ejemplo A y con diamantes en el ejemplo B.)

Los pasos finales del proceso son, al menos hasta donde se llegará aquí, escribir ecuaciones de regresión y, después, utilizando las ecuaciones, calcular los valores predichos de Y o Y', dados los valores de X. Las dos ecuaciones se presentan en la última línea de la tabla 32.1. Primero observe la ecuación de regresión para $r = .00$: $Y' = 3 + (0)X$. Evidentemente ello significa que todas las Y predichas son iguales a 3, la media de Y. Cuando $r = 0$, la mejor predicción es la media, como ya se indicó antes. Cuando $r = 1.00$, en el otro extremo, el lector notará que es posible realizar una predicción exacta: simplemente se suma a, la constante, a las puntuaciones de X. Cuando $r = .90$, la predicción es menos que perfecta y se predicen los valores de Y' calculados con la ecuación de regresión. Por ejemplo, para predecir la primera puntuación Y', se calcula:

$$Y'_1 = -.60 + (.90)(2) = 1.20$$

Las puntuaciones predichas de los conjuntos A y B ya se presentaron en la tabla 32.1. (Véase las columnas denominadas Y'.) Note un aspecto importante: si para el ejemplo A se grafica la X y la Y predicha o los valores de Y', todos los puntos graficados caen en la línea de regresión. Es decir, la línea de regresión de la figura representa el conjunto de valores predichos de Y, dados los valores observados de X y la correlación entre la X y los valores de Y.

Ahora se pueden calcular los valores predichos de Y. A mayor correlación, habrá mayor precisión en la predicción. La precisión de las predicciones de los dos conjuntos de puntuaciones se demuestra con claridad al calcular las diferencias entre los valores originales de Y y los valores predichos de Y, o $Y - Y' = d$, y calculando, después, las sumas de cuadrados de tales diferencias, que se consideran *residuales*. En la tabla 32.1 se calcularon los dos conjuntos de residuales y sus sumas de cuadrados (véase las columnas denominadas d). Los dos valores de Σd^2, 1.90 para A y 10.00 para B, son bastante diferentes, así como las gráficas en la figura 32.1 son bastante diferentes: el valor del conjunto B, $r = .00$, es mucho mayor que el del conjunto A, o $r = .90$. Es decir, a mayor correlación, menores serán las desviaciones de la predicción y, por lo tanto, la predicción se vuelve más precisa.

En la próxima sección se examinará una extensión del modelo más simple de regresión. El modelo desarrollado, llamado regresión múltiple, posee una utilidad mucho mayor para la investigación que el modelo simple presentado aquí. Sin embargo, no se debe pensar que el modelo más simple carece de utilidad o de valor, ya que puede, de hecho, proporcionar información valiosa de investigación y de tipo práctico. Por ejemplo, Erlich y Lee (1978) demostraron cómo el modelo de regresión simple puede utilizarse con algunos estadísticos adicionales (banda o intervalo de confianza), para determinar la responsabilidad de los consejos y políticas educativas.

Regresión lineal múltiple

El método de la regresión lineal múltiple extiende las ideas presentadas en la sección anterior a más de una variable independiente. A partir del conocimiento de los valores de dos o más variables independientes, X_1, X_2,..., X_k, se desea predecir una variable dependiente, Y. Anteriormente en este libro se mencionó la gran necesidad de evaluar la influencia de diversas variables sobre una variable dependiente. Evidentemente, es posible predecir a partir de la aptitud verbal, por ejemplo, el rendimiento en lectura, o a partir del conservadurismo, las actitudes hacia los diferentes grupos étnicos. No obstante, sería más poderoso si se pudiera predecir a partir de la aptitud verbal junto con otras variables que se sabe o se piensa que influyen en la lectura —por ejemplo, la motivación de logro y la actitud hacia el trabajo escolar—. En teoría no existe límite alguno para el número de variables que se pueden utilizar, aunque existen límites prácticos. A pesar de que en el siguiente ejemplo se utilizan únicamente dos variables independientes, los principios se aplican de la misma forma a cualquier número de variables independientes.

Un ejemplo

Tome uno de los problemas que acaban de mencionarse. Suponga que se tienen las puntuaciones de rendimiento en lectura (RL), de aptitud verbal (AV) y de motivación de

TABLA 32.2 *Ejemplo ficticio: puntuación de rendimiento en lectura (Y), aptitud verbal (X_1) y motivación de logro (X_2)*

Y	X_1	X_2	Y'	$Y - Y' = d$
2	2	4	3.0305	-1.0305
1	2	4	3.0305	-2.0305
1	1	4	2.3534	-1.3534
1	1	3	1.9600	-.9600
5	3	6	4.4944	.5056
4	4	6	5.1715	-1.1715
7	5	3	4.6684	2.3316
6	5	4	5.0618	.9382
7	7	3	6.0226	.9774
8	6	3	5.3455	2.6545
3	4	5	4.7781	-1.7781
3	3	5	4.1010	-1.1010
6	6	9	7.7059	-1.7059
6	6	8	7.3125	-1.3125
10	8	6	7.8799	2.1201
9	9	7	8.9504	.0496
6	10	5	8.8407	-2.8407
6	9	5	8.1636	-2.1636
9	4	7	5.5649	3.4351
10	4	7	5.5649	4.4351
Σ: 110	99	104		0
M: 5.50	4.95	5.20		
Σ^2: 770.0	625.0	600.0		81.6091

logro (ML), de 20 alumnos de segundo año de secundaria. Se desea predecir el rendimiento en lectura, Y, a partir de la aptitud verbal, X_1, y la motivación de logro, X_2. O se desea calcular la regresión del rendimiento en lectura en *ambas* variables, tanto en la aptitud verbal como en la motivación de logro. Si las puntuaciones de aptitud verbal y motivación de logro fueran puntuaciones estandarizadas, podrían promediarse, tratar los promedios como una variable independiente compuesta y calcular los estadísticos de regresión como se realizó antes. No podría resultar muy incorrecto, sin embargo, existe una mejor forma.

Suponga que X_1 (aptitud verbal), X_2 (motivación de logro) y Y (rendimiento en lectura), las puntuaciones de los 20 sujetos y sus sumas, medias y sumas de cuadrados de los datos en bruto aparecen en la tabla 32.2. (Por el momento haga caso omiso de las columnas de Y' y de d.) Es necesario calcular las sumas de cuadrados de *desviación*, los productos cruzados de desviación, las desviaciones estándar y las correlaciones entre las tres variables, pues éstos son los estadísticos básicos que se calculan para casi cualquier conjunto de datos. (Se presentan en la tabla 32.3.) Los cálculos no se hacen aquí debido a que su mecánica se explicó en capítulos previos. El lector debe realizarlos y notar que los resultados obtenidos probablemente serán ligeramente diferentes de los reportados antes. Tales diferencias se deben a errores de redondeo: un problema siempre presente en el análisis multivariado. De hecho, los resultados de este problema, obtenidos con una calculadora de escritorio, difieren ligeramente de los obtenidos por computadora. Las sumas de cuadrados y los productos cruzados se presentan en la diagonal (de la parte superior izquierda a la parte inferior derecha) y encima de ella, y las correlaciones se presentan por debajo de la diagonal. Las r de interés primordial son las de las dos variables independientes con la variable dependiente, r_{y1} y r_{y2}, .6735 y .3946, respectivamente. Con la eliminación de estos cálculos de rutina, ahora es posible concentrarse en los conceptos básicos de la regresión múltiple. La ecuación fundamental de la regresión es:

$$Y' = a + b_1X_1 + \ldots + b_kX_k \qquad (32.4)$$

Los símbolos tienen el mismo significado que los de la ecuación de regresión simple, con la excepción de que hay k variables independientes y k coeficientes de regresión. De cualquier manera, la a y las b deben calcularse a partir del conocimiento de las X y Y, cuyos cálculos son los más complejos del análisis de regresión múltiple. Para sólo dos variables independientes se utilizan las fórmulas algebraicas incluidas en los libros de estadística (véase Cohen y Cohen, 1983; Draper y Smith, 1981; Neter, Wasserman y Kutner, 1983; Pedhazur, 1996). Una vez que se tienen las b, el cálculo de a es directo. El problema es el

TABLA 32.3 *Sumas de cuadrados y productos cruzados de desviación, coeficientes de correlación y desviaciones estándar (datos de la tabla 34.2)*[a]

	y	x_1	x_2
y	165.00	100.50	39.00
x_1	*.6735*	134.95	23.20
x_2	*.3946*	*.2596*	59.20
s	2.9469	2.6651	1.7652

[a] Los datos de la tabla son los siguientes: la primera línea da, de forma sucesiva, Σy^2, la suma de cuadrados de las puntuaciones de desviación para Y, el producto cruzado de las desviaciones de X_1 y Y, o Σx_1y y finalmente Σx_2y. Las cifras en la segunda y tercera líneas, sobre la diagonal o encima de ella, son Σx_1^2, Σx_1x_2, y (en la esquina de la parte baja derecha) Σx_2^2. Las cifras en *itálicas* (*cursivas*) por *debajo* de la diagonal son los coeficientes de correlación. Las desviaciones estándar se presentan en la última línea.

cálculo de las b cuando hay más de dos variables independientes. Sólo se explicarán las ideas generales que subyacen a los cálculos, ya que los detalles nos desviarían del tema de interés central. Se le pide al lector que consulte cualquiera de las referencias que se mencionan arriba.

Lo que se tiene, en efecto, es un conjunto de ecuaciones lineales, una ecuación para cada variable independiente. El objetivo de la determinación de las b de la ecuación 32.4 consiste en encontrar aquellos valores de b que minimicen las sumas de cuadrados de los residuales. Éste es el *principio de los mínimos cuadrados*. El cálculo proporciona el método de diferenciación para llevarlo a cabo. Si se utiliza, produce un conjunto de ecuaciones lineales simultáneas llamadas ecuaciones *normales* (que no tienen relación con las distribuciones normales). Una forma conveniente de estas ecuaciones contiene los coeficientes de correlación entre todas las variables independientes, y entre las variables independientes y la variable dependiente y un conjunto de ponderaciones llamadas *pesos beta*, β_j, que se explicarán posteriormente (son como los pesos b). Las ecuaciones normales para el problema anterior son:

$$\begin{aligned} r_{11}\beta_1 + r_{12}\beta_2 &= r_{y1} \\ r_{12}\beta_1 + r_{22}\beta_2 &= r_{y2} \end{aligned} \tag{32.5}$$

donde β_j es igual a los pesos beta; r_{12} es igual a las correlaciones entre las variables independientes, y r_{yj} a las correlaciones entre las variables independientes y la variable dependiente, Y. (Observe que $r_{12} = r_{21}$, y que $r_{11} = r_{22} = 1.00$, y también que la ecuación 32.5 puede extenderse para cualquier número de variables independientes.)

Quizá la mejor manera —y con seguridad la más elegante— de resolver las ecuaciones para β_j sea por medio del álgebra de matrices. Por desgracia, el conocimiento del álgebra de matrices no puede darse como un hecho. Por lo tanto, la solución real de las ecuaciones, utilizando el álgebra de matrices, debe omitirse; pero la solución para las dos variables independientes se obtiene algebraicamente sin el uso de matrices. Sin embargo, cualquier tamaño mayor que éste seguramente requeriría el empleo de un programa computacional, ya que la cantidad de cálculos se incrementa de manera exponencial. Se mostrará cómo se obtiene la solución con un poco de álgebra. Para utilizar la ecuación normal dada anteriormente, se necesitará calcular r_{12}, r_{y1} y r_{y2}. Recuerde que tanto r_{11} como r_{22} son iguales a 1.00. A partir de examinar la tabla 32.3, se obtiene $r_{12} = .259562$, $r_{y1} = .6735$ y $r_{y2} = .394604$. Sustituyendo estos valores en la ecuación normal, se obtiene:

$$1.000000\beta_1 + .259562\beta_2 = .6735$$

$$.259562\beta_1 + 1.00000\beta_2 = .394604$$

Tome cada ecuación normal y despéjela para que β_1 esté a un lado del signo de igual y el resto quede del otro lado:

$$\beta_1 = .6735 - .259562\beta_2$$

$$\beta_1 = 1.5202688 - 3.8526441\beta_2$$

Ahora iguale las β_1 entre sí y despéjelas para obtener β_2.

$$.6735 - .259562\beta_2 = 1.5202688 - 3.852641\beta_2$$

$$3.5930821\beta_2 = .8467688$$

$$\beta_2 = \frac{.8467688}{3.5930821} = .235666 \approx .2357$$

Despejando para β_1 se sustituye el valor de β_2 en la ecuación: $\beta_1 = .6735 - .259562\beta_2$. Por lo tanto,

$$\beta_1 = .6735 - .259562 \times .235666 = .6123$$

La solución para la situación de dos variables presentada arriba produce los siguientes pesos beta: $\beta_1 = .6123$ y $\beta_2 = .2357$. Los pesos b o los pesos no estandarizados de la regresión se obtienen a partir de la siguiente fórmula:

$$b_j = \beta_j \frac{s_y}{s_j} \tag{32.6}$$

donde s_j es igual a las desviaciones estándar de las variables uno y dos (véase tabla 32.3) y s_y es igual a la desviación estándar de Y. Sustituyendo en la ecuación 32.6 se obtiene:

$$b_1 = (.6123)\left(\frac{2.9469}{2.6651}\right) = .6771$$

$$b_2 = (.2357)\left(\frac{2.9469}{1.7652}\right) = .3934$$

Para obtener la constante de la intersección, se extiende la ecuación 32.3 a dos variables independientes:

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2$$

$$a = 5.50 - (.6771)(4.95) - (.3934)(5.20) = .1027$$

Un método alternativo para encontrar los coeficientes no estandarizados de la regresión es por medio de ecuaciones normales. Note aquí que dichas ecuaciones normales darán directamente los pesos de regresión. Las ecuaciones normales dadas arriba serían utilizadas para obtener los coeficientes de regresión.

$$nb_0 + \Sigma x_1 b_1 + \Sigma x_2 b_2 = \Sigma y$$

$$\Sigma x_1 b_0 + \Sigma x_1^2 b_1 + \Sigma x_1 x_2 b_2 = \Sigma x_1 y$$

$$\Sigma x_2 b_0 + \Sigma x_1 x_2 b_1 + \Sigma x_2^2 b_2 = \Sigma x_2 y$$

Observe que para dos variables independientes hay tres ecuaciones normales y que algunas veces se utiliza b_0 para representar el término o constante de la intersección. Se tienen tres ecuaciones arriba con tres parámetros desconocidos para estimarse: b_0, b_1 y b_2. Los cálculos para resolver los pesos de la regresión son demasiado laboriosos para realizarse a mano y, por lo tanto, no se presentarán los cálculos aquí.

Por último, se escribe la ecuación de regresión completa:

$$Y' = a + b_1X_1 + b_2X_2$$

$$Y' = .1027 + .6771X_1 + .3934X_2$$

Sustituyendo los valores observados de X_1 y X_2 de la tabla 34.2, se obtienen los valores predichos de Y. Por ejemplo, calcule las Y predichas para el quinto y vigésimo sujetos:

$$Y'_5 = .1027 + (.6771)(3) + (.3934)(6) = 4.4944$$

$$Y'_{20} = .1027 + (.6771)(4) + (.3934)(7) = 5.5649$$

Estos valores y los otros 18 se presentan en la cuarta columna de la tabla 32.2. La quinta columna de la tabla presenta las desviaciones a partir de la regresión, o los residuales, $Y_i - Y'_i = d_i$. Por ejemplo, los residuales para Y_5 y Y_{20} son:

$$d_5 = Y_5 - Y'_5 = 5 - 4.4944 = .5056$$

$$d_{20} = Y_{20} - Y'_{20} = 10 - 5.5649 = 4.4351$$

Observe que una desviación es pequeña y la otra es grande. Los residuales se presentan en la última columna de la tabla 32.2. La mayoría de ellos son relativamente pequeños; casi la mitad son positivos y la otra mitad negativos.

Ahora puede calcularse la suma de cuadrados debida a la regresión; sin embargo, debe considerarse la regresión de Y sobre X_1 y X_2. Se eleva al cuadrado cada uno de los valores de Y' de la cuarta columna de la tabla 32.2 y se suman:

$$(3.0305)^2 + \cdots + (5.5649)^2 = 688.3969$$

Ahora, se utiliza la fórmula acostumbrada para la suma de cuadrados de desviación (véase capítulo 13):

$$\sum y^2 = 688.3969 - \frac{(110)^2}{20} = 83.3969$$

De manera similar, calcule la suma de cuadrados de los residuales:

$$\sum d^2 = (-1.0305)^2 + \cdots + (4.4351)^2 = 81.6091$$

Note que éste es un "buen" ejemplo de los errores que se acumulan debido al redondeo. La verdadera suma de cuadrados de la regresión, calculada por medio de una computadora, es 83.3909, un error de .006. No obstante, también debe notarse que aunque los residuales se calcularon a partir de las Y predichas calculadas a mano, la suma de cuadrados de los residuales es exactamente igual a la producida por medio de la computadora, 81.6091.

Para verificar, calcule:

$$SC_{reg} + SC_{res} = SC_t$$

$$83.3969 + 81.6091 = 165.0060$$

La regresión y las sumas de cuadrados residuales por lo común no se calculan de esta manera. Aquí se calcularon tan sólo para mostrar cuáles son estas cantidades. Si se hubieran utilizado las fórmulas empleadas comúnmente, entonces quizá no se habría visto con claridad que la suma de cuadrados de regresión es la suma de cuadrados de los valores de Y', calculados por medio de la ecuación de regresión. Quizá tampoco se hubiera visto con claridad que la suma de cuadrados residual es la suma de cuadrados calculada con las d de la quinta columna de la tabla 32.2. Recuerde también que la a y las b (o las β) de la ecuación de regresión se calcularon para satisfacer el principio de los mínimos cuadrados; es decir,

para minimizar las d o errores de predicción o, más bien, para minimizar la suma de cuadrados de los errores de predicción. Para sintetizar, la suma de cuadrados de regresión expresa aquella porción de la suma de cuadrados total de Y debida a la regresión de Y, la variable dependiente, sobre X_1 y X_2, las variables independientes. La suma de cuadrados residual expresa la porción de la suma de cuadrados total de Y que *no* se debe a la regresión.

Tal vez el lector se pregunte: ¿Para qué tomarse la molestia con este procedimiento complicado para determinar los pesos de regresión? ¿Es necesario recurrir a un procedimiento de mínimos cuadrados? ¿Por qué no sólo promediar los valores de X_1 y X_2, y llamar a las medias de los valores individuales de X_1 y X_2 las Y predichas? La respuesta es que funcionaría bastante bien. De hecho, en este caso funcionaría muy bien, casi tan bien como el procedimiento total de regresión. Pero quizá *no* funcionaría tan bien. El problema radica en que realmente no se sabe cuándo funcionará bien y cuándo no. El procedimiento de regresión generalmente "funciona", siempre y cuando las cosas se mantengan iguales. Siempre minimiza los errores cuadrados de predicción. Note que en ambos casos se utilizan ecuaciones lineales y que sólo difieren los coeficientes:

$$\textit{Ecuación de regresión: } Y' = a + b_1X_1 + b_2X_2$$

$$\textit{Ecuación media: } Y' = \frac{1}{2}X_1 + \frac{1}{2}X_2$$

De las innumerables formas posibles para ponderar X_1 y X_2, ¿cuál debe elegirse si no se utiliza el principio de los mínimos cuadrados? Por supuesto, es concebible que se tenga conocimiento previo o alguna razón para X_1 y X_2. Las X_1 pueden ser las puntuaciones en alguna prueba que haya demostrado ser altamente exitosa a nivel predictivo. X_2 puede ser un predictor exitoso también, pero no tanto como X_1. De esta manera, se toma la decisión de darle un gran peso a X_1, por ejemplo, cuatro veces más que a X_2. La ecuación sería: $Y' = 4X_1 + X_2$; lo cual podría funcionar bien. El problema es que en pocas ocasiones se tiene este conocimiento previo, y aun cuando se tenga, suele resultar bastante impreciso. ¿Cómo se puede llegar a decidir darle cuatro veces más peso a X_1 que a X_2? Es posible efectuar una suposición con cierto fundamento. No obstante, el método de regresión no es una suposición; se trata de un método preciso basado en los datos y en un principio matemático poderoso. Es en tal sentido que los pesos calculados de regresión son "mejores".

Las sumas de cuadrados de regresión y residuales se calculan de forma más directa de lo indicado antes. Las fórmulas son:

$$sc_{reg} = b_1\Sigma x_1y + \ldots + b_k\Sigma x_ky \tag{32.7}$$

$$sc_{res} = sc_t - sc_{reg} \tag{32.8}$$

En el presente caso, (32.7) se convierte en:

$$sc_{reg} = b_1\Sigma x_1y + b_2\Sigma x_2y$$

Esto se calcula fácilmente sustituyendo los dos valores de b calculados arriba y los productos cruzados dados en la tabla 32.3.

$$sc_{reg} = (0.6771)(100.50) + (0.3934)(39.00) = 83.3912$$

$$sc_{res} = 165.0 - 83.3912 = 81.6088$$

Con los errores por redondeo, éstos son los valores calculados de forma directa, a partir de la cuarta y quinta columnas de la tabla 32.2. (Note los valores "más precisos" dados por

una computadora: $sc_{regM} = 83.3909$ y $sc_{res} = 81.6091$, que evidentemente da un total $sc_t = \Sigma y^2 = 165.0$.)

El coeficiente de correlación múltiple

Si se calcula el coeficiente común de correlación producto-momento entre los valores predichos Y', y los valores observados de Y, se obtiene un índice de la magnitud de la relación entre un compuesto de los mínimos cuadrados de X_1 y X_2, por un lado, y Y, por el otro. Este índice se llama *coeficiente de correlación múltiple*, R. A pesar de que en el presente capítulo casi siempre se simboliza como R por cuestiones de brevedad, una manera más satisfactoria de hacerlo es con subíndices: $R_{y.12...k}$ o, en este caso, $R_{y.12}$. La teoría de la regresión múltiple parece ser especialmente elegante cuando se considera el coeficiente de correlación múltiple. Se trata de uno de los vínculos que une los diversos aspectos de la regresión múltiple y del análisis de varianza. La fórmula de R que expresa el primer enunciado de este párrafo es:

$$R = \frac{\Sigma yy'}{\sqrt{\Sigma y^2 \Sigma y'^2}} \tag{32.9}$$

Se calcula su cuadrado:

$$R^2 = \frac{(\Sigma yy')^2}{\Sigma y^2 \Sigma y'^2} \tag{32.10}$$

Utilizando los valores de Y y Y' de la tabla 32.2, se obtiene: $R^2 = .5054$ y $R = \sqrt{.5054} = .7109$. El cálculo de estos valores es un buen ejercicio. Ya se tiene $\Sigma y^2 = 165$. Entonces se calcula:

$$\Sigma y'^2 = \Sigma Y'^2 - \frac{(\Sigma Y')^2}{N} = 688.3969 - \frac{(110)^2}{20} = 83.3969$$

y

$$\Sigma yy' = \Sigma YY' - \frac{(\Sigma Y)(\Sigma Y')}{N} = 688.3939 - \frac{(110)(110)}{20} = 83.3939$$

Puede demostrarse algebraicamente que $\Sigma y'^2$ es igual a $\Sigma yy'$. La diferencia de .003 se debe a los errores de redondeo.

Entonces, R es la correlación más alta posible entre un compuesto lineal de los mínimos cuadrados de las variables independientes y la variable dependiente observada. La R^2, análoga a r^2, indica la porción de varianza de la variable dependiente, Y, debida a las variables independientes en conjunto. R, a diferencia de r, varía únicamente de 0 a 1.00; no posee valores negativos.

Otras dos conclusiones importantes se obtienen calculando las correlaciones de los *residuales, d_i, de la tabla 32.2, con X_1 y X_2, por un lado, y con Y, por el otro. Las correlaciones de los residuales con X_1 y X_2 son ambas cero, lo cual* no sorprende cuando se sabe que,

por definición, los residuales son aquella parte de Y que no está explicada por X_1 y X_2. Es decir, cuando los valores de Y' se restan de los valores de Y, aquella porción debida a la regresión de Y sobre X_1 y X_2 se toma de ellos. Cualquier cosa que quede, entonces, no se relaciona con X_1 ni con X_2. Si el lector se tomara la molestia de calcular la correlación entre el vector de d —un vector es un solo conjunto de medidas, ya sea en una columna o en un renglón— y el vector de X_1 o de X_2 se podrá observar que esto es cierto. No se debe subestimar la importancia de realizar dichos cálculos y de ponderar su significado. Ello es especialmente importante como ayuda para comprender la regresión múltiple y otras técnicas multivariadas. Puede cometerse un serio error al permitir que la computadora haga todo por nosotros, especialmente con programas computacionales. En lo que se refiere a los estadísticos más sencillos como r y las diversas sumas de cuadrados, es mejor escribir programas relativamente simples para una microcomputadora, almacenarlos en discos flexibles y utilizarlos cuando así se requiera. Una importante implicación para la investigación de esta generalización también se analizará posteriormente cuando se sinteticen y expongan ejemplos reales de investigación.

La correlación de los residuales, d_i, de la tabla 32.2 con los valores originales de Y, también ayuda a aclarar algunos puntos. Esta correlación es: $r_{dy} = .7033$, y su cuadrado es: $r_{dy}^2 = (.7033)^2 = .4946$. Si este último valor se suma a la R^2 calculada anteriormente, el resultado es interesante: $R^2 + r_{dy}^2 = .5054 + .4946 = 1.0000$; lo cual siempre será cierto: "1.0000" representa la varianza total de Y. La varianza de Y debida a la regresión de las Y sobre X_1 y X_2 es .5054. Se puede calcular la varianza de Y que no se debe a la regresión de Y sobre X_1 y X_2: $1.0000 - .5054 = .4946$ que es, evidentemente, el valor de r_{dy}^2 que acaba de calcularse de manera directa. El significado de r_{dy}^2 se interpreta de dos formas. El cálculo directo de la correlación muestra que los residuales constituyen la parte de la varianza de Y que no se debe a la regresión de Y a partir de X_1 y X_2. En el presente caso, el 51 por ciento ($R^2 = .51$) de la varianza del rendimiento en lectura (Y) de los 20 alumnos se explica por una combinación lineal de los mínimos cuadrados de la aptitud verbal (X_1) y la motivación de logro (X_2). Pero el 49 por ciento de la varianza se debe a otras variables y al error. Después de analizar formas más comunes para calcular R y R^2, se considerará nuevamente la interpretación de la proporción o porcentaje de R^2.

En síntesis, R^2 es un estimado de la proporción de la varianza de la variable dependiente Y, explicado por las variables independientes X_j. R, el coeficiente de correlación múltiple, es la correlación producto-momento entre la variable dependiente y otra variable producida por una combinación de los mínimos cuadrados de las variables independientes. Su cuadrado se interpreta de manera análoga al cuadrado de un coeficiente de correlación ordinario. Sin embargo, difiere del coeficiente ordinario en que únicamente toma valores entre 0 y 1. La R no es tan útil ni tan interpretable como la R^2, y de aquí en adelante se utilizará la R^2 casi de manera exclusiva en los análisis presentados.

La interpretación de la proporción o del porcentaje de R^2 se vuelve más clara si se utiliza una fórmula de suma de cuadrados:

$$R^2 = \frac{sc_{reg}}{sc_t} \tag{32.11}$$

donde sc_t es, como siempre, la suma de cuadrados total de Y, o $\sum_{yt}^2$. Sustituyendo la suma de cuadrados de regresión calculada antes por medio de la fórmula 32.7, y la suma de cuadrados total de la tabla 32.3, se obtiene:

$$R^2 = \frac{83.3912}{165.000} = .5054$$

Y R^2 parece ser esa parte de la suma de cuadrados de Y asociada con la regresión de Y a partir de las variables independientes. Como sucede con todas las proporciones, al multiplicarla por 100 se convierte en un porcentaje.

La fórmula 32.11 proporciona otro vínculo con el análisis de varianza. En el capítulo 13, al explicar los fundamentos del análisis de varianza, se presentó una fórmula para calcular η, la *razón de correlación* (fórmula 13.4). Se eleva al cuadrado dicha fórmula:

$$\eta^2 = \frac{sc_e}{sc_t}$$

dónde sc_e es igual a la suma de cuadrados entre grupos, y sc_t es igual a la suma de cuadrados total. sc_e es la suma de cuadrados debida a la variable independiente. sc_{reg} es la suma de cuadrados debida a la regresión. Ambos términos se refieren a la suma de cuadrados de una variable dependiente, debida a una variable independiente o a variables independientes.

R y R^2 quizá estén infladas, lo que sucede con frecuencia. Por lo tanto, R^2 debe interpretarse de manera conservadora. Si la muestra es grande, por ejemplo, mayor de 200, entonces existen pocas razones para preocuparse; sin embargo, si la muestra es pequeña, resulta sensato reducir algunos puntos a la R^2 calculada. Para hacerlo, se utiliza una *fórmula de encogimiento*:

$$R_c^2 = 1-(1-R^2)\left(\frac{N-1}{N-n-1}\right)$$

donde R_c^2 es igual a la R^2 encogida o corregida; N es el tamaño de la muestra; n es el número total de variables en el análisis. Usando esta fórmula, la R^2 en el ejemplo se reduce a .45. Cuando se compara R_c^2 con R^2, se percibe qué tanto está inflada la R^2 por el error del azar. A partir de dicha fórmula también se observa el efecto que tiene un tamaño pequeño de muestra sobre el valor de R_c^2. Las muestras pequeñas tienden a producir valores inestables de R^2, lo cual se determina por medio de la fórmula de encogimiento antes expresada.

Pruebas de significancia estadística

Anteriormente se estudió la regresión simple de Y a partir de X. Para probar la significancia estadística de la regresión simple se evalúa la significancia del coeficiente de correlación entre X y Y, r_{xy}, refiriéndose a una tabla apropiada. Algunos de los libros reconocidos que contienen tablas utilizadas en análisis estadísticos son los de Beyer (1990) y Burlington y May (1970). Con el avance en los programas estadísticos computacionales y su fácil acceso, los investigadores utilizan cada vez menos las revisiones de tablas. Los programas computacionales ahora son capaces de calcular y dar el resultado de la probabilidad del error tipo I, junto con el estadístico de prueba, haciendo innecesaria la consulta de las tablas. Sin embargo, desde un punto de vista educativo, los estudiantes necesitan aprender el manejo de las tablas para comprender el resultado que ofrece la computadora. Las pruebas de significancia estadística en la regresión múltiple, aunque son más complejas, se basan en la idea relativamente simple de la comparación de varianzas (o cuadrados medios), como en el análisis de varianza. Las mismas preguntas planteadas muchas veces antes, deben plantearse de nuevo: ¿puede esta R^2 haber surgido por azar? ¿Se aleja lo suficiente de lo esperado por el azar como para considerarla "significativa"? Preguntas similares pueden plantearse acerca de los coeficientes individuales de regresión. En este capítulo y en el siguiente, se utilizarán casi exclusivamente las pruebas F, pues se ajustan bastante bien tanto al análisis de regresión como al análisis de varianza y, además, son simples tanto

a nivel conceptual como a nivel computacional. Es posible realizar análisis sobre cada coeficiente de regresión. Si la prueba t de un coeficiente de regresión es significativa, ello indica que el peso de regresión difiere significativamente de cero, lo cual, a su vez, significa que la variable con la que está asociado contribuye de manera significativa a la regresión. La prueba t para coeficientes de regresión individuales se presenta en la siguiente sección. Primero se utilizará la prueba F para determinar si el modelo completo de regresión resulta estadísticamente significativo.

Una forma se expresa por medio de las ecuaciones 32.12a y 32.12b

$$F = \frac{sc_{reg}/gl_1}{sc_{res}/gl_2} \tag{32.12a}$$

$$F = \frac{sc_{reg}/k}{sc_{res}/(N-k-1)} \tag{32.12b}$$

donde sc_{reg} es la suma de cuadrados debida a la regresión; sc_{res} es igual a la suma de cuadrados residual o de error; k es igual al número de variables independientes; N es igual al tamaño de la muestra. Si se definen gl_1 y gl_2, los grados de libertad para el numerador y el denominador de la razón F, en la ecuación 32.12a, se obtiene la ecuación 32.12b. Tales fórmulas son importantes a causa de que se utilizan para probar la significancia de cualquier problema de regresión múltiple. Con los valores calculados antes para el ejemplo de la tabla 32.2, ahora se calcula:

$$F = \frac{83.3912/2}{81.6091/(20-2-1)} = \frac{41.6956}{4.8005} = 8.686$$

Note que la idea expresada por esta fórmula pertenece a la misma familia de ideas que el análisis de varianza. El numerador es el cuadrado medio debido a la regresión, análogo al cuadrado medio entre grupos, y el denominador es el cuadrado medio que *no* se debe a la regresión, el cual se utiliza como un término de error, análogo al cuadrado medio dentro de grupos o varianza del error. Nuevamente, el principio básico es siempre el mismo: la varianza debida a la regresión de Y a partir de $X_1, X_2, \ldots, X_k$, o, en el análisis de varianza, debida a los efectos experimentales, se evalúa en contra de la varianza debida presuntamente al error o al azar. Dicho concepto básico, elaborado con profundidad en capítulos previos, se expresa de la siguiente forma:

$$\frac{\text{varianza de regresión}}{\text{varianza de error}} : \frac{\text{varianza experimental}}{\text{varianza de error}}$$

Otra fórmula para F es:

$$F = \frac{R^2/k}{(1-R^2)/(N-k-1)} \tag{32.13}$$

donde k y N son iguales que arriba. Para el mismo ejemplo:

$$F = \frac{.5054/2}{(1-.5054)/(20-2-1)} = \frac{.2527}{.0291} = 8.684$$

el cual es igual al valor de F obtenido con la ecuación 32.12, incluyendo los errores debidos al redondeo. Con 2 y 17 grados de libertad, es significativo al nivel .01. Esta fórmula resulta particularmente útil cuando los propios datos de investigación aparecen sólo en forma de coeficientes de correlación. En tal caso, quizá no se conozcan las sumas de cuadrados requeridas por la ecuación 32.12. Gran parte del análisis de regresión puede realizarse usando únicamente la matriz de correlaciones entre todas las variables, independientes y dependientes. Dicho análisis va más allá del enfoque de este libro. No obstante, el lector y el estudiante de investigación deben estar conscientes de esa posibilidad (Pedhazur, 1996).

Pruebas de significancia de los coeficientes individuales de regresión

La significancia de los pesos o coeficientes individuales de regresión interesa a muchos investigadores, pues les indican cuáles variables independientes, en un sentido estadístico, realizan la principal contribución para explicar la variable dependiente. Por ejemplo, el encargado de admisiones de una importante universidad tiene a su disposición un número de variables que son pertinentes para predecir o explicar el éxito en la universidad. No obstante, con un análisis real, algunas de estas variables tendrían una contribución considerablemente menor que otras. Por ejemplo, McWhirter (1997) fue capaz de identificar qué variables predecían la soledad íntima y la soledad social de estudiantes universitarios.

La fórmula para probar la significancia de los pesos o coeficientes individuales de regresión es:

$$t_i = \frac{b_i}{s_{bi}}$$

donde b_i es el coeficiente de regresión y s_{bi} es el error estándar de la variable i. La fórmula anterior parece lo suficientemente simple; sin embargo, el cálculo del error estándar resulta complejo. La mejor manera para obtener el error estándar es por medio de un programa computacional o, en caso necesario, del álgebra matricial. Esta prueba t se conduce con grados de libertad igual a n – (el número de coeficientes de regresión en la ecuación de regresión).

Interpretación de los estadísticos de regresión múltiple

La interpretación de los estadísticos de la regresión múltiple llega a ser compleja y difícil. De hecho, la interpretación de los estadísticos del análisis multivariado es, en general, considerablemente más difícil que la interpretación de los estadísticos univariados estudiados previamente. Por lo tanto, se profundizará en cierto grado en la interpretación de los estadísticos del ejemplo.

Significancia estadística de la regresión y de R^2

La razón F de 8.684, calculada antes, indica que la regresión de Y sobre X_1 y X_2, expresada por $R^2_{y.12}$, es estadísticamente significativa. La probabilidad de que una razón F tan grande

como ésta, ocurra debido al azar, es menor al .01 (en realidad es aproximadamente .003), lo cual quiere decir que la relación entre Y y una combinación de los mínimos cuadrados de X_1 y X_2 quizá no ocurrió debido al azar.

La $R = .71$ se interpreta de forma similar a un coeficiente de correlación ordinario, excepto que los valores de R oscilan entre 0 y 1.00, a diferencia de la r, que va de −1.00 a 1.00, pasando por cero. Sin embargo, $R^2 = .71^2 = .51$ es más útil y tiene mayor importancia: significa que el 51 por ciento de la varianza de Y se explica o "determina" por X_1 y X_2, en combinación. Se le denomina, en concordancia, *coeficiente de determinación*. Una denominación alternativa para este estadístico es *CMC*, que son las siglas de "correlación múltiple cuadrada".

Contribuciones relativas de X a Y

Permítase el planteamiento, un tanto tímido, de una pregunta más difícil: ¿cuáles son las contribuciones relativas de X_1 y de X_2, de la aptitud verbal y la motivación de logro, a Y, el rendimiento en lectura? El enfoque restringido de este libro no permite el examen que merecen las respuestas a esta pregunta. El problema de la contribución relativa de las variables independientes a una variable o variables dependientes es uno de los más complejos y difíciles de los análisis de regresión. Parece que en realidad no existe ninguna solución satisfactoria, al menos cuando se correlacionan las variables independientes. No obstante, el problema no puede soslayarse. Sin embargo, el lector debe tener en cuenta que hay que tener mucha reserva en relación con la explicación anterior y las posteriores. Los problemas técnicos y sustantivos de la interpretación del análisis de regresión múltiple se tratan en dos o tres de las referencias presentadas en el ejercicio 1 de la sección de sugerencias de estudio, en el capítulo 33.

Se pensaría que los coeficientes de regresión, b o β, proporcionarían medios preparados para identificar las contribuciones relativas de las variables independientes a una variable dependiente; y así es, pero sólo de forma aproximada y en algunas ocasiones de forma confusa. Anteriormente se mencionó que el coeficiente de regresión b se llama *pendiente*. La pendiente de la línea de regresión está en relación con el porcentaje de unidades b de Y para una unidad de X. En el problema A de la tabla 32.1, por ejemplo, $b = .90$. Así, como se indicó antes, con el cambio de una unidad en X se predice un cambio de .90 en Y. No obstante, en la regresión múltiple, una interpretación tan directa como ésta no es tan fácil, debido a que existe más de una b. Sin embargo, se puede decir, *para los propósitos pedagógicos presentes*, que si X_1 y X_2 tienen aproximadamente la misma escala de valores —en el ejemplo de la tabla 32.2 los valores de X_1 y X_2 están en el rango aproximado del 1 al 10— las b son pesos que muestran un aproximado de la importancia relativa de X_1 y X_2. En el presente caso, la fórmula de regresión es:

$$Y' = .1027 + .6771X_1 + .3934X_2$$

Se puede decir que X_1, la aptitud verbal, tiene mayor peso que X_2, la motivación de logro, lo cual es verdad para este caso, pero quizá no siempre sea así, especialmente con más variables independientes.

Los coeficientes de regresión, por desgracia para los propósitos de interpretación, no permanecen estables, pues cambian con diferentes muestras y con la suma o resta de variables independientes en el análisis (véase Dillon y Goldstein, 1984; Howell, 1997; Pedhazur, 1996). No existe una forma absoluta para interpretarlos. Si todas las correlaciones entre las variables independientes son iguales o cercanas a cero, entonces se simplifica mucho la interpretación. Pero muchas, o la mayoría, de las variables que se correlacionan con la

TABLA 32.4 *Ejemplos de regresión múltiple con y sin correlaciones entre las variables independientes*

A			B		
1	2	Y	1	2	Y
1.00	.50	.87	1.00	0	.87
.50	1.00	.43	0	1.00	.43
.87	.43	1.00	.87	.43	1.00
$R^2_{y.12} = .76$			$R^2_{y.12} = .94$		

variable dependiente también se correlacionan entre sí. El ejemplo de la tabla 32.3 indica lo siguiente: la correlación entre X_1 y X_2 es .26, una correlación modesta. Sin embargo, dichas intercorrelaciones con frecuencia son más altas; y cuanto más altas sean (hasta cierto punto), más inestable será la situación de interpretación.

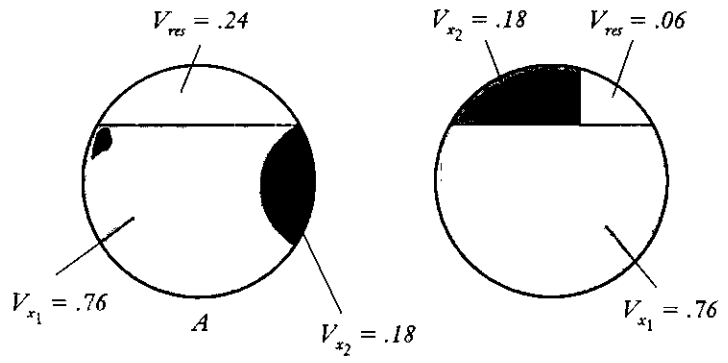
La situación predictiva ideal ocurre cuando las correlaciones entre las variables independientes y la variable dependiente son altas, y cuando las correlaciones entre las variables independientes son bajas. Este principio es importante. A mayor intercorrelación entre las variables independientes, habrá mayor dificultad en la interpretación. Entre otras cuestiones, se tiene una gran dificultad para establecer la influencia relativa de las variables independientes sobre la variable dependiente. Examine las dos matrices de correlación ficticias de la tabla 32.4 y sus R^2 acompañantes. En las dos matrices las variables independientes, X_1 y X_2, tienen una correlación de .87 y .43, respectivamente, con la variable dependiente, Y. Pero las correlaciones entre las variables independientes difieren en los dos casos. En la matriz A, $r_{12} = .50$, una correlación importante. En la matriz B, sin embargo, $r_{12} = 0$.

El contraste entre las R^2 es notorio: .76 para A y .94 para B. Puesto que en B, X_1 y X_2 no están correlacionadas, entonces, cualesquiera correlaciones que tengan con Y contribuyen de forma directa a la predicción y a la R^2. Cuando las correlaciones entre las variables independientes son exactamente iguales a cero, como en la matriz B, entonces resulta fácil calcular la R^2; simplemente es la suma de cuadrados de las r, entre cada variable independiente y la variable dependiente: $(.87)^2 + (.43)^2 = .94$. Cuando las variables independientes están correlacionadas, como sucede en la matriz A ($r_{12} = .50$), una parte de la varianza común de Y y X_1 también es compartida por X_2. En síntesis, X_1 y X_2 son redundantes, hasta cierto punto, al predecir Y. En la matriz B no existe dicha redundancia.

La situación se aclara, quizás, por medio de la figura 32.2. Sean los círculos la varianza total de Y, y que dicha varianza total sea 1.00. Entonces, las porciones de la varianza de Y, explicadas por X_1 y X_2, pueden representarse. En ambos círculos, el sombreado gris claro indica la varianza explicada para X_1 o V_{X1}; y el sombreado gris oscuro, para X_2 o V_{X2}. (Las varianzas restantes después de V_{X1} y V_{X2} son las varianzas residuales, denominadas así en la figura.) En B, V_{X1} y V_{X2} no se sobreponen. Sin embargo, en A, V_{X1} y V_{X2} sí se sobreponen. Sencillamente, debido a que $r_{12} = 0$ en B y $r_{12} = .50$ en A, el poder predictivo de las variables independientes es mucho mayor en B que en A. Por supuesto, ello se ve reflejado por las R^2: .76 en A y .94 en B.

Aunque se trata de un ejemplo ficticio y artificial, tiene la virtud de mostrar el efecto de correlación entre las variables independientes y, por lo tanto, ilustra el principio enunciado antes. También refleja la dificultad para interpretar los resultados de la mayoría de los análisis de regresión, ya que en muchas investigaciones las variables independientes

FIGURA 32.2



$R_{y \cdot 12}^2 = .76$
$R_{y \cdot 12}^2 = V_{x_1} + V_{x_2}$
$= .76 + 0 = .76$

$R_{y \cdot 12}^2 = .94$
$R_{y \cdot 12}^2 = V_{x_1} + V_{x_2}$
$= .76 + .18 = .94$

están correlacionadas. Y cuando se añaden más variables independientes, la interpretación se torna aún más compleja y difícil. Un problema central es: ¿cómo se clasifican los efectos relativos de las diferentes X sobre Y? La respuesta también es compleja. Existen varias formas de hacerlo, algunas más satisfactorias que otras; aunque ninguna lo es completamente. Quizás la forma más satisfactoria, por lo menos según la opinión y experiencia de los autores, se logra calculando *correlaciones semiparciales cuadradas* (también llamadas *correlaciones por partes*), las cuales se calculan con la fórmula:

$$SP^2 = R^2_{y.12\ldots k} - R^2_{y.12\ldots(k-1)}$$

o en el presente caso, para B:

$$SP^2 = R^2_{y.12} - R^2_{y.1} - .94 - .76 = .18$$

que indica la contribución de X_2 a la varianza de Y, *después de que* X_1 *se ha tomado en cuenta.* El mismo cálculo para A produce: .76 – .76 = 0, lo cual indica que X_2 no contribuye en lo absoluto a la varianza de Y, después de que X_1 se ha tomado en cuenta. (En realidad existe un ligero incremento que surge únicamente con un gran número de decimales.)

Se refiere al lector a Howell (1997), Dillon y Goldstein (1984) y Pedhazur (1996) para un análisis sobre los problemas involucrados. Kerlinger y Pedhazur (1973) también tratan el problema con considerable detalle y lo relacionan con ejemplos de investigación.

Otros problemas analíticos y de interpretación

Una cantidad de problemas del análisis de regresión múltiple no se tratan en este libro con el detalle que merecen. Sin embargo, es necesario mencionar algunos de ellos a causa de la creciente importancia de la regresión múltiple en la investigación del comportamiento.

Uno de ellos, ya mencionado, es el problema de los pesos de la regresión. En este capítulo y en el siguiente, la exposición se centra en los pesos b, ya que en la mayor parte de los usos de investigación de la regresión se predice con las puntuaciones por renglón o de desviación, y las b se utilizan con dichas puntuaciones. Beta o los pesos β, por otro lado, se utilizan con las puntuaciones estándar. Se les llama *coeficientes de regresión parcial estandarizados*. "Estandarizados" significa que serían utilizados si todas las variables estuvieran en forma de puntuación estándar. "Parcial" significa que los efectos de variables, distintas de aquella donde se aplica el peso, se mantienen constantes. Por ejemplo, $\beta_{y.123}$ o β_1, en un problema de tres variables (independientes), es el peso estandarizado de la regresión parcial, el cual expresa el cambio en Y, debido al cambio en X_1, manteniendo constantes las variables dos y tres. Un segundo significado, utilizado en el trabajo teórico, es que b es el peso de regresión poblacional que β estima. Aquí se omite ese significado. Las β se traducen en b con la fórmula:

$$b_j = \beta_j \frac{s_y}{s_j}$$

donde s_y es igual a la desviación estándar de Y y s_j la desviación estándar de la variable j. Los pesos b también son coeficientes parciales de regresión, pero no están en forma estandarizada.

Otro problema es que en cualquier regresión dada, R, R^2 y los pesos de la regresión serán iguales, sin importar el orden de las variables. No obstante, si se suma o se resta una o más variables de la regresión, dichos valores cambiarán. Y los pesos de regresión pueden cambiar de muestra a muestra. En otras palabras, no existe una calidad absoluta respecto a ellos. Por ejemplo, no puede decirse que debido a que las aptitudes verbal y numérica tienen pesos de regresión de .60 y .50 en un conjunto de datos, tendrán los mismos valores en otro conjunto.

Anteriormente en este libro se dijo: "El diseño es la disciplina de los datos." El diseño de investigación y el análisis de datos surgen a partir de las demandas de los problemas de investigación. De nuevo, el orden de aparición de las variables independientes en la ecuación de regresión se determina por el problema de investigación y el diseño de la investigación, el cual, a su vez, se determina por el problema de investigación.

Aunque el orden de anotación de las variables y los cambios en los pesos de regresión que ocurren con muestras que difieren son problemas difíciles, es necesario recordar que los pesos de regresión finales no cambian con órdenes diferentes de anotación. Ésta es una verdadera compensación, especialmente útil en la predicción. Por ejemplo, en muchos problemas de investigación, la contribución relativa de las variables no es una consideración importante. En tales casos, la ecuación total de la regresión y sus pesos de regresión se requieren principalmente para predecir y para evaluar la naturaleza general de la situación de regresión.

Sin embargo, cuando el investigador desea encontrar la contribución de cada variable independiente, deben utilizarse los pesos beta (pesos de regresión estandarizados). Dichos pesos beta se han escalado, de tal manera que se comparan entre sí de manera directa. Los pesos de regresión no estandarizados reflejan la escala de medición utilizada para medir esa variable. Por lo tanto, los pesos de regresión no estandarizados no pueden compararse de forma directa. Además, la significancia de los pesos beta es igual a la significancia del cambio en R^2, cuando una variable independiente se anota al final, dentro de la ecuación de regresión.

Otro aspecto importante es que, por lo general, existe una utilidad limitada en la añadidura de variables a una ecuación de regresión. Debido a que muchas variables de

investigación del comportamiento se correlacionan, opera el principio ilustrado por los datos de la tabla 32.4 y analizado anteriormente, disminuyendo la utilidad de variables adicionales. Si se encuentran tres o cuatro variables independientes que estén altamente correlacionadas con una variable dependiente, y no altamente correlacionadas entre sí, entonces se tiene suerte. Pero se vuelve cada vez más difícil encontrar otras variables dependientes que no sean, en efecto, redundantes con las primeras tres o cuatro. Si $R^2_{y.123} = .50$, entonces es poco probable que $R^2_{y.1234}$ sea mucho mayor que .55, y $R^2_{y.12345}$ tal vez no será mayor que .56 o .57. Se tiene una ley de regresión de ganancias reducidas. Cuando se agregan variables independientes, se nota cuánto agregan a R^2 y se prueba su significancia estadística. La fórmula para hacerlo, similar a la fórmula 32.13, es:

$$F = \frac{(R^2_{y.12.k_1} - R^2_{y.12.k_2})/(k_1 - k_2)}{(1 - R^2_{y.12.k_1})/(N - k_1 - 1)}$$

donde k_1 es el número de variables independientes de la R^2 mayor, k_2 es el número de variables independientes de la R^2 menor, y N es igual al número de casos. Tal fórmula se utilizará posteriormente. A pesar de que una F calculada como ésta puede ser estadísticamente significativa, en especial con una muestra grande, el incremento real de R^2 llega a ser muy pequeño. En un estudio realizado por Layton y Swanson (1958), la añadidura de una sexta variable independiente produjo una razón F estadísticamente significativa, ¡pero el incremento real de R^2 fue de .0147! La diferencia entre las R^2 en el numerador es el coeficiente de correlación semiparcial elevado al cuadrado.

Anteriormente se dijo que R, R^2 y los coeficientes de regresión permanecen iguales si las mismas variables se anotan en diferente orden. Sin embargo, no debe pensarse que ello significa que no importa el orden en que se anotan las variables en la ecuación de regresión. Por el contrario, el orden de anotación suele ser muy importante. Cuando las variables independientes están correlacionadas, la cantidad relativa de varianza de la variable dependiente explicada o a la que contribuye cada variable independiente, cambia drásticamente con un orden diferente de anotación de las variables. Con los datos A de la tabla 32.4, por ejemplo, si se revierte el orden de X_1 y X_2, sus contribuciones relativas cambian marcadamente. Con el orden original, X_2 no contribuye en nada a R^2; mientras que con el orden revertido, X_2 se convierte en X_1 y contribuye en un 19 por ciento [$r^2 = (.43)^2 = .19$] a la R^2 total, y la X_1 original, que se convierte en X_2, contribuye un 57 por ciento (.19 + .57 = .76). El orden de las variables, aunque no produce ninguna diferencia en la R^2 final y, por lo tanto, en la predicción general, constituye un problema importante en investigación.

Sin embargo, la multicolinealidad o las variables independientes correlacionadas no siempre son indeseables. En algunos casos, cuando se utiliza la regresión múltiple para establecer la validez de una medida o escala, las variables independientes correlacionadas son de gran utilidad. Las variables independientes que tienen correlaciones de cero o cercanas a cero con la variable dependiente, pero una alta correlación con otra variable independiente, en realidad llegan a incrementar la cantidad de varianza compartida por las variables dependiente e independiente. Este tipo de variable independiente se llama *variable supresora*. Algunos investigadores, como el doctor Leonard Helmers de Nueva Orleans,[1] se refieren a ellas como variables "de recorte". Dichas variables poseen el efecto de eliminar, suprimir o recortar la varianza irrelevante en las otras variables independientes. Suponga que se desea desarrollar una ecuación de regresión para predecir habilidades mecánicas. Se podría utilizar la puntuación de la persona en una prueba de desempeño de

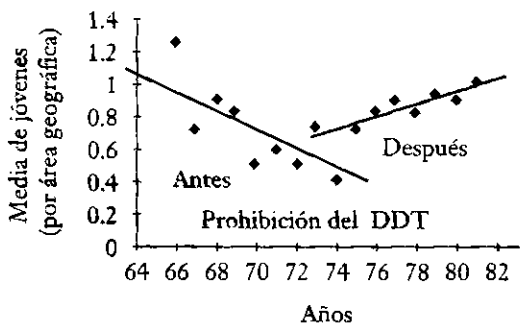
[1] Comunicación personal. El doctor Helmers fue Director de Investigación de ASI Marketing, Inc., en Hollywood, California.

habilidades mecánicas como variable dependiente, y se seleccionaría una prueba escrita sobre aptitud mecánica como variable independiente (predictora). Tal vez se desearía incluir una variable supresora, como la comprensión lectora, la cual sería un candidato para funcionar como variable supresora, ya que muy probablemente no se correlaciona con las habilidades mecánicas, pero sí con la *prueba escrita de aptitud mecánica* (Mechanical Aptitude Test), pues dicha prueba requiere de lectura. Así, las dos variables independientes *aptitud mecánica* y *comprensión lectora*, pueden correlacionarse, y la prueba de desempeño mecánico (Mechanical Performance Test) puede correlacionarse con la prueba de aptitud mecánica; sin embargo, la prueba de comprensión lectora no estaría correlacionada con la prueba de desempeño mecánico.

En otras situaciones, un investigador quizá no esté consciente de que una variable supresora se haya utilizado en el análisis. ¿Entonces cómo se puede saber si en este caso tiene una variable supresora? Bueno, si se cuenta con un programa computacional como el SPSS o el SAS (Sistema de Análisis Estadístico), el resultado generado por estos programas para computadora puede, bajo un escrutinio cuidadoso, utilizarse para detectar la presencia de una variable supresora. El primer paso consiste en determinar qué variables independientes poseen un peso beta distinto a cero. Si se encuentra una y el valor absoluto de la correlación simple entre la variable dependiente y su variable independiente es considerablemente menor que el peso beta asociado con esa variable independiente, es posible obtener una variable supresora. Además, si el peso beta para esa variable independiente es diferente a cero, y la correlación simple entre la variable dependiente y la variable independiente tiene un signo opuesto al del peso de beta, entonces ésta es una señal de que la variable independiente puede ser una variable supresora. En algunos análisis de investigación, como aquellos encontrados en la investigación de mercado, las variables supresoras se obtienen del análisis y se calcula nuevamente la ecuación de regresión.

En la literatura se encuentran ejemplos de variables supresoras. Hadfield, Littleton, Steiner y Woods (1998) utilizaron la regresión múltiple para analizar las habilidades pedagógicas de estudiantes y para probar la hipótesis de que el conocimiento de contenido matemático sería el correlato más significativo con la eficacia de la microenseñanza. [Se utilizaron las calificaciones en una filmación para medir la eficacia de la enseñanza. Las variables independientes fueron el conocimiento de contenido pedagógico, el conocimiento de contenido matemático, ansiedad ante las matemáticas y habilidad espacial.] Los autores encontraron que las puntuaciones del conocimiento de contenido matemático y de la ansiedad ante las matemáticas actuaron como variables supresoras. Con estas variables dentro de la ecuación de regresión hubo un incremento del 25 por ciento en la varianza

FIGURA 32.3



investigación del comportamiento se correlacionan, opera el principio ilustrado por los datos de la tabla 32.4 y analizado anteriormente, disminuyendo la utilidad de variables adicionales. Si se encuentran tres o cuatro variables independientes que estén altamente correlacionadas con una variable dependiente, y no altamente correlacionadas entre sí, entonces se tiene suerte. Pero se vuelve cada vez más difícil encontrar otras variables dependientes que no sean, en efecto, redundantes con las primeras tres o cuatro. Si $R^2_{y.123} = .50$, entonces es poco probable que $R^2_{y.1234}$ sea mucho mayor que .55, y $R^2_{y.12345}$ tal vez no será mayor que .56 o .57. Se tiene una ley de regresión de ganancias reducidas. Cuando se agregan variables independientes, se nota cuánto agregan a R^2 y se prueba su significancia estadística. La fórmula para hacerlo, similar a la fórmula 32.13, es:

$$F = \frac{(R^2_{y.12.k_1} - R^2_{y.12.k_2})/(k_1 - k_2)}{(1 - R^2_{y.12.k_1})/(N - k_1 - 1)}$$

donde k_1 es el número de variables independientes de la R^2 mayor, k_2 es el número de variables independientes de la R^2 menor, y N es igual al número de casos. Tal fórmula se utilizará posteriormente. A pesar de que una F calculada como ésta puede ser estadísticamente significativa, en especial con una muestra grande, el incremento real de R^2 llega a ser muy pequeño. En un estudio realizado por Layton y Swanson (1958), la añadidura de una sexta variable independiente produjo una razón F estadísticamente significativa, ¡pero el incremento real de R^2 fue de .0147! La diferencia entre las R^2 en el numerador es el coeficiente de correlación semiparcial elevado al cuadrado.

Anteriormente se dijo que R, R^2 y los coeficientes de regresión permanecen iguales si las mismas variables se anotan en diferente orden. Sin embargo, no debe pensarse que ello significa que no importa el orden en que se anotan las variables en la ecuación de regresión. Por el contrario, el orden de anotación suele ser muy importante. Cuando las variables independientes están correlacionadas, la cantidad relativa de varianza de la variable dependiente explicada o a la que contribuye cada variable independiente, cambia drásticamente con un orden diferente de anotación de las variables. Con los datos A de la tabla 32.4, por ejemplo, si se revierte el orden de X_1 y X_2, sus contribuciones relativas cambian marcadamente. Con el orden original, X_2 no contribuye en nada a R^2; mientras que con el orden revertido, X_2 se convierte en X_1 y contribuye en un 19 por ciento [$r^2 = (.43)^2 = .19$] a la R^2 total, y la X_1 original, que se convierte en X_2, contribuye un 57 por ciento (.19 + .57 = .76). El orden de las variables, aunque no produce ninguna diferencia en la R^2 final y, por lo tanto, en la predicción general, constituye un problema importante en investigación.

Sin embargo, la multicolinealidad o las variables independientes correlacionadas no siempre son indeseables. En algunos casos, cuando se utiliza la regresión múltiple para establecer la validez de una medida o escala, las variables independientes correlacionadas son de gran utilidad. Las variables independientes que tienen correlaciones de cero o cercanas a cero con la variable dependiente, pero una alta correlación con otra variable independiente, en realidad llegan a incrementar la cantidad de varianza compartida por las variables dependiente e independiente. Este tipo de variable independiente se llama *variable supresora*. Algunos investigadores, como el doctor Leonard Helmers de Nueva Orleans,[1] se refieren a ellas como variables "de recorte". Dichas variables poseen el efecto de eliminar, suprimir o recortar la varianza irrelevante en las otras variables independientes. Suponga que se desea desarrollar una ecuación de regresión para predecir habilidades mecánicas. Se podría utilizar la puntuación de la persona en una prueba de desempeño de

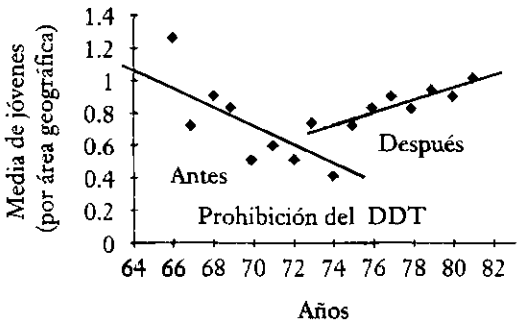
[1] Comunicación personal. El doctor Helmers fue Director de Investigación de ASI Marketing, Inc., en Hollywood, California.

habilidades mecánicas como variable dependiente, y se seleccionaría una prueba escrita sobre aptitud mecánica como variable independiente (predictora). Tal vez se desearía incluir una variable supresora, como la comprensión lectora, la cual sería un candidato para funcionar como variable supresora, ya que muy probablemente no se correlaciona con las habilidades mecánicas, pero sí con la *prueba escrita de aptitud mecánica* (Mechanical Aptitude Test), pues dicha prueba requiere de lectura. Así, las dos variables independientes *aptitud mecánica* y *comprensión lectora*, pueden correlacionarse, y la prueba de desempeño mecánico (Mechanical Performance Test) puede correlacionarse con la prueba de aptitud mecánica; sin embargo, la prueba de comprensión lectora no estaría correlacionada con la prueba de desempeño mecánico.

En otras situaciones, un investigador quizá no esté consciente de que una variable supresora se haya utilizado en el análisis. ¿Entonces cómo se puede saber si en este caso tiene una variable supresora? Bueno, si se cuenta con un programa computacional como el SPSS o el SAS (Sistema de Análisis Estadístico), el resultado generado por estos programas para computadora puede, bajo un escrutinio cuidadoso, utilizarse para detectar la presencia de una variable supresora. El primer paso consiste en determinar qué variables independientes poseen un peso beta distinto a cero. Si se encuentra una y el valor absoluto de la correlación simple entre la variable dependiente y su variable independiente es considerablemente menor que el peso beta asociado con esa variable independiente, es posible obtener una variable supresora. Además, si el peso beta para esa variable independiente es diferente a cero, y la correlación simple entre la variable dependiente y la variable independiente tiene un signo opuesto al del peso de beta, entonces ésta es una señal de que la variable independiente puede ser una variable supresora. En algunos análisis de investigación, como aquellos encontrados en la investigación de mercado, las variables supresoras se obtienen del análisis y se calcula nuevamente la ecuación de regresión.

En la literatura se encuentran ejemplos de variables supresoras. Hadfield, Littleton, Steiner y Woods (1998) utilizaron la regresión múltiple para analizar las habilidades pedagógicas de estudiantes y para probar la hipótesis de que el conocimiento de contenido matemático sería el correlato más significativo con la eficacia de la microenseñanza. [Se utilizaron las calificaciones en una filmación para medir la eficacia de la enseñanza. Las variables independientes fueron el conocimiento de contenido pedagógico, el conocimiento de contenido matemático, ansiedad ante las matemáticas y habilidad espacial.] Los autores encontraron que las puntuaciones del conocimiento de contenido matemático y de la ansiedad ante las matemáticas actuaron como variables supresoras. Con estas variables dentro de la ecuación de regresión hubo un incremento del 25 por ciento en la varianza

FIGURA 32.3



explicada. No obstante, ambas variables tuvieron una baja correlación con la variable dependiente, la eficacia de la enseñanza, pero estaban correlacionadas con las otras variables independientes.

El estudio de Leichtman y Erickson (1979) sobre los determinantes cognitivos, demográficos y de interacción de las habilidades para asumir roles en niños de cuarto grado, desarrollaron una ecuación de regresión donde cinco variables predijeron el 36 por ciento de la varianza de las puntuaciones del asumir roles. Estas variables fueron la puntuación de la escala de *vocabulario del WISC*, los errores de la prueba de emparejamiento de figuras familiares (Matching Familiar Figures Test), el sexo, el vecindario y la dominancia de mano. Se encontró que la puntuación de la escala de *vocabulario del WISC* era una variable supresora, la cual se correlacionaba con las otras variables independientes, pero no con la variable dependiente: la puntuación respecto a asumir roles.

Ejemplos de investigación

El DDT y las águilas calvas

Una de las diversas controversias sobre el deterioro del ambiente ocasionado por intereses comerciales, y la oposición y las protestas de los grupos ambientalistas se ha centrado en el uso del DDT. Uno de los efectos de rociar DDT ha sido la disminución de especies de aves. Por ejemplo, la reproducción de la población del águila calva fue seriamente afectada. En diciembre de 1972, la Agencia de Protección Ambiental (Enviromental Protection Agency) prohibió el uso del DDT. Grier (1982) en un estudio sobre el efecto que tuvo la prohibición en la reproducción de las águilas calvas, reportó el número promedio de águilas jóvenes por área geográfica, durante los años 1966 a 1981. Sus análisis de regresión (y de otros tipos) de los promedios de reproducción (medias) antes y después de la prohibición, mostraron que las dos pendientes, o coeficientes b, difirieron de forma significativa. De 1966 a 1974, $b = -.07$, lo cual indica un decremento en la reproducción a lo largo de esos años, pero de 1973 a 1981, $b = .07$, lo que indica un incremento. (Ambas b fueron estadísticamente significativas.) El método para comparar pendientes de manera estadística se presenta en Pedhazur (1996), Howell (1997) y Lee y Little (1996). Las regresiones simples se calculan utilizando los años como variable independiente y las tasas de reproducción como variable dependiente. La correlación antes de la prohibición del DDT fue de –.74, pero después de la prohibición fue de .80 (cálculos aproximados de los autores de este libro). Se graficaron dos regresiones en la figura 32.3. La gráfica refleja la regresión de la media de águilas jóvenes por área geográfica durante los años 1966 a 1974 (antes de la prohibición del DDT) y durante los años 1975-1981 (después de la prohibición). La regresión para los datos antes de la prohibición se calculó desde 1974, ya que se podía esperar que el efecto de la prohibición no se manifestara durante un año, aproximadamente. Grier realizó este cálculo desde 1973. La marcada diferencia entre las dos relaciones o pendientes es drástica.

Sesgo por exageración en exámenes de autoevaluación

¿Dicen la verdad, con frecuencia, los solicitantes de un empleo, respecto a sus capacidades? Los contratantes se preocupan cada vez más respecto al hecho de si las credenciales presentadas por un solicitante a un empleo son verdaderas. Anderson, Warner y Spencer (1984) utilizaron la regresión múltiple para responder esta pregunta. Los participantes en

TABLA 32.7 *Cantidad de varianza explicada por el examen de autoevaluación y la escala de exageración, sobre el desempeño en mecanografía*

Variable independiente	R^2	ΔR^2	
Autoevaluación	.07	.07	$p < .05$
Escala de exageración	.23	.16	$p < .05$
Interacción	.25	.02	$p > .05$

el estudio fueron 351 solicitantes de empleo para un puesto en el estado de Colorado. Se les pidió a los participantes que indicaran su nivel de experiencia con cierto tipo de tareas de trabajo. Algunas de las tareas presentadas a los solicitantes estaban diseñadas para fines de la investigación. Los investigadores crearon una escala sobre el sesgo por exageración para determinar el grado en que los solicitantes exageraban sobre sí mismos. Se utilizó un análisis de regresión múltiple para determinar la validez de esta escala de medición. A los solicitantes para puestos administrativos se les pidió que indicaran cuántas palabras por minuto podían mecanografiar, además de completar la escala de sesgo por exageración. Entonces, los investigadores usaron la prueba de mecanografía como variable dependiente en una regresión múltiple con la escala de sesgo por exageración y el examen de autoevaluación. Los resultados de la investigación se presentan en la tabla 32.7. La correlación entre la prueba de mecanografía y el examen de autoevaluación fue de .27 ($r^2 = .073$) y .41 ($r^2 = .168$) para la escala de exageración. En lo que se refiere a la explicación de la variación en las puntuaciones de la prueba de mecanografía, la escala de exageración incrementó la posibilidad de predicción. A pesar de que Anderson *et al.* no utilizaron una variable supresora para incrementar la R^2 en su estudio, podrían haberlo hecho. Un posible candidato para el trabajo de variable supresora quizás habría sido la comprensión lectora. Puesto que se requiere de la comprensión lectora para leer el cuestionario de autoevaluación, y a que probablemente esté correlacionada con el desempeño real en mecanografía, quizá habría sido una variable supresora.

Análisis de regresión múltiple e investigación científica

La regresión múltiple está cercana al corazón de la investigación científica. También es fundamental para la estadística y la inferencia, y está íntimamente relacionada con métodos matemáticos básicos y poderosos. Además, desde el punto de vista del investigador, es útil y práctica: realiza su trabajo analítico de manera exitosa y eficiente. Mediante la explicación de estas fuertes y radicales declaraciones, es posible aclarar lo que se ha aprendido.

El científico está relacionado, básicamente, con proposiciones del tipo "si p, entonces q", las cuales "explican" fenómenos. Cuando se dice "si incentivo positivo, entonces alto rendimiento", hasta cierto punto se está "explicando" el rendimiento. Pero ello no es suficiente. Aun cuando estuviera apoyada por una gran cantidad de evidencia empírica, esta proposición no va demasiado lejos en la explicación del rendimiento. Además de otras proposiciones *si-entonces* de tipo similar, el científico debe plantear preguntas más complejas. Él podría preguntar, por ejemplo, en qué condiciones es válida la proposición "si incentivo positivo, entonces alto rendimiento". ¿Es verdadera tanto para niños estadounidenses negros como para niños estadounidenses blancos? ¿Es verdadera para niños con baja y alta inteligencia? Para probar tales preguntas y para impulsar el conocimiento, los científicos escriben proposiciones del tipo *si p, entonces q, bajo las condiciones r, s y t*, donde p es una variable independiente; q una variable dependiente, y r, s y t otras variables inde-

pendientes. También se pueden escribir, evidentemente, otras proposiciones —por ejemplo, *si p y r, entonces q*—. En este caso *p* y *r* son dos variables independientes, las cuales se requieren para *q*.

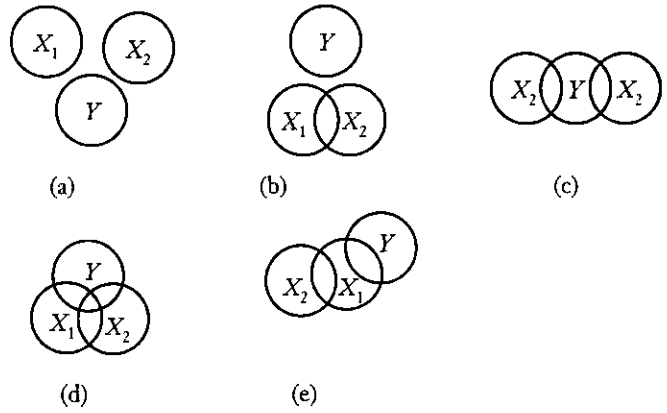
El punto central aquí es que la regresión múltiple puede manejar dichos casos de manera exitosa. En la mayor parte de la investigación del comportamiento existe, por lo general, una variable dependiente, aunque no se restringe teóricamente a una sola. Como consecuencia, la regresión múltiple constituye un método general para analizar muchos datos de investigación del comportamiento. Ciertos otros métodos se consideran casos especiales de regresión múltiple. El más relevante es el análisis de varianza, donde todos sus tipos se conceptualizan y logran con análisis de regresión múltiple.

Anteriormente se mencionó que todo control se refiere a control de la varianza. El análisis de regresión múltiple se considera como un método refinado y poderoso del "control" de varianza. Esto lo logra de la misma forma en que lo hace el análisis de varianza: estimando la magnitud de las diferentes fuentes de influencia sobre Y, de las diferentes fuentes de varianza de Y, a través del análisis de las interrelaciones de todas las variables. Indica qué cantidad de Y presuntamente se debe a $X_1, X_2, \ldots, X_k$. Da cierta idea sobre las cantidades relativas de influencia de las X; y facilita pruebas de significancia estadística de influencias combinadas de las X sobre Y, y de las influencias separadas de cada X. En síntesis, el análisis de regresión múltiple constituye una técnica eficiente y poderosa de comprobación de hipótesis y para realizar inferencias. Lo es debido a que ayuda a los científicos a estudiar, con relativa precisión, interrelaciones complejas entre variables independientes y una variable dependiente; Y, por lo tanto, los ayuda a "explicar" el presunto fenómeno representado por la variable dependiente.

RESUMEN DEL CAPÍTULO

1. La regresión múltiple constituye un método para estudiar los efectos y la magnitud de los efectos de más de una variable independiente, sobre una variable dependiente.
2. La regresión simple incluye una variable independiente y una variable dependiente.
3. A través del método de mínimos cuadrados, la regresión múltiple implica encontrar los mejores pesos de regresión que maximicen la relación entre una combinación lineal de las variables independientes y la variable dependiente.
4. R es la correlación múltiple. Es la correlación entre los valores reales de la variable dependiente y los valores que se predicen de la variable dependiente.
5. El cuadrado de la correlación múltiple, R^2, es un estadístico utilizado para determinar la calidad de la ecuación de regresión encontrada a través de los datos empíricos.
6. Los cálculos de la regresión múltiple son complicados. Se recomienda el uso de un programa computacional.
7. El cuadrado de la correlación múltiple, o coeficiente de determinación, se usa en pruebas estadísticas para determinar si la ecuación de regresión está explicando una cantidad significativa de la varianza.
8. Un problema de la regresión múltiple es que las variables independientes pueden estar correlacionadas. Cuando así sucede, conducen a estimados inestables de los coeficientes de regresión y a dificultades de interpretación.
9. Puede probarse la significancia estadística de la ecuación de regresión completa, así como de cada peso individual de regresión.
10. Las pruebas de los pesos individuales de regresión informarán al investigador qué variable está contribuyendo a la explicación de la variable dependiente.

FIGURA 32.4



SUGERENCIAS DE ESTUDIO

1. Lea uno o más de los siguientes estudios que utilizaron regresión múltiple. Tome nota de las variables utilizadas, de los programas computacionales utilizados y de las conclusiones obtenidas a partir de los resultados.

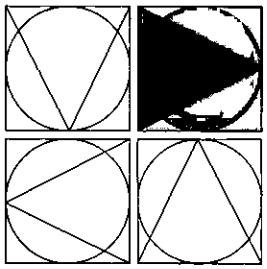
   Abel, M. H. (1998). Interaction of humor and gender in moderating relationships between stress and outcomes. *Journal of Psychology, 132*, 267-276.

   Connelly, C. D. (1998). Hopefulness, self-esteem, and perceived social support among pregnant and nonpregnant adolescents, *Western Journal of Nursing Research, 20*, 195-209.

   Ho, R. (1998). The intention to give up smoking: Disease versus social dimensions. *Journal of Social Psychology, 138*, 368-380.

   Stalenheim, E. G., Eriksson, E., von Knorring, L. y Wide, L. (1998). Testosterone as a biological marker in psychopathy and alcoholism. *Psychiatry Research*, 77, 79-88.

2. Dados los diagramas de Venn en la figura 32.4, una variable dependiente, Y, y dos variables independientes, X_1 y X_2,
   a) Determine cuál posee una variable supresora.
   b) ¿Cuál es ideal para una regresión múltiple?
   c) Indique cuál(es) exhibe(n) multicolinealidad.
   d) ¿Cuál produciría una ecuación de regresión inútil?
   e) ¿Cuál tiene mayor probabilidad de producir pruebas estadísticas no significativas en la ecuación de regresión?
   f) ¿Cuál tiene mayor probabilidad de producir pruebas estadísticas no significativas en los coeficientes individuales de regresión?



CAPÍTULO 33

REGRESIÓN MÚLTIPLE, ANÁLISIS DE VARIANZA Y OTROS MÉTODOS MULTIVARIADOS

- ANÁLISIS DE VARIANZA DE UN FACTOR Y ANÁLISIS DE REGRESIÓN MÚLTIPLE
- CODIFICACIÓN Y ANÁLISIS DE DATOS
- ANÁLISIS FACTORIAL DE VARIANZA, ANÁLISIS DE COVARIANZA Y ANÁLISIS RELACIONADOS
- ANÁLISIS DISCRIMINANTE, CORRELACIÓN CANÓNICA, ANÁLISIS MULTIVARIADO DE VARIANZA Y ANÁLISIS DE RUTA
- REGRESIÓN DE CRESTA, REGRESIÓN LOGÍSTICA Y ANÁLISIS LOGARÍTMICO LINEAL
  Tablas de contingencia de múltiples factores y análisis log-lineal
- ANÁLISIS MULTIVARIADO E INVESTIGACIÓN CIENTÍFICA

Un examen detallado muestra que las bases conceptuales de los diferentes modelos de análisis de datos son iguales o similares. La simetría de las ideas fundamentales tiene un gran atractivo estético, y en ninguna parte es más interesante y atractiva que en la regresión múltiple y en el análisis de varianza. Anteriormente, cuando se explicaron los fundamentos del análisis de varianza, se resaltó la similitud entre los principios y estructuras del análisis de varianza y los denominados métodos de correlación. Ahora se vincularán los dos métodos y, en el proceso, se mostrará que el análisis de varianza puede realizarse utilizando la regresión múltiple. Además, el enlace de los dos métodos producirá felizmente ganancias inesperadas. Se observará, por ejemplo, que ciertos problemas analíticos que no pueden tratarse con el análisis de varianza —o al menos difíciles de tratar cuando no verdaderamente inapropiados— se conceptualizan y abordan con facilidad por medio del uso juicioso y flexible de la regresión múltiple y sus variantes. Por limitaciones de espacio y como el propósito del libro no es enseñar los mecanismos de los modelos ni los métodos estadísticos,

la exposición será breve: algunas de las cosas que se digan deberán creerse con base en la confianza del lector. No obstante, aun con un nivel de exposición así se verá que ciertos problemas difíciles asociados con el análisis de varianza se manejan de manera natural y fácil con el análisis de regresión múltiple. Algunos de estos problemas asociados incluyen el análisis de covarianza, los datos del pretest y del postest, inequidad del número de casos en las casillas (o diseños factoriales), variables dependientes categóricas, tablas de contingencia de múltiples factores y el manejo tanto de datos experimentales como no experimentales.

Análisis de varianza de un factor y análisis de regresión múltiple

Suponga que se llevó a cabo un experimento con tres métodos de presentación de materiales verbales a niños de tercer año de secundaria. La variable dependiente es la comprensión, medida a través de una prueba objetiva sobre los materiales. Suponga también que los resultados son los que se presentan en la tabla 33.1. Evidentemente hay que realizar un análisis de varianza. Los resultados del análisis de varianza se indican en la parte final de la tabla. Se invita al lector a que realice los cálculos de los ejemplos de este capítulo, lo cual es sumamente necesario para lograr la cabal comprensión de importantes aspectos que se tratarán aquí. Por ejemplo, realice los cálculos del análisis de varianza de la tabla 33.1 y de la regresión múltiple del problema de la tabla 33.2; estudie y pondere los resultados de ambos análisis. Es importante que no se realicen por medio de un programa computacional, ya que es probable que no se comprendan. Siempre que sea posible, debe trabajar los problemas con detenimiento. Si se sucumbe a la tentación de utilizar uno de los paquetes para las grandes computadoras o para microcomputadoras, se debe ser cauteloso, pues algunas ocasiones la calidad de los programas estadísticos para las microcomputadoras (y para las grandes) es cuestionable. La razón F en la tabla 33.1 es 18, la cual con 2 y 12 grados de libertad, es significativa al nivel .01. El efecto del tratamiento experimental es claramente significativo: $\eta^2 = sc_e/sc_t = 90/120 = .75$. La relación entre el tratamiento experimental y la comprensión es fuerte.

Ahora, es posible transferir el pensamiento del marco de referencia del análisis de varianza al marco de referencia de la regresión múltiple. ¿Es posible obtener $b^2 = .75$ de forma "directa"? La variable independiente *métodos* se considera como la pertenencia en

TABLA 33.1 *Datos ficticios y resultados del análisis de varianza de un factor con tres grupos experimentales*

A_1	A_2	A_3
4	7	1
5	8	2
6	9	3
7	10	4
8	11	5

Fuente	*gl*	*sc*	*cm*	*F*
Entre grupos	2	90.0	45.0	18.0 ($p < .01$)
Dentro de grupos	12	30.0	2.5	
Total	14	120		

TABLA 33.2 *Distribución de los datos y cálculos de regresión (datos de la tabla 33.1)*

	Y	X_1	X_2
	4	1	0
A_1	5	1	0
	6	1	0
	7	1	0
	8	1	0
A_2	7	0	
	8		1
	9	0	1
	10	0	1
	11	0	1
A_3	1	0	
	2	0	0
	3	0	0
	4	0	0
	5	0	0
Σ:	90	5	5
M:	6	.3333	.3333
Σ^2:	660	5	5

los tres grupos experimentales, A_1, A_2 y A_3, que se expresa con los números 1 y 0: si un sujeto es miembro de A_1, se le asigna un 1; si se trata de un miembro de A_2 o de A_3, se le asigna un 0. O se pueden asignar números 1 a los sujetos pertenecientes a A_2 y 0 a los miembros de los otros dos grupos. Los resultados serán básicamente los mismos. De hecho, si se utilizan cualesquiera dos números diferentes, por ejemplo 1 y 10 o 31 y 5, o cualquier serie de dos números aleatorios, los resultados básicos serán los mismos. Sin embargo, la asignación de los números 1 y 0 posee ventajas interpretativas que se mencionarán posteriormente (véase Cohen y Cohen, 1983), y casi siempre funcionan mejor con los programas estadísticos computacionales. La distribución de los datos del análisis de regresión de los resultados de la tabla 33.1 se presenta en la tabla 33.2. Considere como un solo conjunto de puntuaciones las 15 medidas de la variable dependiente en la columna denominada Y. Haga lo mismo con las "puntuaciones" de X_1 y X_2, con excepción de que los números 1 y 0 indican la pertenencia al grupo. A los miembros de A_1 se le asignaron números 1 en la columna X_1; mientras que a los miembros de A_2 y A_3 se les asignó 0 (segunda columna). A los miembros de A_2 se les asignó 1 en la tercera columna, X_2; mientras que a los miembros de A_1 y A_3 se les asginó 0. Se podría plantear la pregunta: ¿en qué parte de la tabla está A_3? Al codificar grupos experimentales, sólo hay $k - 1$ vectores codificados (columnas), donde k es igual al número de tratamientos experimentales (en este caso $k = 3$). Expresado de otra forma, sólo hay un vector codificado (columnas) para cada grado de libertad. Recuerde, de la exposición previa sobre el análisis de varianza, que los grados de libertad entre grupos eran $k - 1$. En tal caso existen tres tratamientos, A_1, A_2 y A_3, y $k = 3$. Por lo tanto son $k - 1 = 2$ vectores codificados. Los vectores de los números 1 y 0 se llaman *variables prototipo o dummy* (véase Suits, 1967, para mayores detalles sobre las variables dummy). Para encontrar una exposición completa sobre las variables codificadas en el análisis de regresión múltiple, consulte los capítulos 6 y 7 de Kerlinger y Pedhazur (1973) o Pedhazur (1966), donde se expresan los tres tratamientos experimentales de forma completa.

TABLA 33.3 *Sumas de cuadrados y productos cruzados (datos de la tabla 33.2)*[a]

	x_1	x_2	y
x_1	3.3333	−1.6667	0
x_2		3.3333	15.0000
y			120.0000

[a] Los valores sobre la diagonal son las sumas de cuadrados de desviación: Σx_1^2, Σx_2^2 y Σy_t^2. Los tres valores restantes que están por arriba de la diagonal son los productos cruzados de desviación: $\Sigma x_1 x_2$, $\Sigma x_1 y$ y $\Sigma x_2 y$.

Ahora realice un análisis de regresión múltiple con los datos de la tabla 33.3, tal como se hizo en el capítulo 32. Las sumas de cuadrados y los productos cruzados necesarios para el análisis se incluyen en la tabla 33.3. Por ejemplo:

$$\Sigma x_1^2 = (1^2 + 1^2 + \ldots + 0^2) - \frac{5^2}{15} = 5 - 1.6667 = 3.3333$$

$$\Sigma x_2 y = (0)(4) + (0)(5) + \ldots + (0)(5) - \frac{(5)(90)}{15} = 45 - 30 = 15$$

$$\Sigma x_1 x_2 = (1)(0) + (1)(0) + \ldots + (0)(0) - \frac{(5)(5)}{15} = 0 - 1.6667 = -1.6667$$

Para calcular la regresión y las sumas de cuadrados residuales se utilizan las fórmulas 32.7 y 32.8 del capítulo 32 (presentadas aquí con la numeración de este capítulo):

$$sc_{reg} = b_1 \Sigma x_1 y - b_2 \Sigma x_2 y \qquad (33.1)$$

$$sc_{res} = sc_t - sc_{reg} \qquad (33.2)$$

En la tabla 33.3 se incluyen todos los valores anteriores, con excepción de b_1 y b_2, los coeficientes de regresión, y a, la constante de la intersección. Existen varias formas para calcular las b, pero están fuera del alcance de la presente exposición. Por lo tanto, deberán aceptarse por confianza: $b_1 = 3$ y $b_2 = 6$. La constante de la intersección a se calcula de la siguiente manera:

$$a = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2 \qquad (33.3)$$

$$= 6 - (3)(.3333) - (6)(.3333) = 3$$

Las sumas de los productos cruzados se muestran en la tabla 33.3: $\Sigma x_1 y = 0$ y $\Sigma x_2 y = 15$. Sustituyendo 33.1 en 33.2 se obtiene:

$$sc_{reg} = (3)(0) + (6)(15) = 90$$

$$sc_{res} = 120 - 90 = 30$$

Para calcular R^2 se utiliza la fórmula 32.11 del capítulo 32 (con un nuevo número):

$$R^2 = \frac{sc_{reg}}{sc_t} = \frac{90}{120} = .75$$

$$R = \sqrt{.75} = .8660 \qquad (33.4)$$

Por último, se calcula la razón F por medio de la fórmula 32.13, (con un nuevo número):

$$F = \frac{R^2/k}{(1 - R^2)/(N - k - 1)} \tag{33.5}$$

donde k es igual al número de variables independientes y N es igual al número de casos. Sustituyendo:

$$F = \frac{.75/2}{(1 - .75)/(15 - 2 - 1)} = \frac{.375000}{.020833} = 18$$

Se puede tomar prestada otra fórmula de F del capítulo previo:

$$F = \frac{sc_{reg}/gl_1}{sc_{reg}/gl_2} = \frac{sc_{reg}/k}{sc_{res}/(N - k - 1)} = \frac{90/2}{30/(15 - 2 - 1)} = \frac{45}{2.5} = 18$$

Después se verifica esta razón F en la tabla respectiva (véase Kerlinger y Pedhazur, 1973, apéndice D o C de este libro), con $gl = 2, 12$. La cifra para $p = .05$ es 3.88, y para $p = .01$ es 6.93. Puesto que la F de 18 calculada antes es mayor que el valor 6.93 de la tabla, la regresión es estadísticamente significativa, y la R^2 también lo es. Note que aun cuando A_3 no fue codificada —no posee vector codificado propio— su media es fácilmente recuperada al sustituir los 0 de X_1 y X_2.

Aun cuando se ha demostrado que el análisis de regresión múltiple logra lo mismo que el análisis de varianza, ¿puede decirse que existe alguna ventaja real en el uso del método de regresión? En realidad, los cálculos son más complicados. Entonces, ¿por qué hacerlo? La respuesta es que con el tipo de datos del ejemplo anterior no existe una ventaja práctica, más allá de la belleza estética y la aclaración conceptual. Pero cuando los problemas de investigación son más complejos —por ejemplo, cuando se involucran interacciones, covariables (como puntuaciones de pruebas de inteligencia), variables nominales (como género, clase social), variables continuas y componentes no lineales (X^2, X^3)— el procedimiento de la regresión tiene ventajas definitivas. De hecho, muchos problemas analíticos de investigación que el análisis de varianza no puede manejar con facilidad, o en lo absoluto, pueden resolverse con bastante facilidad mediante el análisis de regresión múltiple. El análisis factorial de varianza, el análisis de covarianza y, de hecho, todas las formas del análisis de varianza, también se pueden llevar a cabo con el análisis de regresión. Como no es propósito de los autores la enseñanza de la estadística ni de los mecanismos de análisis, se refiere al lector a las explicaciones adecuadas, como las que han sido citadas previamente en este capítulo. Sin embargo, en la siguiente sección se explicará la naturaleza de métodos sumamente importantes de codificación de variables y su utilidad en el análisis.

Codificación y análisis de datos

Antes de extender la explicación sobre la regresión múltiple y el análisis de varianza, es necesario conocer algo sobre las diferentes formas de codificación de los tratamientos experimentales, para el análisis de regresión múltiple. Un código es un conjunto de símbolos asignados a un conjunto de objetos por diversas razones. En el análisis de regresión múltiple, la codificación es la asignación de números a los miembros de una población o muestra,

para indicar la pertenencia a un grupo o subconjunto, de acuerdo con una regla determinada por un medio independiente. Cuando alguna característica o aspecto de los miembros de una población o muestra se define de forma objetiva, entonces es posible crear un conjunto de pares ordenados, cuyo primer miembro constituye la variable dependiente, Y, y el segundo miembro es el indicador numérico de los subconjuntos o de la pertenencia al grupo.

En la explicación anterior sobre la codificación de los tratamientos experimentales en la regresión múltiple análoga al análisis de varianza de un factor, se utilizaron los números 1 y 0. Los vectores de los 1 y de los 0 están correlacionados. Por ejemplo, en la tabla 33.3 la suma de los productos cruzados, $\sum x_1 x_2$, es –1.6667, y $r_{12} = -.50$. Dichos códigos 1 y 0 o *dummy* funcionan bastante bien. También es posible utilizar otras formas de codificación. Una de ellas, la codificación de los efectos, consiste en asignar {1, 0, –1} o {1, –1} a los tratamientos experimentales. Aunque se trata de un método útil, sólo se comentará brevemente.

Para aclarar algunos aspectos, la codificación de la tabla 33.2, una regresión múltiple análoga al análisis de varianza de un factor de los datos de la tabla 33.1, con tres grupos experimentales o tratamientos, se presenta en la tabla 33.4. Bajo el encabezado "Dummy" se presenta la codificación dummy de la tabla 33.4, utilizando sólo dos sujetos por grupo experimental. Como existen dos grados de libertad, o $k - 1 = 3 - 1 = 2$, hay dos vectores de columna denominados X_1 y X_2. Ya se explicó la asignación de la codificación dummy: un "1" indica que un sujeto es miembro del grupo experimental junto al que se ha colocado el 1, y un 0 indica que el sujeto no es miembro del grupo experimental.

Bajo la columna de "Efectos", la codificación parece ser {1, 0, –1}. La codificación de los efectos es prácticamente la misma que la codificación dummy —de hecho, se denominó codificación dummy— con excepción de que a un grupo experimental, por lo general al último, siempre se le asignan los números –1. Si las n de los grupos experimentales son iguales, entonces las sumas de las columnas de los códigos son iguales a cero. Sin embargo, los vectores no carecen de correlación de forma sistemática. La correlación entre las dos columnas bajo "Efectos" en la tabla 33.4, por ejemplo, es 50. (Contraste esto con la correlación entre las columnas de los códigos "Dummy": $r = -.50$.)

Cada uno de estos sistemas de codificación tiene sus propias características. Dos de las características de la codificación dummy se comentaron en la sección previa. Por otro lado, una de las características de la codificación de efectos es que la constante de la intersección, a, generada por el análisis de regresión múltiple, igualará a la gran media, o M_1,

TABLA 33.4 *Ejemplos de codificación dummy, de efectos y ortogonal de los tratamientos experimentales*[a]

	Dummy		*Efectos*		*Ortogonal*	
Grupos	X_1	X_2	X_1	X_2	X_1	X_2
A_1	1	0	1	0	0	2
	1	0	1	0	0	2
A_2	0	1	0	1	–1	–1
	0	1	0	1	–1	–1
A_3	0	0	–1	–1	1	–1
	0	0	–1	–1	1	–1
		$r_{12} = -.50$		$r_{12} = .50$		$r_{12} = .00$

[a] En la codificación dummy, A_3 es un grupo control. En la codificación ortogonal, A_2 se compara con A_3, y A_1 se compara con A_2 y A_3, o $(A_2 + A_3)/2$.

de Y. Para los datos de la tabla 33.2 la constante de la intersección es 6.00, que es la media de todas las puntuaciones de Y.

La tercera forma de codificación es la ortogonal; también se denomina codificación de "contrastes", aunque algunas codificaciones de contrastes pueden no ser ortogonales. Como su nombre lo indica, los vectores codificados son ortogonales o no relacionados. Si el principal interés de un investigador son los contrastes específicos entre medias más que en la prueba F en general, la codificación ortogonal puede proporcionar los contrastes requeridos. En cualquier conjunto de datos es posible realizar una cantidad de contrastes. Por supuesto, ello es particularmente útil en el análisis de varianza. La regla es que sólo se hacen los contrastes que son ortogonales entre sí, o independientes. Por ejemplo, en la tabla 33.4 la codificación del último conjunto de vectores es ortogonal: cada uno de los vectores totaliza cero y la suma de sus productos es cero, o

$$(0 \times 2) + (0 \times 2) + (-1)(-1) + \ldots + (1)(-1) = 0$$

r_{12} también es igual a cero.

Suponga que en lugar de la codificación dummy de la tabla 33.2, ahora se utiliza la codificación ortogonal y que también se decide probar A_2 contra A_3, o $M_{A2} - M_{A3}$, y que también se prueba A_1 contra A_2 y A_3, o $M_{A1} - (M_{A2} + M_{A3})/2$. Entonces X_1 se codifica (0, −1, 1) y X_2 se codifica (2, −1, −1), como lo indica la codificación ortogonal de la tabla 33.4. El lector interesado en el análisis de varianza puede seguir dichas posibilidades al consultar a Cohen y Cohen (1983) o a Kerlinger y Pedhazur (1973).

Sin importar qué codificación se utilice, R^2, F, las sumas de cuadrados, los errores estándar de estimación y las Y predichas serán iguales (las medias de los grupos experimentales). La constante de la intersección, los pesos de regresión y las pruebas t de los pesos b serán diferentes. Estrictamente hablando, no es posible recomendar un método sobre otro; cada uno tiene sus propósitos. En un inicio, quizá sea pertinente que el estudiante utilice el método más simple, la codificación dummy o los números 1 y 0. No obstante, poco después se debe usar la codificación de los efectos. Al final, se puede probar y dominar la codificación ortogonal. Antes de utilizar la codificación ortogonal en cualquier grado, el estudiante debe estudiar el tema de la comparación de medias (véase Hays, 1994).

El uso más simple de la codificación constituye la indicación de variables nominales, en particular las de tipo dicotómico. Algunas variables son dicotomías "naturales": género, escuela pública-escuela privada, con convicción-sin convicción, voto a favor-voto en contra. Todas ellas pueden calificarse (1, 0) y los vectores resultantes se analizan como si fueran vectores de puntuación continua. Sin embargo, la mayoría de las variables son continuas, o lo son potencialmente, aun cuando pueden ser siempre tratadas como dicotómicas. En cualquier caso, el uso de vectores (1, 0) para las variables dicotómicas en la regresión múltiple es sumamente útil.

Con variables nominales que no son dicotómicas aun se pueden utilizar los vectores (1, 0). Tan sólo se crea un vector (1, 0) para cada subconjunto, pero sólo uno de una categoría o partición. Suponga que la categoría A se divide en A_1, A_2, A_3, por ejemplo, protestante, católico, judío. Entonces, se crea un vector para protestantes, a cada uno de los cuales se les asigna un 1, a los católicos y judíos se les asigna 0. Se crea otro vector para católicos: a cada católico se le asigna un 1; a los protestantes y judíos se les asigna 0. En efecto sería redundante crear un tercer vector para los judíos. El número de vectores es $k - 1$, donde k es igual al número de subconjuntos de la partición o categoría.

Aunque algunas veces es conveniente o necesaria, la partición de una variable continua en una dicotomía o tricotomía descarta información. Si, por ejemplo, un investigador dicotomiza inteligencia, etnocentrismo, cohesión de grupos o cualquier otra variable que

pueda medirse con una escala que tenga intervalos inclusive aproximadamente iguales, se estaría descartando información potencialmente valiosa. Reducir un conjunto de valores con un rango relativamente amplio a una dicotomía, implica reducir su varianza y, por lo tanto, su posible correlación con otras variables. Por lo tanto, una buena regla del análisis de datos de investigación es: no reducir variables continuas a variables discretas (dicotomías, tricotomías, etcétera), a menos que se esté obligado a hacerlo debido a circunstancias o a la naturaleza de los datos (seriamente sesgados, bimodales, etcétera).

Análisis factorial de varianza, análisis de covarianza y análisis relacionados

Es por medio del análisis factorial de varianza, el análisis de covarianza y las variables nominales que se empiezan a apreciar las ventajas del análisis de regresión múltiple. Aquí se hará algo más que comentar el uso de los vectores codificados en el análisis factorial de varianza. Explicaciones excepcionalmente completas se encuentran en el extenso trabajo de Pedhazur (1996). Sin embargo, se explicará la razón básica de por qué el análisis de regresión múltiple con frecuencia resulta más conveniente que el análisis factorial de varianza.

La dificultad subyacente en investigación y análisis consiste en que las variables independientes de interés estén correlacionadas. Sin embargo, el análisis de varianza supone que no están correlacionadas. Si, por ejemplo, se tienen dos variables experimentales independientes, y los sujetos se asignan aleatoriamente a las casillas de un diseño factorial, se puede suponer que las dos variables independientes no están correlacionadas, por definición. Además, es apropiado usar un análisis factorial de varianza. Sin embargo, si se tienen dos variables independientes no experimentales y las mismas dos variables experimentales, no es posible asumir que las cuatro variables independientes no estén correlacionadas. A pesar de que existen formas para analizar tales datos con un análisis de varianza, éstos son engorrosos y un tanto forzados. Más aún, si existen n desiguales en los grupos, el análisis de varianza se vuelve todavía más inapropiado, a causa de que n desiguales también introducen correlaciones entre las variables independientes. Por otro lado, el procedimiento analítico de la regresión múltiple toma conocimiento, por así decirlo, de las correlaciones entre las variables independientes, así como entre las variables independientes y la variable dependiente. Esto significa que la regresión múltiple puede analizar —de forma separada o conjunta— tanto los datos experimentales como los no experimentales, de manera efectiva. Además, las variables continuas y categóricas pueden utilizarse de manera conjunta.

Cuando los sujetos se han asignado aleatoriamente a las casillas de un diseño factorial, y lo demás permanece igual, no se deriva demasiado beneficio del uso de la regresión múltiple. Pero cuando las n de las casillas son desiguales, y se desea incluir una, dos o más variables de control —como inteligencia, género y clase social— o cuando el análisis involucra el uso de variables continuas, entonces debe utilizarse la regresión múltiple. Este punto es de gran importancia. En el análisis de varianza, añadir variables de control resulta difícil y absurdo. Sin embargo, en la regresión múltiple la inclusión de tales variables es fácil y natural: cada una es sencillamente otro vector de puntuaciones, ¡otra X_j!

Análisis de covarianza

El análisis de covarianza (y no el análisis estructural de covarianza, que se estudiará posteriormente) es un ejemplo particularmente bueno del valor de un modelo de regresión

múltiple, ya que es difícil y pesado desde el marco conceptual del análisis de varianza, y fácil y completamente comprendido y logrado en un marco de referencia de regresión. Lo que el análisis de covarianza hace en su aplicación tradicional (véase Hays, 1994) es probar la significancia de las diferencias entre medias, después de tomar en cuenta o controlar diferencias individuales iniciales respecto a una *covariable*, es decir, una variable que está correlacionada con la variable dependiente. (Dicha correlación se toma en cuenta.) No obstante, en el modelo de regresión múltiple la influencia de la covariable está controlada tal y como si hubiera cualquier variable independiente, cuya influencia sobre la variable dependiente tuviera que controlarse. La covariable puede ser un pretest o una variable cuya influencia debe ser "eliminada" estadísticamente.

Estudios a gran escala realizados por Prothro y Grigg (1960) y McClosky (1964) encontraron que el grado de acuerdo de las personas acerca de aspectos sociales se incrementa conforme el aspecto se vuelve más abstracto. Suponga que un científico en política considera que el autoritarismo está muy involucrado en esta relación, que cuanto más autoritaria sea la persona, mayor acuerdo mostrará con afirmaciones sociales abstractas. Para estudiar la relación entre lo abstracto y el grado de acuerdo, el investigador tendrá que controlar el *autoritarismo*. En otras palabras, el científico político está interesado en estudiar la relación entre qué tan abstracto es un aspecto y las afirmaciones, por un lado, y el grado de acuerdo con dichos aspectos y las afirmaciones, por el otro. Hasta aquí el interés no se centra en el autoritarismo ni en el grado de acuerdo. El interés principal es *controlar la influencia del autoritarismo sobre el grado de acuerdo. El autoritarismo es la covariable.*

El científico político diseña tres tratamientos experimentales, A_1, A_2 y A_3, que son materiales con diferentes niveles de abstracción. Se obtienen las respuestas de 15 sujetos que han sido asignados aleatoriamente a los tres grupos experimentales, cinco en cada grupo. Antes de que el experimento empiece, el investigador aplica la escala *F* (de autoritarismo) a los 15 sujetos y utiliza tales medidas como la covariable. El objetivo es controlar la posible influencia del autoritarismo sobre el grado de acuerdo. Se trata de un análisis bastante directo de un problema de covarianza, donde se prueba la significancia de las diferencias entre las tres medias del grado de acuerdo, después de corregir las medias de la influencia del autoritarismo y tomando en cuenta la correlación entre autoritarismo y el grado de acuerdo. Ahora se realiza el análisis de covarianza usando el análisis de regresión múltiple.

En la tabla, los datos se presentan en la forma acostumbrada para el análisis de covarianza. En el análisis de covarianza se realizan análisis de varianza separados con las puntuaciones *X*, con las puntuaciones *Y* y con los productos cruzados de las puntuaciones de *X* y *Y*, *XY*. Después, mediante un análisis de regresión, se calculan las sumas de cuadra-

TABLA 33.5 *Análisis de un problema ficticio de covarianza con tres grupos experimentales y una covariable*

Tratamientos					
A_1		A_2		A_3	
X	*Y*	*X*	*Y*	*X*	*Y*
12	12	6	9	12	15
11	12	9	9	10	12
10	11	11	13	4	9
12	10	14	14	4	8
10	12	2	5	8	11

dos y los cuadrados medios de los errores de estimación del total y dentro de grupos y, finalmente, los ajustados entre grupos. Puesto que el interés aquí no es el procedimiento usual del análisis de covarianza, no se realizan estos cálculos. En su lugar, se procede de inmediato al enfoque de regresión múltiple para el análisis.

En la tabla 33.6 se presentan los datos de la tabla 33.5 ordenados para el análisis de regresión múltiple. Como siempre, hay un vector para la variable dependiente, Y. Un segundo vector, X_1, es la covariable. Los dos vectores restantes, X_2 y X_3, representan los tratamientos experimentales A_1 y A_2. (No es necesario tener un vector para A_3 ya que sólo existe un vector por cada grado de libertad, y únicamente hay dos grados de libertad.)

Un análisis de regresión produce: $R^2_{y.123} = .8612$ y $R^2_{y.1} = .7502$. Para probar la significancia de las diferencias entre las medias de A_1, A_2 y A_3, después de realizar el ajuste para el efecto de X_1, la varianza de Y debida a la covariable se resta de la varianza total explicada por la regresión de Y a partir de las variables X_1, X_2 y X_3: $R^2_{y.123} - R^2_{y.1}$. Después se prueba el producto:

$$F = \frac{(R^2_{y.123} - R^2_{y.1})/(k_1 - k_2)}{(1 - R^2_{y.123})/N - k_1 - 1)} \tag{33.6}$$

donde k_1 es igual al número de variables independientes asociadas con $R^2_{y.123}$, la R^2 mayor, y es igual al número de variables independientes asociadas con $R^2_{y.1}$, la R^2 menor. Así, sustituyendo los valores se obtiene:

$$F = \frac{(.8612 - .7502)/(3 - 1)}{(1 - .8612)/(15 - 3 - 1)} = \frac{.0555}{.0126} = 4.405$$

que, con 2 y 11 grados de libertad, es significativa al nivel .05. (Note que un análisis de varianza ordinario de tres grupos, sin tomar en cuenta la covariable, produce una razón F no significativa.) $R^2_{y.23}$, o la varianza de Y explicada por la regresión a partir de las variables dos y tres (los tratamientos experimentales), después de permitir la correlación de la variable

TABLA 33.6 *Análisis de covarianza de los datos ficticios de la tabla 33.5, ordenados para el análisis de regresión múltiple*[a]

	Y	X_1	X_2	X_3
A_1	12	12	1	0
	12	11	1	0
	11	10	1	0
	10	12	1	0
	12	10	1	0
A_2	9	6	0	1
	9	9	0	1
	13	11	0	1
	14	14	0	1
	5	2	0	1
A_3	15	12	0	0
	12	10	0	0
	9	4	0	0
	8	4	0	0
	11	8	0	0

[a] Y = variable dependiente; X_1 = covariable; X_2 = tratamiento A_1; X_3 = tratamiento A_2.

1 y Y, es de .1110. Aunque no se trata de una relación fuerte, especialmente si se compara con la correlación masiva entre la covariable, el autoritarismo y Y ($r^2_{1y} = .75$), sí tiene consecuencias. Evidentemente, lo abstracto de los aspectos influye en las respuestas de acuerdo: cuanto más abstracto es el asunto, habrá mayor acuerdo. Es poco probable que el autoritarismo tenga una correlación con Y de .87. El ejemplo fue alterado deliberadamente para demostrar cómo una fuerte influencia, como la covariable X, puede controlarse y la influencia de las variables restantes (en este caso los tratamientos experimentales) puede evaluarse. Note que la fórmula 33.6 puede utilizarse en cualquier análisis de regresión múltiple; no está limitada al análisis de covarianza o a otros métodos experimentales.

Entonces, el análisis de covarianza se considera simplemente como una variante del tema del análisis de regresión múltiple, en cuyo caso es más fácil de conceptualizar que el procedimiento más bien elaborado del análisis de varianza —en especial si existe más de una covariable (véase Bruning y Kinte, 1987 o Li, 1957)—. La covariable o no es más que una variable independiente. Además, una variable considerada como covariable en un estudio puede ser considerada fácilmente como una variable independiente en otro estudio.

Análisis discriminante, correlación canónica, análisis multivariado de varianza y análisis de ruta

La correlación canónica y el análisis discriminante se dirigen a dos importantes preguntas de investigación: ¿Cuál es la relación entre dos conjuntos de datos con independientes y dependientes? ¿Cómo asignar mejor a los individuos a los grupos, con base en numerosas variables? El análisis de correlación canónica se refiere a la primera pregunta y el análisis discriminante a la segunda. Como se esperaría por el nombre, el análisis multivariado de varianza es la contraparte multivariada del análisis de varianza: se evalúa la influencia de k variables independientes experimentales sobre m variables dependientes. El análisis de ruta constituye más un apoyo gráfico y heurístico que un método multivariado. Como tal, es muy útil, especialmente como ayuda para aclarar y conceptualizar problemas multivariados.

Análisis discriminante

Una función *discriminante* es similar a una ecuación de regresión con una variable dependiente categórica. Sin embargo, cada una posee un propósito diferente. Esta variable dependiente por lo común se presenta en forma de pertenecia al grupo. No obstante, en la regresión múltiple la combinación lineal del predictor o variables independientes sirve para estimar la variable dependiente, la cual en regresión es una medida continua. La mayoría de los investigadores utilizan la regresión múltiple para estimar los valores de la variable dependiente con propósitos de selección. Es decir, si un valor predicho para un conjunto de valores de la variable independiente excede cierto límite, entonces se toma una decisión. El análisis discriminante está involucrado con la clasificación y no necesariamente con la selección. Dado un perfil de puntuaciones en la variable independiente, el análisis discriminante ayuda al investigador a determinar a qué grupo pertenece ese individuo. Algunos investigadores naturalistas han aplicado el método para ayudar a clasificar hallazgos antropológicos de huesos o animales. De la misma manera que en la regresión múltiple, se asume que las variables independientes son continuas, pero que la variable dependiente es categórica. En las situaciones más elementales, la variable dependiente discreta o categórica tiene sólo dos categorías. El problema que la función

discriminante intenta resolver es encontrar un conjunto de coeficientes o pesos, U_i, para las variables independientes (también llamadas variables discriminantes). Se buscan tales pesos para probar si una combinación particular de las variables independientes se asemeja a los miembros de la *categoría 1* o si se asemeja más a los de la *categoría 2*. El objetivo principal consiste en pesar y combinar linealmente las variables independientes, de tal manera que las categorías estén forzadas a ser estadísticamente lo más diferentes que sea posible.

El análisis discriminante responde dos preguntas principales: primero, indica si el conjunto de variables independientes sirve o no para distinguir entre los dos grupos o categorías. La segunda pregunta sólo es importante si la respuesta a la primera pregunta es "sí". La segunda se refiere a la clasificación; indica a qué grupo o categoría debe pertenecer un solo individuo. En otras palabras, la función discriminante separa a los miembros del grupo al máximo. Indica a qué grupo probablemente pertenece cada miembro. Además, también puede hacer una prueba para determinar cuál de las variables independientes explica la diferencia entre los grupos. En síntesis, si se tienen dos o más variables independientes y a los miembros de, por ejemplo, dos grupos, la función discriminante ofrece la "mejor" predicción, en el sentido de los mínimos cuadrados, de la pertenencia "correcta" a un grupo de cada miembro de la muestra.

Algunos investigadores han establecido que el análisis discriminante de dos grupos es igual a la regresión múltiple, con excepción de que la variable dependiente, Y, es dicotómica en lugar de continua. Otros han llegado a aseverar que puede utilizarse cualquier codificación binaria de la variable dependiente (codificación dummy). Sin embargo, ello no es exactamente verdadero. Lindeman, Merenda y Gold (1980) demostraron que los pesos de regresión de la regresión múltiple, b_i, son proporcionales a los pesos de la función discriminante, u_i, si la variable dependiente se codifica como $n_2/(n_1 + n_2)$ para los miembros del grupo 1 y $-n_1/(n_1 + n_2)$ para los miembros del grupo 2.

El análisis discriminante lineal, tal y como lo formuló Fisher (1936), constituye un método apropiado cuando la variable dependiente es categórica. Las variables independientes o predictoras deben medirse en una escala de intervalo. Para probar si existe una diferencia estadísticamente significativa entre grupos, las variables independientes deben distribuirse de manera normal, con varianzas y covarianza iguales. Para utilizar el análisis discriminante de forma adecuada con fines de clasificación, se formulan otras suposiciones acerca de los datos. Un supuesto es que se debe considerar que el perfil de cada individuo tiene la misma posibilidad de estar en cada grupo o categoría. También se debe suponer que el costo de una mala clasificación para cada individuo es el mismo. Estos supuestos, necesarios para la función discriminante, no siempre se cumplen. Como resultado de esto, en años recientes muchos investigadores se alejaron del análisis discriminante a favor de la regresión logística.

Correlación canónica

No es un paso conceptual demasiado grande aquel que va desde el análisis de regresión múltiple con una variable dependiente, hasta el análisis de regresión múltiple con más de una variable dependiente. No obstante, a nivel de cálculos sí es un paso considerable; por lo tanto, no se proporcionarán los cálculos. El análisis de regresión de datos con k variables independientes y m variables dependientes se llama *análisis de correlación canónica*. Se trata de un método que fue desarrollado por Hotelling (1935, 1936). La idea principal es que se forman dos compuestos lineales, por medio de un análisis de mínimos cuadrados: uno para las variables independientes X_i, y otro para las variables dependientes Y_j. La correlación entre estos dos compuestos es la correlación canónica; y, como R, será la máxima correla-

ción posible, dados los conjuntos particulares de datos, debe quedar claro que lo que hasta ahora se ha denominado análisis de regresión múltiple es un caso especial del análisis canónico. En vista de las limitaciones prácticas que tiene el análisis canónico, sería mejor afirmar que el análisis canónico es una generalización del análisis de regresión múltiple.

La correlación canónica puede tener uno o más de los siguientes objetivos:

1. Probar si dos conjuntos de variables están correlacionados o no.
2. Hallar dos conjuntos de pesos o coeficientes, de tal manera que la correlación entre los dos conjuntos esté en su máximo.
3. Buscar en cada conjunto las variables que hagan la mayor contribución a la correlación entre los conjuntos.
4. Predecir los valores en un conjunto de variables, usando valores incluidos en el otro conjunto.

De éstos, quizás el tercer punto sea el más interesante y útil. Por ejemplo, se podría desear determinar cuáles variables basadas en el rendimiento tendrían la mayor relación con un conjunto de medidas de desempeño. La correlación canónica es capaz de brindar dicha información. Después de todo, si se tiene un conjunto de variables basadas en el rendimiento, tales como una batería de pruebas de rendimiento, no todas las pruebas son iguales. Además, no se puede esperar, de manera razonable, que todas realicen la misma contribución. Por lo tanto, resulta lógico considerar a la correlación canónica como un método de correlación de pasos sucesivos que selecciona dos variables, una de cada conjunto de variables, que tengan la relación más fuerte sobre cualquier otro par, después de lo cual, continuará buscando el siguiente mejor par.

En lo que se refiere a los supuestos que deben formularse al aplicar la correlación canónica a los datos, éstos no son tan importantes si no se hacen inferencias acerca del estadístico canónico. Si se utiliza tan sólo con propósitos descriptivos, no se debe asumir que los datos provienen de una distribución multinormal o que provienen de una población con varianzas y covarianza comunes. Sin embargo, si se van a hacer inferencias, como en una prueba de significancia estadística, entonces deben cumplirse tales supuestos. Además, tanto las variables independientes como las dependientes deben medirse con una escala de intervalo, o un conjunto se mide con escala de intervalo y el otro en una escala dicótoma.

De forma similar que en la regresión múltiple y en el análisis discriminante, el objetivo de la correlación canónica aquí es encontrar pesos o coeficientes. La diferencia radica en que hay dos conjuntos en lugar de uno: un conjunto para las variables independientes (también llamadas *predictoras*) y otro conjunto para las variables dependientes (también llamadas *criterios*). Los pesos para ambos conjuntos de variables se encuentran para maximizar la correlación entre los dos conjuntos. Así que, a diferencia de la regresión múltiple y el análisis discriminante, la correlación canónica es capaz de producir más de un conjunto de pesos para las variables independientes y dependientes. Sin embargo, el primer conjunto de pesos sería el que explica la mayor cantidad de varianza. Cada combinación lineal de variables (hay una para cada conjunto de variables) con frecuencia se llama *variante canónico* (véase Lindeman, Merenda y Gold, 1980).

Ejemplo de investigación

Bedini, Williams y Thompson (1995) utilizaron la correlación canónica para estudiar la relación entre el desgaste en el empleo y el estrés del rol terapéutico. Dicho estudio trató únicamente con especialistas en recreación terapéutica. Las medidas de desgaste: *agotamiento emocional*, *despersonalización* y *realización personal*, se utilizaron como variables dependientes; mientras que las medidas de estrés del rol: ambigüedad del rol y conflicto del rol se

consideraron como variables independientes. Los investigadores encontraron una función que explicó la relación entre los dos conjuntos de variables. Esta función explicó casi el 36 por ciento de la varianza explicada entre los dos conjuntos. Análisis adicionales determinaron que aproximadamente el 53 por ciento de la varianza de la extenuación se explica por las variables del estrés del rol. Los resultados sugieren que la gente que experimenta estrés del rol tiene mayor probabilidad de sufrir desgaste: experimentar agotamiento emocional, despersonalización y un sentido de baja realización personal.

Análisis multivariado de varianza

Como puede sospecharse, el análisis de varianza posee su contraparte multivariada, el *análisis multivariado de varianza*, el cual permite a los investigadores evaluar los efectos de k variables independientes sobre m variables dependientes. Al igual que su compañero univariado, que ya se examinó con cierto detalle anteriormente, debe utilizarse con datos experimentales. El análisis multivariado de varianza, o MANOVA, es un método íntimamente relacionado con el análisis discriminante con grupos múltiples. La similitud se presenta sólo en su estructura y no necesariamente respecto a dónde debe utilizarse y cuáles sean los supuestos. Tal como la versión univariada del análisis de varianza presentada en un capítulo previo, los diseños utilizados de forma univariada (una variable dependiente), pueden utilizarse con múltiples variables dependientes. En otras palabras, cada participante del estudio se mide más de una vez, de tal manera que la persona tiene, por lo menos, dos medidas dependientes. En algunos casos, pueden ser dos o más variables dependientes diferentes. En otros casos, puede ser la misma variable medida en diferentes momentos, lo cual a menudo se denomina *análisis de varianza de medidas repetidas*. Algunos investigadores han llamado y analizado inapropiadamente los datos de su estudio como un ANOVA de medidas repetidas, cuando debían haberlo realizado utilizando un MANOVA. La razón de ello es que se debe cumplir el requisito de que el componente de error de las puntuaciones sea independiente. Éste es el supuesto de homogeneidad de la varianza, que en muchas ocasiones es difícil de cumplir, especialmente si las variables dependientes no son verdaderas medidas repetidas. Existen pruebas estadísticas disponibles para demostrar dicho supuesto (véase Kirk, 1995).

Sin embargo, el MANOVA posee unos cuantos supuestos propios que podrían cuestionarse. Las varianzas dentro de grupos, medidas para las variables dependientes para cada uno de los grupos en el análisis, deben ser iguales. También, sería necesario suponer que las variables dependientes se distribuyen de forma multivariada normal. Las pruebas sobre la normalidad multivariada no están lo suficientemente avanzadas. La determinación casi siempre se realiza con series de pruebas poco sistemáticas.

El análisis multivariado de varianza funciona mejor cuando se cumplen los supuestos y también cuando existe una alta correlación entre las variables dependientes. Si la correlación entre las variables dependientes es baja o cercana a cero, el investigador no logrará nada utilizando el MANOVA. En este caso, es posible calcular ANOVAS separados con cada variable dependiente utilizada como una sola medida de resultado. Si éste fuera el caso, el investigador necesitará ajustar el nivel del error tipo I para compensar posibles errores de este tipo. En el otro extremo, si la correlación entre las variables dependientes es 1.00 o cercana a este valor, entonces se sabe que las dos están midiendo esencialmente lo mismo y son redundantes. Siendo así, sólo se necesita calcular un ANOVA para una de esas variables dependientes.

Se evitarán mayores explicaciones aquí, sólo se dirá que, como en todos o en la mayoría de los análisis multivariados, los resultados del análisis multivariado de varianza algunas veces son difíciles de interpretar. Esto se debe a las dificultades mencionadas antes

para evaluar la importancia relativa de las variables en esta influencia sobre una variable dependiente que, como en el análisis de regresión múltiple, con frecuencia están compuestos en el análisis multivariado de varianza, en la correlación canónica y en el análisis discriminante. Si un efecto de interacción resulta estadísticamente significativo, el proceso para determinar qué variables están involucradas en el efecto de interacción puede ser muy laborioso. Bray y Maxwell (1982) y Pedhazur (1996) ofrecen muy buenas explicaciones sobre el análisis multivariado de varianza. Note también que si existen covariables involucradas, entonces se trata de un MANCOVA.

El estudio de Nemeroff (1995) sobre enfermedad y percepción del contagio utilizó un diseño donde los datos recopilados pueden analizarse por medio de un MANOVA. Tal estudio se llevó a cabo para examinar la forma en que la gente reacciona ante los individuos que padecen una enfermedad contagiosa. A los participantes se les dieron crayolas y cuatro hojas en blanco. Se les pidió que dibujaran el germen de la gripe para diferentes personas específicas: para sí mismos, un amigo, un extraño y una persona que les disgustara. Los dibujos fueron calificados en diversas dimensiones por jueces entrenados. Dichas puntuaciones de evaluación sirvieron como variables dependientes. Las dimensiones *activo, grande* y *complejo* se combinaron en una sola medida: *intensidad*. La dimensión *activo* se calificó con base en qué tan activo o pasivo parecía el germen. La *complejidad* se refería a la cantidad de detalle que el participante ponía a los dibujos. Las tres variables individuales restantes eran *abstracción, alcance* y *felicidad*. La *abstracción* se refería a qué tan personificada era la apariencia del dibujo. El *alcance* se refería a qué tan contenido aparecía el germen, y la *felicidad* medía la percepción de los jueces respecto a qué tan agradable o feliz era la apariencia del germen. La variable independiente en tal estudio era la fuente de los individuos, por ejemplo, el novio, el propio sujeto, un extraño, etcétera.

Por medio del uso del MANOVA, Numeroff encontró que la gente percibe el germen de la gripe de forma diferente cuando proviene de una fuente distinta de contagio. Por ejemplo, los gérmenes propios difirieron de los gérmenes de los extraños en su intensidad. Los gérmenes del novio(a) se percibían como menos enojados, por el color, que los gérmenes de personas no agradables, los cuales resultaron más amenazantes. Se encontró que el germen menos amenazante era el del novio(a).

Análisis de ruta

El desarrollo del análisis de ruta se acredita a Wright (1921). El objetivo era desarrollar un modelo causal para la genética y la biología utilizando correlaciones. Como ya se explicó antes, las correlaciones no implican causalidad. No obstante, Wright fue capaz de utilizarlas de ese modo, ya que sus estudios fueron realizados bajo controles muy estrictos. Wright distinguió entre efectos directos e indirectos utilizando correlaciones y regresión. Una variable X puede tener un efecto directo sobre la variable Y, pero un efecto indirecto sobre la variable Z. Tal efecto se establece al examinar los pesos estandarizados de la regresión (correlaciones) entre X y Y, X y Z y Y y Z. Si las correlaciones entre 'X y Y' y 'Y y Z' son altas, pero la correlación entre X y Z es mínima, entonces se tiene un efecto directo entre X y Y y entre Y y Z, y uno indirecto entre X y Z. La contribución de Wright también incluyó una forma específica del uso de las reglas de trazo en los diagramas de ruta para realizar los cálculos necesarios.

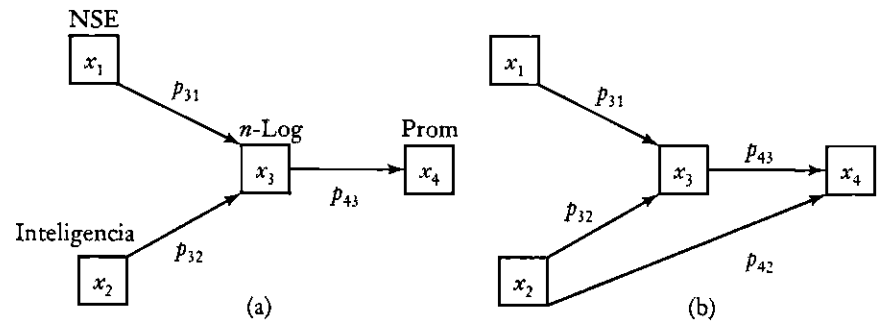
Un redescubrimiento del trabajo de Wright ocurrió en sociología entre la mitad de los años sesenta y el inicio de los setenta. Entonces dichos métodos se volvieron populares con los investigadores en psicología y educación en los años setenta y ochenta. Bentler (1986) ofrece una buena perspectiva histórica de esta transición. Sin embargo, los datos de las ciencias del comportamiento y sociales no son muy similares a los datos que Wright

recolectó y utilizó. Por ende, llamarlo un modelo "causal" es engañoso. Como ya se mencionó, Wright tenía controles muy estrictos para las variables genéticas y de crianza; pero el nivel de control en los estudios en ciencias sociales y del comportamiento resulta mucho menor. Blalock (1972) menciona los requisitos para realizar un análisis de ruta, donde los resultados sean útiles. En la actualidad el análisis de ruta aún funciona como una herramienta de investigación útil para el desarrollo de un modelo conceptual que pueda ser probado de manera empírica. Aunque el término "modelo causal" persiste, en realidad no es causal. Esto será así también cuando se considere el último capítulo sobre el modelamiento de la ecuación estructural. Un libro que ofrece una buena cobertura sobre el análisis de ruta es el de Loehlin (1998).

El análisis de ruta es una forma del análisis de regresión múltiple aplicado que utiliza diagramas de ruta para guiar la conceptualización del problema o para probar hipótesis complejas. Con su uso se calculan las influencias directas e indirectas de las variables independientes sobre la variable dependiente. Dichas influencias se reflejan en los llamados coeficientes de ruta, que en realidad son coeficientes de regresión (beta, b o b). Más aún, es posible probar la congruencia de diferentes modelos de ruta, con los datos observados (véase Pedhazur, 1996). A pesar de que el análisis de ruta ha sido y es un importante método analítico y heurístico, es dudoso que continúe utilizándose como ayuda para probar la congruencia que tienen los modelos con los datos obtenidos. Más bien, su valor será el de un método heurístico que ayude a la conceptualización y a la formación de hipótesis complejas. Sin embargo, la comprobación de dichas hipótesis quizá se realizará con herramientas analíticas más poderosas y más apropiadas para tales comprobaciones. El método que se cubre en el capítulo 35 es, en la actualidad, el mejor método para utilizarse en el análisis y comprobación de hipótesis que surjan a partir de modelos de análisis de ruta. Ahora se estudiará un ejemplo para dar una idea general sobre el método.

Considere los dos modelos, a y b, de la figura 33.1. Suponga que se está tratando de "explicar" el rendimiento o promedio, x_4, en la figura, o Prom. Una persona considera que el modelo a es "correcto"; sin embargo, una segunda persona considera que el modelo b es "correcto". En efecto, el modelo a dice que tanto el nivel socioeconómico NSE, como la inteligencia influyen en x_3, que representa la necesidad de logro (n-Log), y que x_3 influye en x_4, Prom o rendimiento. ¡Qué bien! En otras palabras, la primera persona considera que el modelo a expresa mejor las relaciones entre las cuatro variables. Por otro lado, la segunda persona considera que el modelo b es una mejor representación. Éste añade una influencia directa de x_2, inteligencia, sobre x_4, rendimiento (note las rutas de x_2 a x_4 y de x_2

FIGURA 33.1 x_1 = nivel socioeconómico (NSE); x_2 = inteligencia; x_3 = n-Log o necesidad de logro; x_4 = Prom o promedio de calificaciones (rendimiento)



a x_3 y a x_4). ¿Qué modelo es "correcto"? En el análisis de ruta se prueban los dos modelos utilizando el método del capítulo 35.

Regresión de cresta, regresión logística y análisis logarítmico lineal

Regresión de cresta

La regresión de cresta constituye un método reconocido por los estadísticos aplicados y por los investigadores de las ciencias de ingeniería. Sin embargo, no ha sido popular en la investigación de la ciencia psicológica o del comportamiento. Los creadores del método, Arthur Hoerl y Robert Kennard, publicaron su artículo monumental en 1970. A pesar de la actitud de la psicología hacia su método, el impacto del artículo de Hoerl y Kennard ha sido tan grande que el Instituto para la Información Científica (Institute for Scientific Information) lo denominó como una "cita clásica" (Hoerl, 1995).

La psicología casi siempre asocia el artículo de Price (1977) como la introducción de la psicología a la regresión de cresta. No obstante, el manuscrito bien escrito e informativo de Simon (1975) sobre el tema precedió al de Price por dos años. Además, Bolding y Houston (1974) desarrollaron un programa computacional para realizar la regresión de cresta. Simon (1975) demostró cómo podía usarse el método en estudios de factor humano, donde no se cumplían todos los requisitos de un experimento verdadero. Dichos estudios incluían una o más variables predictoras que estaban *altamente* correlacionadas. Las variables correlacionadas se debían al fracaso para completar un experimento verdadero o a la falta de habilidad del investigador para seleccionar o controlar condiciones experimentales relevantes. Simon llama a este tipo de estudios "no diseñados". Keith (1988) y Price (1977) los denominan "investigación no experimental". Hoerl y Kennard (1970) los consideran "estudios no ortogonales".

La desaprobación de la regresión de cresta por parte de la investigación psicológica quizá se deba en parte a las críticas de Rosenboom (1979) acerca del método. La mayoría de los métodos bayesianos o las técnicas de estimación del sesgo requieren de la intervención y juicio humanos, en lugar de un método estrictamente matemático y analítico, tal como el de los mínimos cuadrados. La regresión de cresta es uno de dichos métodos, y como tal se ha considerado deshonesto. Aun los autores de reconocidos libros sobre estadística, como Draper y Smith (1981) afirman que el método fue muy polémico en los años setenta. Sin embargo, como se estableció en un artículo de Frank y Friedman (1993), la regresión de cresta constituye claramente el mejor método de análisis de regresión en muchas condiciones no experimentales. Diversos autores coinciden con Keith (1988) en que la regresión múltiple es el método de elección cuando se trata de estudios no experimentales. No obstante, la regresión múltiple se debilita con rapidez cuando las variables predictoras están altamente correlacionadas o son colineales. Esto se debe al hecho de que la regresión múltiple, tal como se maneja actualmente en la mayoría de los programas estadísticos, utiliza el método de los mínimos cuadrados. El lector debe notar que el segundo autor de este libro está tomando una posición más bien extrema al respecto. En estudios de investigación donde las variables predictoras están ligera o moderadamente correlacionadas (alrededor de .50 o menos), el uso de un método como la regresión de cresta podría no ser necesario. De hecho, Keith (1999) ha señalado que la regresión múltiple produce estimados bastante estables de los coeficientes de regresión cuando el nivel de colinealidad es moderado.

El problema con los mínimos cuadrados ordinarios (MCO)

Uno de los propósitos del análisis de regresión múltiple es obtener un conjunto de coeficientes o pesos no sesgados, que tengan una mínima cantidad de error de la variable, así como un ajuste razonable con un conjunto existente de datos. Un método popular para hacerlo es el de los mínimos cuadrados ordinarios (MCO), el cual se trata en todo libro de texto de estadística elemental. Aquí lo abordamos en un capítulo anterior. Este método es directo, determinado matemáticamente y no requiere del juicio humano. Además implica la estimación de los coeficientes de regresión, con la limitante de que la suma de cuadrados de la diferencia entre la medida resultante predicha y observada sea la mínima:

$$\Sigma(Y'_i - Y_i)^2 = \text{mínimo para la ecuación } Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_n x_n + e$$

En esta ecuación las X son las variables predictoras y la Y es la variable dependiente o criterio. Cuando las variables predictoras son matemáticamente independientes (por ejemplo, las correlaciones entre las X son iguales a *cero*), los coeficientes de regresión estimados son representaciones razonables de los verdaderos coeficientes de regresión, dentro de los límites de las fluctuaciones de muestreo. Cuando las variables predictoras están altamente correlacionadas, los coeficientes individuales de regresión calculados con el método de MCO con frecuencia resultan insatisfactorios. La matriz de variables predictoras altamente correlacionadas se llama *mal condicionada*. Las cualidades predictivas de dicha ecuación generada por medio de los mínimos cuadrados son razonablemente precisas para los datos utilizados para generar la ecuación. No obstante, la aplicación de la misma ecuación de regresión a un nuevo conjunto de datos da como resultado valores predichos pobres para la medida resultante. Es decir, la validación cruzada de la ecuación de regresión es extremadamente pobre. Además, los efectos relativos de los coeficientes individuales de regresión no pueden evaluarse. En otras palabras, los coeficientes de regresión obtenidos por medio de los mínimos cuadrados en variables predictoras correlacionadas quizá no tengan sentido cuando se evalúan en términos del mundo real. Hoerl y Kennard (1970), retomados por Simon (1975), afirmaron que una o más de las siguientes características pueden presentarse en un ajuste de mínimos cuadrados de variables predictoras correlacionadas:

1. los coeficientes de regresión se vuelven demasiado grandes en su valor absoluto,
2. algunos coeficientes llegan a tener el signo equivocado,
3. los coeficientes son inestables; otro conjunto de datos para las mismas variables producirá valores resultantes diferentes, y
4. los pesos individuales de regresión sobrestiman o subestiman el efecto de una variable en particular.

Dos variables altamente correlacionadas resultarán en coeficientes donde una variable recibe un peso grande y la otra recibe un peso pequeño e insignificante. Una explicación más completa sobre los peligros de los MCO se encuentra en Newman (1976).

Hoerl y Kennard (1970) desarrollaron una alternativa al método de regresión múltiple convencional con variables predictivas correlacionadas. El método se creó para permitir a los investigadores evaluar variables en problemas de ingeniería química, donde sería impráctico abandonar variables o crear variables compuestas. Este método, llamado *regresión de cresta* produce una mejor ecuación de predicción que la que se obtendría utilizando los mínimos cuadrados y es mejor debido a que los coeficientes estimados se acercan más a los coeficientes verdaderos, en promedio. Los signos de los coeficientes son más precisos; los coeficientes son más estables, con una mayor probabilidad de repetirse con un

nuevo conjunto de datos, y la medida estimada resultante puede lograrse con un error de cuadrados medios más pequeño. A través de la regresión de cresta los valores eigen se vuelven menos discrepantes.

En esencia, el análisis de regresión de cresta es idéntico a la regresión de MCO, con excepción de que un número pequeño, k, se ha añadido a la diagonal de la matriz de correlación de las variables predictoras. La añadidura de dicho número k a la diagonal hace a la matriz menos mal condicionada. También tiene el efecto de disminuir el error de los cuadrados medios cuando se comparan con los mínimos cuadrados. El mejor valor k es aquel donde los coeficientes de regresión se estabilizan y donde las sumas de cuadrados residuales son bajas. Para encontrar este valor el investigador debe probar diversos valores de k. Un número de investigadores objetaron dicho método *ad hoc* para seleccionar k. Como Simon (1975) demostró, el valor k puede agregarse directamente a la diagonal de la matriz de correlación y, después, someterse a un programa computacional de regresión, o el investigador puede añadir casos dummy a los datos originales (datos aumentados) y seleccionar la opción de intersección cero del programa computacional. El BMDP (Dixon, 1990) tiene un subprograma en su paquete estadístico para realizar la regresión de cresta. Lee (1980) demuestra cómo se puede efectuar lo anterior con programas de regresión que no tienen una opción de intersección cero. Se han realizado diversos estudios para desarrollar una forma más analítica para determinar k. La explicación de dichos estudios ocuparía demasiado espacio en la presente obra. Se pide al lector que consulte el artículo de Simon o el de Draper y Smith (1981).

El precio que se paga por utilizar la regresión de cresta es que los coeficientes de regresión ya no están sin sesgo y que la suma de cuadrados residual (SCR) ya no es mínima. Sin embargo, los beneficios de una regresión de cresta ejecutada apropiadamente llegan a contrarrestar tales desventajas. No obstante, como todos los métodos estadísticos, la regresión de cresta debe emplearse de forma correcta. Draper y Smith (1981, p. 322) lo exponen de manera más directa con sus palabras:

> La selección de la regresión de cresta no constituye una panacea milagrosa; se trata de una solución de mínimos cuadrados restringida por la suma de alguna información externa acerca de los parámetros. El uso a ciegas de la regresión de cresta sin esta consideración resulta peligroso y engañoso. Si la información externa es sensata y no se contradice con los datos, entonces la regresión de cresta también es sensata.

Ejemplo de investigación

Bee y Beronja (1986) utilizaron la regresión de cresta en un estudio de estudiantes universitarios con un área de estudios principal indeterminada. Los investigadores reunieron los resultados de una prueba de ingreso a la universidad (ACT) junto con el desempeño académico universitario y medidas de personalidad, como motivación y hábitos de trabajo académico. La meta consistía en desarrollar una ecuación de regresión que pudiera predecir el desempeño académico universitario (promedio de calificaciones) utilizando variables de personalidad, variables de experiencia del programa (por ejemplo, nivel de dificultad de los cursos en el área de estudio principal) y las puntuaciones de una prueba de ingreso a la universidad (ACT). Tales variables explicativas o independientes eran colineales. Cuando Bee y Beronja ajustaron los datos por medio del método de regresión ordinaria, encontraron que ninguna de las variables explicativas estaba en relación significativamente con el desempeño académico. No obstante, los estimados de la regresión de cresta proporcionaron resultados muy diferentes. La regresión de cresta encontró que las variables ACT de matemáticas, los hábitos de trabajo académico, la motivación para el éxito y la dificultad de los cursos de matemáticas estaban relacionadas de manera significativa con el desempeño

académico. Bee y Beronja encontraron que $k = .4$ producía los mejores resultados en la regresión de cresta. Ninguno de los pesos de regresión encontrados madiante los mínimos cuadrados ordinarios fue estadísticamente significativo ($p > .05$). Sin embargo, cuatro de los pesos de regresión que fueron determinados por medio de la regresión de cresta resultaron significativos.

Regresión logística

Ya se explicó anteriormente en el libro la regresión múltiple y el análisis de función discriminante. Por lo general, si se tiene una variable dependiente categórica, la recomendación era realizar un análisis de función discriminante, el cual, no obstante, sólo es efectivo si las variables cumplen ciertos supuestos. En algunos estudios de las ciencias sociales y del comportamiento, las variables independientes o predictoras son categóricas o nominales. Cuando esto sucede, el análisis de función discriminante empieza a perder su eficacia en términos de su buen ajuste con los datos. Si se utiliza la regresión múltiple, la ecuación con mejor ajuste quizá sea inaccesible y puede tratarse de una ecuación que no produzca información útil. Después de todo, la regresión múltiple tradicional supone que los datos se miden en una escala de intervalo (o algo cercano a ello) y que sigue una distribución normal.

Un método que va ganando gran popularidad en años recientes es la regresión logística. Su desarrollo parece muy novedoso a los investigadores en el campo de la psicología, la sociología y la educación. La vida y las ciencias médicas lo han utilizado por un periodo mucho más largo. El retraso de las ciencias sociales en el uso de este método es irónico. Durante los años sesenta, los psicólogos y los investigadores sociales del Instituto de Investigación Social (Institute for Social Research, ISR) de la Universidad de Michigan, habían desarrollado los métodos denominados *análisis de clasificación múltiple* (ACM) y el *análisis multivariado de escala nominal* (AMN), ahora conocido como regresión logística. En realidad, la psicología tuvo un temprano encuentro con el método, el cual permaneció latente durante años, con excepción de aquellos afiliados o que conocían el Instituto de Investigación Social.

Andrews, Morgan, Sonquist y Klem (1973) describen una técnica para examinar las relaciones entre variables independientes y una variable dependiente, que es similar a la regresión múltiple. Ellos señalan el problema implicado cuando las variables independientes o predictoras se miden en una escala nominal y ofrecen una solución. El análisis de clasificación múltiple (ACM), como ellos lo llaman, incluye datos tanto nominales como no nominales en las variables predictoras. En la variable dependiente o criterio, los datos se miden con una escala de intervalo o dicotómica. El *análisis multivariado de escala nominal* (Andrews y Messenger, 1973) es una extensión del ACM. Permite el uso de variables dependientes de escala nominal con más de dos categorías. De acuerdo con la nomenclatura actual, el ACM se denomina *regresión logística;* y el AMN se conoce como *regresión logística policotómica* (véase Dixon, 1990). Se utilizarán los términos más populares en la explicación de este método.

La *regresión logística*, por lo tanto, constituye una técnica para ajustar una superficie de regresión con los datos, en la cual la variable dependiente es dicotómica. En psicología educativa es posible clasificar a los estudiantes con un funcionamiento mental *alto* o *bajo*; o, en referencia a la terapia, se puede clasificar a los pacientes como *con mejoría* y *sin mejoría*, *exitosa* o *no exitosa*. En cada estudio que utiliza una variable dependiente dicotómica, la regresión logística constituye un candidato viable como método de análisis. Sin embargo, se podría plantear la pregunta: "¿Cuál método se debe utilizar: el análisis discriminante o la regresión logística?" Existe controversia respecto a la comparación del análisis discrimi-

nante con la regresión logística. Press y Wilson (1978) reportaron aquellas situaciones donde el análisis discriminante funciona bastante bien; es decir, cuando los datos cumplen los supuestos. No obstante, cuando no se cumplen los supuestos, la regresión logística representa una forma superior de análisis. La regresión logística requiere que se cumplan menos supuestos; pero no constituye una panacea para los datos recolectados con diseños de investigación cuestionables. El análisis discriminante produce con facilidad una probabilidad de éxito que cae fuera del rango del 0 al 1, lo cual no es aceptable. Por otro lado, la regresión logística no produce probabilidades más allá de entre 0 y 1. Ambos producen estimados de regresión y ambos son capaces de clasificar individuos. En la regresión logística el investigador obtiene una ganancia adicional: el coeficiente o peso de regresión puede transformarse en *razones de probabilidad*, un estadístico útil que se describió en el capítulo 10. Es útil al ofrecer al investigador ideas respecto a lo que está sucediendo dentro de los datos.

Con una comparación directa entre el análisis discriminante y la regresión logística, esta última funciona mejor cuando las variables no son normales. Además, la regresión logística no se ve tan afectada, como el análisis discriminante, cuando se incluyen variables sin significado dentro del análisis. Lo anterior implica variables dicotómicas o variables que se han sometido a codificación dummy. El uso de variables dicotómicas es muy común en la investigación de las ciencias del comportamiento. Así, si los propios datos de investigación contienen variables categóricas o de escala nominal, o si existe alguna duda razonable respecto a alguna de las variables, quizá sea mejor utilizar la regresión logística en lugar del análisis discriminante.

Podría realizarse una comparación similar entre la regresión logística y la regresión múltiple ordinaria utilizando variables independientes y dependientes dicotómicas. Con la regresión múltiple los valores predichos caerían fuera del rango 0-1. Además, si las varianzas calculadas de la variable dependiente fluctuaran con valores de las variables dependientes, como tener más números 1 que 0 para un nivel e igual número de 0 y 1 en otro, el análisis producirá una varianza grande. Como resultado, se violarán los supuestos de homogeneidad de varianza y de normalidad. Sin embargo, existen situaciones donde la regresión múltiple con una variable dependiente dicotómica daría buenos resultados (véase Cox y Wermuth, 1992).

Un ejemplo de investigación

Los datos de este ejemplo son cortesía de Dorothy Scattone de la Universidad del Sur de Mississippi. Su estudio trató de las percepciones de dos grupos diferentes de asiáticos, hacia las discapacidades físicas y mentales. Un grupo de asiáticos había nacido y crecido en un país asiático; mientras el otro grupo se componía de asiáticos nacidos en Estados Unidos. Los participantes fueron 215 estudiantes universitarios que respondieron un número de preguntas respecto a ciertas discapacidades como el síndrome de Down o su nivel de aceptación hacia personas con asma o cicatrices faciales, etcétera. Las variables se midieron en una escala de calificación de 5 puntos, donde el 5 significaba una alta aceptación y el 1 una baja aceptación.

Los resultados de este análisis indican cuáles variables discriminaron entre los asiáticos nacidos en Estados Unidos y los nacidos en el extranjero; además señalan las respectivas probabilidades. Respecto a la variable *tartamudeo*, se les pidió a los participantes que indicaran su nivel de aceptación de una persona con problemas de habla, específicamente con tartamudeo, donde el 1 implicaba la no aceptación y el 5 una aceptación total. Con la aceptación total, el sujeto afirma que está de acuerdo con tener a la persona como un miembro de la familia a través del matrimonio. Puesto que esta variable fue significativa, indica que los asiáticos nacidos en Estados Unidos y los nacidos en el extranjero difieren

en su respuesta. La razón de probabilidad para las variables es una forma diferente de hablar sobre las probabilidades. La ventaja que la razón de probabilidad tiene sobre la prueba de significancia consiste en que la primera casi no se ve afectada por el tamaño de la muestra. Para la variable *tartamudeo* la razón de probabilidad fue .4516, lo cual indica que los asiáticos nacidos en el extranjero tienen .4516 veces menos probabilidades de aceptar a las personas con tartamudeo, que los asiáticos nacidos en Estados Unidos o, en otras palabras, los asiáticos nacidos en Estados Unidos tienen 2.2 veces más probabilidades de aceptar a una persona con dicha incapacidad del habla, que los asiáticos nacidos en el extranjero (1/.4516).

Además de esta información, el análisis de regresión logística también proporciona una medida sobre qué tan precisa es la ecuación de regresión, en términos de clasificación. Con tales datos, la ecuación de regresión logística fue capaz de clasificar correctamente al 76.74 por ciento de los casos. La ecuación resultó, de forma general, más precisa al predecir a los asiáticos nacidos en el extranjero (92.02 por ciento) que a los asiáticos nacidos en Estados Unidos (28.85 por ciento). Existen otros estadísticos, como la prueba Wald, que se asocian con un resultado de regresión logística; aunque no se estudiarán aquí; en su lugar, se referirá al lector a libros como el de Hosmer y Lemeshow (1989) o el de Shoukri y Edge (1996).

Tablas de contingencia de múltiples factores y análisis log-lineal

Es adecuado presentar esta sección después de la regresión logística a causa de que el tema que se inicia a continuación trata de manera exclusiva con datos categóricos. En la regresión logística se tiene una variable dependiente dicotómica y variables independientes categóricas y de intervalo. En las tablas de contingencia de múltiples factores se trata sólo con datos categóricos. Los análisis de las tablas de contingencia de múltiples factores son importantes porque muchos de los datos utilizados por los científicos sociales y del comportamiento son categóricos. El empleo de los métodos tradicionales de análisis de varianza y regresión múltiple para analizar datos categóricos no funciona bien en muchos casos.

En el capítulo 10 se estudiaron las tablas de contingencia unidimensional y bidimensional. En aquel momento, se introdujo de forma breve el concepto de las tablas de múltiples factores. Tradicionalmente muchos investigadores estudiaban las tablas de contingencia de múltiples factores observando una serie de tablas de dos factores. Los cálculos son relativamente directos y el investigador puede, por lo general, llegar a alguna conclusión razonable respecto de los datos. También, tienen menor probabilidad de tener casillas con pocos casos o vacías. Uno de los eventos más probables en tablas grandes es la presencia de casillas con pocos casos o vacías, lo cual puede afectar los resultados del análisis. Como consecuencia, muchos investigadores reducen el número de categorías para eliminar este problema. Sin embargo, las series de tablas de contingencia de dos factores, utilizadas para analizar tablas de contingencia de factores múltiples, no permiten captar la existencia de efectos de interacción de orden superior entre las dimensiones. Glick, DeMorest y Hotze (1988) ofrecen un buen ejemplo de cómo analizar una tabla de contingencia de tres factores de forma correcta con una serie de análisis de dos factores. También fueron capaces, a través de una progresión de afirmaciones y análisis lógicos, de llegar a una conclusión acerca del término de interacción de tres factores. Posteriormente, en esta sección, se hará un nuevo análisis de sus datos a la luz de las tablas de contingencia de factores múltiples o del análisis logarítmico lineal. Además, las asociaciones entre las variables difieren más a través del análisis de dos factores que a través de factores múltiples, ya que este último

toma en consideración las otras variables implicadas. Además, el uso de tablas de sólo dos factores no permite la comparación simultánea de todas las asociaciones de pares.

Recuerde que en el análisis de datos categóricos del capítulo 10 se estableció la diferencia entre los valores observados y los valores esperados. Ambos representan frecuencias dentro de cada casilla de una tabla de contingencia. Si las frecuencias esperadas coinciden con las frecuencias observadas, entonces se diría que no hubo relación entre las dos variables categóricas, lo cual sucede así porque las frecuencias esperadas se calculan bajo condiciones de lo que se podría esperar si no hubiera relación entre las dos variables. Por lo tanto, si las frecuencias observadas coinciden con las frecuencias esperadas, es posible concluir que los datos reunidos no encontraron relación alguna. De forma subsecuente, si no hubo coincidencia, entonces se afirma que las dos variables están relacionadas. El análisis de las tablas de contingencia de factores múltiples opera de una forma muy similar. El investigador especifica un modelo que incluya a las variables, como sin interacciones entre tres factores o sólo interacciones específicas entre dos factores. Una vez especificado el modelo, se generan las frecuencias esperadas. Si las frecuencias observadas coinciden con las frecuencias esperadas, entonces se sabe que el modelo elegido se ajusta a los datos observados y los elementos del modelo explican los valores observados. En las tablas de factores múltiples, una de las metas consiste en encontrar las variables que se relacionan con otras variables. Hallar los valores esperados para probar si los valores observados coinciden es más demandante a nivel de los cálculos que en las tablas simples de dos factores. A pesar de que no se realizarán los cálculos, se puede dirigir al lector hacia referencias útiles que demuestran con claridad la operación. Uno de los algoritmos más populares para encontrar los valores esperados fue desarrollado por Deming y Stephan (1940). Una descripción de dicho método puede encontrarse en su artículo original. También se encuentran en la reimpresión de Dover de 1964 del libro de Deming de 1943 o en Dillon y Goldstein (1984), quienes ofrecen un ejemplo del cálculo claro y fácil de seguir. Algunas veces el método de Deming y Stephan se denomina *ajuste iterativo proporcional*. De acuerdo con su nombre, el ajuste iterativo proporcional requiere de un estimado inicial de las frecuencias esperadas y después, a través de un número de pasos, se ajustan. En la primera iteración se obtienen los estimados que van a utilizarse en la siguiente iteración. Tal como sucede en la regresión logística, las iteraciones cesan una vez que dos iteraciones sucesivas producen estimados muy cercanos entre sí.

Uno de los beneficios producidos por el método log-lineal es el de la parsimonia. Es decir, con el método logístico lineal el investigador especifica un modelo de términos al cual ajustarse, de manera muy similar a lo que se hace en el análisis de varianza o en la regresión múltiple. El investigador intenta obtener el mejor ajuste posible con el menor número de términos. Considere que m_{ij} represente la frecuencia esperada en la casilla (i,j) de una tabla de contingencia de dos factores. Nombre a una de las variables A y a la otra B. A tendría i categorías y B tendría j categorías. El modelo para esta tabla de contingencia se puede expresar como:

$$\log_e(m_{ij}) = \mu + \mu_{A(i)} + \mu_{B(j)} + \mu_{AB(i,j)}$$

Algunas ocasiones la ecuación se escribe sin los subíndices i y j, los cuales se utilizan para indicar el número de categorías en cada variable. Si la ecuación se escribe sin los subíndices, entonces las categorías están implícitas. Para simplificar, se escribirán las ecuaciones sin referencia directa al número de categorías en cada variable. Se podrá notar que tal ecuación se asemeja a aquella utilizada en el análisis de varianza. Sin embargo, entre las diferencias, el lector debe recordar que en la regresión múltiple o en el análisis de varianza la ecuación se desarrolla para predecir o explicar la variación de un individuo a otro. En otras palabras,

el análisis de varianza y la regresión múltiple hacen un estimado de la variable dependiente para cada caso o individuo. En el log-lineal o tablas de contingencia, la predicción se hace respecto a la frecuencia o categoría de la casilla y no respecto al individuo. A pesar de que ciertos escritores (Bakeman y Robinson, 1994; Fienberg, 1980; Howell, 1997; Kennedy, 1992) sobre tablas de contingencia de factores múltiples y sobre análisis log-lineal hacen analogías con la regresión múltiple y el análisis de varianza, ellos enfatizan esta importante diferencia.

En el análisis log-lineal, si se tuvieran tres variables categóricas, se escribiría como

$$\log_e(m_{ijl}) = \mu + \mu_A + \mu_B + \mu_C + \mu_{AB} + \mu_{AC} + \mu_{BC} + \mu_{ABC}$$

Note que en cada modelo existen términos principales y las interacciones. Cuando todas las posibles combinaciones de los términos están explicadas en la ecuación, el modelo se denomina *saturado*. El estadístico de bondad de ajuste es un cero perfecto que indica que el modelo se ajusta con los datos observados; es decir, los valores observados se ajustan con los valores esperados. Sin embargo, un investigador puede ajustar diferentes modelos con este método de análisis. La guía dictada por la teoría o por información previa sirve para eliminar algunos de los términos. Si se encuentra que los datos observados se ajustan con los valores esperados generados por el nuevo modelo, entonces se ha ajustado exitosamente un modelo más parsimonioso. Los modelos más parsimoniosos con frecuencia se denominan *modelos reducidos* o *modelos no saturados*. Lo que un modelo bien ajustado no saturado indica es que no se necesitan todos los términos del modelo saturado para obtener un ajuste adecuado de los valores observados con los valores esperados.

Los términos de interacción del modelo log-lineal son conocidos como *términos de orden superior*. A mayor número de términos en la interacción, más alto será el término. En el modelo de tres factores presentado anteriormente, el término μ_{ABC} es el término de orden superior; mientras que μ es el término de orden inferior. Esta breve especificación es importante para la explicación de la diferencia entre los modelos jerárquicos y los no jerárquicos. Algunos expertos en el análisis log-lineal para tablas de contingencia han establecido que el modelo jerárquico es el más útil y que los resultados de los modelos no jerárquicos son cuestionables (Bakeman y Robinson, 1994; Howell, 1997). La siguiente explicación se restringirá solamente a los modelos jerárquicos, en los cuales se observan los términos de orden superior como un compuesto de términos de orden inferior. Para calcular μ_{AB} se necesitaría calcular también μ, μ_A y μ_B, los cuales son todos los términos de orden inferior. Así, en los modelos jerárquicos los términos de orden superior se incluyen sólo si los términos de orden más bajo también se incluyen en el modelo. Los modelos no jerárquicos no tienen esta restricción y, como tales, generan resultados difíciles de interpretar.

Puede haber un gran número de modelos no saturados. Conforme se incrementa el número de variables categóricas, también lo hace el número de modelos. En algunos casos se vuelve muy difícil probar todos los modelos. De hecho, el investigador no debe intentar probar todos los modelos con la esperanza de hallar uno que se ajuste a los datos. El modelo debe basarse en la teoría o en una combinación de teoría y hallazgos previos. Tome como ejemplo el estudio de Glick, DeMorest y Hotze (1988), quienes encontraron una interacción de tres factores sin utilizar el llamado análisis log-lineal. Dicho término de interacción de tres factores, μ_{ABC}, en términos log-lineales sería el término de orden superior para los datos. Por lo tanto, el modelo se especificaría con el término de tres factores. O, si se deseara extenderse en el análisis e incluir una cuarta variable categórica, definitivamente se incluiría en el modelo un término para la interacción de tres factores. La decisión sobre cuáles términos se incluirán en el modelo para comprobarlo se denomina *especificación*.

La especificación del modelo para un análisis log-lineal en tablas de contingencia utiliza una notación especial. Se utilizan letras mayúsculas del alfabeto, encerradas en corchetes, para representar el efecto de cada variable de forma separada. Por ejemplo, cuando se habla del efecto de A en una tabla de tres factores, se escribiría $[A]$. En un modelo jerárquico, si se establece $[AB]$, se está refiriendo al modelo:

$$\log_e(m_{ij}) = \mu + \mu_A + \mu_B + \mu_C + \mu_{AB}$$

Si se anota $[A][BC]$ se refiere al modelo:

$$\log_e(m_{ij}) = \mu + \mu_A + \mu_B + \mu_C + \mu_{BC}$$

Si se anota $[ABC]$, entonces se estaría hablando del modelo saturado:

$$\log_e(m_{ij}) = \mu + \mu_A + \mu_B + \mu_C + \mu_{AB} + \mu_{AC} + \mu_{BC} + \mu_{ABC}$$

Bakeman y Robinson (1994) emplean un sistema de notación que difiere ligeramente de éste. El programa computacional que acompaña su libro es muy fácil de usar y parece estar tan bien desarrollado como algunos de los programas disponibles comercialmente. Sin embargo, su programa no imprime los corchetes "[ ]".

El estadístico de bondad de ajuste que se revisó en el capítulo 10 tiene un nombre formal que no se mencionó, que es necesario en este capítulo principalmente porque se presentará otro estadístico de bondad de ajuste. El estadístico revisado en el capítulo 10 es la chi cuadrada de Pearson. Su fórmula es:

$$\chi^2 = \sum\left[\frac{(f_o - f_e)^2}{f_e}\right]$$

Otro estadístico que es casi idéntico a la χ^2 de Pearson es la de razón de probabilidad χ^2. Para distinguir entre ambas, la chi cuadrada de la razón de probabilidad por lo común se anota como G^2. Como lo explica Wickens (1989), las dos son casi idénticas respecto a su aproximación a una distribución de chi cuadrada. La decisión sobre cuál debe usarse es una cuestión de preferencia. Wickens sí menciona que la χ^2 de Pearson es más familiar y más clara a nivel intuitivo, que la razón de probabilidad. Algunos programas computacionales calculan ambas. La ventaja que la chi cuadrada de la razón de probabilidad (G^2) tiene sobre la chi cuadrada de Pearson es a nivel del cálculo. La fórmula de la razón de probabilidad no utiliza las frecuencias esperadas de forma directa. La chi cuadrada de la razón de probabilidad para una tabla de dos factores se presenta como:

$$G^2 = 2\left(\sum f_{0ij} \log f_{0ij} - \sum R_i \log R_i - \sum C_j \log C_j + N \log N\right)$$

donde

$$\sum f_{0ij} \log f_{0ij}$$

es la suma de los valores observados por el logaritmo del valor observado en cada casilla de la tabla de contingencia.

$$\sum R_i \log R_i$$

TABLA 33.7 *Marginales y frecuencia de casillas*

	C_1	C_2	C_3	
R_1	$x_{11} = 34$	$x_{12} = 26$	$x_{13} = 75$	$x_{1+} = 135$
R_2	$x_{21} = 20$	$x_{22} = 30$	$x_{23} = 82$	$x_{2+} = 132$
	$x_{+1} = 54$	$x_{+2} = 56$	$x_{+3} = 157$	$x_{++} = 267$

es la suma del total de renglón por el logaritmo de los totales de renglón,

$$\Sigma C_j \log C_j$$

es la suma de los totales de columna por el logaritmo de los totales de columna. La "N" en $N \log N$ es el conteo de la frecuencia total. El uso de dichos símbolos sólo es efectivo cuando se habla de tablas bidimensionales. Con tablas de tres factores o de factores múltiples, los investigadores usarían una notación diferente. Los valores de casilla observados para una tabla de tres factores se escribirían x_{ijk}. Lo que se designaría como total de renglón y totales de columna en una tabla de dos factores, ahora se llaman *totales marginales*. Para una tabla de tres factores, se anotarían como x_{+jk}, x_{i+k}, x_{ij+}. El gran total o el número total de conteos es x_{+++}. Estas notaciones también son útiles para una tabla de contingencia de dos factores. La tabla 33.7 muestra la relación de los componentes de contingencia y la notación.

Es posible reescribir la ecuación de bondad de ajuste para una tabla de contingencia de dos factores como

$$G^2 = 2(\Sigma x_{ij} \log x_{ij} - \Sigma x_{i+} \log x_{i+} - \Sigma x_{+j} \log x_{+j} + x_{++} \log x_{++})$$

Para los datos del ejemplo en la tabla 33.7,

$$\begin{aligned} G^2 &= 2\,(34 \log 34 + 26 \log 26 + 75 \log 75 + 20 \log 20 + 30 \log 30 \\ &\quad + 82 \log 82 - 135 \log 135 - 132 \log 132 - 54 \log 54 - 56 \log 56 \\ &\quad - 157 \log 157 + 267 \log 267) \\ &= 2\,(119.8996 + 84.7105 + 323.8116 + 59.9147 + 102.0359 + 361.3510 \\ &\quad - 662.2121 - 644.5299 - 215.4051 - 225.4197 - 793.8306 + 1491.7954) \\ &= 2\,(2.1213) = 4.2426 \end{aligned}$$

Si se hubiera calculado χ^2 en lugar de G^2, el valor de χ^2 sería 4.1943. El cálculo de este valor requirió del cálculo adicional de los valores esperados, fe_{ij}. Para este ejemplo serían $fe_{11} = 27.3$, $fe_{12} = 28.31$, $fe_{13} = 79.38$, $fe_{21} = 26.7$, $fe_{22} = 27.69$ y $fe_{23} = 77.62$. Como se puede ver, G^2 y χ^2 no producen el mismo valor. Sin embargo, ambos valores están evaluados con el mismo número de grados de libertad y también con la misma tabla de chi cuadrada.

Ejemplo de investigación

Puesto que Glick *et al.* (1988) presentaron sus datos en su artículo, éstos se utilizarán para ilustrar la aproximación logarítmica lineal de las tablas de contingencia de factores múltiples. Glick *et al.* estudiaron tres variables: *obediencia* (O), *distancia* (D) y *pertenecia al grupo* (G). La variable *obediencia* tenía dos categorías: *obedeció* o *se rehusó*. La *pertenecia al grupo* implicaba si el solicitante de un favor era o no-miembro del mismo grupo que el participante. Esto es, ¿tenía el solicitante una apariencia física similar al participante? Si la respuesta era "sí",

entonces se consideraba un cómplice *dentro del grupo*; si la respuesta era "no", entonces se trataba de un cómplice *fuera del grupo*. La variable *distancia* medía tres distancias: cerca, medio y lejos. Los investigadores hipotetizaron que las personas estarían dispuestas a obedecer si el cómplice *fuera del grupo* estaba más lejos de ellas que los cómplices *dentro del grupo*. Los investigadores suponían básicamente una interacción de tres factores. Se puede escribir el modelo log-lineal de Glick *et al.* como

$$\log_e(m_{ij}) = \mu + \mu_C + \mu_D + \mu_G + \mu_{OD} + \mu_{OG} + \mu_{DG} + \mu_{ODG}$$

Si se puede obtener un ajuste adecuado entre los valores esperados y los valores observados *sin* el término de interacción de los tres factores, ello indicaría que los datos no pudieron justificar la interacción de los tres factores. Siendo así, los datos no apoyarían la hipótesis de los investigadores, lo cual significaría que el efecto de la distancia interpersonal sobre la obediencia no es diferente en los miembros *dentro del grupo* y *fuera del grupo*.

Al utilizar el programa computacional *ILOG* de Bakeman y Robinson (1994) que acompaña su libro de texto, se obtuvieron los resultados presentados en la tabla 33.8.

Aquí se percibe lo que sucedió con la G^2 cuando se ajustó el modelo saturado. El modelo saturado se ajusta perfectamente a los datos observados. Después se elimina el término de interacción de los tres factores del modelo; esto es, el modelo se prueba:

$$\log_e(m_{ij}) = \mu + \mu_O + \mu_D + \mu_G + \mu_{OD} + \mu_{OG} + \mu_{DG}$$

Si el estadístico G^2 no es significativo, se sabe que se ha encontrado por lo menos un modelo que se ajusta bien a los datos observados. Cuando el modelo se ajusta, la G^2 que se obtiene es 12.4 con 2 grados de libertad. Si se consulta una tabla de χ^2 para $\alpha = 0.05$ y $gl = 2$ el valor crítico es 5.99. Puesto que 12.4 es mayor que 5.99, entonces se tiene una prueba χ^2 estadísticamente significativa, y esto indica que el modelo no se ajusta. Al observar la tabla 33.8 se han listado los resultados de la prueba estadística para cada modelo. Todos los modelos reducidos probados son estadísticamente significativos, lo cual indica que los modelos que se probaron no se ajustan a los datos observados. Como el único modelo que se ajusta a los datos es el modelo saturado, entonces se llega a la misma conclusión que encontraron Glick *et al.*: existe una interacción de tres factores entre las variables.

Análisis multivariado e investigación científica

A pesar de que la revisión de los métodos multivariados ha sido más bien superficial, se debe hacer un alto para ubicarlos dentro del esquema de la investigación para evaluarlos. Por ejemplo, ¿se debe abandonar el análisis de varianza sólo porque la regresión múltiple puede lograr todo lo que hace el análisis de varianza, y más? Tal vez implicaciones como éstas ya se han captado por el lector. ¿En realidad el análisis de regresión múltiple no es

TABLA 33.8 *Análisis de los datos de Glick et al.*

Modelo	G^2	gl	Sig.	*Término eliminado*	ΔG^2	Δgl
[ODG] (saturado)	0.0	0	$p > 0.05$	—		
[DG][OD][OG]	12.4	2	$p < 0.005$	ODG	12.4	2
[OD][OG]	13.0	4	$p < 0.05$	DG	0.6	2
[OG][D]	20.2	6	$p < 0.005$	OD	7.2	2
[DG][O]	26.8	7	$p < 0.001$	OG	6.6	1

adecuado para los datos experimentales debido a que es uno de los llamados métodos de correlación (lo cual es sólo en parte)? Otras preguntas importantes pueden y deben plantearse y responderse, especialmente en este punto del desarrollo de la investigación de las ciencias del comportamiento. Quizás estemos en un momento de una importante transición. Desde que Fisher inventó y expuso el análisis de varianza en los años veinte y treinta, el método, o más bien el modelo, ha ejercido una gran influencia en la investigación del comportamiento, particularmente en la psicología. ¿Estamos a punto de superar esta etapa? ¿Hemos entrado en una "etapa multivariada"? Si es así, existiría una influencia enormemente importante sobre el tipo y calidad de investigación realizada por psicólogos, sociólogos y educadores en este nuevo siglo. Evidentemente no es posible manejar todas estas preguntas en un libro de texto. Pero al menos se debe abrir la puerta al estudiante.

¿El análisis de varianza debe ser suplantado por el análisis de regresión múltiple? Los autores no creen que deba ser así. ¿Pero es esto una mera atadura sentimental a algo que se ha encontrado interesante y satisfactorio? Tal vez. Pero hay más cosas que hacer que eso. No tiene mucho sentido utilizar la regresión múltiple en la situación de un problema ordinario de análisis de varianza: la asignación aleatoria de los sujetos a los tratamientos experimentales; número de sujetos iguales o proporcionales en las casillas; una, dos o tres variables independientes. Otro argumento para el análisis de varianza es su utilidad en la enseñanza. El análisis de regresión múltiple, aunque elegante y poderoso, carece de la calidad heurística estructural del análisis de varianza. No existe nada tan efectivo en la investigación de la enseñanza y el aprendizaje como el dibujo de paradigmas de los diseños utilizando la partición analítica del análisis de varianza.

La respuesta es que ambos métodos deben enseñarse y aprenderse. Las demandas adicionales sobre el maestro y el estudiante son inevitables, de la misma manera en que el desarrollo, crecimiento y uso de la estadística inferencial anteriormente en el siglo XX hicieron que su enseñanza y aprendizaje resultaran inevitables. No obstante, la regresión múltiple y otros métodos multivariados sin duda sufrirán de falta de comprensión, inclusive alguna oposición, de la misma forma en que la estadística inferencial lo ha sufrido. Aun en la actualidad existen psicólogos, sociólogos y educadores que saben muy poco sobre estadística inferencial o análisis moderno, y quienes inclusive se oponen a su aprendizaje y a su uso. Sin embargo, esto forma parte de la psicología social y patología del tema. A pesar de que sin duda habrá un retraso cultural, la aceptación definitiva de estas poderosas herramientas de análisis probablemente está asegurada.

Los métodos multivariados, como se ha visto, no son tan fáciles de usar y de interpretar como los métodos univariados. Lo anterior se debe no sólo a su complejidad; se debe más bien a la complejidad de los fenómenos con los que trabajan los científicos del comportamiento. Uno de los atrasos de la investigación educativa, por ejemplo, ha sido que la enorme complejidad de una escuela o de un salón de clases no puede manejarse adecuadamente por los métodos demasiado simples que se utilizan. Algunos científicos consideran que ellos nunca pueden captar el mundo "real" con sus métodos de observación y de análisis. Se trata de individuos atados a la simplificación de las situaciones y problemas que estudian. Nunca pueden "ver el todo", de la misma manera que ningún ser humano es capaz de observar y comprender la totalidad de cierto fenómeno. Pero los métodos multivariados captan la realidad psicológica, sociológica y educativa mejor que métodos más simples, y permiten que los investigadores manejen porciones mayores de sus problemas de investigación. En la investigación educativa, los días del experimento de métodos simples con un grupo experimental y un grupo control están contados. En la investigación sociológica, la reducción de gran cantidad de datos valiosos a frecuencias y cruces de porcentajes disminuirá en relación con el cuerpo total de la investigación sociológica.

Lo más importante de todo, el futuro sano de la investigación del comportamiento depende del desarrollo sano de las teorías psicológicas, sociológicas y de otros tipos, para ayudar a explicar las relaciones entre los fenómenos del comportamiento. Por definición, las teorías forman conjuntos interrelacionados de constructos o variables. Evidentemente los métodos multivariados están bien adaptados para comprobar formulaciones teóricas bastante complejas, pues su campo natural es el análisis simultáneo de diversas variables. De hecho, el desarrollo de la teoría del comportamiento debe ir de la mano, e inclusive depender de la asimilación, la maestría y del uso inteligente de los métodos multivariados. En los próximos dos capítulos se ofrecerá una visión multivariada diferente para realizar investigación en las ciencias sociales y del comportamiento.

RESUMEN DE CAPÍTULO

1. Este capítulo examina las diferencias y similitudes entre el análisis de varianza y la regresión múltiple.
2. La regresión múltiple puede, en esencia, hacer todos los análisis que el ANOVA, y aún más.
3. El análisis de varianza posee una estructura que es intuitivamente atractiva para los investigadores.
4. En el presente capítulo se analizan las diferencias entre los diferentes tipos de codificación: dummy, de efectos y ortogonal. Cada una produce la misma R^2, pero los coeficientes individuales de las variables son diferentes.
5. La diferencia entre el análisis de varianza y el análisis de covarianza reside en que en el ANOVA las variables independientes no están correlacionadas. En el análisis de covarianza, por lo menos una variable está correlacionada con las otras variables independientes.
6. Esta variable correlacionada se llama *covariable*. Sirve para eliminar su varianza de la variable dependiente, antes de que se prueben las variables independientes.
7. El análisis de covarianza se maneja con facilidad por medio de la regresión múltiple.
8. El análisis discriminante resulta similar a la regresión múltiple, con algunas excepciones. En el análisis discriminante la variable dependiente es categórica. Además ofrece un estadístico que indica qué tan bien la función discriminante clasifica observaciones.
9. En la correlación canónica, en lugar de una variable dependiente como en la regresión múltiple y en el análisis discriminante, existen más variables dependientes. El objetivo es encontrar dos conjuntos de coeficientes que maximicen la varianza entre los dos conjuntos de variables.
10. El análisis multivariado de varianza o MANOVA es el equivalente multivariado del análisis de varianza. En el ANOVA univariado el análisis se hace para una variable dependiente a la vez. En el análisis multivariado de varianza, se consideran al mismo tiempo múltiples variables dependientes.
11. Los MANOVA, al igual que todos los métodos multivariados, pueden llevar a resultados que sean difíciles de interpretar.
12. El análisis de ruta utiliza los pesos estandarizados de la regresión para estudiar los efectos directos e indirectos de unas variables sobre otras. Se utiliza mejor como un modelo conceptual a probarse.
13. El análisis de ruta implica el dibujo de un diagrama de ruta que muestre cómo se relacionan las variables.

14. La regresión de cresta fue utilizada primero en ingeniería química por Arthur Hoerl. Posteriormente se convirtió en una herramienta para otras áreas.
15. Los mínimos cuadrados ordinarios son el método estadístico utilizado por la mayoría de los programas computacionales de regresión múltiple. No obstante, cuando las variables independientes están altamente correlacionadas entre sí, los mínimos cuadrados presentan problemas en términos de estimación.
16. La regresión de cresta agrega un sesgo a la ecuación y, por ende, estabiliza los coeficientes de regresión. La regresión de cresta constituye un tema polémico en las ciencias del comportamiento.
17. La regresión logística es el modelo popular y alternativo al análisis discriminante. No tiene tantas restricciones como las impuestas al análisis discriminante.
18. Con menos restricciones, la regresión logística puede manejar una diversidad de problemas. Tal como el análisis discriminante, la regresión logística posee una variable dependiente categórica.
19. Las tablas de contingencia de factores múltiples se manejan utilizando el análisis log-lineal.
20. La idea principal que está detrás del análisis log-lineal para las tablas de contingencia de factores múltiples es encontrar el modelo apropiado que explicará la variación de los valores observados.
21. Existen modelos jerárquicos y no jerárquicos en el análisis log-lineal. El jerárquico es el más útil. Los modelos no jerárquicos están sujetos a problemas de interpretación.

SUGERENCIAS DE ESTUDIO

1. Por desgracia son escasos los tratamientos elementales de regresión múltiple completamente satisfactorios, en especial si se espera un tratamiento de regresión concomitante al análisis de varianza. Tal vez no sea posible un tratamiento elemental satisfactorio de un tema tan complejo. Las siguientes referencias sobre regresión múltiple y otros métodos multivariados pueden resultar de utilidad. Algunas de ellas también se incluyen en la sección de referencias, debido a que se citaron en el presente capítulo.

Kerlinger, F. y Pedhazur, E. (1973). *Multiple regression in behavioral research*. Nueva York: Holt, Rinehart and Winston. [Un texto que intenta incrementar la comprensión de la regresión múltiple y sus usos en la investigación por medio de la presentación de una exposición lo más simple posible, así como diversos ejemplos con números sencillos. También incluye un programa computacional completo de regresión múltiple en el apéndice.]

Pedhazur, E. (1996). *Multiple regression in behavioral research: Explanation and prediction* (4a. ed.). Orlando, Florida: Harcourt Brace. [Es la revisión del libro de Kerlinger y Pedhazur. Sin embargo, es más detallado y profundo. Bastante recomendable.]

Stevens, J. P. (1996). *Applied multivariate statistics for the social sciences* (3a. ed.). Mahwah, Nueva Jersey: Lawrence Erlbaum. [Un libro sobre estadística multivariada fácil de leer. Incluye notas respecto a los resultados en computadora de conocidos programas estadísticos.]

Tabachnick, B. y Fidell, L. (1996). *Using multivariate statistics* (3a. ed.). Nueva York: HarperCollins. [Muestra la estadística multivariada desde un punto de vista de

resultados computacionales. Muy útil para quienes desean aprender estadística multivariada y también sobre los programas computacionales utilizados para realizar los análisis.]

Una vez que el estudiante y el investigador manejan los elementos del análisis de regresión múltiple y que han tenido alguna experiencia con problemas reales, las siguientes referencias proporcionan una guía sofisticada en el uso del análisis de regresión múltiple y, más importante aún, en la interpretación de los datos.

Cohen, J. y Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2a. ed.). Mahwah, Nueva Jersey: Lawrence Erlbaum. [Un tratamiento excelente de la regresión múltiple. Muestra cuantos problemas donde se utilizó el análisis de varianza, podrían haberse analizado por medio de regresión múltiple. También muestra la forma en que la regresión múltiple se utiliza para estudiar causalidad.]

Daniel, C. Wood, F. S. y Gorman, J. W. (1980). *Fitting equations to data: Computer analysis of multifactor data* (2a. ed.). Nueva York: Wiley. [Resume, por medio de ejemplos, la forma en que estos estadísticos se aproximaron al análisis de datos de investigación, donde los investigadores no siguieron los requisitos estándares del diseño estadístico de experimentos. Utiliza muchos análisis y explicaciones generados por computadora.]

Draper, N. y Smith, H. (1981). *Applied regression analysis* (2a. ed.). Nueva York: John Wiley & Sons. [Un clásico en el campo del análisis de regresión. Requiere de alguna sofisticación matemática, pero es un libro útil y citado con frecuencia.]

Kleinbaum, D. G., Kupper, L. L., Muller, K. E. y Nizam, A. (1997). *Applied regression analysis and other multivariate methods.* (3a. ed.). Belmont, California: Duxbury. [Esta obra aclara la confusión entre multivariado y multivariable. Se mencionó brevemente en el capítulo 2. Vale la pena leerlo.]

Mendenhall, W. (1968). *An introduction to linear models and the design and analysis of experiments.* Belmont, California: Wadsworth. [Sin duda uno de los mejores libros escritos, que presenta, sin sofisticación, el uso de la regresión múltiple (modelo lineal general) en lugar del análisis de varianza. Requiere de conocimiento sobre álgebra de matrices. Se trata de un libro que ya no se imprime.]

Neter, J. Wasserman, W. y Kutner, M. H. (1996). *Applied linear regression models* (3a. ed.). Burr Ridge, Illinois: Irwin. [Este libro es similar al escrito por Woodward, Bonett y Brecht, citado más adelante. Cubre bien el uso de la regresión. Requiere de conocimiento sobre álgebra de matrices.]

Los siguientes textos son fundamentales: enfatizan las bases teóricas y matemáticas de los métodos multivariados.

Carroll, J. D. y Green, P. (1997). *Mathematical tools for applied multivariate analysis* (3a. ed.). Nueva York: Academic Press. [Un libro sobresaliente sobre las bases matemáticas del análisis multivariado. Altamente recomendable.]

Kenny, D. (1979). *Correlation and causality.* Nueva York: John Wiley & Sons. [Merece muchas horas de estudio.]

Wickens, T. D. (1994). *The geometry of multivariate statistics.* Mahwah, Nueva Jersey: Lawrence Erlbaum. [Presenta los procedimientos de la estadística multivariada de forma geométrica. Ayuda al estudiante a conceptualizar las relaciones multivariadas.]

2. Suponga que un psicólogo social tiene dos matrices de correlación:

$$\begin{array}{c} \\ X_1 \\ X_2 \\ Y \end{array}\begin{array}{c} \begin{array}{ccc} X_1 & X_2 & Y \end{array} \\ \begin{bmatrix} 1.00 & 0 & .70 \\ 0 & 1.00 & .60 \\ .70 & .60 & 1.00 \end{bmatrix} \end{array} \qquad \begin{array}{c} \\ X_1 \\ X_2 \\ Y \end{array}\begin{array}{c} \begin{array}{ccc} X_1 & X_2 & Y \end{array} \\ \begin{bmatrix} 1.00 & .40 & .70 \\ .40 & 1.00 & .60 \\ .70 & .60 & 1.00 \end{bmatrix} \end{array}$$

A B

a) ¿Cuál matriz, la A o la B, producirá la R^2 mayor? ¿Por qué?
b) Calcule la R^2 de la matriz A.
[Respuestas: ***a)*** la matriz A; ***b)*** $R^2 = .85$.]

3. A continuación se presentan tres conjuntos de datos ficticios simples, ordenados para un análisis de varianza. Ordene los datos para un análisis de regresión múltiple y realice tanto del análisis de regresión como le sea posible. Utilice codificación dummy (1, 0), como en la tabla 33.2. Los coeficientes b son: $b_1 = 3$; $b_2 = 6$.

A_1	A_2	A_3
7	12	5
6	9	2
5	10	6
9	8	3
8	11	4

Imagine que A_1, A_2 y A_3 son tres métodos para cambiar las actitudes raciales y que la variable dependiente es una medida del cambio en donde las puntuaciones más altas indican mayor cambio. Interprete los resultados. [Respuestas: a = 4; R^2 = .75; F = 18, con gl = 2, 12; sc_{reg} = 90; sc_t = 120. Observe que estos datos ficticios son en realidad las puntuaciones de la tabla 33.2 con un 1 que se agrega a cada puntuación. Compare los diversos estadísticos de regresión y de análisis de varianza anteriores, con aquellos calculados con los datos de la tabla 33.2.]

4. Utilizando los datos de la tabla 34.2 en el capítulo 34, calcule las sumas de cada par de X_1 y X_2. Correlacione dichas sumas con las puntuaciones de Y. Compare el cuadrado de esta correlación con $R^2_{y.12} = .51$ ($r^2 = .70^2 = .49$). Puesto que los valores son bastante cercanos, ¿por qué no debemos simplemente utilizar los promedios de las variables independientes sin molestarnos con la complejidad del análisis de regresión múltiple?

5. A continuación se presenta una lista de varios estudios interesantes que han utilizado regresión múltiple, análisis de ruta y análisis discriminante de forma efectiva. Lea uno o dos de ellos cuidadosamente.

Abel, M. H. (1998). Interaction of humor and gender in moderating relationships between stress and outcomes. *Journal of Psychology, 132*, 267-276. [Utiliza la regresión múltiple para estudiar los efectos moderadores del humor sobre el estrés y la ansiedad.]

Bachman, I. y O' Malley, P. (1977). Self-esteem in young men: A longitudinal analysis of the impact of educational and occupational attainment. *Journal of Personality*

and Social Psychology, 35, 365-380. [Un estudio educativo sobresaliente que utiliza el análisis de ruta. Los resultados son contrarios a lo esperado.]

Fischer, C. (1975). The city and political psychology. *American Political Science Review, 69*, 559-571. [Utiliza el análisis de ruta para estudiar el sentido de eficacia política.]

Frederick, C. M. y Morrison, C. S. (1998). A mediational model of social physique anxiety and eating disordered behaviors. *Perceptual and Motor Skills, 86*, 139-145. [Desarrolla un modelo de ruta muy simple y fácil de entender, en relación con la ansiedad sobre el físico, los rasgos del trastorno de alimentación y el comportamiento del trastorno de alimentación.]

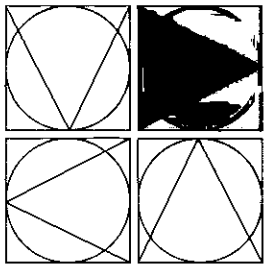
Leith, K. P. y Baumeister, R. F. (1998). Empathy, shame, guilt and narratives of interpersonal conflicts: Guilt prone people are better at perspective taking. *Journal of Personality, 66*, 11-39. [Utiliza análisis multivariado de varianza, análisis de covarianza y análisis de ruta para estudiar a personas con tendencia a sentirse culpables.]

Marjoribanks, K. (1972). Ethnic and environmental influences on mental abilities. *American Journal of Sociology, 78*, 323-337. [Un uso interesante de la suma y resta de las R^2 para evaluar la influencia relativa de variables, especialmente del ambiente y la etnicidad.]

Onwuegbuzie, A. J. (1997). The teacher as researcher: The relationship between research anxiety and learning style in a research methodology course. *College Student Journal, 31*, 496-506. [Este estudio utiliza la regresión múltiple para determinar algunas de las características de los maestros ansiosos por la investigación, en términos del tipo de estilo de aprendizaje.]

Ronis, D. L., Antonakos, C. L. y Lang, W. P. (1996). Usefulness of multiple equations for predicting preventive oral health behaviors. *Health Education Quarterly, 23*, 512-527. [Analiza los resultados de un estudio que utiliza la correlación canónica. Los investigadores encontraron tres funciones que explicaban tres conductas de salud oral.]

Vincke, J. y Bolton, R. (1997). Beyond the sexual model: Combining complementary cognitions to explain and predict unsafe sex among gay men. *Human Organization, 56*, 38-46. [Emplea tanto el análisis multivariado de varianza como el análisis discriminante para evaluar los placeres y peligros de las prácticas sexuales sin protección.]



CAPÍTULO 34

ANÁLISIS FACTORIAL

Muchos investigadores consideran el análisis factorial como la reina de los métodos analíticos, debido a su poder, elegancia y cercanía al corazón del propósito científico. Sin embargo, se trata de un método que no está libre de controversia. A pesar de que se trata de un método poderoso, no constituye una panacea para estudios mal diseñados o sin diseño. Comrey (1978) señaló que el análisis factorial ha sido un tema de gran discusión y crítica. No obstante, a pesar de las críticas, el aumento de su uso continúa. En este capítulo se explorará lo que es el análisis factorial, y por qué y cómo se realiza. También se explorarán las dificultades que un investigador llega a encontrar si no tiene cuidado al utilizar este poderoso método. Durante la exploración también se examinará investigación pasada y actual, donde el análisis factorial ha sido la metodología central.

El análisis factorial sirve a la causa de la parsimonia científica. Reduce la multiplicidad de las pruebas y medidas a una mayor simplicidad. En efecto, indica qué pruebas o medi-

das van juntas —las que virtualmente miden lo mismo— y qué tanto es así. Por lo tanto, reduce el número de variables con las que el científico debe enfrentarse. También ayuda al científico a ubicar e identificar unidades o propiedades fundamentales que subyacen a pruebas y medidas. Suponga que un investigador ha medido 20 variables en un grupo de personas. Se calculan las correlaciones entre las variables y se sintetizan en forma de una matriz de correlación. Cuando un investigador examina una tabla de correlación entre variables, resulta muy difícil interpretar lo que en realidad está sucediendo. Por lo general, es muy difícil encontrar un patrón de correlaciones interpretable. El análisis factorial está diseñado para tomar esas correlaciones y encontrar algún orden entre ellas. El método está diseñado para encontrar lo que las variables tienen en común. Aun cuando en el presente capítulo la explicación sobre el análisis factorial se centrará en el uso de los coeficientes de correlación, el análisis factorial no se limita tan sólo a matrices de correlación. Sin embargo, en las ciencias sociales, del comportamiento y educativas, la correlación es el índice utilizado con mayor frecuencia en un análisis factorial.

Un *factor* es un constructo, una entidad hipotética, una variable latente que se asume como el fundamento de pruebas, escalas, reactivos y, de hecho, de medidas de casi cualquier clase. Para aquellos investigadores en ciencias del comportamiento que estén desarrollando escalas o pruebas, el análisis factorial les sirve para ofrecer evidencia de la ausencia o presencia de validez. Se ha encontrado una serie de factores que son fundamento de la inteligencia, por ejemplo: destreza verbal, habilidad numérica, razonamiento abstracto, razonamiento espacial, memoria, etcétera. De forma similar, también se han aislado e identificado factores de aptitud, actitud y personalidad. ¡Inclusive se han encontrado factores de naciones y personas!

Fundamentos

Breve historia

El desarrollo del método denominado análisis factorial se atribuye a Charles Spearman. En 1904, Spearman publicó un artículo de 93 páginas que cubría su teoría sobre la inteligencia y el desarrollo del método para confirmar su teoría de que un factor común explicaba toda la inteligencia humana. Ese único factor se denomina *g*, o factor general. Spearman analizó tablas de correlación entre pruebas psicológicas y demostró que había un factor común a todas las pruebas. Las varianzas restantes se atribuían a la prueba específica. Algunas ocasiones la teoría de Spearman se conoce como la teoría de los dos factores. Spearman también vinculó su teoría con la neuropsicología al afirmar que el factor *g* abarcaba toda la corteza humana o sistema nervioso. Muchos investigadores realizaron diversos trabajos respecto a la teoría de Spearman. Algunos de los nombres más notables se mencionan en Ferguson (1971) o Carroll (1993). El concepto de *g* resulta polémico. A pesar de que muchos autores han desarrollado estudios empíricos para demostrar que no existe, su uso y referencia permanecen. Uno de los principales antagonistas de la teoría de Spearman fue L. L. Thurstone (1947), quien proporcionó evidencia inicial de que había factores de inteligencia, a los que llamó "habilidades mentales primarias". Aunque Thurstone proporcionó evidencia en contra de Spearman, más tarde se demostró que su teoría de inteligencia también era cuestionable.

A pesar de tales raíces, comienzos e intenciones iniciales del análisis factorial, los métodos creados como resultado del trabajo de Spearman y de Thurstone continúan siendo importantes. En sus esfuerzos por "probar" que el otro estaba equivocado, surgieron

los métodos modernos y creativos del análisis factorial. Los investigadores continúan utilizando dichos métodos hoy en día. La contribución de Thurstone al campo es monumental. La mayor parte del análisis factorial es el resultado directo de su trabajo. El profesor Andrew Comrey de la UCLA le dijo en una ocasión al segundo autor de este libro, que Thurstone (1947) estableció las bases para cerca del 90 por ciento del análisis factorial moderno. Desde entonces, los investigadores intentan definir el restante 10 por ciento. Thurstone tomó de manera efectiva el método del análisis factorial de Spearman y lo mejoró. Thurstone fue responsable del desarrollo del método centroide. Antes del advenimiento de la computadora de alta velocidad, los análisis factoriales se realizaban a mano. El método centroide ofreció una muy buena aproximación al método más poderoso desarrollado por Hotelling (1933). El método del factor principal de Hotelling se adaptaba mejor para las computadoras; aunque era bastante laborioso para ser realizado mediante cálculos manuales. Con el propósito de facilitar la interpretación de los resultados analíticos del factor, Thurstone también desarrolló el método de rotación y el concepto de estructura simple. La estructura simple es uno de los desarrollos clave dentro de la metodología analítica del factor.

En este capítulo y en el siguiente se procurará hacer las menores referencias posibles al álgebra matricial. Sin embargo, la explicación del análisis factorial requiere considerar las matrices. Algunas explicaciones se facilitan al utilizarlas, en especial cuando se exponga el análisis factorial confirmatorio y el modelamiento de la ecuación estructural.

Un ejemplo hipotético

Considere que se aplican seis pruebas a un gran número de alumnos de primero de secundaria. Se sospecha que las seis pruebas están midiendo no seis, sino un menor número de variables. Las pruebas son *vocabulario*, *lectura*, *sinónimos*, *números*, *aritmética* (prueba estandarizada), *aritmética* (prueba elaborada por el profesor). Los nombres de estas pruebas indican su naturaleza. Se les denomina, respectivamente, *V*, *L*, *S*, *N*, *AE*, *AEP*. (Aunque las últimas dos pruebas son de aritmética, tienen un contenido diferente. Se asume que existe una buena razón para incluir ambas en la pequeña batería de pruebas.) Una vez que las pruebas se aplican y se evalúan, se calculan los coeficientes de correlación de cada prueba con todas las demás. Se ordenan las *r* en una matriz de correlación (comúnmente llamada matriz **R**). La matriz se presenta en la tabla 34.1.

Recuerde que una matriz es cualquier arreglo rectangular de números (o símbolos). Las matrices de correlación siempre son cuadradas y simétricas; ello se debe a que la mitad que se encuentra por debajo de la diagonal principal (de la parte izquierda superior a la

TABLA 34.1 *Matriz R: coeficientes de correlación entre seis pruebas*

		V	*L*	*S*	*N*	*AE*	*AEP*
	V		.72	.63	.09	.09	.00
Conglomerado I	*L*	.72		.57	.15	.16	.09
	S	.63	.57		.14	.15	.09
	N	.09	.15	.14		.57	.63
	AE	.09	.16	.15	.57		.72
	AEP	.00	.09	.09	.63	.72	

Conglomerado II

parte inferior derecha) es igual a la mitad superior de la matriz. Es decir, los coeficientes de la mitad inferior son idénticos a los de la mitad superior, con excepción de su arreglo. (Note que el renglón superior es igual a la primera columna, el segundo renglón es igual a la segunda columna y así sucesivamente). Si se intercambian los renglones y las columnas de la matriz de correlación, la matriz resultante será idéntica a la matriz original. Cuando así sucede, se sabe que la matriz es simétrica. También, cuando se intercambian los renglones con las columnas, la matriz resultante se denomina de *transposición*. Si se tiene una matriz llamada **A**, la de transposición se llama $\mathbf{A}^{\mathrm{T}}$. Este concepto se utilizará más adelante.

El problema al que hay que enfrentarse se expresa en dos preguntas: ¿cuántas variables o factores subyacentes existen? ¿Cuáles son los factores? Se presume que son unidades subyacentes a los desempeños de las pruebas, que se reflejan en los coeficientes de correlación. Si dos o más pruebas están altamente correlacionadas, entonces las pruebas comparten varianza; tienen varianza de factor común, y están midiendo algo en común.

La primera pregunta, en este caso, es fácil de contestar. Existen dos factores, lo cual está indicado por los dos grupos de *r* circuladas y llamadas *conglomerado I* y *conglomerado II* en la tabla 34.1. Observe que *V* se correlaciona con *L* en .72; *V* con *S* en .63; y *L* con *S* en .57. *V*, *L* y *S* parecen medir algo en común. De manera similar, *N* se correlaciona con *AE* en .57 y con *AEP* en .63; y *AE* se correlaciona con *AEP* en .72. *N*, *AE* y *AEP* miden algo en común. Las pruebas en el *conglomerado I*, aunque correlacionadas entre sí, no están muy correlacionadas con las pruebas en el *conglomerado II*. De la misma manera, aunque *N*, *AE* y *AEP* se correlacionan entre sí, no están altamente correlacionadas con las pruebas *V*, *L* y *S*. En efecto, lo que miden en común las pruebas del *conglomerado I*, no es lo mismo que miden en común las pruebas del *conglomerado II*. Parece haber dos conglomerados o factores en la matriz. El lector debe notar que en esta presentación se utilizan sobresimplificaciones ocasionales y ejemplos poco realistas. La matriz **R** de la tabla 34.1 no es realista. Todas las pruebas estarían correlacionadas positivamente, y quizá los dos factores surgirían. Además, aunque los grupos asemejan factores, *no* son factores. Sin embargo, por simplicidad y por razones pedagógicas, se arriesgó a efectuar tales sobresimplificaciones.

Al estudiar la matriz **R** se determina que existen dos factores detrás de estas pruebas. Respecto a la segunda pregunta (¿cuáles son los factores?) casi siempre es más difícil responder. Cuando se pregunta cuáles son los factores, se busca nombrarlos. Se buscan *constructos* que expliquen las unidades subyacentes o las varianzas de factor común de los factores. Se pregunta qué es lo que tienen en común las pruebas *V*, *L* y *S*, por un lado, y las pruebas *N*, *AE* y *AEP*, por el otro. *V*, *L* y *S* son pruebas de vocabulario, lectura y sinónimos. Las tres involucran palabras, en un sentido amplio. Quizás el factor subyacente sea *habilidad verbal*. Se denomina al factor *verbal* o *V*. Las pruebas *N*, *AE* y *AEP* involucran operaciones numéricas o aritméticas. Suponga que se denomina a este factor *aritmética*. Un amigo señala que la prueba *N* en realidad no involucra operaciones aritméticas, pues consiste principalmente de la manipulación no aritmética de números, lo cual se pasó por alto a causa de la insistencia por dar un nombre a la unidad subyacente. De cualquier manera, ahora se llama al factor *numérico* o *N*. No existe ninguna inconsistencia: las tres pruebas involucran números, manipulación y operación numérica.

Ya se respondieron las dos preguntas: existen dos factores y se denominan *verbal*, *V*, y *numérico*, *N*. No obstante, debe señalarse rápida y urgentemente que ninguna de las preguntas se contesta por completo en la verdadera investigación analítica factorial. Lo anterior sucede en especial en las investigaciones iniciales en un campo. El número de factores puede cambiar en investigaciones subsecuentes, utilizando las mismas pruebas. Una de las pruebas de *V* puede tener también alguna varianza en común con otro factor, por ejemplo, *K*. Si una prueba que mide *K* se añade a la matriz, quizá surja un nuevo factor. Tal vez más importante, el nombre de un factor puede ser incorrecto. Investigación subsecuente utili-

zando estas pruebas *V* y otras pruebas, demuestra que *V* ahora ya no es común a todas las pruebas. Entonces, el investigador debe encontrar otro constructo, otra fuente de varianza del factor común. En síntesis, los nombres de los factores son tentativos; son hipótesis a comprobarse en análisis factoriales posteriores y en otros tipos de investigación.

Matrices factoriales y cargas factoriales

Si una prueba mide sólo un factor, se dice que es *factorialmente pura*. Dependiendo del grado en que una prueba mida un factor, se dice que está *cargada* o *saturada* con el factor. En realidad el análisis factorial no está completo a menos que se conozca si una prueba es factorialmente pura o qué tan saturada está con un factor. Si una medida no es factorialmente pura, por lo general se busca saber qué otros factores la conforman. Algunas medidas son tan complejas que es difícil decir exactamente qué miden. Un buen ejemplo son las calificaciones del profesor, o los promedios de calificaciones, que pueden consistir de un número de dimensiones del desempeño del estudiante. Si una prueba contiene más de un factor, se dice que es *factorialmente compleja*.

Algunas pruebas y medidas son factorialmente muy complejas. La prueba de inteligencia de Stanford-Binet, la prueba de inteligencia de Otis y la escala *F* (de autoritarismo) son algunos ejemplos. Un hecho apreciado de la investigación científica es tener medidas puras de las variables. Si una medida de habilidad numérica no es factorialmente pura, entonces, ¿cómo se puede confiar en que una relación entre habilidad numérica y rendimiento académico, por ejemplo, es en realidad la relación que se piensa que es? Si la prueba mide tanto habilidad numérica como razonamiento verbal, las relaciones que se estudian con su ayuda resultarán dudosas.

Para resolver éstos y otros problemas se requiere de un método objetivo que determine el número de factores, las pruebas que pesan en los diversos factores y la magnitud de las cargas. Existen varios métodos analíticos factoriales para lograr tales propósitos. Posteriormente se analizará uno de ellos.

Uno de los resultados finales de un análisis factorial es la llamada *matriz factorial*, que es una tabla de coeficientes que expresa la relación entre las pruebas y los factores subyacentes. En la tabla 34.2 se presenta la matriz factorial producida por el análisis factorial de los datos de la tabla 34.1, con el método de factores principales, uno de los diversos métodos disponibles, y con la subsecuente rotación factorial (que se explicará más adelante). Los datos de la tabla se denominan pesos o *cargas factoriales*. Pueden escribirse como a_{ij}, lo que significa la carga *a* de la *prueba i* sobre el *factor j*. En la segunda línea, .79 es la carga

TABLA 34.2 *Matriz factorial de los datos de la tabla 34.1, solución rotada*[a]

Pruebas	*A*	*B*	b^2
V	*.83*	.01	.70
L	*.79*	.10	.63
S	*.70*	.10	.50
N	.10	*.70*	.50
AE	.10	*.79*	.63
AEP	.01	*.83*	.70

[a] Véase el texto para identificar las pruebas. Las cargas "significativas" están en *itálicas*. Véase también los pies de la tabla 34.5.

factorial de la prueba *R* sobre el factor *A*. Algunos de los analistas de factores llaman a los factores de la solución final I, II, ..., o I′, II′, etcétera. En este capítulo se denominan I, II, ..., a los factores sin rotación; y *A*, *B*,..., a los factores con rotación (solución final). En la cuarta línea, .70 es la carga factorial de la prueba *N* en el factor *B*. La prueba *AE* tiene las siguientes cargas: .10 en el factor *A* y .79 en el factor *B*.

Las cargas factoriales no son difíciles de interpretar. Oscilan entre −1.00 y +1.00, como los coeficientes de correlación. Además se interpretan de manera similar. De hecho, expresan las correlaciones *entre las pruebas y los factores*. Por ejemplo, la prueba *V* tiene las siguientes correlaciones con los factores *A* y *B*, respectivamente: .83 y .01. Evidentemente, la prueba *V* tiene una fuerte carga en *A*, pero ninguna en *B*. Las pruebas *V*, *L* y *S* pesan en *A* pero no en *B*. Las pruebas *N*, *AE* y *AEM* pesan en *B* pero no en *A*. Todas las pruebas parecen ser "puras".

Las cifras de la última columna se llaman *comunalidades* o h^2. Son las sumas de cuadrados de las cargas factoriales de una prueba o variable. Por ejemplo, la comunalidad de la prueba *R* es $(.79)^2 + (.10)^2 = .63$. La comunalidad de una prueba o variable es su varianza del factor común, lo cual se explicará más adelante cuando se exponga la teoría de factores.

Antes de continuar, debe señalarse nuevamente que este ejemplo no es realista. Las matrices de factores rara vez presentan una imagen tan clara. De hecho, la matriz factorial de la tabla 34.2 ya era "conocida". El autor primero escribió la matriz presentada en la tabla 34.3. Si esta matriz se multiplica por sí misma, entonces se obtiene la matriz de la tabla 34.1 (con valores en la diagonal). En tal caso, lo que se necesita para obtener **R** es multiplicar cada renglón por cada uno de los otros renglones. Por ejemplo, se multiplica el renglón *V* por el renglón *L*: (.90)(.80) + (.00)(.10) = .72; el renglón *V* por el renglón *S*: (.90)(.70) + (.00)(.10) = .63; el renglón *S* por el renglón *AE*: (.70)(.10) + (.10)(.80) = .15; etcétera. Después, la matriz **R** resultante se analizó factorialmente. Esta operación de multiplicación de la matriz surge de la llamada *ecuación básica del análisis factorial:* $\mathbf{R} = \mathbf{F}\mathbf{F}^{\mathrm{T}}$, que indica, de forma sucinta y en símbolos de matriz, lo que se expuso de forma más elaborada anteriormente. Algunas veces dicha ecuación fundamental se escribe $\mathbf{R} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$ o $\mathbf{R} = \mathbf{P}\Phi\mathbf{P}^{T} + \mathbf{U}$. La última ecuación es la más general de las tres. Un conocimiento profundo del análisis factorial requiere de un buen entendimiento del álgebra matricial.

Resulta instructivo comparar la tabla 34.2 con la tabla 34.3. Observe las discrepancias, que son pequeñas. Es decir, el método analítico factorial falible no puede reproducir de forma perfecta la "verdadera" matriz factorial; sólo la estima. En tal caso el ajuste es cercano debido a la simplicidad deliberada del problema. Los datos reales no son tan complacientes. Además, nunca se conoce la matriz factorial "verdadera". Si así fuera, no habría necesidad de realizar el análisis factorial. Por lo general, se estima la matriz factorial a partir de la matriz de correlación. La complejidad y falibilidad de los datos de investigación con frecuencia hacen que dicha estimación sea un asunto difícil.

TABLA 34.3 *Matriz factorial original de la cual se derivó la matriz R de la tabla 34.1*

Pruebas	A	B	h^2
V	*.90*	.00	.81
L	*.80*	.10	*.65*
S	*.70*	.10	.50
N	.10	*.70*	*.50*
AE	.10	*.80*	*.65*
AEP	.00	*.90*	.81

Un poco de teoría factorial

En el capítulo 28 se escribió una ecuación que expresa las fuentes de varianza en una medida (o prueba):

$$V_t = V_{co} + V_{esp} + V_e \tag{34.1}$$

donde V_t es igual a la varianza total de una medida; V_{co} significa la varianza del factor común o la varianza que dos o más medidas comparten en común; V_{esp} equivale a la varianza específica o la varianza de la medida que no es compartida por cualquier otra medida, es decir, la varianza de esa medida y de ninguna otra; V_e es igual a la varianza del error.

La varianza del factor común V_{co} se dividió en dos fuentes de varianza, A y B, que son dos factores (véase ecuación 28.11):

$$V_{co} = V_A + V_B \tag{34.2}$$

V_A podría ser la varianza de la habilidad verbal, y V_B podría ser la varianza de la habilidad numérica, lo cual es razonable si se piensa en las sumas de cuadrados de las cargas factoriales de cualquier prueba:

$$h_i^2 = a_i^2 + b_i^2 + \ldots + k_i^2 \tag{34.3}$$

donde $a_i^2, b_i^2, \ldots$ son los cuadrados de las cargas factoriales de la prueba i, y h_i^2 es la comunidad de la prueba i. Pero $h_i^2 = V_{co}$. Por lo tanto, $V(A) = a^2$ y $V(B) = b^2$, y la ecuación 34.2 se vincula con operaciones analíticas factoriales reales.

Sin embargo, por supuesto, quizás haya más de dos factores. La ecuación generalizada es

$$V_{co} = V_A + V_B + \ldots + V_k \tag{34.4}$$

Sustituyendo en la ecuación 34.1 se obtiene

$$V_t = V_A + V_B + \ldots + V_k + V_{esp} + V_e \tag{34.5}$$

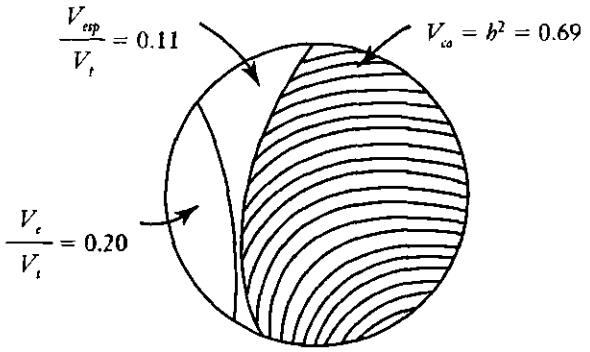
Dividiendo entre V_t, se encuentra la representación proporcional:

$$\underbrace{\overbrace{\frac{V_t}{V_t} = 1.00 = \frac{V_A}{V_t} + \frac{V_B}{V_t}}^{h^2} + \ldots + \frac{V_k}{V_t} + \frac{V_{esp}}{V_t}}_{r_{tt}} + \frac{V_e}{V_t} \tag{34.6}$$

Las partes h^2 y r_{tt} de la ecuación se han denominado de la misma manera que en el capítulo 28. Esta ecuación tiene belleza. Une fuertemente la teoría de medición y la teoría de factores. h^2 es la proporción de la varianza total que es varianza del factor común. r_{tt} es la proporción de la varianza total que es varianza confiable. V_e/V_t es la proporción de la varianza total que es varianza del error. En el capítulo 28 una ecuación como ésta permitió ligar la confiabilidad y la validez. Ahora demuestra la relación entre la teoría de factores y la teoría de medición. Se observa, brevemente, que el *principal problema del análisis factorial consiste en determinar los componentes de varianza de la varianza del factor común total.*

Considere la prueba V en la tabla 34.2. Un vistazo a la ecuación 34.6 indica, entre otras cuestiones, que la confiabilidad de una medida siempre es mayor que, o igual que su comunidad. Entonces, la confiabilidad de la prueba V es, por lo menos, .70. Suponga que $r_{tt} = .80$. Como $V_t/V_t = 1.00$, se pueden especificar todos los términos:

FIGURA 34.1



$$\frac{V_t}{V_t} = 1.00 = \overbrace{(.83)^2 + (.01)^2}^{h^2 = .69} + \overbrace{.11}^{V_{sp}} + \overbrace{.20}^{V_e}$$
$$r_{tt} = .80$$

Entonces, la prueba *V* posee una alta proporción de varianza de factor común y una baja proporción de varianza específica.

Las proporciones se ven claramente en un diagrama circular. Sea el área del círculo de la figura 34.1 igual a la varianza total o 1.00 (100 por ciento del área). Las tres varianzas se han indicado al separar las áreas del círculo. V_{co} o h^2, por ejemplo, es el 69 por ciento, V_{esp} es el 11 por ciento y V_e es el 20 por ciento de la varianza total.

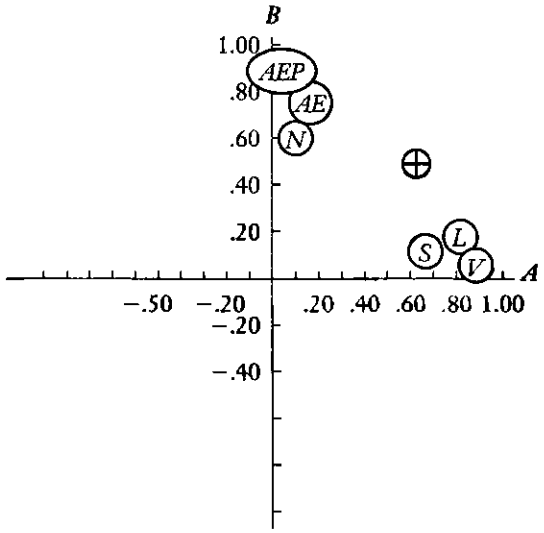
Una investigación analítica factorial que incluya a *V* informaría principalmente sobre V_{co}, la varianza de factor común. Indicaría la proporción de la varianza total de la prueba que es varianza del factor común y ofrecería indicios sobre su naturaleza al indicar qué otras pruebas comparten la misma varianza del factor común, y cuáles no.

Representación gráfica de factores y cargas factoriales

El estudiante del análisis factorial debe aprender a pensar de forma espacial y geométrica, para captar la naturaleza esencial del modelo factorial. Existen varias formas adecuadas para lograrlo. Una tabla de correlaciones se representa por medio del uso de vectores y de los ángulos entre ellos. Aquí se utiliza un método más común. Se trata a las cifras en los renglones de la matriz factorial como coordenadas y se les grafica en un espacio geométrico. En la figura 34.2 se graficaron las cifras de la matriz factorial de la tabla 34.2.

Los dos factores, *A* y *B*, se colocan entre sí en un ángulo recto y se denominan *ejes de referencia*. Los valores apropiados de la carga del factor se señalan en cada uno de los ejes. Entonces cada una de las cargas de prueba se trata como una coordenada y se grafica. Por ejemplo, las cargas de la prueba *L* son (.79, .10). Se recorre hasta .79 sobre el eje *A* y se sube hasta .10 sobre el eje *B*. Este punto se indicó en la figura 34.2 por medio de una letra encerrada en un círculo, la cual indica la prueba. Las coordenadas de las otras cinco pruebas se grafican de manera similar.

FIGURA 34.2



La estructura factorial ahora se percibe con claridad. Cada prueba está altamente cargada en un factor, pero no en el otro. Todas son medidas relativamente "puras" de sus factores respectivos. Un séptimo punto se indicó en la figura 34.2 por medio de una *cruz* encerrada en un círculo, para ilustrar una supuesta prueba que mide ambos factores. Sus coordenadas son (.60, .50), lo cual significa que la prueba está cargada en ambos factores: .60 en *A* y .50 en *B*. No es "pura". Note que estructuras factoriales de esta simplicidad y claridad, donde 1) los factores son ortogonales (los ejes forman ángulos rectos entre sí), 2) las cargas de la prueba son sustanciales y "puras" (casi ninguna prueba se cargó con dos o más factores), y 3) sólo dos factores no son comunes. De nueva cuenta, el lector debe estar consciente de que el ejemplo no contiene datos reales.

La mayoría de los estudios de análisis de factores publicados reportan más de dos factores. Se han reportado cuatro, cinco e inclusive nueve, diez o más factores. La representación gráfica de dichas estructuras factoriales en una sola gráfica, por supuesto, no es posible. Por costumbre, los analistas factoriales grafican dos factores a la vez, aunque es posible graficar tres al mismo tiempo. No obstante, debe admitirse que es difícil visualizar o recordar estructuras *n*-dimensionales complejas. Por lo tanto, se visualizan estructuras bidimensionales y se generalizan a *n* dimensiones de forma algebraica. Un aspecto afortunado de los programas computacionales de análisis factorial es que dicha graficación de factores es fácilmente posible. Comrey (véase Comrey y Lee, 1992) desarrolló un programa gráfico para computadora que permite al usuario graficar dos factores al mismo tiempo.

Extracción y rotación de factores, puntuaciones factoriales y análisis factorial de segundo orden

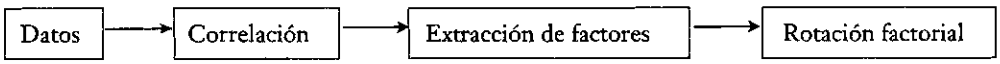
El análisis factorial moderno, tal y como lo define Thurstone (1947), implica cierto número de pasos. El primero consiste en la extracción de factores. Muchos de los métodos producen

factores que no son interpretables. Por consiguiente, dichos factores "sin rotación" se rotan con propósitos de interpretación. Existe un número de métodos de factores extraídos a partir de una matriz de correlación, que incluyen: factores principales, centroide, de probabilidad máxima, residual mínimo, de imagen, poder vectorial y alfa. No es posible analizar aquí todos estos métodos. El propósito de este texto es una comprensión básica y elemental, por lo que la explicación se limitará a uno de los métodos. El método más utilizado actualmente y que está fácilmente disponible en las instalaciones computacionales es el método de factores principales.

El lector puede preguntarse: ¿por qué no usar un método de grupo comparativamente simple, como el modelo de inspección utilizado con anterioridad, en lugar de un método complejo como el de factores principales? Los métodos de grupo pueden utilizarse (véase Lee y MacQueen, 1980) y se recomienda utilizarlos. Dependen de los grupos identificados y de los presuntos factores, por medio de encontrar grupos interrelacionados de coeficientes de correlación u otras medidas de relación. Resulta sencillo localizar los grupos en la tabla 34.1. Sin embargo, en la mayor parte de las matrices **R** no es tan fácil identificar los grupos. Se requieren métodos más objetivos y precisos.

En esta sección se examinan los principales pasos involucrados en el análisis factorial. No será posible presentar todos los detalles necesarios para hacerlo de forma completa. En su lugar, se espera brindarle al lector la esencia del análisis factorial y referirlo a muy buenos libros de texto para los detalles. Se presentarán buenos repasos de dichos textos en la sección de sugerencias de estudio.

Los pasos principales de los estudios que utilizan el análisis factorial se resumen de la siguiente forma:



El problema de la comunalidad y del número de factores

Antes de elegir qué método utilizar para la extracción de los factores, el investigador debe decidir qué va a poner en las casillas diagonales de la matriz de correlación como estimados comunales y cuántos factores va a extraer. Comrey (1978) señala que éstas son las dos decisiones más difíciles que un investigador debe tomar al hacer un análisis factorial. Lo que se utilice como estimados comunales y el número de factores a extraer, puede tener un fuerte impacto en la solución final (véase Comrey y Lee, 1992; Lee y Comrey, 1979; Lee, 1979). Si se conocen y utilizan las comunalidades correctas en el análisis factorial, se obtendrá el número correcto de factores, utilizando el método de factores principales que se describe en la siguiente sección. Por lo tanto, cuando un investigador usa el método de factores principales, los estimados comunales juegan un papel importante en la determinación de la solución factorial obtenida, y deben elegirse de manera cuidadosa. Los programas computacionales definitivamente han hecho que las soluciones factoriales por computadora sean más fáciles. Sin embargo, una de sus principales desventajas radica en que existen determinaciones dadas automáticamente por los programas para computadora, en términos de la forma en que se realiza la extracción factorial. Hubbard y Allen (1989) compararon las soluciones factoriales obtenidas por medio de dos populares programas computacionales, usando los valores predeterminados. Ellos encontraron soluciones muy diferentes. Algunos programas utilizan la regla del valor eigen, donde las unidades se ubican en la diagonal como estimados de las comunalidades y donde se extraen todos los factores que tienen un valor eigen igual o mayor que 1.00. Algunas veces dicho método se conoce como método de *componentes principales truncados*. Lo anterior tiene un

atractivo tanto intuitivo como matemático, ya que parece presentar una solución para ambos problemas. Sin embargo, Comrey y Lee (1992) han alertado en contra del uso indiscriminado de este método. Tiende a inflar demasiado las comunalidades y las cargas factoriales; entonces, las distorsiones se ven amplificadas por la rotación factorial. Aun así, continúa siendo uno de los procedimientos más utilizados. Las pruebas psicológicas y educativas que se desarrollaron gracias al uso de este método, como la *escala de estimación de habilidades sociales* (Social Skills Rating Scale) (Gresham y Elliot, 1990), deben interpretarse con precaución. Comrey y Lee (1992), así como Gorsuch (1983) han alertado sobre la realización del análisis factorial "a ciegas" y la posterior interpretación de los datos como verdaderos. Otro popular estimado de comunalidad es la correlación múltiple al cuadrado, R^2. La investigación realizada por Guttman (1956) demostró que dicho estadístico es el límite inferior para los estimados de las comunalidades y, como tal, podría subestimar los verdaderos valores de las comunalidades. Otros autores recomiendan el uso de la mayor correlación de la variable con otras variables, como el estimado inicial de la comunalidad.

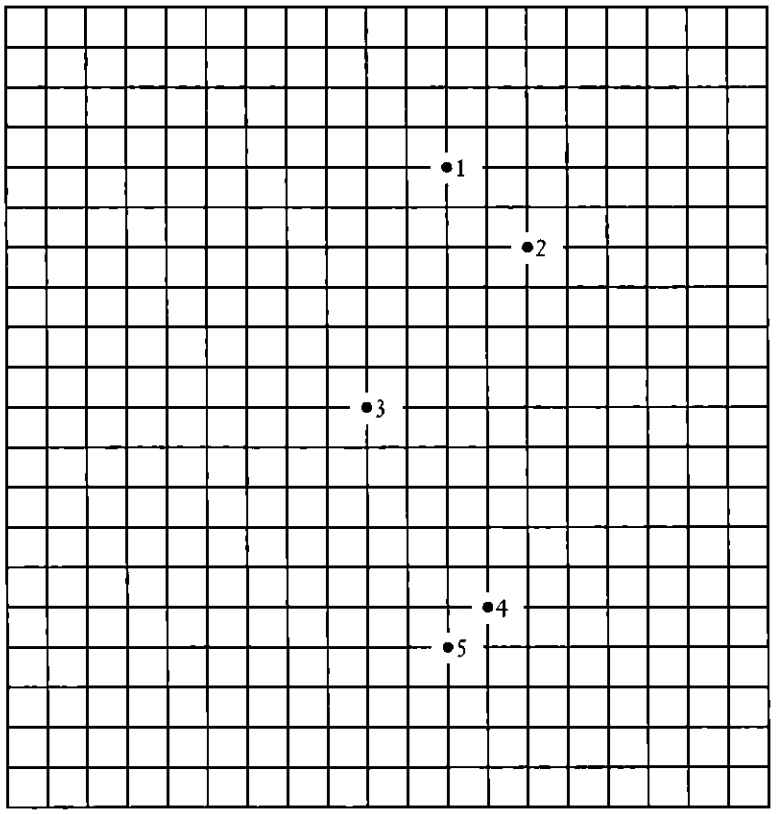
El método de factores principales

El método de factores principales es matemáticamente satisfactorio a causa de que produce una solución matemáticamente única de un problema factorial. Tal vez la principal característica de su solución es que extrae una cantidad máxima de varianza conforme se calcula cada factor. En otras palabras, el primer factor extrae la mayor cantidad de varianza, el segundo la siguiente mayor cantidad de varianza, y así sucesivamente. El primer factor consiste de pesos o coeficientes que maximizarán las correlaciones cuadradas entre las variables y el factor. Entonces la contribución del primer factor se remueve de la matriz de correlación. Entonces esta "nueva" matriz de correlación se usa para encontrar los coeficientes de un factor que maximice las correlaciones cuadradas entre las variables y el segundo factor. Cada factor subsecuente extraído tendrá cada vez menos varianza que el anterior a él. La extracción de factores cesa cuando la varianza se torna insignificante, o cuando el proceso de extracción alcanza el número de factores establecido por el investigador. Cada factor extraído consistirá en coeficientes que no están correlacionados con los coeficientes de los otros factores. En otras palabras, cada factor es independiente de los otros factores.

Es difícil demostrar la lógica del método de factores principales sin demasiadas matemáticas. No obstante, se puede lograr una cierta comprensión intuitiva del método al enfocarlo de manera geométrica. Conciba las pruebas o variables como puntos en un espacio m-dimensional. Las variables que están alta y positivamente correlacionadas deben estar cercanas entre sí, y lejos de las variables con las que no se correlacionan. Si tal razonamiento es correcto, debe haber conjuntos de puntos en el espacio. Cada uno de esos puntos puede ser ubicado en el espacio si se insertan ejes adecuados en éste, es decir, un eje para cada dimensión de las m dimensiones. Entonces, cualquier ubicación de un punto es su identificación múltiple obtenida por medio de la lectura de sus coordenadas en los ejes m. El problema factorial consiste en proyectar los ejes a través de conjuntos vecinos de puntos, para así ubicar aquellos ejes que "explican" tanta varianza de las variables como sea posible.

Es posible demostrar lo anterior con un ejemplo bidimensional simple. Suponga que se tienen cinco pruebas y que están situadas en un espacio bidimensional, tal como se indica en la figura 34.3. Cuanto más cerca estén dos puntos, tendrán mayor relación. El problema es determinar: 1) cuántos factores hay, 2) cuáles pruebas están cargadas en cuáles factores y 3) la magnitud de las cargas de las pruebas.

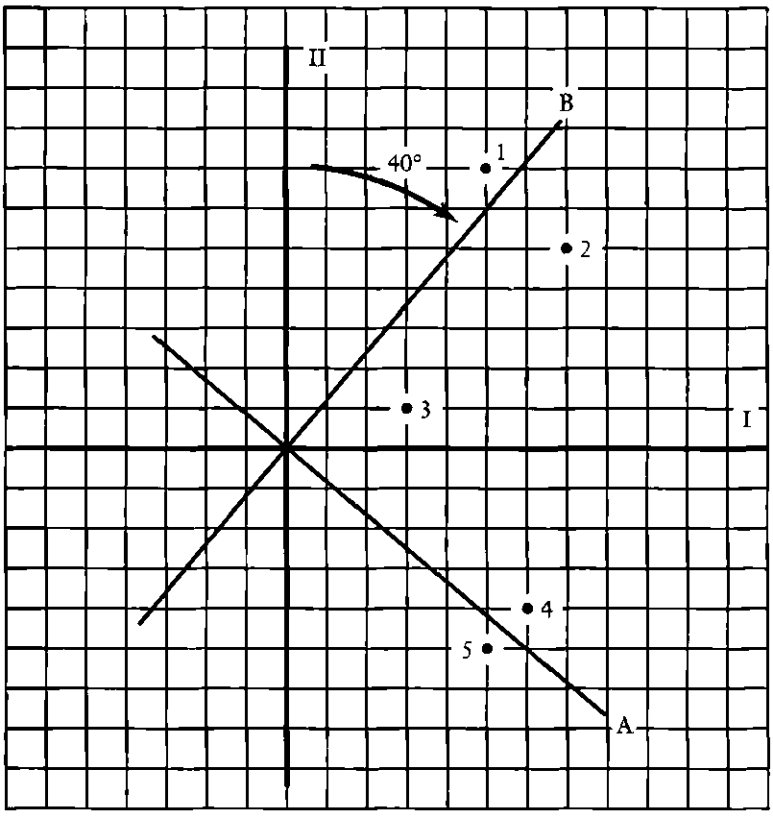
FIGURA 34.3



Ahora el problema se va a resolver de dos maneras diferentes, cada una interesante e instructiva. Primero, se resuelve directamente a partir de los propios puntos. Es menester seguir las siguientes instrucciones. Trace una línea vertical tres unidades a la izquierda del punto 3. Dibuje una línea horizontal debajo del punto 3. Designe estos ejes de referencia I y II. Ahora lea las coordenadas de cada punto, por ejemplo, el punto 2 es (.70, .50), el punto 4 es (.60, –.40). Escriba una "matriz factorial" con estos cinco pares de valores.

Después los ejes se rotan ortogonalmente y en sentido de las manecillas del reloj, de tal manera que el eje I quede entre los puntos 4 y 5. En efecto, el eje II queda entre los puntos 1 y 2. (Se recomienda el uso de un transportador; la rotación debe ser de aproximadamente 40°.) A estos "nuevos" ejes rotados se les llama *A* y *B*. Ahora se corta una tira de papel de cuadrícula pequeña. (Los puntos se grafican en papel cuadriculado.) Considere la base de cada cuadrado como .10 (.10 = ¼ de pulgada; 10 unidades de hecho son iguales a 1.00). Utilizando la tira como instrumento de medición, se miden las distancias de los puntos con los nuevos ejes. Por ejemplo, el punto 2 debe estar cerca de (.22, .83), y el punto 5 debe estar cerca de (.71, –.06). (No importa si existen pequeñas discrepancias.) Los ejes originales (I y II) y los rotados (*A* y *B*) y los cinco puntos se presentan en la figura 34.4.

FIGURA 34.4



Ahora se escriben ambas matrices factoriales, sin rotar y rotadas. Éstas se presentan en la tabla 34.4. El problema está resuelto: hay dos factores. Los puntos (pruebas) 1 y 2 tienen cargas altas del factor *B*, los puntos 4 y 5 tienen cargas altas del factor *A*, y el punto 3 tiene cargas bajas de ambos factores. Se respondieron las tres preguntas planteadas originalmente.

TABLA 34.4 *Matrices sin rotación y rotada, problema de la distancia de los puntos*[a]

Puntos	Sin rotación		Puntos	Rotada	
	I	II		A	B
1	.50	.70	1	−.07	*.86*
2	.70	.50	2	.22	*.83*
3	.30	.10	3	.17	.27
4	.60	−.40	4	*.72*	.08
5	.50	−.50	5	*.71*	−.06

[a] Las cargas con rotación sustanciales están en *itálicas*.

TABLA 34.5 *Matrices factoriales sin rotación y rotadas, la matriz R de la tabla 36.1*[a]

	Sin rotación		Rotadas		
Pruebas	I	II	*A*	*B*	h^2
V	.60	–.58	*.83*	.01	.70
L	.63	–.49	*.79*	.10	.63
S	.56	–.43	*.70*	.10	.50
N	.56	.43	.10	*.70*	.50
AE	.63	.49	.10	*.79*	.63
AEP	.60	.58	.01	*.83*	.70

[a] Las cargas significativas > 2.30 están en *itálicas*.

Este procedimiento es análogo a los problemas factoriales psicológicos. Las pruebas se consideran como puntos en un espacio factorial *m*-dimensional. Las cargas factoriales son las coordenadas. El problema es introducir marcos de referencia o ejes apropiados y después "leer" las cargas factoriales. Por desgracia, en problemas reales no se conoce el número de factores (la dimensionalidad del espacio factorial y, por lo tanto, el número de ejes) o la ubicación de los puntos en el espacio, los cuales es necesario determinar a partir de los datos.

La descripción anterior es figurativa. No se "leen" las cargas factoriales a partir de los ejes de referencia; se calculan utilizando métodos bastante complejos. El método de factores principales en realidad involucra la solución de ecuaciones lineales simultáneas. Las raíces obtenidas de las soluciones se denominan *valores eigen*. También se obtienen *vectores eigen*; después de la transformación adecuada, se convierten en cargas factoriales. Así se resolvió la matriz **R** ficticia de la tabla 34.1 y produjo la matriz factorial que se presentó después en la tabla 34.5. La mayoría de los programas de análisis computacionales usan las soluciones de factores principales. El estudiante que tenga la expectativa de utilizar el análisis factorial con cualquier profundidad, debe estudiar el método cuidadosamente y, por lo menos, entender lo que hace. No hay nada tan riesgoso y autoderrotante como usar ciegamente programas computacionales. En especial, éste es el caso en el análisis factorial.

Rotación y estructura simple

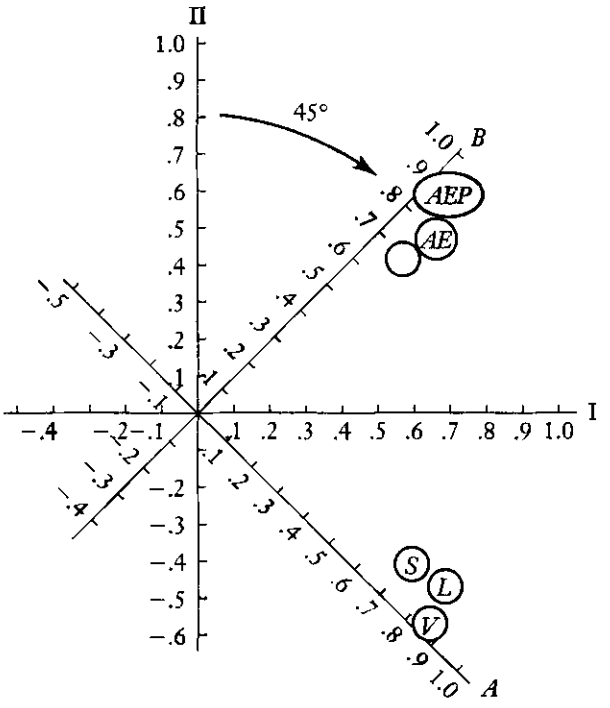
La mayoría de los métodos de extracción factorial producen resultados de tal forma que son difíciles o imposibles de interpretar. Lo anterior se percibe al observar los factores sin rotación de la tabla 34.4. Thurstone (1947, pp. 508-509) comentó que era necesario rotar las matrices factoriales si se deseaba interpretarlas de manera adecuada. Señaló que las matrices factoriales originales son arbitrarias en el sentido de que es posible encontrar un infinito número de marcos de referencia (ejes), para reproducir cualquier matriz **R** dada (véase Thurstone, 1947, p. 93). Una matriz de factores principales y sus cargas explican la varianza del factor común de las puntuaciones de la prueba; pero en general no proporcionan estructuras con un significado científico. Son las configuraciones de las pruebas o variables en el espacio factorial las que tienen una importancia fundamental. Para descubrir tales configuraciones de manera adecuada, deben rotarse los ejes de referencia arbitrarios. En otras palabras, se supone que existen posiciones únicas y "mejores" de los ejes, "mejores" formas de ver las variables en el espacio *n*-dimensional.

No existe aquí la intención de materializar constructos, variables o factores. Los factores son meras estructuras o patrones generados por covarianzas de mediciones. Lo que se quiere decir con "mejores maneras de ver las variables" es la manera más parsimoniosa y más simple. Una "mejor" forma se predice a partir de la teoría y de las hipótesis. O una "mejor" forma se descubre a partir de una estructura tan clara y fuerte que casi obligue a creer en su validez y "realidad".

Entre las importantes contribuciones de Thurstone, la invención de las ideas de la estructura simple y de la rotación de los ejes factoriales son, tal vez, las más importantes. Con ellas, él estableció lineamientos relativamente claros para lograr soluciones analíticas factoriales con significado psicológico e interpretables. En la tabla 34.2 se reportó una matriz factorial obtenida a partir de la matriz **R** de la tabla 34.1. Ésta era la matriz ***rotada*** **final y no la matriz producida originalmente por el análisis factorial.** La matriz ***sin rotación***, originalmente producida por medio del método de factores principales, se presenta en la parte izquierda de la tabla 34.5. Los factores rotados se reproducen en la parte derecha de la tabla. También se muestran las comunalidades (h^2) que son iguales en las dos matrices.

Si se intenta interpretar la matriz sin rotación en la parte izquierda de la tabla, se enfrenta un problema. Se puede decir que todas las pruebas se cargan de forma sustancial sobre un factor I general; y que el segundo factor, II, es bipolar. (Un *factor bipolar* es aquel que tiene cargas sustanciales positivas y negativas.) Lo anterior equivale a decir que todas las pruebas miden lo mismo (factor I), pero que las tres primeras miden el aspecto negativo de lo que sea que miden las otras tres (factor II). Pero aparte de la naturaleza ambigua

FIGURA 34.5



de una interpretación como ésta, se sabe que los ejes de referencia, I y II, y en consecuencia las cargas factoriales, son arbitrarios. Observe la gráfica factorial de la figura 34.2. Existen dos grupos de pruebas claramente definidos, pegados cerca de los ejes *A* y *B*. Aquí *no* hay un factor general, tampoco hay un factor bipolar. El segundo problema principal del análisis factorial, por lo tanto, consiste en descubrir una solución única y convincente o la posición de los ejes de referencia.

Si se grafican las cargas de I y II se "observa" la estructura original sin rotación. Esto se hizo en la figura 34.5. Ahora se giran los ejes de tal manera que I queda lo más cerca posible de los puntos *V, L* y *S*, al mismo tiempo, II queda lo más cerca posible a los puntos *N, AE* y *AEP*. Una rotación de 45° será adecuada. Entonces se obtiene, en esencia, la estructura de la figura 34.2. Es decir, las nuevas posiciones con los ejes rotados y las posiciones de las seis pruebas son iguales a las posiciones de los ejes y pruebas de la figura 34.2. La estructura simplemente se inclina a la derecha. Si se gira la figura, de tal manera que la *B* del eje *B* apunte directamente hacia arriba, esto se hace evidente. Ahora es posible leer las nuevas cargas factoriales rotadas sobre los ejes rotados. Puesto que los ejes se mantienen en un ángulo recto de 90°, se denomina rotación *ortogonal*.

Este ejemplo, aunque poco realista, puede ayudar al lector a comprender que los analistas factoriales buscan las unidades que presuntamente subyacen al desempeño de las pruebas. Concebido de forma espacial, buscan las relaciones entre variables "afuera" en el espacio factorial multidimensional. Por medio del conocimiento de las relaciones empíricas entre pruebas u otras medidas, exploran el espacio factorial con los ejes de referencia hasta que encuentran las unidades o relaciones entre relaciones —si es que existen—.

Las cargas significativas ($\geq .30$) están en itálicas. Note que los vectores *A* y *B* están invertidos en esta tabla. Las h^2 calculadas a partir de los valores sin rotación y con rotación difieren ligeramente, a causa de los errores de redondeo; por ejemplo, $.60^2 + .58^2 = .70$, y $.83^2 + .01^2 = .69$. En la tabla se utilizaron los valores exactos obtenidos por computadora (y en la tabla 34.2).

Para dirigir las rotaciones, Thurstone estableció cinco principios o reglas de la estructura simple. Las reglas son aplicables tanto para las rotaciones ortogonales como para las oblicuas; aunque Thurstone enfatizó el caso oblicuo. (Las rotaciones oblicuas son aquellas donde los ángulos entre los ejes son agudos y obtusos.) Los principios de la estructura simple son los siguientes:

1. Cada renglón de la matriz factorial debe tener por lo menos una carga cercana a cero.
2. Por cada columna de la matriz factorial debe haber por lo menos tantas variables con cargas iguales o cercanas a cero como factores.
3. Por cada par de factores (columnas) debe haber diversas variables con cargas en un factor (columna), pero no en el otro.
4. Cuando haya cuatro o más factores, una gran proporción de las variables debe tener cargas insignificantes (cercanas a cero) en cualquier par de factores.
5. Por cada par de factores (columnas) de la matriz factorial debe haber sólo un pequeño número de variables con cargas sustanciales (diferentes de cero), en ambas columnas.

En efecto, dichos criterios demandan variables que sean lo más "puras" posibles; es decir, que cada variable esté cargada en el menor número de factores posibles y que haya *la mayor cantidad de ceros posibles en la matriz factorial rotada*. De esta forma, es posible lograr la interpretación más simple posible de los factores. En otras palabras, la rotación para lograr la estructura más simple es una manera bastante objetiva de lograr la simplicidad de variables o reducir la complejidad de las variables.

Para comprender lo antes expuesto, imagine una solución ideal en donde la estructura simple sea "perfecta". Podría verse de la siguiente forma, por ejemplo, en una solución de tres factores, la cual se presenta en la siguiente página.

Las X indican cargas factoriales sustanciales, los 0 indican cargas cercanas a cero. Por supuesto, dichas estructuras factoriales "perfectas" son poco frecuentes. Es más probable que algunas de las pruebas tengan cargas en más de un factor. Aun así, se han logrado buenas aproximaciones a la estructura simple, especialmente en estudios analíticos factoriales bien planeados y bien ejecutados. Comrey (1978) señala que la estructura simple funciona bien si el estudio está bien diseñado, con varios factores bien definidos y donde cada uno se mide a través de varias medidas factoriales puras que están distribuidas de manera normal y con alta confiabilidad. Sin embargo, Comrey también afirma que estudios sin diseño o mal diseñados tendrán variables complejas; y como resultado, la solución no se ajustará muy bien a la estructura simple.

Antes de abandonar el tema de la rotación factorial, debe señalarse que existe una serie de métodos de rotación. Los dos tipos principales de rotación son los llamados "ortogonal" y "oblicuo". Las rotaciones *ortogonales* mantienen la independencia de los factores; es decir, los ángulos entre los ejes se mantienen a 90°. Por ejemplo, si se rotan los factores I y II de forma ortogonal, se giran juntos ambos ejes, manteniendo el ángulo recto entre ellos. Esto quiere decir que la correlación entre los factores es cero. La rotación realizada en la figura 34.5 es ortogonal. Si se tuvieran cuatro factores, se rotarían I y II, I y III, I y IV, II y III, etcétera, manteniendo ángulos rectos entre cada par de ejes. Algunos investigadores prefieren hacer una rotación ortogonal. Otros insisten en que la rotación ortogonal no es realista, que los factores reales por lo general están correlacionados, y que las rotaciones deben ajustarse a la "realidad" psicológica.

Las rotaciones donde los ejes factoriales permiten formar ángulos agudos u obtusos se denominan *oblicuas*. Por supuesto, oblicuo significa que los factores están correlacionados. No cabe duda de que las estructuras factoriales pueden ajustarse mejor con ejes oblicuos y que los criterios de la estructura simple se cumplen mejor. Algunos investigadores objetan los factores oblicuos debido a la posible dificultad al comparar estructuras factoriales de un estudio a otro. Se deja el polémico tema con dos señalamientos. Primero, el tipo de rotación parece ser una cuestión de gusto. Segundo, el lector necesita comprender ambos tipos de rotación al grado de que pueda interpretar ambos tipos de factores, y ser particularmente cuidadoso al enfrentarse con los resultados de soluciones oblicuas, los cuales contienen particularidades y sutilezas que no están presentes en las soluciones ortogonales.

La rotación factorial que se ha estudiado hasta ahora constituye el modelo gráfico. Antes de la existencia de las computadoras digitales de alta velocidad, los investigadores que realizaban análisis factoriales usaban dicho método gráfico o manual de rotación. La naturaleza imprecisa de las rotaciones gráficas por medio de aproximaciones visuales, era una de las principales críticas al método. Sin embargo, conforme las computadoras se volvieron más precisas y confiables a la mitad de los años cincuenta, emergió un número de métodos analíticos de rotación, donde las rotaciones se realizaban utilizando una fórmula matemática. El más popular fue el desarrollado por Henry Kaiser (1958) llamado Varimax. Virtualmente cada paquete computacional que lleva a cabo análisis factoriales en la actualidad, utiliza este método de rotación. En muchos de tales programas el método Varimax se utiliza de forma automática. En la esquela de defunción de Kaiser, Jensen y Wilson (1994) afirmaron que el artículo de Kaiser de 1958 es el tercer artículo más citado en la literatura psicológica. Varimax funciona muy bien en la aproximación de la estructura simple en estudios analíticos factoriales bien diseñados. Si un investigador está buscando un factor general, el método Varimax no funciona muy bien. La meta del método Varimax consiste en dispersar la mayor cantidad de varianza a través de los factores y, al mismo

Pruebas	A	B	C
1	X	0	0
2	X	0	0
3	X	0	0
4	0	X	0
5	0	X	0
6	0	X	0
7	0	0	X
8	0	0	X
9	0	0	X

tiempo, tratar de obtener la estructura simple. Si el investigador incluye demasiados factores al emplear el método Varimax, entonces un resultado posible es el incremento artificial de los factores pequeños. Por lo tanto, Varimax se vuelve sensible al número de factores utilizados en la rotación.

Si un investigador está interesado en encontrar un factor general, el mejor método ortogonal disponible es el *criterio de tándem de Comrey* (1967), que consiste en dos pasos, cada uno de los cuales se basa en un principio diferente.

> Principio 1: Si dos variables están correlacionadas, deben aparecer en el mismo factor (criterio 1).
> Principio 2: Si dos variables no están correlacionadas, no deben aparecer en el mismo factor (criterio 2).

El *criterio* o *principio 1* es el más interesante si se busca un factor general. El *criterio 1* intenta dispersar la varianza desde los factores más grandes hasta los más pequeños, mientras satisface el *principio 1* al mismo tiempo. Si existe un factor general, las variables que estén correlacionadas entre sí, serán retenidas lo más posible en el mismo factor, en lugar de dispersarse alrededor.

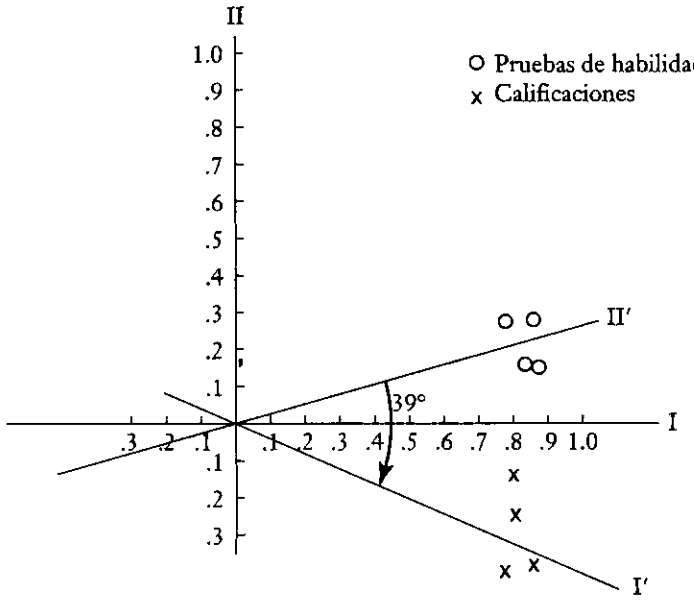
Del mismo modo en que existen muchos métodos de extracción de factores, también hay muchos métodos de rotación, además de los que ya se han mencionado. Existen diferentes métodos tanto para la rotación ortogonal como para la oblicua. El lector puede consultar a Gorsuch (1983) y a Comrey y Lee (1992), donde encontrará una lista.

Análisis factorial de segundo orden

El análisis factorial de segundo orden es un método sumamente importante, aunque rechazado, para el análisis de datos complejos y la comprobación de hipótesis. Cuando se rotan los datos de forma oblicua, existen correlaciones entre los factores. Antes en el presente capítulo se mencionó la ecuación fundamental del análisis factorial. Una versión de ella era $\mathbf{R} = \mathbf{P}\,\Phi\,\mathbf{P}^T + \mathbf{U}$. La matriz Φ contiene la correlación entre factores. En las rotaciones ortogonales, la matriz Φ no se utiliza debido a que los factores no están correlacionados. Para realizar un análisis factorial de segundo orden de la manera tradicional, dicha matriz Φ se analiza factorialmente. Por completar, la matriz $\mathbf{P}$ es la matriz del patrón factorial; contiene las cargas factoriales. $\mathbf{P}^T$ es su transposición y $\mathbf{U}$ es la matriz que contiene la singularidad de cada variable.

En un estudio provocativo de análisis factorial y de correlación canónica, sobre la redundancia presente en puntuaciones de pruebas de estudiantes, Lohnes y Marshall (1965)

FIGURA 34.6



extrajeron dos factores de 21 pruebas de habilidad y rendimiento. Las cargas sin rotación de ocho de sus medidas, cuatro pruebas de habilidad y cuatro calificaciones (inglés, aritmética, estudios sociales, ciencia) se graficaron en la figura 34.6. Los ejes se rotaron de forma oblicua, de tal manera que quedaran entre los dos grupos de cargas. Existe un ángulo agudo de aproximadamente 39° entre los ejes rotados, ahora denominados I′ y II′. Cualquier ángulo entre los ejes que no sea de 90° implica la existencia de correlación entre los factores. En este caso, la correlación es bastante alta, aproximadamente de .78.

Imagine la situación anterior multiplicada por seis, ocho o 10 factores: habría un conjunto de correlaciones entre los factores. Si se analizan factorialmente estas correlaciones se tiene un análisis factorial de segundo orden, que es un método para encontrar los factores que subyacen a los factores. El famoso componente *g* de las pruebas de inteligencia es, evidentemente, un factor de segundo orden o de orden mayor. Siempre que se analizan factorialmente grandes cantidades de pruebas de habilidad, las correlaciones entre las pruebas son, por lo general, positivas. Si se analizan factorialmente surge un patrón como éste, aunque más complejo, tal como sucedió en la figura 34.6. Si se calculan las correlaciones entre los factores y se analizan factorialmente de nuevo, puede surgir un solo factor, quizás *g*.

Para obtener información adicional sobre el análisis factorial de segundo orden y de orden mayor véase Gorsuch (1983). El investigador puede realizar el proceso para encontrar factores oblicuos de primer orden y después analizar factorialmente la matriz de correlación de factores; aunque existe un método alternativo. Con la disponibilidad de programas para computadora, tales como LISREL, EQS y AMOS, el investigador puede realizar un análisis factorial de orden superior de forma bastante fácil y en un solo paso. Comrey y Lee (1992) demuestran cómo realizar un análisis factorial de segundo orden utilizando el programa EQS.

Puntuaciones factoriales

Mientras que el análisis factorial de segundo orden está más orientado hacia la investigación básica y teórica, otra técnica de análisis factorial, las llamadas puntuaciones o medidas factoriales, es eminentemente práctica, pero no sin importancia teórica. Las *puntuaciones factoriales* son medidas de los individuos en los factores. Suponga que, como lo hicieron Lohnes y Marshall (1965), se encuentran dos factores detrás de 21 medidas de habilidad y de calificación. En lugar de utilizar las 21 puntuaciones de los grupos de niños en investigación, ¿por qué no utilizar sólo dos puntuaciones calculadas a partir de los factores? Lohnes y Marshall recomiendan hacerlo así, y señalan la redundancia en las puntuaciones usuales de los alumnos. En efecto, dichas puntuaciones factoriales son promedios ponderados: ponderados de acuerdo a las cargas factoriales.

A continuación se presenta un ejemplo sobresimplificado. Suponga que los datos de la matriz factorial de la tabla 34.2 fueran reales y que se desea calcular las puntuaciones factoriales *A* y *B* de un individuo. Las puntuaciones en bruto de un individuo en las seis pruebas son, por ejemplo: 7, 5, 5, 3, 4 y 2. Se multiplican tales puntuaciones por las cargas factoriales relacionadas, primero para el factor *A* y después para el factor *B*, de la siguiente manera:

$$A\text{: } FA = (.83)(7) + (.79)(5) + (.70)(5) + (.10)(3) + (.10)(4) + (.01)(2) = 13.98$$

$$B\text{: } FB = (.01)(7) + (.10)(5) + (.10)(5) + (.70)(3) + (.79)(4) + (.83)(2) = 7.99$$

Las "puntuaciones factoriales" del individuo son $FA = 13.98$ y $FB = 7.99$. Por supuesto, es posible calcular las "puntuaciones factoriales" de otros individuos de manera similar.

Ésta no es la mejor forma para calcular las puntuaciones factoriales. Comrey y Lee (1992), Gorsuch (1983) y Harman (1976) presentan métodos alternativos para calcular las puntuaciones factoriales. Ellos también explican las ventajas y desventajas de cada método. Sin embargo, el ejemplo aquí presentado fue inventado solamente para trasmitir la idea de dichas puntuaciones como sumas o promedios ponderados, donde los pesos son las cargas factoriales. En cualquier caso, aunque el método no fue utilizado de forma extensa en el pasado, tiene un gran potencial para la investigación compleja del comportamiento. En lugar de utilizar muchas puntuaciones separadas, se utiliza un menor número de puntuaciones factoriales. Un excelente ejemplo real es el descrito por Mayeske (1970), quien participó en un nuevo análisis de los datos del reporte de gran influencia de Coleman, Campbell, Hobson, *et al.* (1966) llamado *Equality of Educational Opportunity.*

Ejemplos de investigación

La mayoría de los estudios analíticos factoriales han factorizado las pruebas y escalas de inteligencia, aptitud y personalidad, donde se han intercorrelacionado y analizado factorialmente las propias pruebas y escalas. El ejemplo de Thurstone que se explicó anteriormente es un excelente ejemplo; de hecho, es un clásico. También se pueden factorizar las personas o sus respuestas. De hecho, las variables incluidas en las matrices de correlación y factoriales pueden ser pruebas, escalas, personas, reactivos, conceptos o cualquier cuestión que pueda estar intercorrelacionada. Los estudios descritos más adelante han sido seleccionados no para representar investigaciones analíticas factoriales en general, sino más bien para familiarizar al estudiante con los diferentes usos del análisis factorial.

Las escalas de personalidad de Comrey

El trabajo de Comrey (1970) sobre investigación de la personalidad constituye uno de los mejores ejemplos sobre el uso del análisis factorial. Las escalas de personalidad de Comrey, también conocidas como el CPS (por sus siglas en inglés), conforman un inventario de rasgos de personalidad factorizados. Dicha taxonomía de rasgos de personalidad se desarrolló durante un periodo de 15 años. Originalmente se inspiró en las discrepancias existentes entre reconocidos autores de pruebas de personalidad y teóricos de la personalidad. Lo que inició como un esfuerzo para resolver las diferencias entre los teóricos de la personalidad finalizó con el surgimiento de las escalas de personalidad de Comrey. Las CPS comparten algunas de las características de las otras pruebas; pero difieren de ellas.

Para obtener una solución factorial estable y control de la jerarquía factorial, Comrey desarrolló una unidad de medición llamada "dimensión del reactivo homogéneo factorizado" (DRHF), que se desarrolló para resolver algunos de los problemas experimentados en los reactivos factorizados. Un problema se refiere a la falta de confiabilidad asociada con reactivos únicos. La otra se refiere a la extracción de factores de reactivos que sean factores de bajo nivel, consistentes de reactivos que sean similares en su redacción u otra características distinguible. Tales factores de reactivos de bajo nivel, por lo general, no ofrecen al investigador mucha información sobre el rasgo de personalidad subyacente. La DRHF no es más que la suma de las puntuaciones de los reactivos que definen la DRHF. Puesto que la DRHF es una suma, es más confiable que cualquier reactivo solo.

Un análisis factorial de la DRHF produce ocho factores. Los nombres de estos factores son *confianza* contra *defensividad*; *disciplina* contra *falta de compulsión*; *conformidad social* contra *rebelión*; *actividad* contra *falta de energía*; *estabilidad emocional* contra *neurosis*; *extraversión* contra *introversión*; *fuerza mental* contra *sensibilidad*; y *empatía* contra *egocentrismo*.

Información adicional sobre los pasos y procedimientos del desarrollo de las CPS puede encontrarse en Comrey y Lee (1992), Comrey (1980) y Comrey (1988). De los años setenta a los años noventa Comrey y sus colegas han validado dicha estructura de ocho factores en varias culturas diferentes y países diferentes. La investigación actual realizada por Comrey sobre las CPS demuestra todos los pasos correctos que toma un investigador al realizar un estudio analítico factorial.

Estudio factorial de Thurstone sobre la inteligencia

Thurstone y Thurstone (1941), en su trabajo monumental sobre los factores de inteligencia y su medición, analizaron factorialmente 60 pruebas además de las variables *edad cronológica*, *edad mental* y *sexo*. El análisis se basó en las respuestas de 710 alumnos de primer año de secundaria a 60 pruebas y reveló, esencialmente, el mismo conjunto de los llamados factores primarios, que se habían encontrado en estudios previos.

Los Thurstone eligieron las tres mejores pruebas de cada uno de siete de los 10 factores primarios. Seis de estas pruebas parecían tener estabilidad a distintos niveles de edad, para usos escolares prácticos. Entonces, ellos revisaron y administraron las pruebas a 437 estudiantes de segundo grado de secundaria. El principal propósito del estudio consistía en verificar la estructura factorial de las pruebas. En otras palabras, ellos predijeron que los mismos factores primarios de inteligencia puestos en las 21 pruebas surgirían de un nuevo análisis factorial, en una nueva muestra de niños.

Inteligencia fluida y cristalizada

Uno de los problemas más activos, importantes y polémicos de interés científico y práctico del comportamiento es la naturaleza de las habilidades mentales. Diferentes teorías con diversas cantidades de tipos de evidencia que las soporten han sido propuestas por algunos de los psicólogos más sobresalientes del siglo: Spearman, Thurstone, Burt, Thorndike,

Guilford, Cattell y otros. No cabe ninguna duda respecto a la gran importancia científica y práctica del problema. Se ha aludido, aunque sólo brevemente, al trabajo y pensamiento de Thurstone y Guilford. Ahora se describirá, también de manera breve, uno de los muchos estudios analíticos factoriales de Cattell (1963).

Puede mostrarse que el famoso factor general de inteligencia, *g*, es un factor de segundo orden que aparece en la mayoría de las pruebas de habilidad mental. En efecto, Cattell considera que existen dos *g*, o dos aspectos de *g*: el cristalizado y el fluido. La *inteligencia cristalizada* se exhibe por medio de desempeños cognitivos donde "hábitos de juicio hábiles" se han fijado o cristalizado, debido a la aplicación anterior de la habilidad de aprendizaje general a dichos desempeños. Los reconocidos factores verbal y numérico son ejemplos de ello. Por otro lado, la *inteligencia fluida* se exhibe por medio de desempeños caracterizados más por la adaptación a situaciones nuevas, la aplicación "fluida" de la habilidad general, por así decirlo. Dicha habilidad es más característica del comportamiento creativo que la inteligencia cristalizada. Si los factores se analizan factorialmente y las correlaciones entre factores que se encuentren se factorizan a su vez (análisis factorial de segundo orden), entonces tanto la inteligencia cristalizada como la fluida deben surgir como factores de segundo orden.

Cattell administró la prueba de habilidades primarias de Thurstone y un número de sus propias pruebas de habilidad mental y personalidad a 277 niños de segundo grado de secundaria, después analizó factorialmente las 44 variables y rotó los 22 factores obtenidos (probablemente demasiados) a la estructura simple. Las correlaciones entre dichos factores fueron, a su vez, factorizadas, produciendo ocho factores de segundo orden. (Recuerde que las rotaciones oblicuas producen factores que están correlacionados.) A pesar de que Cattell incluyó un número de variables de personalidad, aquí se enfocan sólo los primeros dos factores: inteligencia fluida e inteligencia cristalizada. Él pensó que las pruebas de Thurstone debían cargarse en un factor general, puesto que miden habilidades cognitivas cristalizadas; y que sus propias pruebas relacionadas con la cultura debían cargarse en otro factor, debido a que miden habilidad fluida. Y así sucedió. Los dos conjuntos de cargas factoriales se indican en la tabla 34.6, junto con los nombres de las pruebas. Los dos factores estuvieron también correlacionados positivamente ($r = .47$), como se predijo.

Dicho estudio demuestra el poder de una inteligente combinación de teoría, construcción de pruebas y análisis factorial. Similar a la también inteligente conceptualización

TABLA 34.6 *Parte de la matriz factorial de segundo orden (estudio de la inteligencia fluida y cristalizada de Cattell)*[a]

	$F_1(gf)$	$F_2(gc)$
Pruebas de Thurstone:		
Verbal	.15	*.46*
Espacial	*.32*	.14
Razonamiento	.08	*.50*
Numérica	.05	*.59*
Fluidez	.07	.09
Pruebas de Cattell:		
Series	*.35*	*.43*
Clasificación	*.63*	−.02
Matrices	*.50*	.10
Topología	*.51*	.09

[a] *gf* = factor general fluido; *gc* = factor general cristalizado. Las itálicas fueron añadidas por el autor (FNK). Éstos son sólo dos de los ocho factores de Cattell.

y análisis de factores divergentes, convergentes y de otros tipos ya mencionados, realizada por Guilford, se trata de una significativa contribución al conocimiento psicológico de un tema extremadamente complejo e importante. Sin embargo, para obtener una visión completa, también debe leerse el artículo de Humphreys (1967) que critica la teoría de Cattell.

Análisis factorial confirmatorio

Los modelos de análisis factorial descritos anteriormente hasta este punto representan procedimientos tradicionales que, por lo general, ahora se conocen como "análisis factorial exploratorio" o AFE. Se han creado métodos más novedosos con una base más fuerte sobre la teoría de comprobación de hipótesis, los cuales se conocen como "análisis factorial confirmatorio". Sin embargo, existen pequeñas variantes del análisis factorial confirmatorio. Existen los iniciales, basados en el AFE y los más nuevos basados en teoría estadística más estricta. Los métodos más nuevos serán referidos como AFC.

Anteriormente en este capítulo se enfatizó que los métodos analíticos factoriales exploratorios son muy poderosos cuando se utilizan para la comprobación de hipótesis. Es decir, cuando se desarrollan hipótesis tanto acerca de los factores que se busca encontrar en cierto dominio como acerca de las variables que los miden. Se eligen diversas variables para cada factor hipotetizado que debe proporcionar medidas relativamente puras de ese factor. Se reúnen datos para una muestra grande y se analizan factorialmente para ver qué tan bien los factores obtenidos y las variables cargadas en ellos corresponden a la estructura factorial originalmente hipotetizada. Con base en el primer análisis, se efectúan revisiones en la hipótesis y en las variables designadas para medir cada factor, y se repite el estudio. Dicho proceso se repite de manera pragmática hasta que la estructura factorial que emerge corresponde razonablemente bien a la estructura factorial hipotetizada con anticipación.

Se trata de un método que representa el tipo anterior de análisis factorial confirmatorio donde el número de factores que surgen a partir del análisis no está restringido a un número preconcebido. Si el número "correcto" resulta ser el que se hipotetizó, eso está bien, pero no se especifica con antelación. Además, se permite que las cargas caigan donde sea, en lugar de ser forzadas a conformar lo más posible un patrón especificado previamente. En particular, grandes números de parámetros no se fuerzan hacia cero. La proporción de la varianza atribuida a factores únicos para cada variable surge como un resultado final del análisis, en lugar de ser un parámetro, *per se*, a estimarse. Así, en el AFE se evalúa qué tan bien se ajusta la solución obtenida a un patrón factorial preconcebido; aunque sin el uso de poderosas técnicas de optimización que fuercen un ajuste que puede tener un débil sostén con datos nuevos.

Algunos de esos métodos iniciales AFE del análisis factorial confirmatorio se denominan soluciones *de rotación forzada o soluciones procusteanas** (véase Comrey y Lee, 1992). Dicho método es susceptible de adoptar diversas formas. Se puede rotar una matriz no rotada de la manera más cercana posible a una matriz meta, ya sea ortogonal u oblicua. La matriz meta podría ser una matriz hipotética basada en la teoría o en las expectativas desarrolladas a partir de investigación previa. Por lo general, se utilizan métodos de mínimos cuadrados para encontrar la matriz de transformación que logrará las rotaciones de-

*Nota del revisor técnico: Como lo mencionan Nunnally y Bernstein en su libro *Teoría psicométrica* de esta misma editorial: "El nombre Procusto viene de un posadero de la mitología griega que tenía una cama que se ajustaría al tamaño de cualquiera. Si el visitante era demasiado pequeño para la cama, Procusto alargaba al visitante en un potro. Si un visitante era demasiado alto para caber en la cama, Procusto acortaba la longitud de las piernas del visitante para que se adaptara a la cama" (pág. 630).

seadas. El estudio de Verba y Nie (1972) constituye un ejemplo de un análisis factorial confirmatorio realizado con el uso de tal método.

Esos antiguos métodos de análisis factorial confirmatorio tal vez se volverán menos populares, a medida que los métodos más novedosos sean capaces de realizar el mismo tipo de tarea en la mayoría de los casos y, además, proporcionen una prueba estadística de bondad de ajuste, así como indicaciones sobre cómo mejorar el modelo.

Los métodos más nuevos, que aquí se denominan AFC, se basan en el trabajo de Lawley (1940), quien introdujo el método de probabilidad máxima al análisis factorial y posteriormente lo desarrolló (véase Lawley y Maxwell, 1971). Gorsuch (1983) señala que el AFC tiene sus raíces en el método de análisis factorial de probabilidad máxima de Maxwell-Lawley. Dichos métodos nuevos por lo común destacan por la ausencia de rotación factorial. El AFC es, en realidad, un caso especial de un conjunto más general de métodos de análisis estadístico que se conocen como *análisis estructural de covarianza*. En esta sección se proporciona una introducción al AFC y se muestra cómo difiere del AFE.

Según diversos autores, en su forma actual, el AFC se atribuye a Joreskog y sus colegas (Joreskog, 1967, 1969, 1970; Joreskog y Goldberger, 1972), aunque Bock y Bargmann (1966) sugirieron algo similar en una fecha anterior. El procedimiento de Beck y Bargmann requiere que el investigador especifique todos los parámetros en la matriz de cargas factoriales, la matriz de correlaciones entre los factores y la matriz de varianza única. Con el empleo de tales matrices especificadas previamente, se crea una matriz de correlación estimada o de covarianza. Después esta matriz se compara con la correlación muestra o matriz de covarianza por medio de un estadístico de bondad de ajuste, tal como la prueba de Bartlett (véase Morrison, 1967). Dicho procedimiento por lo general resulta difícil de llevar a cabo, a menos que el investigador sepa qué valores iniciales de la matriz serían apropiados. Bernstein (1988) se refiere a lo anterior como la "solución forzada débil".

Joreskog (1969) reconoció que se podían estimar tan sólo algunos de los parámetros, pero no todos. El desarrollo de Joreskog permite al investigador especificar algunos de los parámetros y permite que los otros se estimen a partir de los datos. Estas modificaciones hacen del método de Joreskog un importante avance respecto de los procedimientos previos. Los cálculos para efectuar un análisis factorial confirmatorio se realizan por medio de un programa computacional. En la actualidad, los programas de elección son LISREL, EQS y AMOS, los cuales se utilizan para el análisis estructural de covarianza. Cada uno usa un algoritmo ligeramente distinto para realizar los cálculos. Con la aparición de cada nuevo programa, se facilita más su uso.

No obstante, como con el AFE, el uso apropiado de métodos de AFC como el LISREL, EQS y AMOS requiere que el investigador posea un conocimiento amplio del área que se estudia y de lo que representa una "buena" hipótesis respecto a la estructura factorial subyacente. No se trata de un método que pueda aplicarse exitosamente a un gran cuerpo de datos con muchas variables, donde el investigador no tiene idea alguna sobre cuál puede ser la estructura factorial subyacente.

Anteriormente en este capítulo se explicó la matriz de correlación. Dicha matriz o tabla contiene el coeficiente de correlación de cada variable con cada una de las otras variables. En notación de matriz lo anterior se escribe **R**. Para comprender el AFC, resulta necesario volver a examinar la ecuación fundamental del análisis factorial presentada anteriormente:

$$\mathbf{R} = \mathbf{P}\,\mathbf{\Phi}\,\mathbf{P}^{\mathrm{T}} + \mathbf{U}$$

La matriz o tabla **P** contiene las cargas factoriales. También se conoce como la matriz de patrón factorial. Si se vuelve a escribir esta matriz, de tal manera que sus renglones estén

intercambiados con sus columnas, dicha matriz transformada se denominaría la transposición de **P**, y se escribiría $\mathbf{P}^{\mathbf{T}}$. La matriz **Φ** indica qué tanto están correlacionados los datos entre sí. **U** representa la cantidad de singularidad dentro de cada variable.

La meta consiste en encontrar los valores de **P**, **Φ** y **U** que mejor reproduzcan la matriz **R** de correlación. La matriz de correlación reproducida se simboliza **R´**. Por ende, se tiene la ecuación de análisis factorial:

$$\mathbf{R'} = \mathbf{P}\,\mathbf{\Phi}\,\mathbf{P}^{\mathbf{T}} + \mathbf{U}$$

Una solución aceptable para **P**, **Φ** y **U** sería aquella donde **R´** y **R** difieran por una muy pequeña cantidad. En otras palabras, los valores encontrados dentro de estas matrices son tales que **R** y **R´** tienen un buen ajuste. Uno de los índices más populares de bondad de ajuste es el índice de ajuste normado de Bentler y Bonnet (1980). Existe una cantidad de estadísticos de bondad de ajuste que se emplean regularmente (véase Comrey y Lee, 1992, capítulo 12).

Para desarrollar el modelo del AFC, el investigador debe elegir cuáles valores dentro de las matrices **P**, **Φ** y **U** se van a fijar y cuáles se van a estimar. Se pueden imponer restricciones sobre ciertos valores, tales como especificar un rango de valores dentro de los cuales deben de caer. Dichos valores también se denominan *parámetros*.

Los modelos de ecuación estructural para el análisis factorial confirmatorio, tal como se implementan por medio del LISREL y EQS, son nuevos procedimientos analíticos factoriales poderosos que con frecuencia representan el método de elección en una situación dada. Sin embargo, no son los únicos modelos que podrían emplearse, y en muchos casos se preferirían otros métodos.

Cualquiera que sea el método que los investigadores encuentren más apropiado para sus datos, resulta claro que el análisis factorial ha experimentado una revolución importante en los años recientes. Actualmente están disponibles nuevos métodos poderosos que expanden de manera formidable la capacidad de los trabajadores de investigación para examinar las implicaciones de sus datos. No obstante, debe enfatizarse que los métodos más nuevos deben considerarse como complementarios más que como sustitutos de los antiguos métodos del AFE, los cuales deben continuar siendo los procedimientos más efectivos para enfrentar situaciones de análisis con muchos datos. Por consiguiente, en el futuro previsible los estudiantes del análisis factorial necesitarán familiarizarse tanto con el método AFE como con el método AFC.

Ejemplo de investigación usando el análisis factorial confirmatorio

El capítulo de Keith (1997) brinda diversos ejemplos de investigación donde el AFC se aplica a problemas dentro de la psicología escolar. Este sobresaliente artículo es de fácil lectura, está bien escrito y es muy recomendable para quienes buscan buenos ejemplos donde se utilice el análisis factorial confirmatorio. Keith prueba la estructura de diversas

TABLA 34.7 *Estructura teórica del PIAK*

Inteligencia fluida	Inteligencia cristalizada	Recuerdo retardado
Aprendizaje de claves (rebus) visual-fonéticas	Definiciones	Evocación retrasada de claves (rebus)
Pasos lógicos	Comprensión auditiva	Evocación auditiva retardada
Códigos misteriosos	Significados dobles	
Memoria para diseño con cubos	Rostros famosos	

pruebas psicológicas para determinar si sus pretensiones son verdaderas, por medio del uso del análisis factorial confirmatorio. Él ofrece una explicación completa del modelo que va a probarse y de los estadísticos de bondad de ajuste proporcionados por el análisis factorial confirmatorio. En una demostración como ésta, Keith probó las pretensiones de la *prueba de inteligencia para adultos de Kaufman* (PIAK) (Kaufman Adult Intelligence Test), que consiste de 10 subescalas. Cuatro de ellas están diseñadas para medir la parte fluida de la inteligencia (*gf*); otras cuatro miden la parte cristalizada de la inteligencia (*gc*), y dos pruebas fueron diseñadas para medir la *evocación retardada*. Los componentes fluido y cristalizado de la inteligencia se mencionaron anteriormente cuando se explicó la teoría de Cattell sobre la inteligencia, la cual posteriormente fue modificada por John Horn (véase Cattell, 1987; Horn y Cattell, 1966) y ahora se llama teoría de la inteligencia de Horn-Cattell. Las partes cristalizada y fluida de la inteligencia constituyen sólo dos partes de la teoría completa. La evocación retardada mide la memoria de material aprendido en las primeras partes de la prueba, de quien responde la prueba. La tabla 34.7 presenta la estructura teórica o constructo del PIAK.

Con la ecuación fundamental del análisis factorial $\mathbf{R} = \mathbf{P}\,\Phi\,\mathbf{P}^{T} + \mathbf{U}$, los parámetros dentro de estas matrices pueden diseñarse de la siguiente manera:

$$\Phi = \begin{array}{c} \\ \text{F1} \\ \text{F2} \\ \text{F3} \end{array} \begin{array}{c} \begin{array}{ccc} \text{F1} & \text{F2} & \text{F3} \end{array} \\ \begin{bmatrix} 1.00 & & \\ * & 1.00 & \\ * & * & 1.00 \end{bmatrix} \end{array}$$

$$\mathbf{P} = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \\ X_7 \\ X_8 \\ X_9 \\ X_{10} \end{array} \begin{array}{c} \begin{array}{ccc} \text{F1} & \text{F2} & \text{F3} \end{array} \\ \begin{bmatrix} * & 0 & 0 \\ * & 0 & 0 \\ * & 0 & 0 \\ * & 0 & 0 \\ 0 & * & 0 \\ 0 & * & 0 \\ 0 & * & 0 \\ 0 & * & 0 \\ 0 & 0 & * \\ 0 & 0 & * \end{bmatrix} \end{array}$$

$$\mathbf{U} = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \\ X_7 \\ X_8 \\ X_9 \\ X_{10} \end{array} \begin{array}{c} \begin{array}{cccccccccc} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 & X_9 & X_{10} \end{array} \\ \begin{bmatrix} * & & & & & & & & & \\ 0 & * & & & & & & & & \\ 0 & 0 & * & & & & & & & \\ 0 & 0 & 0 & * & & & & & & \\ 0 & 0 & 0 & 0 & * & & & & & \\ 0 & 0 & 0 & 0 & 0 & * & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & * & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & * & & \\ * & 0 & 0 & 0 & 0 & 0 & 0 & 0 & * & \\ 0 & 0 & 0 & 0 & 0 & * & 0 & 0 & 0 & * \end{bmatrix} \end{array}$$

Los asteriscos en cada matriz indican los parámetros que se van a estimar por medio de los datos. En la matriz Φ los asteriscos son las correlaciones entre los factores. En la matriz **P**, los asteriscos son las cargas factoriales; y en la matriz **U** los asteriscos representan las correlaciones entre las variables. En la matriz Φ los valores de la diagonal se fijan en 1.00. En la matriz **P** la solución deseada tendría una estructura más simple. Las variables sin marca están forzadas a tener valores de cero. En la matriz **U** existe una varianza única para cada una de las 10 variables. También tienen asteriscos debido a que serán estimadas por medio de los datos. En este modelo en particular, Keith afirma que el *aprendizaje de claves* (*rebus*) *visual-fonéticas* (variable 1) puede correlacionarse con la *evocación retardada de claves* (*rebus*) *visual-fonéticas* (variable 9) y que la *evocación auditiva retardada* (variable 10) estaría relacionada con la *comprensión auditiva* (variable 6). Tales correlaciones también se estiman a partir de los datos y, por lo tanto, dichos valores reciben asteriscos en la matriz **U**. Cuando se habla de los "datos", se refiere a las puntuaciones de los participantes en las 10 variables y a la matriz de intercorrelaciones para esas 10 variables.

Una vez establecido el modelo y los parámetros a estimar, es posible escribir los comandos de control adecuados para programas computacionales como el EQS, LISREL y AMOS. Los problemas de esta naturaleza son demasiado laboriosos para resolverse a mano.

Keith (1997) presenta el modelo y los estimados en forma de diagrama de ruta, y éstos como se vio anteriormente, son modelos conceptuales y visuales útiles para problemas en el análisis factorial confirmatorio y en el modelamiento de ecuación estructural.[1]

Keith encontró que los estadísticos de bondad de ajuste indican que el modelo PIAK se ajusta a los datos observados; además presenta seis de dichos estadísticos de bondad de ajuste y ofrece una explicación sobre ellos. La prueba chi cuadrada que se ha revisado en capítulos previos sirve como un estadístico de bondad de ajuste; sin embargo, se ve afectada por cambios en el tamaño de la muestra. Existen otros que son más adecuados.

El anterior representa sólo uno de dichos modelos demostrados por Keith. Su artículo continúa demostrando diversas variaciones distintas de modelos que llegan a comprobarse. Keith afirma que hay muchos más modelos y problemas que el AFC es capaz de realizar. De la misma manera en que Comrey y Lee (1992) demuestran cómo probar una estructura factorial hipotética en dos muestras separadas de manera simultánea, Keith indica cómo probar las similitudes de los factores a través de diferentes pruebas de inteligencia; por ejemplo, la *escala Wechsler de inteligencia para niños* (o WISC) y la *batería de Kaufman de evaluación para niños* (o K-ABC).

Análisis factorial e investigación científica

El análisis factorial tiene dos propósitos básicos: explorar áreas de variables para identificar los factores que presuntamente subyacen a las variables, y, como en todo trabajo científico, probar hipótesis sobre las relaciones entre variables. El primer propósito es muy reconocido y bastante bien aceptado. El segundo propósito no es tan reconocido ni tan aceptado.

Para conceptualizar el primer propósito —el exploratorio o reductivo— se debe tener en cuenta la validez de constructo y las definiciones constitutivas. El análisis factorial se concibe como una herramienta para la validez de constructo. Recuerde que en el capítulo 28 la validez se definió como la varianza del factor común. Puesto que la principal preocupación del análisis factorial es la varianza del factor común, por definición está firmemente relacionada con la teoría de medición. De hecho, tal relación fue expresada anteriormente

[1] El modelamiento de la ecuación estructural es un conjunto de términos alternativos para el análisis estructural de covarianza.

en la sección denominada "Un poco de teoría factorial", donde se anotaron las ecuaciones para aclarar la teoría analítica factorial. (Véase, en especial, la ecuación 34.6.)

Recuerde también que la validez de constructo busca el "significado" de un constructo a través de las relaciones entre el constructo y otros constructos. En capítulos iniciales de la presente obra, cuando se explicaron los tipos de definiciones, se aprendió que los constructos podían definirse de dos maneras: por medio de definiciones operacionales, y a través de definiciones constitutivas, las cuales son aquellas que definen constructos con otros constructos. En esencia, lo anterior es lo que hace el análisis factorial. Puede denominársele como un método de significado constitutivo, ya que permite al investigador estudiar los significados constitutivos de los constructos y, por consiguiente, su validez de constructo.

Las medidas de tres variables, por ejemplo, pueden tener algo en común. Este algo es, en sí mismo, una variable, presuntamente una entidad más básica que las variables utilizadas para aislarla e identificarla. A dicha nueva variable se le da un nombre; en otras palabras, se construye una entidad hipotética. Entonces, para indagar sobre la "realidad" de la variable es posible diseñar sistemáticamente una medida de ella y probar su "realidad" correlacionando los datos obtenidos con la medida con los datos de otras medidas teóricamente relacionadas con ella. El análisis factorial ayuda a verificar las expectativas teóricas.

Parte de los aspectos básicos de la vida de cualquier ciencia son sus constructos. Continúan usándose los constructos antiguos y constantemente se inventan nuevos. Note algunos de los constructos generales que son directamente pertinentes a la investigación educativa y del comportamiento: el aprovechamiento, la inteligencia, el aprendizaje, las aptitudes, las actitudes, la habilidad de solución de problemas, las necesidades, los intereses, la creatividad, el conformismo. Ahora considere algunas de las variables más específicas, importantes en la investigación del comportamiento: la ansiedad ante las pruebas, la habilidad verbal, el tradicionalismo, el pensamiento convergente, el razonamiento aritmético, la participación política y la clase social. Claramente una gran porción del esfuerzo de la investigación científica del comportamiento debe ser dedicado a lo que podría denominarse *investigación del constructo* o *validación del constructo*, y ello requiere del análisis factorial.

Cuando aquí se habla de relaciones, se refiere a las relaciones entre constructos: inteligencia y aprovechamiento, autoritarismo y etnocentrismo, reforzamiento y aprendizaje, clima organizacional y desempeño administrativo: todas las cuales son relaciones entre constructos demasiado abstractos o variables latentes. Dichos constructos, por lo común, deben definirse operacionalmente para poder estudiarse. Los factores son variables latentes, por supuesto, y el principal esfuerzo analítico factorial científico en el pasado se ha realizado para identificar factores y para utilizarlos ocasionalmente para medir variables de investigación. Raras ocasiones se han llevado a cabo intentos deliberados para evaluar los efectos de variables latentes sobre otras variables. No obstante, con los recientes avances y desarrollos en el pensamiento y metodología multivariados, resulta claro que ahora es posible evaluar la influencia de las variables latentes entre sí. Dicho desarrollo importante se analizará e ilustrará en el capítulo 35 sobre el análisis de estructuras de covarianza. Ahí se verá que los científicos llegan a obtener índices de las magnitudes y significancia estadística sobre los efectos de variables latentes en otras variables latentes. Si así ocurre, entonces, el análisis factorial se vuelve aún más importante en la identificación de variables o factores latentes, y el científico debe tener mucho cuidado con la interpretación de los datos, en los cuales se está evaluando la influencia de variables latentes.

Entonces, muchas áreas de investigación muy bien pueden ir precedidas por explotaciones analíticas factoriales de las variables del área. Lo anterior no significa que se junte una cantidad de pruebas y se aplique a cualquier muestra que parezca estar disponible. Las

investigaciones analíticas factoriales, tanto exploratorias como de comprobación de hipótesis, deben plantearse cuidadosamente. Es necesario controlar las variables que tengan alguna influencia: sexo, educación, clase social, inteligencia, etcétera. Las variables no se incluyen en un análisis factorial tan sólo por incluirlas. Deben tener un propósito legítimo. Si, por ejemplo, no es posible controlar la inteligencia por medio de la selección de la muestra, se incluye una medida de inteligencia (verbal quizás) en la batería de medidas. Al identificar la varianza de la inteligencia, en cierto sentido se controla la inteligencia. Se puede saber si las propias medidas están contaminadas por sesgos de respuesta, al incluir medidas del sesgo de respuesta, en el análisis factorial.

El segundo propósito principal del análisis factorial es la prueba de hipótesis. Ya se sugirió un aspecto de la comprobación de hipótesis: es factible incluir pruebas o medidas dentro de baterías analíticas factoriales de forma deliberada, para probar la identificación y naturaleza de los factores. El diseño de dichos estudios fue bien establecido por Comrey, Thurstone, Cattell, Guilford y otros. Primero, los factores son "descubiertos". Se infiere su naturaleza a partir de las pruebas que están cargadas en ellos. Dicha "naturaleza" se establece como una hipótesis. Se construyen y se aplican nuevas pruebas con nuevas muestras de sujetos. Los datos se analizan factorialmente. Si los factores surgen tal como *se predijo*, la hipótesis, hasta este punto, se confirma; parecería que los datos poseen "realidad". Pero ciertamente con ello no termina el asunto. Aún se deben probar, entre otras cosas, las relaciones de los factores con otros factores. Aún se deben ubicar los factores, como constructos, en una red nomológica de constructos.

Un uso menos reconocido del análisis factorial, descrito por Fruchter (1966), implica la comprobación de hipótesis experimentales. Por ejemplo, se hipotetiza que cierto método de enseñanza de lectura cambia los patrones de habilidad de los alumnos, de tal manera que la inteligencia verbal no es una influencia tan poderosa como lo es en otros métodos de enseñanza. Se puede planear un estudio experimental para probar esta hipótesis. Los efectos de los métodos de enseñanza se evalúan por medio de los análisis factoriales de un conjunto de pruebas aplicadas antes y después del uso de los diferentes métodos. Woodrow (1938) comprobó una hipótesis similar cuando aplicó un conjunto de pruebas antes y después de la práctica en siete pruebas: sumar, restar, anagramas, etcétera. Encontró que los patrones de carga factorial *sí* cambiaron después de la práctica.

Al considerar el valor científico del análisis factorial debe prevenirse al lector de no atribuir "realidad" y singularidad a los factores. El riesgo de materialización es muy grande. Resulta sencillo nombrar un factor y después creer que existe una realidad detrás del nombre. Sin embargo, el hecho de asignarle un nombre a un factor no le confiere realidad. Los nombres de los factores son meros intentos de comprender la esencia de los factores. Siempre son tentativos, sujetos a confirmación o desconfirmación posterior. Asimismo, muchas cosas pueden producir factores. Cualquier asunto que introduzca correlación entre variables "crea" un factor. Las diferencias en sexo, educación, antecedentes sociales y culturales y en inteligencia, quizá provoquen la aparición de factores. Los factores también difieren —por lo menos hasta cierto punto— en muestras diferentes. Conjuntos de respuestas o formas de pruebas pueden hacer surgir factores. Aun con estas precauciones, debe señalarse que los factores sí surgen repetidamente en diferentes pruebas, diferentes muestras y diferentes condiciones. Cuando así sucede, se tiene bastante certeza de que existe una variable subyacente que se está midiendo exitosamente.

Como se mencionó al inicio del capítulo, existen críticas serias al análisis factorial. Las principales críticas válidas se centran alrededor de la indeterminación de cuántos factores se deben extraer de una matriz de correlación, y del problema de cómo rotar los factores. Otra dificultad que molesta tanto a los críticos como a los adeptos es lo que puede llamarse el "problema de la comunalidad", o qué cantidades poner dentro de la diagonal de la

matriz **R** antes de factorizar. En un capítulo introductorio estos problemas no pueden analizarse en detalle. Se refiere al lector a la explicación de Cattell (1978), Comrey y Lee (1992), Cureton y D'Agostino (1983), Gorsuch (1983), Guilford (1954), Harman (1976) y Thurstone (1947). Una crítica de otro tipo parece inquietar a los educadores y sociólogos, así como a algunos psicólogos. Ésta adquiere dos o tres formas que parecen reducirse en **desconfianza, algunas veces profunda, combinada con antipatía hacia el método,** a causa de su complejidad y, de manera extraña, a su objetividad.

El argumento expresa algo como esto. El análisis factorial junta demasiadas pruebas dentro de una máquina estadística y arroja factores que tienen poco significado psicológico o sociológico. Los factores son meros artefactos del método. Son promedios que no corresponden a realidad psicológica alguna, especialmente a la realidad psicológica del individuo, que no sea otra que la que está en la mente del analista factorial. Además, no se puede obtener más del análisis factorial de lo que se haya puesto dentro de él.

El argumento se vuelve básicamente irrelevante. Decir que los factores no tienen significado psicológico y que son promedios es tanto verdadero como falso. Si los argumentos fueran válidos, ningún constructo científico tendría significado alguno. Todos son, en cierto sentido, promedios. Todos son inventos del científico. Se trata sencillamente del terreno de la ciencia. El criterio básico de la "realidad" de cualquier constructo, de cualquier factor, es su "realidad" empírica, científica. Si, después de descubrir un factor, se predicen exitosamente relaciones a partir de presuposiciones teóricas e hipótesis, entonces el factor posee "realidad". No existe mayor realidad en un factor que ésta, de la misma forma que no existe mayor realidad en un átomo que sus manifestaciones empíricas.

El argumento que sostiene que sólo se obtiene lo que se pone en un análisis factorial no tiene significado alguno y también es irrelevante. Ningún investigador competente del análisis factorial afirmaría algo más que esto. Pero esto no quiere decir que nada se descubra en el análisis factorial. Todo lo contrario. La respuesta es, por supuesto, que no se obtiene nada más del análisis factorial que lo que se pone en él, pero que no se conoce *todo* lo que se pone en él. Tampoco se conoce cuáles pruebas o medidas comparten varianza del factor común; tampoco se conocen las relaciones entre los factores. Únicamente el estudio y el análisis llegan a indicar estas cosas. Se podría desarrollar una escala de actitud que se piense mide una sola actitud. Naturalmente, un análisis factorial de los reactivos de actitud no genera factores que no estén en los reactivos. Sin embargo, puede mostrar, por ejemplo, que existen dos o tres fuentes de varianza común en una escala que se creía unidimensional. De manera similar, una escala que se creía que estaba midiendo *autoritarismo* puede mostrar, por medio del análisis factorial, que mide *inteligencia*, *dogmatismo* y otras variables.

Si se examina la evidencia empírica más que la opinión, se debe concluir que el análisis factorial constituye una de las herramientas más poderosas diseñadas hasta ahora para el estudio de áreas complejas de interés científico del comportamiento. De hecho, el análisis factorial es uno de los inventos creativos del siglo XX, así como lo son las pruebas de inteligencia, el condicionamiento, la teoría del reforzamiento, la definición operacional, el concepto de aleatoriedad, la teoría de medición, el diseño de investigación, el análisis multivariado, la computadora y las teorías del aprendizaje, la personalidad, del desarrollo, de organizaciones y de la sociedad.

Es adecuado que el capítulo se concluya con algunas palabras de un gran científico, maestro y analista factorial de la psicología, Louis Leon Thurstone (1959, p. 8):

> Como científicos, tenemos fe en que las habilidades y personalidades de la gente no sean tan complejas como la enumeración total de los atributos que pueden listarse. Creemos que estas características se componen de un menor número de factores o elementos pri-

marios que se combinan de varias maneras para formar una larga lista de características. Es nuestra ambición encontrar algunas de dichas habilidades y características elementales.

Todo trabajo científico tiene algo en común: que se intenta comprender la naturaleza de la forma más parsimoniosa. Una explicación de un conjunto de fenómenos o un conjunto de observaciones experimentales logra aceptación sólo en tanto ofrezca control intelectual o comprensión de una variedad relativamente amplia de fenómenos, en términos de un número limitado de conceptos. El principio de la parsimonia es intuitivo para cualquiera que tenga aun la más leve aptitud para la ciencia. La motivación fundamental de la ciencia es el anhelo de la comprensión más simple posible de la naturaleza, y encuentra satisfacción en el descubrimiento de las uniformidades simplificadoras llamadas leyes científicas.

RESUMEN DE CAPÍTULO

1. El análisis factorial examina un conjunto de variables y determina cuáles van juntas. Las variables que se agrupan se denominan *un factor*.
2. Un factor es un constructo, una entidad hipotética o una variable latente que fundamenta las mediciones de cualquier tipo.
3. Charles Spearman desarrolló el análisis factorial, pero fue Louis Thurstone quien lo expandió y mejoró. Thurstone se considera el "padre del análisis factorial moderno".
4. Algunas de las contribuciones de Thurstone incluyen el método centroide de extracción, la rotación factorial y la estructura simple. Tanto la rotación como la estructura simple se emplean para volver más interpretables los factores.
5. La ecuación fundamental del análisis factorial es $\mathbf{R} = \mathbf{P}\,\Phi\,\mathbf{P}^{T} + \mathbf{U}$.
6. La ecuación fundamental muestra cómo la matriz de correlación **R** se parte en una matriz **P** de carga factorial, correlación entre factores, Φ, y singularidad, **U**. **R** son los datos observados. Todas las demás se estiman a partir de los datos.
7. Existe una variedad de métodos de extracción factorial diferentes. El más popular ha sido el método de factores principales.
8. El método de factores principales requiere que el investigador aporte estimados de las comunalidades y que establezca el número de factores a extraer.
9. Los estimados de las comunalidades y el número de factores han sido problemas difíciles de resolver en el análisis factorial. No existe un conjunto de reglas claras para cada uno. Sin embargo, Comrey ha desarrollado un método que no utiliza estimados de las comunalidades.
10. La comunalidad se refiere a la proporción de la varianza total que es varianza del factor común. Una meta del análisis factorial consiste en encontrar los componentes de varianza de la varianza de factor común total.
11. Las puntuaciones factoriales implican la combinación de los valores de aquellas variables que definan al factor. Es una puntuación transformada nueva. Si una batería de 10 pruebas produce tres factores, entonces cada persona que responde las pruebas tendría puntuaciones de tres factores.
12. Actualmente el análisis factorial sigue dos metodologías. El método tradicional ahora se llama *análisis factorial exploratorio* o AFE, y el método más nuevo se denomina *análisis factorial confirmatorio* o AFC.
13. El análisis factorial exploratorio por lo común se utiliza para comprender o descubrir cuáles factores subyacen a los datos. Algunos investigadores que son usuarios experimentados de este método saben cómo probar hipótesis sobre los factores.

14. El análisis factorial confirmatorio sirve para probar hipótesis acerca de la estructura factorial. En el AFC se desarrolla un modelo, basado en la teoría o en hallazgos previos, y luego se prueba contra los datos empíricos.
15. El análisis factorial confirmatorio es sólo un caso especial de un grupo de análisis llamados *análisis estructural de covarianza*.

SUGERENCIAS DE ESTUDIO

1. El estudiante más avanzado encontrará valiosa la siguiente selección de artículos:

Comrey, A. L. (1978). Common methodological problems in factor analitic studies. *Journal of Consulting and Clinical Psychology, 46*, 648-659. [Una revisión no matemática de los problemas encontrados en estudios analíticos factoriales.]

Comrey, A. L. (1985). A method for removing outliers to improve factor analytic results. *Multivariate Behavioral Research, 20*, 273-281. [Muestra cómo detectar y eliminar valores extremos que ejercen un efecto negativo sobre la solución de un análisis factorial.]

Comrey, A. L. y Montag, I. (1982). Comparison of factor analytic results with two-choice and seven-choice personality item formats. *Applied Psychological Measurement, 6*, 285-289. [Comparación de dos resultados de un análisis factorial de las *escalas de personalidad de Comrey*, donde uno de los análisis se realizó con un formato de respuesta de dos opciones, y otro con un formato de siete opciones. Los resultados indican la superioridad del formato de siete opciones sobre el de dos opciones, para los inventarios de personalidad.]

Dunlap, W. P. y Cornwell, J. M. (1994). Factor analysis of ipsative measures. *Multivariate Behavioral Research, 29*, 115-126. [Explora el análisis factorial con medidas ipsativas. Los autores muestran de forma analítica los problemas fundamentales que las medidas ipsativas imponen al análisis factorial. Tales investigadores recomiendan que el análisis factorial no debe realizarse con datos que sean reconocidos como ipsativos.]

Fleming, J. S. (1981). The use and misuse of factor scores in multiple regression analysis. *Educational and Psychological Measurement, 41*, 1017-1025. [Explora cuándo y dónde puede usarse el análisis factorial junto con la regresión múltiple con propósitos predictivos.]

Lee, H. B. y Comrey, A. L. (1979). Distortions in a commonly used factor analytic procedure. *Multivariate Behavioral Research, 14*, 301-321. [Un estudio que compara el método más popular de extracción y rotación factorial con otros métodos. Muestra cuán distorsionadas se pueden volver algunas soluciones al utilizar los métodos más populares.]

Montanelli, R. y Humphreys, L. (1976). Latent roots of random data correlation matrices with squared multiple correlations on the diagonal: A Monte Carlo study. *Psychometrika, 41*, 341-348. [Excelente método de correlación y regresión aleatorias respecto al problema del número de factores.]

Overall, J. (1965). Note on the scientific status of factors. *Psychological Bulletin, 61*, 270-276. [Un análisis excelente e incluso brillante, sobre nociones básicas del análisis factorial.]

Peterson, D. (1965). Scope and generality of verbally defined personality factors. *Psychological Review, 72*, 48-59. [Muy convincente sobre el problema del número de factores.]

2. Como siempre, no existe un sustituto para el estudio de los usos de los métodos en la investigación real. Por lo tanto, el lector debe consultar dos o tres buenos estudios sobre el análisis factorial. Selecciónelos de los citados en el capítulo o de los siguientes:

   Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Nueva York: Cambridge University Press. [Un análisis y reanálisis de estudios de inteligencia que utilizaron análisis factorial. Ofrece una muy buena historia de psicología de las diferencias individuales.]

   Daniel, L. G. y Siders, J. A. (1994). Validation of teacher assessment instruments: A confirmatory factor analytic approach. *Journal of Personnel Evaluation in Education, 8,* 29-40. [Examen de la validación de constructo del Mississippi Teacher Assessment Instrument utilizado para la certificación de nuevos maestros. Un análisis factorial exploratorio encontró cuatro factores; sin embargo, análisis factoriales confirmatorios no lograron generar un modelo estructural aceptable.]

   Fleming, J. S. y Whalen, D. J. (1990). The personal and academic self-concept inventory: Factor structure and gender differences in high school and college samples. *Educational and Psychological Measurement, 50,* 957-967. [Análisis factorial confirmatorio aplicado a diversos modelos estructurales competentes, del Personal and Academic Self-Concept Inventory, una extensión de las Self-Rating Scales.]

   Isaacson, R. L. McKeachie, W. L., Milholland, J. E. y Lin, Y. G. (1964). Dimensions of student evaluations of teaching. *Journal of Educational Psychology, 55,* 344-351. [Un estudio competente de los factores que subyacen a las evaluaciones de los estudiantes respecto a los instructores. El primer factor es importante.]

   Mitrushina, M. y Satz, P. (1991). Changes in cognitive functioning associated with normal aging. *Archives of Clinical Neuropsychology, 6,* 49-60. [Uso del análisis factorial en pruebas de memoria y psicomotrices para encontrar factores de funcionamiento cognitivo en participantes ancianos. Se calcularon las puntuaciones factoriales y se utilizaron en un análisis de varianza.]

   Thurstone, L. L. (1944). *A factorial study of perception. Psychometric Monographs,* núm. 4. Chicago: University of Chicago Press. [Otro estudio pionero y clásico de Thurstone.]

3. A continuación se presenta una pequeña matriz de correlación ficticia, con los nombres de las pruebas.

		1	2	3	4	5	6
1.	Vocabulario	.70	.22	.20	*.15*	*.25*	
2.	Analogías	.70		.15	.26	.12	.30
3.	Suma	.22	.15		.81	.21	.10
4.	Multiplicación	.20	.26	.81		.31	.29
5.	Recuerdo de nombres propios	.15	.12	.21	.31		.72
6.	Reconocimiento de figuras	.25	.30	.10	.29	.72	

   a) Realice un análisis factorial a simple vista. Es decir, por medio de la inspección de la matriz determine cuántos factores probablemente hay y qué pruebas están en qué factores.

b) Asigne un nombre a los factores. ¿Qué tan seguro está de sus nombres? ¿Qué puede hacer usted para estar más seguro de sus conclusiones?

4. Algunos excelentes libros sobre el análisis factorial que uno desearía leer para obtener información son los siguientes:

Cattell, R. B. (1978). *The scientific use of factor analysis in the behavioral and life sciences.* Nueva York: Plenum. [Se trata de un libro sobresaliente que cubre todo lo que ha sucedido con el análisis factorial desde la publicación del libro de Cattell en 1952. Se compone de dos partes e introduce conceptos matemáticos de manera gradual.]

Comrey, A. L. y Lee, H. B. (1992). *A first course in factor analysis* (2a. ed.). Hillsdale, Nueva Jersey: Lawrence Erlbaum. [En lugar de introducir al estudiante a las matemáticas de las matrices en un capítulo, el libro presenta de forma gradual las matrices y su uso en el análisis factorial. Los temas en este libro suplementan los temas cubiertos en otros libros sobre el análisis factorial. El capítulo 11 resulta especialmente valioso, pues muestra al lector cómo se desarrollaron las *Escalas de Personalidad de Comrey* con el uso del análisis factorial. Se exponen métodos valiosos que normalmente no se encuentran en otros libros de texto.]

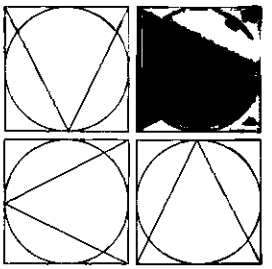
Cureton, E. E. y D'Agostino, R. B. (1983). *Factor Analysis: An applied approach.* Hillsdale, Nueva Jersey: Lawrence Erlbaum. [Un libro que presenta modelos y teorías del análisis factorial exploratorio, sin el empleo de matemáticas avanzadas como el cálculo. Contiene un buen capítulo sobre álgebra matricial y cubre adecuadamente el uso de métodos del principio del eje para el análisis factorial común. También proporciona algunas comparaciones entre diferentes métodos de extracción y rotación factoriales.]

Gorsuch, R. (1983). *Factor analysis* (2a. ed.). Hillsdale, Nueva Jersey: Lawrence Erlbaum. [Se trata de un libro académico, informativo y con autoridad. Una de sus grandes virtudes es que explora de forma profunda los problemas más difíciles y complicados del análisis factorial. Gorsuch no sólo explica las ideas técnicas, sino que también cita contribuciones teóricas e investigaciones empíricas de los problemas. *Sumamente recomendable* como trabajo de referencia para los investigadores del comportamiento.]

Mulaik, S. A. (1972). *The foundations of factor analysis.* Nueva York: McGraw-Hill. [Un tratamiento matemáticamente sofisticado del análisis factorial. Definitivamente se recomienda para quienes poseen un entrenamiento matemático extenso, como del cálculo multivariado. Este libro actualmente ya no se imprime pero puede estar disponible en algunas bibliotecas.]

Rummel, R. J. (1970). *Applied factor analysis.* Evanston, Illinois: Northwestern University Press. [Un libro profundo sobre el análisis factorial exploratorio con un énfasis en ciencias políticas. Requiere de conocimientos sobre matemáticas. Contiene buenos capítulos sobre álgebra matricial y su uso en la explicación del análisis factorial. Se proporciona una buena explicación sobre los diversos modelos factoriales. Sin embargo, no explica de forma extensa los asuntos implicados en el uso e interpretación de las soluciones analíticas factoriales.]

Thurstone, L. L. (1947). *Multiple factor analysis.* Chicago: University of Chicago Press. [Se trata de un trabajo clásico del creador del análisis factorial moderno. Aunque se escribió hace más de 50 años, el material que presenta aún es relevante.]



CAPÍTULO 35

ANÁLISIS ESTRUCTURAL DE COVARIANZA

En esta larga y complicada disertación sobre los fundamentos de la investigación del comportamiento, con frecuencia se ha hablado sobre la importancia de la teoría y de su comprobación. En ciertos momentos se ha enfatizado el propósito de la investigación científica de formular explicaciones de los fenómenos naturales y de someter las implicaciones de las explicaciones a una prueba empírica. En este capítulo se estudiará y se tratará de comprender un sistema analítico altamente desarrollado y conceptualmente sofisticado para modelar y probar teorías científicas del comportamiento: el *análisis estructural de covarianza*. El análisis estructural de covarianza también tiene otro nombre, algunas veces se denomina *modelamiento de ecuaciones estructurales* (MEE). Para entender esta metodología, se consideran de forma importante algunos sistemas matemático-estadísticos y programas computacionales. Existen al menos tres que son reconocidos, y que actualmente están en uso. A la mitad de los años setenta, el LISREL (*Linear Structural Relations*) (*relaciones estructurales lineales*) fue creado y desarrollado por Joreskog y sus colaboradores (Joreskog y Sorbom, 1993) para establecer y analizar estructuras de covarianza. Las primeras versiones de este programa computacional requerían del establecimiento de planteamientos difíciles. Sin embargo, las generaciones posteriores se hicieron mucho más fáciles. Hasta hace poco formaba parte del paquete estadístico SPSS y aún continúa siendo el método preferido para muchos diseñadores. A finales de los

setenta y principios de los ochenta, el programa computacional llamado EQS fue desarrollado por Bentler (1986). Los investigadores interesados en el análisis estructural de covarianza encontraron que el programa de Bentler resultaba más fácil de utilizar. Los planteamientos del programa y los símbolos del modelo eran más fáciles de comprender que los del LISREL. Sin embargo, el LISREL ha incrementado en mucho su número de usuarios con la aparición de la versión 8. Aunque ya es un poco antiguo, el trabajo de Brown (1986) comparó el LISREL y el EQS en términos de la estimación de parámetros para el análisis factorial confirmatorio. Aquí se utilizará el EQS debido a que muchos lo encuentran más fácil de entender, pues utiliza denominaciones estándar; mientras que el modelado LISREL emplea bastantes letras griegas. Sin embargo, una vez que el investigador está familiarizado con la estructura de covarianza o con el modelamiento de ecuación estructural, las diferencias se vuelven menos importantes.

Los investigadores tienen ahora un tercer sistema y programa computacional al cual enfrentarse: es el llamado análisis de estructuras momentáneas (Analysis of Moment Structures, AMOS), que fue publicado por SmallWaters Corporation (Arbuckle, 1995). Ellos tienen una versión de demostración de su programa en su sitio de Internet. La versión más reciente permite que el usuario especifique, vea y modifique el modelo estructural *gráficamente*, por medio del uso de herramientas de dibujo. Cada uno de dichos programas y modelos ha logrado, de manera consistente, que los investigadores trabajen con mayor facilidad el uso del modelamiento de la ecuación estructural o análisis estructural de covarianza.

Por desgracia, aun con programas computacionales mejorados y diversos manuales y libros sobre el tema (véase Schumacker y Lomax, 1996), no es fácil aprender el análisis estructural de covarianza o el modelamiento de ecuación estructural para quienes carecen de conocimientos matemáticos. Debe confesarse que la dificultad radica en explicar el sistema en lenguaje comprensible a aquellos que no saben leer las matemáticas y, al mismo tiempo, cumplir con los propósitos y objetivos de este libro. Por lo tanto, la exposición se limita a presentar y explicar el mero esqueleto matemático del sistema y a explicar cómo y por qué se utiliza. Por fortuna, el tema va íntimamente relacionado con las exposiciones del análisis de regresión múltiple y del análisis factorial de los capítulos 32, 33 y 34.

Estructuras de covarianza, variables latentes y comprobación de la teoría

El análisis estructural de covarianza puede considerarse como una combinación del análisis factorial y del análisis de regresión múltiple. De hecho, Lee y Jennrich (1984) han mostrado cómo emplear el análisis de regresión no lineal para analizar datos de estructuras de covarianza. Su ventaja más importante consiste en que se pueden evaluar los efectos de las variables latentes entre sí y sobre otras variables observadas. Recuerde que una *variable latente* es un constructo o "entidad" hipotética: inteligencia, destreza verbal, habilidad espacial, prejuicio, ansiedad, aprovechamiento. Las variables latentes son, por supuesto, variables no observadas, cuya "realidad" se asume o infiere a partir de variables o indicadores observados. Los factores son variables latentes, constructos que inventamos para explicar el comportamiento observado.

El análisis factorial confirmatorio se presentó en el capítulo 34, y éste constituye una forma de estructura de covarianza. El lector podrá recordar que un diagrama de ruta resultaba muy útil para conceptualizar la apariencia del modelo. En el análisis estructural de covarianza se utilizarán mucho los modelos de ruta. Una vez que el modelo de ruta se

construye de forma correcta, entonces puede usarse el EQS, LISREL o AMOS. A lo largo de este capítulo se utilizarán los modelos de ruta para describir el modelo de estructuras de covarianza.

Existen diez puntos clave que el diseñador del modelo debe considerar al dibujar el diagrama de ruta que se va a analizar con el uso del EQS. Si se siguen estos puntos, entonces el diagrama del análisis de ruta se ajustará a los planteamientos del programa EQS. Se listarán dichos puntos y después se explicará cada uno. Los 10 puntos son relevantes tanto para la estructura de covarianza más simple como para la más compleja.

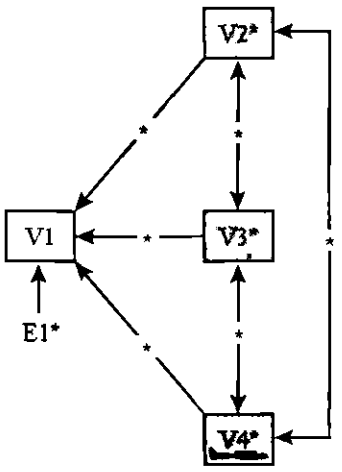
1. Existe una flecha unidireccional a partir de cada variable independiente, señalando hacia la variable dependiente.
2. Cada variable que tiene una flecha unidireccional apuntando hacia sí misma genera una ecuación de regresión lineal en el modelo de covarianza o de ecuación estructural.
3. Existe un asterisco (*) insertado en cada flecha, de las variables independientes a la variable dependiente, lo cual indica que existen parámetros libres a ser estimados para estas rutas.
4. El asterisco identifica un parámetro libre en el modelo.
5. Todas las covarianzas (correlaciones) entre las variables independientes también representan parámetros libres en el modelo. Los parámetros libres de la covarianza se indican por medio de flechas bidireccionales, con un asterisco en medio.
6. Las varianzas de las variables independientes medidas también son parámetros libres. Dichas variables se encuentran subrayadas en sus cajas, con un asterisco junto a su símbolo.
7. Todas las variables independientes poseen varianzas que funcionan como parámetros en el modelo.
8. Las variables dependientes no poseen varianzas que funcionen como parámetros en el modelo.
9. Todas las variables independientes latentes (sin medir) en el modelo deben tener su escala fija en una de dos maneras:
   a) Al establecer el coeficiente de regresión en un valor fijo. Por lo general, se establece en 1.0.
   b) Al fijar su varianza en algún valor conocido, generalmente 1.0.
10. En la mayoría de los casos, los valores E (error de medición) tienen sus coeficientes de regresión fijos en 1.0, y por ello no aparece ningún asterisco en la flecha que apunta hacia la variable dependiente.

En el EQS, el modelamiento requiere de variables independientes y dependientes. Cada una de ellas, o ambas, pueden ser medidas o latentes. En ocasiones, las variables latentes también se denominan *variables sin medir*. En el caso de las variables medidas, se utiliza el símbolo "E" para representar su error de medición. El símbolo "D" se utiliza para representar el error de medición de las variables dependientes latentes. Dependiendo de la numeración de las variables dependientes, aparece un número unido a la "E" o a la "D".

La estructura de covarianza más simple es el análisis de regresión. Si se establece que x_1 sea la variable dependiente, y que x_2, x_3 y x_4 sean las variables independientes, se escribe la ecuación del modelo de la siguiente manera:

$$x_1 = B_2x_2 + B_3x_3 + B_4x_4 + e \qquad (35.1)$$

FIGURA 35.1



Las puntuaciones en esta ecuación aparecen en forma de puntuación de desviación. Lo anterior vuelve innecesario el término de la intersección. Los valores de B son los pesos estandarizados de la regresión. El modelo de regresión de ecuación única mostrado en la ecuación 35.1 puede representarse con el diagrama de ruta de la figura 35.1.

Las variables de datos medidos están representadas en cajas o rectángulos e incluyen números de identificación como V1, V2, V3, V4 o más, según se requiera. Es decir, se utilizan cuadrados o rectángulos para encerrar una variable observada, y no latente. En el caso del análisis de regresión lineal, todas las variables se consideran observadas o medidas. Tal como se mencionó en el capítulo anterior, en la exposición sobre el análisis factorial confirmatorio, se utilizan círculos o elipses para encerrar variables latentes o sin medir. En el ejemplo de regresión existen cuatro variables en la ecuación: x_1, x_2, x_3, x_4. Por lo tanto, el EQS requiere que se nombren V1, V2, V3 y V4. Hay una flecha unidireccional a partir de cada una de las variables independientes V2, V3 y V4, que señalan hacia la variable V1. A partir del punto dos expresado antes, cada variable que tiene una flecha unidireccional apuntando hacia sí misma, genera una ecuación de regresión lineal en el modelo. Aquí sólo hay una variable de este tipo, V1 y, por ende, sólo una ecuación.

Existe un asterisco o estrella (*) incluida en cada una de las flechas que va de V2, V3 y V4 a V1, lo que indica que existen parámetros libres a estimarse en conexión con estas rutas, uno para cada estrella. También hay una flecha unidireccional apuntando desde E1 hacia V1, que indica que la ecuación de regresión contiene una variable de error, E1, que también es una variable independiente en el modelo. El asterisco (*) indica al programa que el valor precedente es un estimado y no un valor fijo. El asterisco también identifica un parámetro libre en el modelo. El conteo de los asteriscos permite al investigador determinar el número total de parámetros libres que se están estimando para el modelo. Se supone que E1 no está correlacionado con V2, V3 y V4; por lo tanto, no aparecen flechas bidireccionales entre estas variables, en el diagrama de ruta. Observe también que el coeficiente para E1 se estableció en 1.0. Se colocó el 1.0 para ser consistentes con el punto clave # 10. Por lo común, el 1.0 está implícito (no escrito).

Otros parámetros libres en el modelo incluyen todas las covarianzas entre las variables independientes medidas, V2, V3 y V4. Las flechas bidireccionales con un asterisco en medio indican parámetros de covarianza libres. Ello añade tres parámetros libres más.

Los parámetros libres adicionales incluyen las varianzas de las variables independientes medidas: V2, V3 y V4. El hecho de que tales varianzas sean parámetros libres se indica en el modelo con el sombreado de V2, V3 y V4 en las cajas, y por la aparición de asteriscos (*) junto con sus símbolos. Resulta sencillo perder de vista el hecho de que las varianzas de estas variables independientes son parámetros en el modelo. El sombreado facilita recordar que el esquema del programa EQS debe contener ya sea estimados o valores fijos para estos parámetros de varianza, lo cual añade tres parámetros libres más (indicados por los asteriscos).

Se debe estimar una varianza más en el modelo debido a que la regla general es que todas las variables independientes tienen varianzas que funcionan como parámetros en el modelo, y que incluyen todas las variables de *error*, así como las variables independientes *medidas*. Sin embargo, las variables dependientes no tienen varianzas que funcionen como parámetros en el modelo. Se colocó un asterisco (*) junto a E1 en el diagrama de ruta de la figura 35.1 para mostrar que su varianza es un parámetro libre en el modelo. Ahora esto da un total de diez asteriscos en el diagrama de ruta, lo cual indica que hay 10 parámetros libres por estimar.

Existe un parámetro adicional en el modelo que está fijo en 1.0 —el coeficiente de regresión para E1—. Redundando con el punto clave # 9, todas las variables independientes sin medir en el modelo deben fijar su escala en una de dos maneras: *a*) al establecer un coeficiente de regresión en un valor fijo, por lo general 1.0, que aquí se hizo para E1; o *b*) al fijar su varianza en un valor conocido, por lo general de 1.0. En la mayoría de los casos, los valores de E tienen sus coeficientes de regresión fijos en 1.0 y, como consecuencia, en el diagrama de ruta no aparece ningún asterisco (*) en la flecha que apunta hacia la variable dependiente, con la cual la variable de error esté asociada.

No es factible fijar tanto el peso de regresión *como* la varianza para una variable E, a causa de que el producto de ambos números debe estar libre para acomodar la cantidad de error, para predecir la variable dependiente de forma correcta. Por lo tanto, se fija uno u otro, pero no ambos.

Existen 10 parámetros libres a estimar en el modelo de regresión de la figura 35.1, a partir de un total de 10 puntos de datos. Los puntos de datos consisten en las varianzas y las covarianzas de las variables medidas, V1, V2, V3 y V4, o $(n(n + 1))/2$, donde n es el número de variables. Es decir, existen seis covarianzas y cuatro varianzas. El número de grados de libertad para la estimación del modelo está dado por el número de puntos de datos menos el número de parámetros libres en el sistema: en este caso, es 10 – 10 o cero.

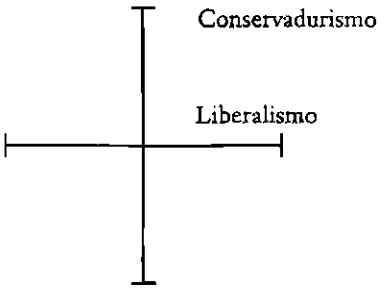
Cuando no existen grados de libertad, se dice que el modelo está "saturado"; es decir, es posible obtener valores para los parámetros libres, que reproducirán los datos de entrada de manera exacta. Por consiguiente, no existe duda sobre si el modelo se ajusta a los datos, y no se requiere de una prueba chi cuadrada ni de otra prueba estadística para saber qué tan bueno es el ajuste, pues el ajuste es perfecto. Por tal razón, los modelos de regresión no se consideran de mucho interés para el análisis estructural de covarianza. En general, quienes deciden utilizar el análisis estructural de covarianza buscan desarrollar un modelo que tenga considerablemente más puntos de datos que parámetros libres a estimar. En tales casos, se vuelve un reto encontrar un modelo no saturado y un conjunto de parámetros que reproduzcan los datos razonablemente bien, es decir, que den un buen ajuste. Sólo este tipo de modelo (uno con el número de grados de libertad mayor que cero) será capaz de ofrecer cualquier información científica con significancia teórica.

Si todo lo antes expuesto continúa pareciendo demasiado abstracto, ahora se examinará un ejemplo de una investigación real. Dicha investigación fue realizada por Kerlinger (1972). Posee las virtudes de la familiaridad y de la simpleza relativa.

Comprobación de hipótesis factoriales alternativas: dualidad contra bipolaridad de las actitudes sociales

Recuerde que existen dos perspectivas generales de las actitudes sociales que por lo general se asocian con el liberalismo y el conservadurismo. Una perspectiva —la más comúnmente adoptada por los científicos y la gente común— señala que los aspectos y la gente liberales y conservadores se oponen entre sí: el conservador está en contra de lo que el liberal apoya, y a la inversa. Esto se expresó antes como una *teoría bipolar*. Implica una dimensión de las actitudes, con los aspectos y la gente liberal en un extremo, y los aspectos y la gente conservadora en el otro.

La teoría, hipótesis o concepción contrastante sobre las actitudes sociales indica, en efecto, que los aspectos e ideas liberales son, en general, diferentes y virtualmente independientes de los aspectos e ideas conservadores. El liberalismo y el conservadurismo, para utilizar los nombres abstractos de las variables latentes, no necesariamente son opuestos entre sí: son dos ideologías separadas e independientes, o son conjuntos de creencias relacionadas que llegan a expresarse como dimensiones ortogonales:



Tal concepción de la estructura de las actitudes sociales es *dualista*.

Las dos "teorías" contrastantes de la estructura de las actitudes sociales pueden expresarse por medio de las dos matrices factoriales *A* y *B*, que se presentan en la tabla 35.1, las cuales pueden denominarse matrices "objetivo", debido a que se establecen para expresar análisis de contraste. Suponga que se han administrado seis escalas de actitudes sociales a una muestra grande heterogénea de individuos. Las escalas 1, 2 y 3 son escalas conservadoras, y las escalas 4, 5 y 6 son escalas liberales. Considere que las respuestas de la muestra a las seis escalas fueron correlacionadas y analizadas factorialmente. Los resultados del análisis factorial de lo que implican las teorías de dualidad y bipolaridad se muestran en la tabla. Los signos + indican cargas sustanciales y positivas, los signos – indican cargas sus-

TABLA 35.1 *Estructuras analíticas factoriales implicadas en la hipótesis dualista (A) y la hipótesis bipolar (B)*[a]

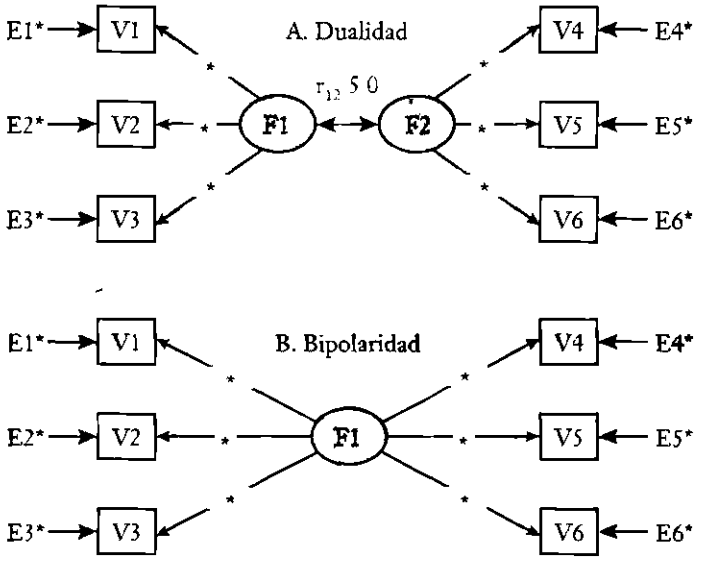
(A) Dualista Escalas	I	II	Tipo	(B) Escalas	Bipolar I	Tipo
1	+	0	*C*	1	+	*C*
2	+	0	*C*	2	+	*C*
3	+	0	*C*	3	+	*C*
4	0	+	*L*	4	–	*L*
5	0	+	*L*	5	–	*L*
6	0	1	*L*	6	2	*L*

[a] + = indica cargas factoriales positivas; - = indica cargas factoriales negativas; 0 = cargas cero; *L* = escalas liberales; *C* = escalas conservadoras.

tanciales y negativas y los ceros indican cargas cercanas a cero. La teoría dualista (*A*) implica, por supuesto, dos factores ortogonales; y la teoría bipolar (*B*) implica un factor con cargas positivas y negativas sustanciales. La *A* y la *B* de la tabla expresan de manera sucinta los dos modelos implicados por las dos teorías. Si se graficaran las "cargas" de la teoría dualista se verían como las de la figura 34.6 del capítulo 34. Los puntos de las cargas de la teoría bipolar se grafican en un solo eje, con las cargas positivas en un extremo del eje y las cargas negativas en el otro extremo.

Como se mencionó con anterioridad y se reitera ahora, a los investigadores que utilizan el análisis estructural de covarianza les gusta desarrollar modelos en diagramas de

FIGURA 35.2 *Medidas observadas (escalas)* $x_1, x_2, \ldots, x_6$; ξ_1, ξ_2: *Xsi 1: conservadurismo (C), Xsi 2: liberalismo (L);* $\lambda_{11}, \lambda_{21}, \lambda_{31}, \ldots$: *lambdas, cargas factoriales;* $\delta_1, \delta_2, \ldots$: *delta 1, delta 2, ...: términos del error*



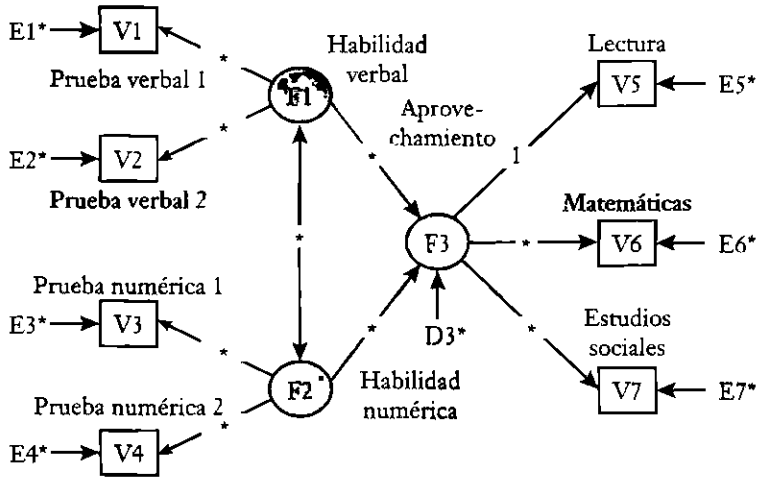
ruta. Los diagramas de ruta de los dos modelos factoriales se presentan en la figura 35.2. (Véase capítulo 33 para encontrar una explicación sobre los diagramas de ruta.) En este momento tan sólo se dibujará el modelo de ruta en términos de la notación del EQS. El principiante encontrará la notación del EQS mucho más fácil de entender que la del LISREL. El modelamiento del EQS sólo utiliza cuatro símbolos: V para variables medidas, F para variables latentes, E para representar el error de medición para las variables V, y D se utiliza para representar el error de medición para las variables latentes. El LISREL resulta más complicado.

A la luz de la naturaleza de este libro, es más congruente para el principiante emplear la notación de modelado dada por el EQS. El modelado del LISREL utiliza bastantes letras griegas para la designación de ciertos componentes o parámetros. Quizás ello asuste a algunos estudiantes y evite que aprendan una metodología de investigación y análisis sumamente importante.

En este ejemplo, $x_1, x_2, \ldots, x_6$ son las variables observadas; x_1, x_2 y x_3 son medidas del conservadurismo; y x_4, x_5 y x_6 son medidas del liberalismo, y se escribirían V1, V2, V3, V4, V5 y V6 en el EQS, donde V1, V2 y V3 miden *conservadurismo* y V4, V5 y V6 miden *liberalismo*. En ambas notaciones las variables observadas están indicadas por medio de cajas (por ejemplo, cuadrados o rectángulos); las variables no observadas latentes o factores, por medio de círculos o elipses. En el modelo dualista F_1 y F_2 se utilizan para representar *conservadurismo* y *liberalismo*. Los términos de error en el EQS se escriben E1, E2, E3, E4, E5 y E6. En el modelo de ruta presentado en la figura 35.3 se encuentra un asterisco junto a cada valor E, lo cual indica que los errores o valores únicos se estimarán por medio de los datos. La correlación entre las variables latentes también se va a estimar, por lo que también se especifica con un asterisco. Puesto que la teoría dice que el conservadurismo y el liberalismo son factores distintos y separados, se predice que $r_{12} = 0$.

El diagrama de la teoría bipolar es más fácil de explicar. Se tienen, evidentemente, las mismas seis x o variables observadas, y los mismos seis términos de error. También existen seis cargas factoriales; sólo existe un factor o F_1. En el modelo dual hay doce cargas factoriales, pero se predice que seis de ellas serán positivas y sustanciales, y el resto se fuerzan a ser iguales a cero. Los valores "forzados" o "fijos" se mantienen durante los

FIGURA 35.3 *Influencia de la habilidad sobre el aprovechamiento (ejemplo ficticio)*



cálculos. En el modelo de bipolaridad existen seis cargas factoriales: tres positivas (las rutas de las flechas están marcadas con +) y tres negativas (marcadas con –). En otras palabras, se predicen las dos matrices factoriales de la tabla 35.1, con excepción de que en la tabla sólo se usan signos + y – en lugar de cargas factoriales.

Para determinar cuál de los dos modelos está más cercano a la "realidad" empírica, es necesario probar cada uno de forma separada y después probar uno en contra del otro. Esto se hace utilizando la información o los datos que se tienen: las correlaciones entre las variables observadas x_1, x_2,..., x_6. Esta matriz de correlaciones, **R**, es una matriz de covarianza. Las variables (o escalas de actitud) 1, 2 y 3 son medidas del *conservadurismo*; y las variables 4, 5 y 6 son medidas del *liberalismo*. La hipótesis dualista predice correlaciones positivas y altas entre 1, 2 y 3; y correlaciones positivas y altas entre 4, 5 y 6. La hipótesis dualista también predice correlaciones de cero o cercanas a cero entre las variables de *C* (1, 2 y 3) y las variables de *L* (4, 5 y 6). Éstas se llaman *correlaciones cruzadas*.

El método que en realidad se utiliza en el análisis estructural de covarianza es el siguiente. Los datos se analizan de acuerdo con el modelo establecido; en este caso, el modelo dualista: dos factores ortogonales (figura 35.2A). A partir de los parámetros estimados por el análisis de los datos, análisis factorial en este caso, se calcula una matriz **R** utilizando los parámetros estimados del modelo teórico, lo cual se hace escribiendo ecuaciones para cada una de las V.

Para ayudar a comprender con mayor claridad lo que se hace y por qué, primero se establecieron las dos teorías en diagramas de ruta. El lector quizá piense que la siguiente explicación es redundante; pero los autores han encontrado que es útil para el aprendizaje y la comprensión de este método. Los investigadores del comportamiento que utilizan el "modelamiento" o "modelamiento causal", como se le llama, usan los diagramas de ruta para ayudar a conceptualizar los problemas de investigación que están estudiando y, casi más importante, para aprender las implicaciones empíricas de las teorías que se ponen a prueba. Resulta sumamente recomendable que los estudiantes traten de representar cualquier problema de investigación bajo estudio, en un diagrama de ruta, pues éste obliga a conceptualizar y obtener las estructuras básicas de los problemas. En cualquier caso, las "teorías" dualista y bipolar sobre las actitudes sociales se han establecido en los dos diagramas de ruta *A* y *B* de la figura 35.3. En dichos diagramas de ruta se acostumbra utilizar cuadrados para las variables observadas y círculos para las variables no observadas o latentes. Las flechas unidireccionales sirven para indicar influencias y las flechas bidireccionales para indicar correlaciones. Por ejemplo, si se realizara un análisis factorial de factores principales del presente problema, se estimarían 12 cargas factoriales: a_{11}, a_{12}, a_{21}, a_{22},..., a_{61}, a_{62}. Sin embargo, el problema en el EQS o en el marco conceptual de las estructuras de covarianza resulta diferente, debido a que ya se ha especificado que seis cargas factoriales o *a* se estimarán; las cargas restantes se fuerzan a ser iguales a cero en la hipótesis dualista. Para asegurarse de que se percibe y comprende la diferencia, se establecieron las dos matrices factoriales en la tabla 35.2. Note que hay 12 cargas factoriales a calcularse en A, con un análisis factorial común, y sólo seis cargas a calcularse en B, la solución de estructuras de covarianza forzada por los ceros, debido a la hipótesis dualista.

La diferencia entre los dos modelos es sorprendente —y muy importante—. En el análisis factorial común se estiman todas las cargas factoriales; pero en el análisis estructural de covarianza sólo se estiman aquellas cargas factoriales relacionadas con las hipótesis. El resto se fuerzan a ser iguales a cero —una estructura simple perfecta—. Para enfatizar los puntos establecidos, los parámetros estimados reales se muestran en la tabla 35.4. Los factores finales rotados de un análisis factorial común se presentan en *a*), y la solución forzada de las estructuras de covarianza se muestra en *b*). Se podría plantear la pregunta: ¿qué les sucede a las cargas factoriales donde se encuentran los ceros en *b*)? El punto es

TABLA 35.2 *Matrices factoriales del a)análisis factorial común y del b)análisis factorial forzado del EQS*[a]

a). Análisis factorial ordinario			*b*). Análisis factorial forzado del EQS			
Variables	I	II	Variables	I	II	Tipo[a]
1	a_{11}	a_{12}	1	a_{11}	0	*C*
2	a_{21}	a_{22}	2	a_{21}	0	*C*
3	a_{31}	a_{32}	3	a_{31}	0	*C*
4	a_{41}	a_{42}	4	0	a_{42}	*L*
5	a_{51}	a_{52}	5	0	a_{52}	*L*
6	a_{61}	a_{62}	6	0	a_{62}	*L*

[a] *C* = conservador, *L* = liberal.

que *b*) expresa la forma "pura" de la hipótesis dualista. Como se dijo anteriormente, la computadora está programada para realizar los cálculos manteniendo intactos los ceros de la tabla 35.2 y de la tabla 35.3. ¿Pero qué pasa con las cargas bastante grandes y negativas, –.44 y –.36 en *a*), el análisis factorial convencional? Ambas son altas, negativas y estadísticamente significativas, en oposición a la hipótesis dualista. Son desviaciones a partir del modelo dualista. Entonces, la pregunta clave es: ¿las desviaciones son lo suficientemente grandes para invalidar la hipótesis la cual incluye ceros? Se retomará este punto en breve.

El modelo de la figura 35.2A requiere cálculos de los términos de error, E. Se calcularon los seis términos de error; aunque aquí no interesa el método de los cálculos. La estimación de las varianzas y covarianzas es mucho más interesante y relevante para la hipótesis dualista, ya que ésta expresa las relaciones entre los factores F1 y F2. Recuerde que la hipótesis dualista incluía la correlación entre los dos factores: sería de cero o cercana a cero. Observe nuevamente la figura 35.2 y note que, de acuerdo a la hipótesis dualista, $r_{12} = 0$. Aunque r_{12} puede forzarse para que sea igual a cero, en su lugar se eligió permitir al EQS estimar la correlación entre los factores, por razones que se explicarán más adelante. Para reflejar esto en la figura 35.2A, se cambiaría r_{12} por un asterisco. Se establece que las varianzas de F1 y F2 sean iguales a 1.00, se estiman las varianzas de E1, E2, ..., E6, y se

TABLA 35.3 *Matrices factoriales obtenidas: a) convencional (con rotación) y b) factores forzados del EQS*[a]

a) Factores convencionales			*b*) Factores forzados del EQS			
Variables	I	II	Variables	I	II	Tipo
1	.69	.19	1	.65	0	*C*
2	.70	.33	2	.87	0	*C*
3	.68	.14	3	.63	0	*C*
4	–.44	.51	4	0	.71	*L*
5	–.05	.64	5	0	.54	*L*
6	–.36	.55	6	0	.63	*L*

[a] El análisis factorial convencional fue el método de factores principales con rotación varimax; el método de EQS fue el de probabilidad máxima. Todas las cargas de *b*) son estadísticamente significativas. La correlación entre los dos factores fue de –.15, que no resultó estadísticamente significativa.

especifica r_{12} como "libre". (Recuerde que cuando un parámetro es "libre", el programa estima su valor.)

En el análisis, $r_{12} = -.157$ no es estadísticamente significativa. Por lo tanto, en efecto, los dos factores son ortogonales, lo cual es consistente con la hipótesis dualista. Tenga en cuenta que la teoría dice que el *conservadurismo* y el *liberalismo* son dimensiones separadas e independientes de las actitudes sociales, lo que quiere decir, evidentemente, que la correlación entre ellas es cero (o cercana a cero).

Sin embargo, la pregunta crucial es: ¿el modelo completo es congruente con los datos? El modelo completo de la hipótesis dualista está expresado en la figura 35.2A. Siguiendo las reglas del EQS, se instruye a la computadora para que estime las seis cargas factoriales, a_{11}, a_{21}, a_{31}, a_{42}, a_{52}, a_{62}; mientras mantiene las limitaciones a cero en la matriz. Además se especifica que se calculen los términos de error de las seis ecuaciones. Además se deben especificar cuáles serán las relaciones entre los dos factores. Por lo tanto, se le debe indicar al EQS qué hacer con las varianzas de los factores F1 y F2; ello se logra al instruir al EQS para que estime r_{12}, la correlación entre F1 y F2. En el análisis factorial tradicional, lo anterior implicaría estimar los valores en la matriz Φ (véase capítulo 34). Siguiendo un procedimiento iterativo, el programa computacional estima los 13 valores que se especificó deben estimarse, utilizando las correlaciones entre las seis variables como datos de entrada (tabla 35.2). También fuerza los ceros de la tabla 35.2 y establece que las varianzas de las variables latentes (o factores) sean iguales a 1.00. Las cargas factoriales se presentan en la tabla 35.3B, y $r_{12} = -.157$. Los seis términos de error son .76, .50, .78, .71, .84 y .77. ¿Estos valores son congruentes con los datos o, de manera alterna, se "ajusta" el modelo dualista con los datos? Antes de responder las preguntas es necesario mencionar que existe una cantidad de otros puntos metodológicos importantes que no se exponen aquí, como los supuestos que subyacen al análisis. Uno de los supuestos es que la distribución de las variables observadas o medidas es normal. Otro supuesto o requisito es la identificación: el problema de la estructura de la covarianza debe establecerse de tal manera que todos los parámetros estimados puedan identificarse. Existen problemas de investigación donde tales supuestos puedan no cumplirse. La idea central detrás de la evaluación de la "bondad de ajuste" de un modelo teórico es simple y poderosa. Emplee los valores de los parámetros estimados y los valores forzados para calcular una matriz de correlación predicha o reproducida o ajustada, $\mathbf{R}^*$. En este caso la matriz $\mathbf{R}^*$ puede generarse multiplicando los renglones de la tabla 35.3: $r_{12}^* = (.65)(.87) + (0)(0) = .57$; $r_{13} = (.65)(.63) + (0)(0) = .41$; $r_{23} = (.87)(.63) + (0)(0) = .55$, etcétera. Entonces esta $\mathbf{R}^*$ se compara con la matriz de correlación obtenida u observada, $\mathbf{R}$, lo cual puede realizarse restando $\mathbf{R}^*$ de $\mathbf{R}$, o $\mathbf{R} - \mathbf{R}^*$. Dicha matriz de diferencias se llama una *matriz residual*. En el análisis estructural de covarianza los residuales casi siempre se analizan con uno de tres modelos de funciones de ajuste:

1. Mínimos cuadrados sin ponderar
2. Mínimos cuadrados generalizados o ponderados
3. Máxima verosimilitud

Como se estudió en el capítulo 34, existe un número de estadísticos diferentes de bondad de ajuste. El más antiguo de ellos —la chi cuadrada— algunas veces se informa, pero no es utilizado como el único estadístico, debido a que su evaluación se basa en el tamaño de la muestra. Conforme el tamaño de la muestra se agranda, pequeñas diferencias se vuelven estadísticamente significativas, e indican una falta de ajuste. Bentler (1980) ofrece una excelente revisión de los estadísticos de bondad de ajuste y sugiere el uso de estadísticos que no dependan del tamaño de la muestra. El programa EQS, desarrollado por Bentler (1986), originalmente utilizaba un estadístico de ajuste de este tipo, llamado

índice de ajuste normado de Bentler-Bonett o IAN, el cual ahora es obsoleto. El índice de ajuste actual de elección es el índice de ajuste comparativo (IAC), y un valor de .95 o mayor es representativo de un buen ajuste entre el modelo y los datos. Los valores del IAC menores de .95 indican al investigador que existe posibilidad de mejorar la manera en que se especifica el modelo. En esencia dice que el modelo no se ajusta muy bien a los datos. Si se obtienen valores alrededor de .95 o mayores, el ajuste de los datos al modelo es bastante bueno y es poco probable que cualquier reespecificación posterior del modelo altere mucho el índice. El LISREL, desarrollado por Joreskog y Sorbom (1993), originalmente utilizaba un estadístico de bondad de ajuste diferente, llamado la Raíz del Cuadrado Medio Residual (RMR por sus siglas en inglés). No obstante, ahora todos los programas populares poseen los estadísticos de ajuste más comunes, tales como el índice de bondad de ajuste y el índice ajustado de bondad de ajuste (véase Comrey y Lee, 1992; o Keith, 1999, para encontrar mayores explicaciones respecto a tales índices).

A partir de los resultados del EQS, el estadístico chi cuadrada para el modelo dualista es 121.253, basado en 8 grados de libertad. El valor de probabilidad para el estadístico chi cuadrada es menor que .001, lo cual indica que los datos no se ajustan al modelo. El índice de ajuste normado Bentler-Bonett fue de .840, lo cual indica que es posible hacer mejoras para lograr un mejor ajuste.

Ahora se revisarán las ideas que subyacen al método. El principio es: *a menor tamaño de los residuales, mejor será el ajuste; a mayor tamaño de los residuales, más pobre será el ajuste.* Si la hipótesis o modelo es válido empíricamente, menos diferencia habrá entre la matriz de covarianza (correlación) generada a partir del modelo, $\mathbf{R}^*$, y la matriz de correlación observada, **R**. Ambas situaciones se reflejan en la matriz de residuales, $\mathbf{R} - \mathbf{R}^*$, y en medidas, como el IAN, que refleja la magnitud de los residuales. Nuevamente, a mayor tamaño de los residuales, más pobre será el ajuste. (Los tres programas computacionales de estructura de covarianza imprimen la matriz residual de manera obligatoria.)

Las implicaciones empíricas de la hipótesis de bipolaridad se describen en la figura 35.2B. Evidentemente existe sólo un factor: F1. Las medidas del conservadurismo V1, V2 y V3 (o x_1, x_2 y x_3) están marcadas con "+"; y las de V4, V5 y V6 (o x_4, x_5 y x_6) están marcadas con "–", lo cual es consistente con la hipótesis de bipolaridad. Es decir, se espera un factor bipolar donde las medidas *conservadoras* tengan signos positivos; y las medidas *liberales*, signos negativos (o a la inversa). Las seis cargas factoriales estimadas por el EQS sobre *un* factor fueron .67, .83, .65, –.25, .12 y –.15. $\chi^2 = 313.143$, basado en 9 grados de libertad. El valor de probabilidad para el estadístico chi cuadrada es menor a .001. El IAN Bentler-Bonett = .586. Tales valores indican que la bondad de ajuste del modelo bipolar fue mucho peor que la del modelo dualista.

Las cargas factoriales son interesantes e informativas. Las de las tres medidas del *conservadurismo*, V1, V2 y V3, son positivas y sustanciales; las de las medidas del *liberalismo* son todas bajas. Evidentemente el modelo de un factor resulta inadecuado: las tres medidas liberales se "pierden". La χ^2 también es significativa, lo que indica una falta de ajuste. Ahora observe los residuales en la mitad superior de la tabla 35.5. Note cuidadosamente que los residuales de r_{45}, r_{46} y r_{56} son sustanciales: .416, .393 y .389. Las correlaciones entre las medidas liberales, V4, V5 y V6 se "perdieron" con la solución de un solo factor, el modelo para la hipótesis bipolar. Parece ser que el modelo de bipolaridad no ha sido muy exitoso. Por otro lado, el modelo dualista se desempeñó mejor en todos los cálculos.

Ahora se realiza una prueba final: se comparan directamente los dos modelos, lo cual se hace por medio de las pruebas de χ^2. La χ^2 para el modelo bipolar fue de 313.143, con nueve grados de libertad, mientras que la χ^2 para el modelo dualista fue de 121.253, con 8 grados de libertad. Recuerde que antes se pidió a la computadora que estimara el valor de r_{12}, aunque estrictamente hablando, se debió haber fijado en cero, o $r_{21} = 0$. Ello se debe a

que el modelo dualista puro predice factores ortogonales. Una de las razones principales para hacer esto fue "agotar" un grado de libertad para comparar las χ^2 de los dos modelos. La prueba directa es $\chi^2 + 2_{bip} - x_{du} - 121.253 = 191.89$. También se restan los grados de libertad: 9 – 8 = 1. Si no se hubiera estimado r_{12}, los grados de libertad para los dos modelos hubieran sido los mismos, haciendo imposible una comparación de χ^2. Se evalúa χ^2 = 191.89, con gl = 1; es altamente significativa, lo cual indica la superioridad de la hipótesis dualista (puesto que la χ^2 del modelo bipolar es significativamente mayor que la χ^2 del modelo dualista). Si no hubiera una diferencia significativa entre las χ^2 de los dos modelos, entonces la hipótesis bipolar sería tan "buena" (o tan "pobre") como la hipótesis dualista. Por consiguiente, no es posible inferir que una hipótesis sea más satisfactoria que la otra. Recuerde que un modelo que es congruente con los datos tendrá una χ^2 no significativa estadísticamente. Sin embargo, si la diferencia entre las χ^2 es significativa, entonces se infiere que el modelo con el valor χ^2 mayor resulta *menos* satisfactorio que el modelo con el valor χ^2 menor. Otra forma de explicarlo es que si la diferencia entre las χ^2 es significativa, prueba la importancia de los parámetros que diferencian a los modelos.

Lo arriba expuesto es difícil de mostrar y explicar, de acuerdo con la manera en que se ha realizado el problema. Un enfoque más elegante es el siguiente. Se establece el modelo dualista de la forma en que se hizo antes. Después se establece el modelo bipolar exactamente de la misma forma, con excepción del término r_{12}. Para el modelo dualista se estima r_{12} igual que antes. Esto producirá una χ^2 con gl = 8. Ahora se establece el modelo bipolar fijando r_{12} = 1.00, con gl = 9, lo cual producirá exactamente los mismos estimados de los parámetros que si se le hubiera indicado al programa que había sólo un factor, excepto que las cargas del único factor aparecerán en dos factores. Puesto que la correlación entre los dos factores es 1.00, el efecto neto es el mismo que con un factor. La prueba de las hipótesis alternativas, $\chi^2_{bip} - \chi^2_{dual}$ será igual a la anterior. No obstante, ahora queda claro que los dos modelos difieren tan sólo en un parámetro: Φ_{21}. Ésta es una de las razones por las que se calcula Φ_{21} o r_{12} en el modelo dualista: para hacer una prueba de hipótesis alternativas debe existir una diferencia en los grados de libertad. Además, un modelo debe ser un subconjunto del otro modelo, lo que significa que ambos modelos estiman los mismos parámetros, con excepción (en este caso) de un parámetro.

Influencias de las variables latentes: el sistema EQS completo

En el ejemplo anterior sobre las actitudes se utilizó sólo una parte de la estructura de covarianza o del sistema de modelamiento de ecuaciones estructurales. Si la solución que se pretendía era el análisis factorial ordinario de primer orden, entonces lo que se hizo era todo lo necesario. Sin embargo, los problemas más interesantes estudian las relaciones entre variables independientes y variables dependientes. Antes de explicar las propiedades formales del sistema, se examinará un ejemplo ficticio simple. Se estableció el diagrama de ruta del ejemplo, para tener algo concreto a qué referirse; dicho diagrama se muestra en la figura 35.3. El ejemplo es un pequeño modelo de *habilidad* y *aprovechamiento*. En efecto, se dice: la *habilidad verbal* y la *habilidad numérica* influyen en el aprovechamiento de forma positiva. Aunque tal vez el ejemplo no sea muy interesante, posee la virtud de ser obvio y fácil de comprender. Aquí no se intenta probar hipótesis alternativas, aun cuando existe una diversidad de posibilidades. Tan sólo se busca transmitir la esencia del sistema.

Observe este sistema y su función desde el punto de vista de la regresión. Primero, considere la parte izquierda de la figura 35.3. Hay cuatro pruebas: *prueba verbal 1*, *prueba*

verbal 2, *prueba numérica 1* y *prueba numérica 2*, V1, V2, V3 y V4 usando notación del EQS. Son variables dependientes medidas. Se calcula la matriz de correlación 4×4, se analiza factorialmente y se obtienen dos factores, F1 y F2, como en la figura 35.3. Las flechas con asteriscos que señalan hacia las variables dependientes, a partir de F1 y F2, contienen las cargas factoriales: a_{11}, a_{21}, a_{32} y a_{42}. Las otras cargas se fijan en cero, tal como se hizo antes con la hipótesis dualista de actitudes:

Pruebas	I	II
1	a_{11}	0
2	a_{21}	0
3	0	a_{32}
4	0	a_{42}

Se pueden considerar las a como coeficientes de regresión. La ecuación de regresión para V1 es:

$$V1 = a_{11}F1 + E1$$

Se busca la regresión de V1 a partir de F1, tal y como se buscó la regresión de y a partir de x, o de y a partir de $x_1, x_2, \ldots$ Es factible pensar que las cargas factoriales a, tienen la misma función que los coeficientes de regresión, b o β, del capítulo 32. El mismo razonamiento se aplica a la parte derecha de la figura 35.3: se escribe la regresión de V5 a partir de F3 como:

$$V5 = a_1F3 + E5$$

En dicha estructura de covarianza existen dos análisis factoriales o sistemas de regresión separados: uno en la parte izquierda, y otro en la parte derecha. Cualquiera de las dos partes puede utilizarse para el análisis factorial confirmatorio o de comprobación de hipótesis, como cuando se probaron las hipótesis dualista y bipolar. No obstante, lo que es más interesante e innovador es plantear y responder preguntas de investigación acerca de la regresión de la variable latente a partir de otra(s) variable(s) latente(s). En efecto, se pregunta sobre las relaciones entre F1, F2 y F3, o la regresión de F3 a partir de F1 y F2, considerando a F1 y F2 como variables independientes, y a F3 como variables dependientes. Esto es lo que hace la ecuación estructural o el análisis estructural de covarianza.

El problema de investigación de la figura 35.3 se expresa como la relación multivariada entre las variables independientes y las variables dependientes. Puede resolverse, por ejemplo, utilizando la correlación canónica, la cual expresaría la relación general entre las V del lado izquierdo (por ejemplo, V1, V2, V3 y V4) y las V del lado derecho (V5, V6 y V7). Pero la correlación canónica no es capaz de refinar las relaciones. Consigue la relación entre dos conjuntos de variables utilizando *todas* las variables. Por lo general, no tiene nada que ver con las variables latentes. El modelo e hipótesis implicadas en la figura 35.3 indican, en efecto, que las variables dependientes medidas de la parte izquierda reflejan dos factores: *habilidad verbal* (V1 = *prueba verbal* 1, y V2 = *prueba verbal* 2) y *habilidad numérica* (V3 = *prueba numérica* 1, y V4 = *prueba numérica* 2). Las variables latentes son *habilidad verbal*, F1 y *habilidad numérica*, F2. Las tres pruebas de aprovechamiento son *lectura*, V5; *matemáticas*, V6; y *estudios sociales*, V7. Se presume que miden un factor, aprovechamiento —en otras palabras, una hipótesis unifactorial—. Note cuidadosamente que las hipótesis que no son satisfactorias, en el sentido de que no son congruentes con los datos, pueden establecerse fácilmente, invalidando al modelo completo. Por ejemplo, las variables V5, V6 y V7, que se dijo medían el reflejo de *un* factor o variable latente, podrían

ser incorrectas. Quizá se necesiten dos factores. Es decir, la figura 35.3 tiene un factor, F3, para el *aprovechamiento*, pero pueden ser en realidad dos factores, F3 y F4. Después de todo, V5 es una prueba de lectura y V6 es una prueba de matemáticas, y se sabe que, por lo general, estos dos factores son diferentes. Si esto es así, entonces el modelo de la figura 35.3 es deficiente al respecto.

Establecimiento de la estructura del EQS

Por último llega la relación crucial: la de F1 y F2, las variables independientes latentes; y F3, la variable dependiente latente. La hipótesis sustantiva establecería qué tanto la *habilidad verbal* F1, y la *habilidad numérica* F2 influyen en el *aprovechamiento* F3.

Dicha hipótesis no es demasiado fascinante, más bien resulta sensible para ejemplificar y explicar. Para probarla, se deben plantear el problema y modelo de la figura 35.3 en proposiciones y estructura del programa EQS. Se trata de un paso complicado y crucial en el EQS. Debido al riesgo de provocar tedio, es menester seguir las ideas y edificarlas en ecuaciones y ecuaciones de matrices, después de descubrir las ecuaciones de una variable individual. Primero, las ecuaciones para el lado izquierdo de la figura 35.3:

```
V1 = .3*F1          + E1
V2 = .3*F1          + E2                    (35.2)
V3 =       + .3*F2          + E3
V4 =       + .3*F2          + E4
```

".3*" indica que existen coeficientes que se estimarán a partir de los datos; éstos son las cargas factoriales. El valor ".3" es un valor inicial arbitrario de "adivinación" para la estimación.

Para la especificación del EQS, se escriben las mismas ecuaciones en forma de matriz:

$$\begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \end{bmatrix} = \begin{bmatrix} a_{11} & 0 \\ a_{21} & 0 \\ 0 & a_{32} \\ 0 & a_{42} \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \end{bmatrix} + \begin{bmatrix} E_1 \\ E_2 \\ E_3 \\ E_4 \end{bmatrix} \qquad (35.3)$$

Donde las *a* se estiman a partir de los datos. (El lector debe hacer una pausa aquí, estudiar la figura 35.4 y las ecuaciones 35.2 y 35.3, e intentar comprender su significado.)

El lado derecho resulta un poco más fácil:

```
V5 = 3*F3 + E5
V6 = 3*F3 + E6                              (35.4)
V7 = 3*F3 + E7
```

En forma de matriz:

$$\begin{bmatrix} V_5 \\ V_6 \\ V_7 \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} F_3 + \begin{bmatrix} E_5 \\ E_6 \\ E_7 \end{bmatrix} \qquad (35.5)$$

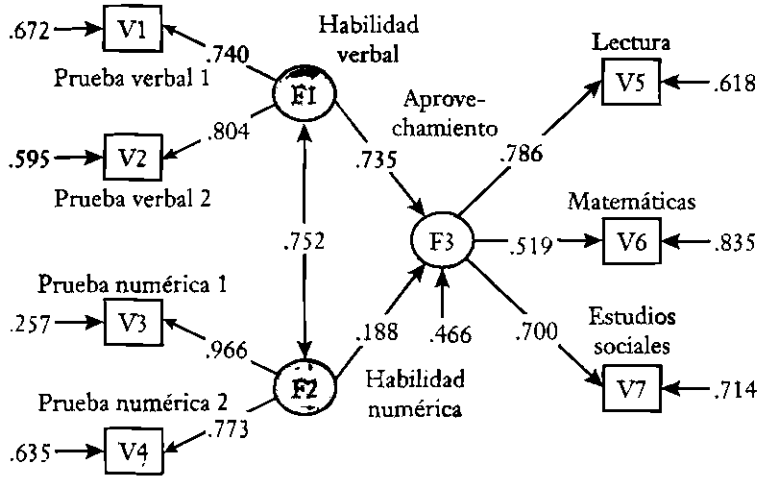
Se sintetizó una matriz de correlación ficticia, de tal manera que la solución del EQS apoye el modelo del diagrama de ruta de la figura 35.3, y las ecuaciones que se escribieron

con base en el diagrama. Los resultados fueron satisfactorios. El estadístico chi cuadrada fue estadísticamente significativo, lo cual indica una posible falta de ajuste. Sin embargo, como se mencionó anteriormente en este capítulo, existen otros índices que quizá sean mejores indicadores del ajuste. El *índice de ajuste normado Bentler-Bonett* (IAN) fue de .94, lo cual constituye un muy buen valor. Por lo común, se considera que cualquier valor de .90 o mayor indica un modelo bien ajustado. Otros índices calculados por el EQS, como el índice de ajuste comparativo (IAC = .95), apoyan la conclusión de que el modelo resultó satisfactorio.

Aunque los parámetros de los análisis factoriales (o análisis de regresión) para las partes izquierda y derecha de la figura 35.3 también son satisfactorios, no se informan. Lo anterior se debe a que el interés aquí es la comprobación del modelo respecto a su congruencia con los datos, en este caso una matriz de correlación. Además, también existe el interés en evaluar las relaciones entre las variables latentes *habilidad verbal* y *habilidad numérica*, por una parte, y el *aprovechamiento*, por la otra. Los valores que expresan estas influencias se presentan en la figura 35.4, la cual es igual a la figura 35.3, con excepción de que se muestran los parámetros estimados.

Anteriormente se dijo que el análisis estructural de covarianza y el programa computacional utilizado para realizar los cálculos complejos necesarios no eran sencillos de aprender. Aun utilizando el modelo más simple del EQS, quizá resulte difícil para el inexperto. Los 10 puntos antes mencionados son importantes para quienes deseen utilizar dicho método extremadamente poderoso. Sin embargo, incluso éstos no son fáciles de comprender. Entonces, ¿para qué molestarse aprendiéndolos? ¿No es posible realizar los análisis factoriales y los análisis de regresión de manera separada, para que el investigador del comportamiento se complique menos? Sí y no. Los análisis factoriales separados de las variables del lado derecho e izquierdo de la figura 35.3 pueden, de hecho, realizarse de forma separada. En efecto, los estudios psicométricos y analíticos factoriales deben realizarse antes de utilizar el EQS o el modelado de ecuación estructural. Pero obviamente el análisis de regresión apenas descrito no puede llevarse a cabo con problemas de investigación complejos que incluyan variables latentes y medidas indirectas y directas. De hecho,

FIGURA 35.4 *Mismo diagrama de ruta de la figura 35.3, con estimados paramétricos*



es posible intentar diversos enfoques para el análisis de los datos. Pero parece no haber una forma simple para estudiar conjuntos de relaciones complejas y para probar la congruencia de modelos teóricos con datos observados. Las ideas del análisis estructural de covarianza son matemática y estadísticamente poderosas, conceptualmente penetrantes y estéticamente satisfactorias. La creación del EQS, LISREL, AMOS y otros programas computacionales similares son logros bastante ingeniosos, productivos y creativos. Constituyen, hasta el momento, el más alto desarrollo del pensamiento analítico y científico del comportamiento; es un desarrollo que une la teoría psicológica y sociológica con el análisis matemático y estadístico multivariado en una síntesis única y poderosa que quizá revolucionará la investigación del comportamiento. Es en este sentido que se dice que el análisis estructural de covarianza es la culminación de la metodología contemporánea.

Estudios de investigación

En el relativamente poco tiempo que el análisis estructural de covarianza y los programas computacionales para realizarlo han funcionado y estado disponibles —desde inicios y mediados de los años setenta— el método se ha utilizado de manera fructífera en diversos campos. Algunos de dichos estudios son reanálisis de datos existentes; otros son estudios que se concibieron teniendo en mente un análisis estructural de covarianza (véase sugerencia de estudio 2). El primer estudio sobre la estructura de las actitudes, estudiado en el presente capítulo, constituye sólo uno de los 12 conjuntos de datos sobre actitudes que se reanalizaron con modelos de ecuación estructural o análisis estructural de covarianza. La mayor parte de la evidencia apoyó la hipótesis dualista (véase Kerlinger, 1980). Joreskog y sus colaboradores reanalizaron los datos de diversos estudios psicológicos y sociológicos (véase Magidson, 1979). El primer estudio descrito a detalle a continuación consiste en un reanálisis estructural de covarianza de los datos de un estudio grande sobre participación política en Estados Unidos.

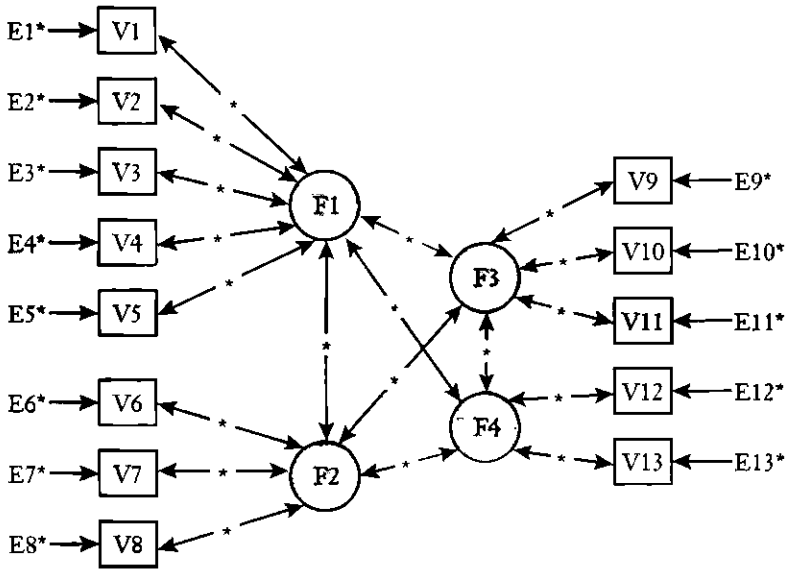
Bentler y Woodward (1978) utilizaron el análisis estructural de covarianza para reanalizar datos del Head Start —con resultados deprimentes—. Encontraron que el programa Head Start no tenía efectos significativos sobre las habilidades cognitivas de los niños que participaron en el programa. Judd y Millburn (1980) estudiaron la estructura de la actitud del público general en Estados Unidos. Utilizaron datos de panel de encuestas realizadas en 1972, 1974 y 1976, investigaron la opinión de Campbell, Converse, Miller y Stokes (1960) de que el público general no posee actitudes sociales estables y significativas. Encontraron que el público sin educación *sí* posee predisposiciones ideológicas consistentes.

Verba y Nie: participación política en Estados Unidos

En un estudio sobre participación política, Verba y Nie (1972) consideraron, a partir de la teoría política, que deberían existir cuatro factores detrás de 13 variables de participación política. Dichos factores y variables se presentan en la figura 35.5. Su estudio consistió en un análisis factorial confirmatorio. Parecían estar correctos en su hipótesis estructural, y se aplaudió su cuidadoso y competente trabajo. Pero el análisis factorial ha sido criticado, entre otras cosas, por su falta de rigor. ¿La hipótesis estructural de Verba y Nie puede someterse a una prueba más rigurosa? Permítase el uso del EQS en el problema. Sin embargo, note que Verba y Nie no utilizaron ecuaciones estructurales. Por consiguiente, los resultados aquí presentados provienen del reanálisis de sus datos.

El modelo de diagrama analítico de ruta que surge a partir de la discusión teórica de Verba y Nie se presenta en la figura 35.5. V1, V2, V3,..., V13 son las variables dependien-

FIGURA 35.5 *Diagrama de ruta (estudio de Verba y Nie)*



* Persuadir a otros sobre cómo votar; 2. Trabajo activo por un partido o candidato; 3. Acudir a junta o debate político; 4. Aportar dinero a un partido o candidato; 5. Participación en clubes políticos; 6. Votó en la elección presidencial de 1964; 7. Votó en la elección presidencial de 1960; 8. Frecuencia de votación en elecciones locales; 9. Trabajar con otros en problemas locales; 10. Formar un grupo para trabajar en problemas locales; 11. Participación activa en organizaciones comunitarias que resuelven problemas; 12. Contactar funcionarios locales; 13. Contactar funcionarios estatales y nacionales.

[b] Factor I: actividades de campaña (variables 1-5); factor II: votación (variables 6, 7 y 8); factor III: actividad de cooperación (variables 9-11); factor IV: contactar (variables 12 y 13).

tes medidas. Los componentes de error, asociados con cada variable dependiente, son E1, E2, E3,..., E13. Existen cuatro factores hipotetizados: F1, F2, F3 y F4; éstas son las variables independientes latentes. Se hipotetiza que los cuatro factores están correlacionados entre sí. Recuerde que cuando los factores están correlacionados, la solución es oblicua. Verba y Nie encontraron los siguientes factores: A-actividad de campaña (variables 1-5); B-votación (variables 6, 7 y 8); C-actividad de cooperación (variables 9-11); D-contactación (variables 12 y 13).

Se dieron instrucciones al EQS para calcular los estimados de los parámetros de la figura 35.5, y después de utilizar los parámetros para calcular una matriz de correlación predicha $\mathbf{R}^*$. Finalmente, para evaluar la adecuación del ajuste del modelo, de los cuatro factores oblicuos de la figura 35.5, se calcularon $\mathbf{R} - \mathbf{R}^*$, las diferencias o residuales y diversos estadísticos de "ajuste".

Los resultados generales apoyan el modelo de Verba y Nie de los cuatro factores oblicuos, a pesar de que la $\chi^2 = 406.648$, con $gl = 59$, es altamente significativa. El alto valor de χ^2 se debe claramente a la enorme N de 3 000 y, por lo tanto, no es una buena medida del ajuste. (Una solución idéntica con una N reducida de 300 produjo una $\chi^2 = 40.54$, que no es significativa.) La raíz del cuadrado medio residual (RCM) fue de .03. Este pequeño índice tan sólo reflejó los generalmente pequeños residuos. El índice de ajuste normado (IAN) Bentler-Bonett calculado por el EQS fue de .961, el cual es muy alto. El índice de

ajuste comparativo (IAC) fue muy alto, .966. Tales índices indican un muy buen ajuste. En síntesis, el ajuste del modelo de la figura 35.5 es bueno. El razonamiento teórico y el procedimiento de medición de Verba y Nie parecen ser adecuados. Ellos contribuyeron de forma significativa a la comprensión del proceso político y a la naturaleza y significado de la participación en el proceso político.

Brecht, Dracup, Moser y Riegel: relación entre la calidad marital y el ajuste psicosocial

En el capítulo anterior se expuso la investigación no experimental, que incluye aquellos estudios sin manipulación de las variables independientes. Por lo general, se estudia la variación entre las variables existentes y, en algunos casos, se implica débilmente una inferencia causal. Dichos estudios prevalecen en la investigación realizada en escenarios aplicados. Los investigadores que realizan investigación en escenarios aplicados por lo general no se dan el lujo de la asignación aleatoria o selección aleatoria. Keith (1999) afirma específicamente que una gran cantidad de estudios de investigación realizados en psicología escolar son de naturaleza no experimental. Otra área donde la investigación no experimental constituye la metodología dominante son las ciencias de la salud, en especial la investigación sobre el cuidado de pacientes. Las bases de datos sobre el cuidado de pacientes son grandes pero complejas y no experimentales. No obstante, es posible obtener información clave a partir de tales datos e investigación.

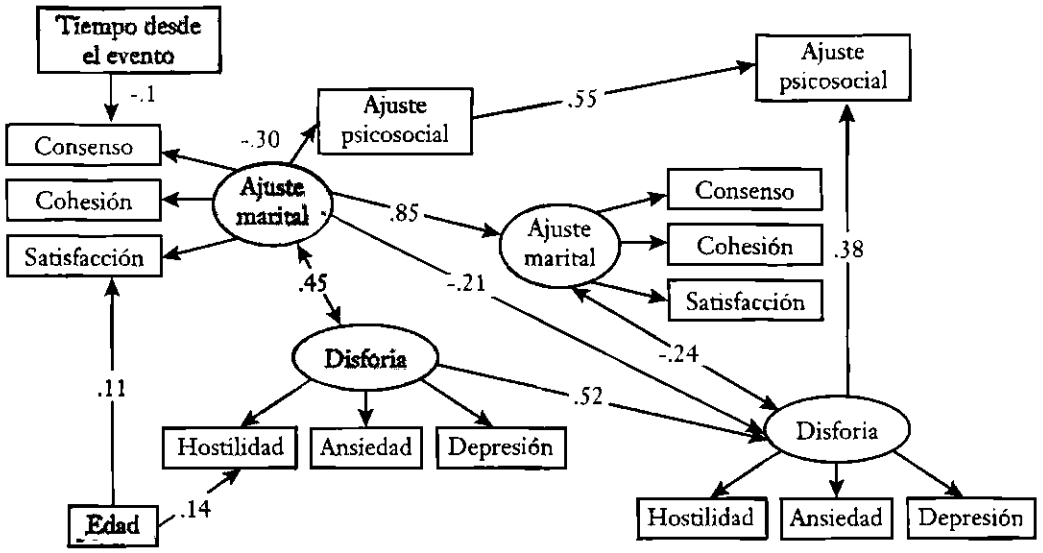
Un método fructífero para analizar los datos no experimentales sobre el cuidado de pacientes es el análisis estructural de covarianza o modelamiento de la ecuación estructural (MEE). Un estudio de este tipo que utilizó apropiada y exitosamente dicho método y, como tal demostró su valor, es el realizado por Brecht, Dracup, Moser y Riegel (1994).

Aquí los investigadores estudiaron el ajuste psicosocial de pacientes con enfermedad cardiaca. Investigación anterior sobre este tema sugería relaciones entre ciertas variables y el ajuste psicosocial, aunque no se ha definido la naturaleza precisa de la relación de las variables y el ajuste. Brecht *et al.* intentaron definir esto al señalar el hallazgo de que algunos pacientes se recuperan más rápido de la cirugía cardiaca que otros, así como de que también experimentan menos tensión emocional.

Bretch *et al.* plantearon la hipótesis de que la calidad de la relación marital, la disforia (ansiedad, depresión y hostilidad), la edad y el tiempo transcurrido desde la cirugía tenían posibles efectos directos e indirectos sobre el ajuste psicosocial. La muestra consistió en 198 pacientes cardiacos masculinos. Se tomaron mediciones de las variables en dos momentos diferentes. La primera fue al inicio del estudio (*momento 1*) y la segunda (*momento 2*) se realizó tres meses más tarde. La variable dependiente primaria era el *ajuste psicosocial*, el cual se midió utilizando la escala de ajuste psicosocial a la enfermedad (Psychosocial Adjustment to Illness Scale, PAIS). Calificaciones altas en la escala indicaban un mal ajuste. La calidad de la relación marital se midió utilizando la escala Spanier de ajuste diádico (Spanier Dyadic Adjustment Scale), y la disforia se midió a través del listado de adjetivos afectivos múltiples (Multiple Affect Adjective Checklist, MAACL).

El modelo completo original de Brecht y sus colaboradores no se incluyó en el artículo. Se trata de una situación común en los artículos publicados cuando el modelo es extenso y el espacio limitado. Por lo común se reporta el modelo final. Durante el proceso de comprobación del modelo se eliminaron rutas estadísticamente no significativas por medio de la prueba Wald, que es una de dos pruebas más populares para encontrar cuáles parámetros son innecesarios en el modelo (para mayores detalles, véase Bentler, 1995; Ullman, 1996). El modelo final de Brecht y colaboradores se presenta en la figura 35.6 y demuestra algunas de las cuestiones que puede realizar el análisis estructural de covarianza. Por un lado, la estructura del modelo puede probarse a través del tiempo. En otras pala-

FIGURA 35.6 *Modelo estructural de covarianza (datos de Brecht* et al.*)*



bras, es capaz de analizar datos complejos reunidos de manera longitudinal. Los datos o modelo del *momento 1* pueden considerarse como una medida de línea base de las relaciones. Los autores también buscaron la correlación entre el error medido a través del tiempo.

El modelo estructural final fue apoyado por los datos. La χ^2, el estadístico de bondad de ajuste, fue 109.41, con 90 grados de libertad. Dicho valor de chi cuadrada no fue significativo. El índice Bentler-Bonett (IAN) fue .95. Ambos estadísticos indican un buen ajuste de los datos empíricos con el modelo hipotetizado. Todos los coeficientes mostrados en la figura 35.6 fueron estimados por el programa EQS y resultaron estadísticamente significativos.

Los hallazgos de Brecht *et al.* implican que el desarrollo de una mejor relación marital promovería un ajuste psicosocial más sano a la enfermedad, si se acepta el concepto de que se está tratando con un modelado "causal". Con tal información los enfermeros o consejeros cardiacos pueden centrarse en el apoyo de una relación marital sana entre los pacientes cardiacos y sus esposas. Ellos son capaces de enseñar a la pareja estrategias para lograr una relación marital positiva. Al hacerlo así se ofrece al paciente la oportunidad de una mejoría significativa en el estrés emocional.

Conclusiones y reservas

Sería erróneo crear la impresión en la mente del lector de que todos los problemas atacados con el análisis estructural de covarianza o modelamiento de la ecuación estructural funcionan tan bien como los descritos en este capítulo, o que debe utilizarse con todos los problemas de investigación multivariada. Todo lo contrario. El propósito de esta sección final del capítulo consiste en intentar colocar el tema en una perspectiva razonable.

Primero se planteará la pregunta más difícil: ¿cuándo debe utilizarse este procedimiento? Como sucede siempre con dichas preguntas, es difícil decir con claridad y sin ambigüedades cuándo debe utilizarse. Un precepto bastante seguro es que *no* debe utilizarse rutinariamente o para análisis o cálculos estadísticos ordinarios. Por ejemplo, no debe emplearse para analizar factorialmente un conjunto de datos para "descubrir" los factores que están detrás de las variables del conjunto. Sencillamente no se ajusta muy bien al análisis factorial exploratorio y quizá sea un "destructor" al comprobar diferencias de medias entre grupos o subgrupos de datos. Si es posible utilizar un procedimiento más simple —como la regresión múltiple, la regresión logística, las tablas de contingencia multifactoriales o el análisis de varianza— y obtener respuestas a preguntas de investigación, entonces no tiene caso el uso del modelado de la ecuación estructural. Es evidente que se intentará un uso inapropiado. Cada vez es más fácil para los investigadores el uso de los programas computacionales LISREL, EQS y AMOS. Muchos de ellos aseguran en sus anuncios que "no se necesita experiencia para realizar el modelamiento de la ecuación estructural". Lo anterior significa que, entre otras cosas, el modelamiento de la ecuación estructural se utilizará con mayor frecuencia. A diferencia de otros procedimientos, el uso del análisis estructural de covarianza requiere de conceptos bastante difíciles, de la comprensión técnica de la teoría de medición, de la regresión múltiple y del análisis factorial. El rápido acceso a los programas computacionales "fáciles de usar" podría conducir al uso inapropiado del análisis estructural de covarianza. Lo mismo sucedió, aunque en menor grado, con el uso del análisis factorial. Aun así, el análisis factorial ha sido integrado "exitosamente" al cuerpo de la metodología de la investigación del comportamiento, aunque con frecuencia se utiliza inadecuadamente (véase Comrey, 1978). La naturaleza del software mismo hace que esto sea casi inevitable. Uno de sus propósitos es facilitar lo que en esencia no es fácil. Por consiguiente, se verá la publicación de muchos estudios que usarán LISREL, EQS, AMOS y otros programas similares, de forma inadecuada. En síntesis, dichos programas de estructuras de covarianza sólo deben utilizarse en una etapa relativamente tardía de un programa de investigación, cuando se requieran pruebas "cruciales" de hipótesis complejas.

El análisis estructural de covarianza se adecua mejor al estudio y análisis de modelos teóricos estructurales complejos, donde se utilicen cadenas complejas de razonamientos para ligar la teoría con la investigación empírica. Bajo ciertas condiciones y limitaciones, el sistema es un medio poderoso de comprobación de explicaciones alternativas de fenómenos de comportamiento. Resolver un problema de estructuras de covarianza de manera adecuada por lo general requiere de gran cantidad de ideas y análisis preliminares —lejos de la computadora y sus programas—.

Otro uso que bien puede dársele a los programas computacionales de estructuras de covarianza es la verificación de resultados complejos, surgidos a partir de otros análisis. En el pasado, por ejemplo, el análisis de ruta se ha utilizado para analizar los datos de muchos problemas de investigación. A pesar de que el análisis de ruta es un modelo útil para los problemas de investigación —es particularmente útil en la conceptualización de los problemas— no puede lograr lo que hacen programas computacionales como el LISREL. Maruyama y Miller (1979) señalaron esto cuando explicaron por qué utilizaron el LISREL para reanalizar los datos de disgregación de Lewis y St. John (1974). Los programas de modelamiento de la ecuación estructural como el EQS y el LISREL con frecuencia tienen la capacidad de establecer limpiamente aspectos de hipótesis de investigación que otros métodos no logran. Aun así, no es una metodología aplicable de manera general. En definitiva, no se trata de una panacea para estudios mal diseñados.

Con frecuencia existen dificultades técnicas al utilizar tales métodos. Ya se han comentado χ^2 grandes y significativas con grandes cantidades de sujetos, y se han sugerido

remedios, especialmente el estudio de residuales y el uso de otros índices de bondad de ajuste, como el IAN de Bentler o el IAC, los cuales no dependen del tamaño de la muestra. Otro remedio consiste en la comprobación de hipótesis alternativas cuando el problema permite dicha comprobación.

Uno de los problemas más difíciles es el de la *identificación*. Un modelo que se prueba requiere estar sobreidentificado. Lo anterior quiere decir que deben existir más puntos de datos, por lo general varianzas y covarianzas, que los parámetros estimados. Si hay n variables dependientes medidas, entonces no puede haber más de t parámetros estimados a partir de los datos, donde $t = n(n + 1)/2$. Si $n = 5$, entonces $t = 5(5 + 1)/2 = 15$, y no más de 15 parámetros pueden estimarse en un modelo. Existen otras condiciones que pueden hacer que un modelo no sea identificable, pero es extremadamente difícil especificarlas de antemano.

La dificultad técnica más común está íntimamente relacionada con la identificación. Por cualquier razón o combinación de razones, el programa computacional quizá no corra y anuncie que "algo" anda mal. ¿Pero qué es? Por otro lado, la ejecución de la computadora puede haberse completado y alguno de los parámetros tal vez no tenga sentido. Por ejemplo, es posible que se reporten varianzas negativas. ¿Por qué? Cualquiera que haya utilizado programas de computación "enlatados" en cualquier grado está familiarizado con los tétricos mensajes que presenta la computadora. Cuando se consulta a un experto, la respuesta es invariable: "Hay algo que está mal en el modelo." ¡Sí, claro! ¿Pero qué es? Y, naturalmente, los modelos teóricos con frecuencia no se ajustan: "¡Hay algo que está mal en el modelo!" Y también a menudo ocurre el análisis por computadora que funciona magníficamente, pero los estadísticos indican que el modelo del investigador no se ajusta. ¿Está mal la teoría? Si se está fuertemente comprometido con una posición teórica, quizá sea difícil admitirlo. En cualquier caso se deben verificar diversas posibilidades. Primero, el modelo no se ajusta porque fue conceptualizado pobre e incorrectamente. Segundo, no se ajusta debido a que el usuario de los programas computacionales cometió un error (o dos o tres) al utilizar el sistema. Tercero, el análisis computacional no funcionará a causa de que existen defectos en los datos (fuerte colinealidad en una matriz de correlación, por ejemplo); y cuarto, el modelo no se ajusta porque la teoría de donde fue derivado está equivocada o es inaplicable.

La medición inadecuada constituye una limitación de gran parte de la investigación del comportamiento. La dificultad técnica para medir variables psicológicas y sociológicas no ha sido apreciada aún por los investigadores en psicología, sociología y educación. No resulta sencillo diseñar pruebas y escalas para medir constructos psicológicos y sociológicos. Tampoco es fácil realizar la investigación psicométrica para establecer la confiabilidad y validez de las medidas utilizadas. Evidentemente es más difícil aun admitir que las propias medidas son deficientes. Consulte la investigación de Comrey sobre personalidad en el capítulo 34. El desarrollo de la *escala de personalidad de Comrey* tomó 15 años de trabajo. Con demasiada frecuencia las medidas de uso común en la investigación del comportamiento se aceptan y utilizan sin cuestionamiento; y pocas veces se cuestionan los supuestos que subyacen a las variables de estudio. Si, por ejemplo, se mide el *autoritarismo*, se asume que parte de la variable latente *autoritarismo* es el *antiautoritarismo* (cualquier cosa que esto sea). Al inicio de este capítulo y en el capítulo 34 se analizó una investigación que surgió a partir del cuestionamiento del supuesto común de que el *conservadurismo* y el *liberalismo* son opuestos lógicos y empíricos. Por desgracia, se ha realizado una serie de estudios —y se han estropeado— debido a que se realizó la medición de actitudes sociales basadas en este supuesto (véase Kerlinger, 1984). De forma similar, otros estudios han sido estropeados, tal vez arruinados, tanto por supuestos incorrectos como por una medición inadecuada.

No es posible construir una metodología analítica, no importa qué tan bien concebida y poderosa sea, para medidas cuya confiabilidad y validez sean insatisfactorias. La validez hecha a partir de supuestos es una amenaza particularmente severa para las conclusiones científicas, debido a que los procedimientos de medición no se cuestionan ni se prueban: se asume su confiabilidad y validez. Surge un análisis factorial pobre de factorizar lo que en efecto es una opción o construcción poco sustentada de pruebas y escalas. De manera similar, existe un uso pobre del análisis estructural de covarianza cuando algunas o todas las medidas utilizadas poseen una pobre base técnica en la teoría psicométrica e investigación empírica. El punto importante debe enfatizarse fuertemente: procedimientos elegantes aplicados a datos pobres, reunidos sin relación con la teoría y con el análisis lógico, no llegan a producir algo con valor científico.

Otra dificultad que enfrentan los usuarios del análisis estructural de covarianza consiste en que el análisis estructural multivariado moderno es bastante diferente de la mayoría de los primeros análisis estadísticos. La preocupación de la estadística clásica era evaluar si las diferencias entre medias observadas (en el análisis de varianza) o las contribuciones conjuntas y separadas de las variables independientes (en un análisis de regresión múltiple) eran estadísticamente significativas. No obstante, en el modelamiento estructural las implicaciones de una teoría se integran en un modelo que refleja la teoría y sus implicaciones: se incluyen variables latentes, se evalúan sus relaciones y efectos, y se somete a la estructura completa de relaciones a una comprobación simultánea. La(s) prueba(s) se basa(n) en la congruencia del modelo hipotetizado con los datos obtenidos. No es de sorprender que los investigadores experimenten fallas lógicas, técnicas y teóricas. De hecho, resulta sorprendente que los modelos sean comprobados exitosamente, dada la complejidad e incluso delicadeza de la tarea.

Sin embargo, no parece existir una alternativa razonable. La ciencia requiere de la formulación de teorías y de su comprobación empírica. La ciencia e investigación del comportamiento se enfrentan con explicaciones psicológicas y sociológicas de complejos fenómenos humanos y sociales. Por lo tanto, requieren tanto de teorías complejas, donde los conjuntos de variables observadas y latentes se relacionen entre sí, *como* de métodos complejos para conceptualizar y analizar los datos que se producen por observaciones y mediciones controladas de los conjuntos de variables. Hasta el momento, el análisis multivariado y el análisis estructural de covarianza parecen ser los caminos más promisorios para lograr las metas de las ciencias del comportamiento. El hecho de que van a plantear muchos problemas metodológicos difíciles, e incluso insolubles, es obvio. El hecho de que producirán avances y beneficios tanto teóricos como prácticos ya se ha demostrado en el presente capítulo.

A pesar de las dificultades y de las reservas mencionadas antes, no cabe duda de que el análisis estructural de covarianza y los programas computacionales que lo implementan son sobresalientes, altamente valiosos y son contribuciones útiles a la investigación científica del comportamiento. Se concluye este capítulo señalando que su uso e influencia tendrá efectos fuertes y saludables en el desarrollo de la teoría psicológica y sociológica y en su comprobación, así como en el avance material de la investigación científica del comportamiento en general.

RESUMEN DEL CAPÍTULO

1. El análisis estructural de covarianza también se denomina modelamiento de ecuación estructural (MEE). Durante algún tiempo se llamó modelamiento causal, aun-

que posteriormente se desechó el término, pues la causalidad no puede inferirse a partir de correlaciones.

2. El análisis estructural de covarianza moderno fue introducido por Bock y Bargmann, y fue desarrollado por Joreskog.
3. El análisis estructural de covarianza se considera la forma más elevada del análisis de datos de las ciencias sociales y del comportamiento. Es una compleja combinación de regresión múltiple y análisis factorial.
4. Los diagramas de ruta con frecuencia se utilizan para desarrollar de forma gráfica el modelo estructural que se va a probar. Existen ciertas reglas a seguir cuando se realizan diagramas de ruta, para facilitar su traducción en análisis de programas computacionales.
5. Existen esencialmente tres tipos de variables en el análisis estructural de covarianza:
   - Variables independientes (medidas o latentes)
   - Variables dependientes (medidas o latentes)
   - Medidas de error
6. Las variables latentes con frecuencia se llaman factores o variables sin medir.
7. El programa computacional utilizado por los autores para realizar el análisis estructural de covarianza es el EQS de Bentler.
8. El EQS se constituye (en opinión del segundo autor de este libro) un método que los principiantes comprenden con mayor facilidad.
9. El LISREL de Joreskog también se usa en muchos estudios reportados en revistas científicas. Se considera que el AMOS es otro programa de empleo sencillo.
10. Los programas computacionales que realizan el análisis estructural de covarianza son capaces de efectuar el análisis factorial confirmatorio.
11. La identificación es uno de los problemas que se encuentran en el modelamiento de estructuras de covarianza. El modelo necesita estar sobreidentificado.
12. Para que el modelo funcione apropiadamente es necesario que haya más puntos de datos que parámetros estimados.
13. El análisis estructural de covarianza se utiliza mejor en etapas tardías de investigación, donde el investigador ya ha reunido suficiente información sobre las relaciones entre las variables.
14. La disponibilidad de programas computacionales como el EQS, LISREL y AMOS, y su creciente facilidad de uso, conllevan la posibilidad de malos estudios de investigación.
15. Sin importar la metodología estadística empleada por el investigador, la validez continúa siendo una meta importante en los estudios de investigación.

SUGERENCIAS DE ESTUDIO

1. El tema del análisis estructural de covarianza presupone el conocimiento del álgebra matricial, del análisis factorial y del análisis de regresión múltiple. Casi todos los artículos de Joreskog resultan difíciles de leer. Sus primeros trabajos están contenidos en el libro de Magidson (1979). A continuación se presenta una lista de referencias que exponen el análisis estructural de covarianza. Algunos son bastante fáciles de leer:

Bentler, P. M. (1980). Causal modeling. *Annual Review of Psychology, 31*, 419-456. [Una de las primeras y claramente escritas exposiciones sobre el análisis estruc-

tural de covarianza. Bentler y otros que realizaban investigación en esta área, en esa época, utilizaban el término causal. Ello fue criticado posteriormente por Freedman (1987) (véase la cita de Freedman más adelante).]

Cliff, N. (1987). Comments on Professor Freedman's paper. *Journal of Educational Statistics, 12*, 158-160. [Una actualización sobre su artículo de 1983, con alguna información proporcionada por el artículo de Freedman.]

Freedman, D. A. (1987). As others see us. A case study of causal modeling methods. *Journal of Educational Statistics, 12*, 101-128. [En este artículo se señaló que no es posible realizar inferencias causales a partir del uso de correlaciones. Esto condujo a que muchos evitaran el uso del término *causal* al tratar con el análisis estructural de covarianza. Freedman también afirma que existe una cantidad de supuestos que son difíciles de verificar y que pueden ser falsos en aplicaciones específicas.]

Hayduk, L. A. (1987). Structural equation modeling with LISREL: Essentials and advances. Baltimore: Johns Hopkins University Press. [Un libro bien escrito con excelentes explicaciones sobre el modelo LISREL para el análisis de ecuaciones estructurales. Sin embargo, no es propio para el principiante debido a que requiere del conocimiento del álgebra matricial.]

Loehlin, J. C. (1998). *Latent variable models: An introduction to factor, path, and structural analysis* (3a. ed.). Mahwah, Nueva Jersey: Lawrence Erlbaum. [Proporciona una buena definición del análisis de ruta y del análisis del rasgo latente. Señala las precauciones que debe tomar el investigador en el análisis estructural de covarianza. Se da un fuerte énfasis a los diagramas de ruta al explicar las ecuaciones estructurales. Los temas cubiertos pueden ser aplicados a cualquiera de los muchos programas computacionales para el análisis estructural de covarianza.]

Ullman, J. B. (1996). Structural equation modeling. En B. G. Tabachnick y L. S. Fidell (eds.), *Using multivariate statistics* (3a. ed.). Nueva York: Harper & Row, 709-811. [Un libro popular; el capítulo 14 está bien escrito y cubre los "derechos y reveses" del modelamiento de la ecuación estructural. Proporciona una buena comparación de los programas computacionales que realizan análisis estructural de covarianza. Con excepción del capítulo 14, el material del libro fue escrito enteramente por Tabachnick y Fidell; Ullman es el único colaborador.]

2. A continuación se presentan seis estudios de investigación que utilizaron de forma benéfica el análisis estructural de covarianza:

Holahan, C. J., Moos, R. H., Holahan, C. K. y Brennen, P. L. (1995). Social support, coping, depressive symptoms in a late-middle-aged sample of patients reporting cardiac illness. *Health Psychology, 14*, 152-163. [Con el uso del LISREL, desarrollaron un modelo predictivo de los síntomas depresivos. El artículo contiene la matriz de correlación de nueve variables observables. Es ideal para estudiantes que desean probar el uso de los programas EQS, LISREL o AMOS.]

Keith, T. Z. (1999). Structural equation modeling in school psychology. En C. R. Reynolds y T. B. Gutkin (eds.), *Handbook of school psychology* (3a. ed.). Nueva York: Willey, 78-107. [Un capítulo sobresaliente escrito por uno de los principales metodólogos de investigación en psicología escolar. Keith ofrece un panorama sobre la forma en que el análisis estructural de covarianza o modelamiento de la ecuación estructural se utiliza para manejar estudios no experimentales complejos en psicología escolar. Es de fácil lectura. Altamente recomendable.]

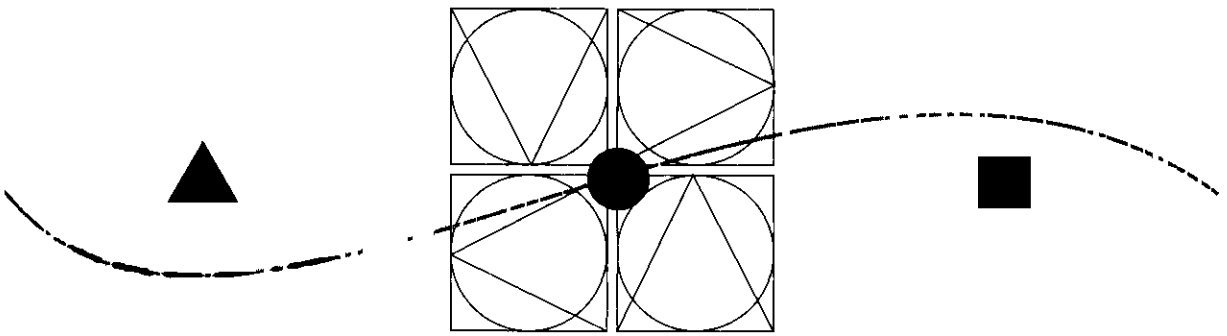
Musil, C. M., Jones, S. L. y Warner, C. D. (1998). Structural equation modeling and its relationship to multiple regression and factor analysis. *Research in Nursing and Health, 21*, 271-281. [Utiliza un modelo conceptual y no técnico para explicar cómo se pueden utilizar las ecuaciones estructurales en la investigación sobre el cuidado del paciente. Los autores muestran cómo el método fue utilizado para estudiar las relaciones entre el estrés, las presiones y la salud física en personas de la tercera edad.]

Nyamathi, A., Stein, J. A. y Brecht, M.L. (1995). Psychosocial predictors of AIDS risk behavior and drug use behavior in homeless and drug addicted women of color. *Health Psychology, 14*, 265-273. [Estos autores desarrollaron un modelo estructural que relaciona recursos personales y sociales, estilos de enfrentamiento, reducción del riesgo y riesgo de SIDA. Obtuvieron un conjunto de factores (variables latentes) y después modelaron dichos factores.]

Wolfle, L. y Robertshaw, D. (1982). Effects of college attendance on locus of control. *Journal of Personality and Social Psychology, 43*, 802-810. [Estudio interesante y bien realizado sobre datos de un estudio longitudinal nacional de la generación 1972 de preparatoria.]

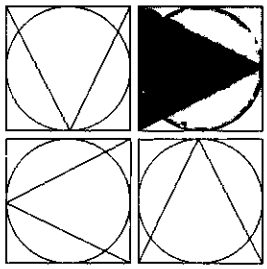
Wyllie, A., Zhang, J. F. y Casswell, S. (1998). Positive responses to televised beer advertisements associated with drinking and problems reported by 18 to 29-year-olds. *Addiction, 93*, 749-760. [Utilizó el modelamiento de la ecuación estructural para estudiar la relación entre respuestas a los anuncios de alcohol y la conducta de beber, y los problemas relacionados con el alcohol. Los investigadores plantearon la hipótesis de que las respuestas positivas a los anuncios de cerveza televisados contribuyen a la cantidad de alcohol consumido en situaciones donde se bebe, lo cual, a su vez, contribuye al nivel de los problemas relacionados con el alcohol. El modelo resultó consistente con la hipótesis.]

Apéndices



Apéndice A

Apéndice B



APÉNDICE A

GUÍA PARA LA ELABORACIÓN DE REPORTES DE INVESTIGACIÓN

El principal medio de comunicación científica es el artículo de investigación. Al paso de los años el formato de dichos reportes se ha estandarizado para cubrir mejor los requerimientos de la comunicación científica. Los convencionalismos de la escritura de un reporte científico se relacionan con la organización del reporte y el estilo de presentación. La elaboración del reporte debe ser tanto breve como clara. Los errores tipográficos, tachaduras y enunciados mal escritos deslucen la presentación del reporte.

Cuando se elabora el reporte de un experimento es necesario que el investigador incluya todo lo que sea relevante al problema de estudio. Deben enfatizarse las bases teóricas del estudio. El lector debe ser capaz de entender la forma en que las predicciones surgen de la teoría. El reporte necesita ser claro en cada detalle en lo que respecta a la manera en que el estudio se realizó. Debe mostrar de forma precisa la manera en que se establecieron las condiciones para permitir la manipulación o el estudio de las variables en el orden demandado por las hipótesis. Debe ser lo suficientemente detallado para permitir que otro investigador independiente replique de manera exacta el estudio. Por último, el reporte requiere establecer qué resultados se obtuvieron y qué interpretación se realizará con ellos, dentro del contexto de la teoría. Un reporte experimental constituye un ciclo completo que inicia en la teoría y termina en la teoría.

Existe una cantidad de estilos populares de elaboración de reportes de investigación. Todos son similares en lo que respecta a lo que el investigador necesita incluir en el artículo. Sin embargo, los detalles dentro de los estilos difieren. Uno de los estilos más populares y "fáciles" es el utilizado por la Asociación Psicológica Americana, el cual con frecuencia se conoce como el "estilo de la APA". A pesar de que originalmente fue desarrollado para las revistas de psicología publicadas por dicha asociación, se ha extendido a revistas que no pertenecen a la APA y también a revistas que no pertenecen al área de la psicología.

Aquí se presentará tal estilo, ya que es el estilo de elaboración que se utiliza con mayor frecuencia en las ciencias sociales y del comportamiento. Inclusive muchas revistas de educación han adoptado este estilo. No obstante, las descripciones serán breves y una de las metas será ofrecer al lector una idea general de la forma en que se organizan los artículos de investigación, lo cual no sustituye al propio manual de la APA (1994). A partir de que surgió como un manual breve, dicho manual ha crecido considerablemente en tamaño y detalle. La edición más reciente, publicada en 1994, consta de 368 páginas. Evidente-

mente, una breve sección en un libro de texto no puede capturar todos los detalles contenidos en el manual completo. La presentación breve incluida aquí será suficiente para brindar al lector un poco de información que ayudará a esa persona cuando consulte el propio manual. Además, existen varias publicaciones dirigidas a ayudar al principiante a aprender y comprender este estilo de presentación (véase Gelfand y Walker, 1990; Hubbach, 1995; Parrott, 1994; Pyrczak y Bruce, 1992). El libro de Hubbach (1995) incluye secciones especiales sobre artículos científicos, el estilo de la APA y ejemplos excelentes.[1]

La estructura del artículo de investigación del estilo utilizado por la APA es:

1. Hoja de presentación con el título
2. Resumen
3. Introducción
4. Método
   a) Participantes
   b) Dispositivos/Materiales
   c) Procedimiento
5. Resultados
6. Discusión
7. Referencias
8. Tablas
9. Figuras

Hoja de presentación

La hoja de presentación es una página separada que contiene el título del estudio. Además contiene el encabezado y el nombre del (de los) autor(es) con su afiliación institucional. El encabezado es una descripción, en una o dos palabras clave, del estudio y aparece en cada página del manuscrito. Si el manuscrito llega a publicarse, el encabezado sirve como una herramienta útil de identificación para el editor.

El propósito del título consiste en proporcionar una descripción en breve del estudio. Para ofrecer la mayor información posible, los títulos por lo general incluyen las variables independientes y dependientes del experimento. Un modelo simple de un título es el siguiente: LAS VARIABLES DEPENDIENTES EN FUNCIÓN DE LAS VARIABLES INDEPENDIENTES. Para anotar un título para un experimento tan sólo se sustituyen la(s) variables(s) dependiente(s) y la(s) variables(s) independientes(s) en el modelo simple, lo cual daría algo como, *La estimulación sexual en función de la cafeína o reducción de la ansiedad por medio del uso de modelamiento videofilmado.* Títulos como "Experimento de psicología" o "Proyecto del curso" NO son aceptables. El título debe contener, por lo general, 15 palabras o menos.

Resumen

El propósito de un resumen es brindar una síntesis del artículo de investigación. Tiene que contener suficiente información para señalarle al lector el propósito y los resultados de la investigación. Debe contener los puntos principales de cada sección del artículo:

- la formulación del problema
- una descripción muy breve del método

[1] El autor agradece a Roberta J. Landi por llamar nuestra atención hacia el libro de Hubbach.

- una definición de todas las abreviaturas y acrónimos
- los resultados más importantes, y
- las conclusiones

El resumen se redacta como un solo párrafo sin separaciones. Al igual que la hoja de presentación, el resumen aparece en una hoja separada. No debe exceder de 15 líneas mecanografiadas a un solo espacio, o de 200 palabras. Tampoco debe incluir ningún dato ni interpretación extensa. Esta sección se intitula "Resumen" y el nombre va centrado en la página.

Introducción

a) La introducción debe iniciar con los *antecedentes* del experimento o estudio. Representa una justificación de la teoría y de las investigaciones previas relevantes al estudio. La introducción le indica al lector la importancia del estudio al ofrecerle una revisión breve sobre la literatura de artículos que son relevantes para el presente estudio de investigación. Si se utiliza el estilo de la APA, ésta es la única sección del artículo que no recibe un título. En otras palabras, en los reportes escritos en el estilo de la APA no aparece un título o encabezado de "Introducción". Se requiere ser preciso al reportar trabajos previos que sean relevantes al estudio de investigación. Cualquier cita textual tiene que escribirse dentro de comillas, anotando la referencia apropiada a la fuente de información. Es importante asegurarse de que los estudios citados sean relevantes al experimento. Las referencias en la introducción (o en cualquier otra sección del reporte) sobre artículos realizados por otros investigadores se realizan anotando el apellido de cada autor y el año de publicación. El año de publicación debe ir entre paréntesis. La referencia completa de cualquier cita se incluye en la sección de "Referencias", que aparece al final del reporte.

Ejemplo

Smith y Martin (1953) reportaron que el desempeño de sus participantes mejoró bajo estas condiciones; mientras que otros observadores notaron un decremento en el desempeño (Burns, 1950; Stevens y White, 1943).

Las referencias específicas respecto a información obtenida de un libro o artículo de revista se indican de la siguiente manera:

Thomas (1983, p. 304) reportó que...
Los resultados de un estudio previo (Carter, 1942, pp. 279-285) condujeron a...

Un investigador nunca debe incluir una referencia que no ha leído personalmente. Para reportar información sobre una fuente secundaria, se cita la fuente secundaria en el texto y se incluye en las referencias.

Ejemplo

En un experimento realizado por Jones y reportado por McGeoch (1952), se encontró que...

b) Una vez que se le han presentado al lector los antecedentes del estudio, la introducción continúa con el propósito o base teórica del experimento. Se establece el

problema específico de estudio junto con una base teórica o literaria de las hipótesis a comprobar y las predicciones y expectativas generales de los resultados de la investigación.

Ejemplo

La disforia puede actuar como una variable importante de confusión... Por lo tanto, se realizó un estudio para examinar las relaciones entre....

c) Después, la hipótesis debe traducirse a términos operacionales. El investigador debe especificar qué variables se manipularán (independientes), en caso de que las haya, y cuáles se observarán (dependientes). Las variables independientes y dependientes deben aclararse *sin* utilizar un enunciado que diga: "La variable independiente fue_____ y la variable dependiente fue_____." En su lugar, se debe decir algo similar a: "En este experimento (estudio), se investigó el número de respuestas correctas a las preguntas de la prueba, en función a la tasa de presentación de las preguntas" o "Se hipotetizó que los efectos indirectos de la relación marital sobre el ajuste psicosocial están mediados por..."

d) Por último, la Introducción se utiliza para definir cualquier término que se utilice por primera vez. Si el investigador desea referirse a las escalas de personalidad de Comrey como EPC, entonces podría escribir una oración como la siguiente:

Las escalas de personalidad de Comrey (Comrey, 1970), que de aquí en adelante serán referidas como EPC, se utilizaron para...

Un término como *aprendizaje* probablemente sea demasiado general para usarse en redacción técnica. Es mejor decir de manera exacta a cuál paradigma de aprendizaje se está refiriendo. Lo mismo se aplica para términos tales como *personalidad* o *ansiedad.*

Existe la tendencia a escribir demasiados detalles en la Introducción. Los detalles del experimento o estudio no deben presentarse en la sección de introducción del reporte. Los detalles se presentan en la sección de Método. Se permite incluir en la Introducción un esbozo o la metodología general seguida; pero sin detalles. Los resultados del estudio no deben aparecer en la Introducción; existe una sección separada para ello.

Método

La sección de Método tiene tres subencabezados: Participantes, Dispositivos y materiales, y Procedimiento. A continuación se describe cada uno de manera separada. La sección de Método, como un todo, describe la experimentación o la conducción del estudio. Tiene que escribirse con suficiente detalle para que otros investigadores puedan, con base en la descripción, repetir de manera *exacta* lo que se hizo.

Participantes

La sección de Participantes consiste en una descripción de las características de los participantes utilizados en el estudio o experimento. Indica quiénes fueron los participantes, cuántos fueron y cualquier detalle que pueda ser relevante. También incluye la manera en que se seleccionaron dichos participantes. Entre las distintas descripciones, la mayoría de

los estudios describe a los participantes en términos del género, edad, nivel educativo, etnia y cualquier otro factor relevante como éstos.

Ejemplo

> Los participantes fueron 48 personas elegidas de una muestra de gente sentada en una fuente ubicada frente al City Hall entre las 10 y 11 AM, un lunes del mes de julio de 1998. Los 18 hombres y 30 mujeres, con edades entre 15 y 35 años, representan cada tercera persona que se detenía en la fuente por un periodo de por lo menos 5 minutos. Otras 10 personas que fueron contactadas se negaron a responder el cuestionario.

Cuando se utilizan grupos de participantes, se debe proporcionar una descripción que indique al lector la forma en que los participantes fueron asignados a los diferentes grupos o condiciones de tratamiento.

Dispositivos y materiales (instrumentación)

Todos los materiales y dispositivos no triviales utilizados en el experimento deben ser descritos en suficiente detalle, para que alguien más pueda establecer una situación idéntica. Si el experimento requirió de lápices, no es necesario definirlos a menos que se trate de lápices raros que tengan un efecto específico en el estudio. Si se utilizan materiales o dispositivos estandarizados, como una prueba de personalidad, por lo común no se incluyen aquí, a menos que tengan algunas características especiales que hayan sido de suma importancia para el experimento. Instrumentos estandarizados ya existentes tales como las escalas de personalidad de Comrey se incluyen en la sección de procedimiento. Si se construyen o utilizan materiales o aparatos nuevos, como aquellos inventados por el investigador, deben describirse de manera completa. Hay que incluir información acerca de los materiales utilizados para medir y/o registrar respuestas. Si se utilizó cierto equipo especial en el experimento, como la "bobina oscilatoria modelo 9 Smith-Johnson" (Smith-Johnson Oscillator Coil Model 9), se requiere informar al lector dónde puede conseguir un recurso como ése.

Procedimiento

La sección de Procedimiento constituye una descripción o explicación de la secuencia de eventos que tuvieron lugar durante la realización del estudio o experimento. En síntesis, indica lo que el (los) experimentador(es) hicieron y lo que sucedió al (los) participante(s). Debe describirse lo que se hizo, en qué orden, durante cuánto tiempo, etcétera.

Ejemplo

> Se obtuvieron los datos de cada participante utilizando un cuestionario sobre la frecuencia del consumo de drogas, durante el cuarto periodo de un día escolar. Se les pidió a los participantes que indicaran la cantidad de alcohol que consumen.

Los métodos estadísticos utilizados para analizar los datos recolectados en el estudio y/o el diseño del estudio de investigación pueden presentarse en la sección de Procedimiento.

Es probable que el investigador descubra que se ha desviado del procedimiento que debía haber seguido; si así sucede, debe describir el procedimiento exactamente como se realizó —y no como tenía que haberse realizado—. Por lo general esto representa un descuido, pero en última instancia llega a disculparse.

Lo que no puede disculparse es la deshonestidad. Las secciones de Procedimiento tienden a complicarse demasiado en experimentos donde existen diversas fases o condiciones. En tal caso, con frecuencia resulta útil adoptar etiquetas para las fases o condiciones. Por ejemplo, con una máquina de enseñanza se podría dividir un ensayo en la "fase de estudio" y la "fase de prueba", con el propósito de hacer referencias posteriores distintivas y simplificadas.

Resultados

En la sección de Resultados se reportan los datos obtenidos en el estudio o experimento, así como los análisis realizados con éstos.

a) Comienza con una descripción de las medidas de la variable dependiente, registradas durante la sesión experimental. Con un ejemplo de la máquina de enseñanza, se registraría el número de errores y el lapso de tiempo ocupado en contestar las preguntas.

b) A continuación, se describen los datos del experimento. Los datos reportados por lo general serán algún tipo de resumen de los datos en bruto. Por ejemplo, tal vez se quiera reportar los resultados en términos de las medias y las desviaciones estándar o los datos en bruto registrados durante la sesión. Se afirmaría: "Se calcularon la media y la desviación estándar del número de errores para cada serie de 20 preguntas."

c) Después se hace referencia a los lugares donde es factible encontrar tales datos. Éstos pueden presentarse ya sea en tablas o en figuras (gráficas). Se les ordena por medio de números y se hace referencia a ellas por su número. Como ejemplo, se podría decir: "La media de los errores de cada serie de 20 preguntas se muestra en la tabla 1 (o figura 1)." Si se tienen diversas variables dependientes, los resultados de cada variable tal vez requieran de una tabla o figura separada. Las reglas para la preparación de tablas y figuras se presentan en una sección posterior de este apéndice.

d) Cuando se hace referencia a una tabla o figura, hay que describir las características importantes de los datos que aparecen en la tabla o figura. En una tabla o figura se presenta mucha información, por lo que es labor del investigador ayudar al lector a comprenderla. Se deben señalar los aspectos importantes, las tendencias generales y cualquier inversión o particularidad que se considere importante; es decir, aquello que parezca ser más que eventos aleatorios.

Es necesario apoyar el análisis de la información con una tabla o figura ofreciendo algunos valores de datos apropiados para ilustrar lo que se desea indicar. Sin embargo, no debe intentarse incluir todos los datos; para eso sirven las tablas y las figuras. Si se tienen 25 páginas de resultados de computadora, dichos resultados deben resumirse y colocarse en una tabla o figura. Si se incluye la hoja de resultados, debe ir como un anexo al documento. Los anexos muy grandes casi siempre son inaceptables si el manuscrito se va a publicar.

Ejemplo

Como se muestra en la figura 1, las funciones para las condiciones con recompensa y sin recompensa inician en el mismo nivel. El número de errores promedio en el primer ensayo fue de aproximadamente 4.5, para ambas condiciones. Después del primer

ensayo, los errores en la condición con recompensa comenzaron a disminuir en una proporción bastante estable, mientras que los errores en la condición sin recompensa permanecieron relativamente constantes. Por ejemplo, en el segundo ensayo se cometieron .50 menos errores en la condición con recompensa, pero para el sexto ensayo dicha diferencia se había incrementado hasta que se cometieron 3.39 menos errores en la condición con recompensa. De forma general, la media del número de errores disminuyó en función de los ensayos en la condición con recompensa; pero no así en la condición sin recompensa.

Una última regla para la elaboración de la sección de Resultados es que *no debe haber discusión* en esta sección. Esto es, no se da una opinión personal o interpretación de los resúmenes de datos. Sólo se presentan los hechos de los hallazgos en la sección de Resultados del reporte. Esto se explica con mayor detalle en la siguiente sección.

Discusión

El propósito de la sección de Discusión consiste en interpretar los resultados y explicar las conclusiones a las que conducen. Es aquí donde se aclara la contribución o valor del experimento o estudio.

a) Por lo común la discusión inicia con una declaración concisa sobre la importancia de los resultados.

Ejemplo

Los resultados del presente estudio coinciden con los de otros estudios que comparan el abuso de alcohol y drogas en jóvenes latinoamericanos.

b) A continuación sigue la interpretación de los resultados. Se hace una inferencia, a partir de las medidas dependientes particulares del experimento, a los procesos psicológicos de interés.

Ejemplo

Los jóvenes asiáticos han presentado de manera consistente menor consumo de drogas debido a que se sienten más amenazados por las consecuencias percibidas por el uso reconocido de drogas.

En apariencia resulta difícil distinguir entre la sección de Resultados y la de Discusión incluidas aquí. En la primera el investigador se adhiere estrictamente a las variables dependientes particulares del estudio. Todas las inferencias, interpretaciones, extrapolaciones y opiniones razonables pertenecen a la sección de Discusión.

Por ejemplo, en la sección de Resultados el estudio de investigación se refiere a la disminución de errores como una función de los ensayos en la condición con recompensa, mientras que los errores permanecen constantes en la condición sin recompensa. En la sección de Discusión, esto podría interpretarse como que la adquisición (o aprendizaje) se llevó a cabo en la condición con recompensa, pero no en la condición sin recompensa. La interpretación de que el aprendizaje se ha visto afectado de manera diferencial involucra una inferencia hecha a partir de los datos de error y, por lo tanto, pertenece a la Discusión y no a la sección de Resultados.

c) Los resultados del experimento o estudio deben, entonces, relacionarse con los resultados de otros estudios sobre problemas iguales o similares, y/o a cualquier teoría relevante con la que se esté familiarizado y que pueda documentarse. Es necesario señalar a qué grado los resultados coinciden o contradicen trabajos previos, qué tanto amplían el cuerpo de conocimientos, qué tanto apoyan o contradicen la teoría, etcétera. La relación de los resultados con otros resultados o teorías también requiere ser analizada. Si existe acuerdo, es suficiente con establecer de manera exacta en qué coinciden. En caso de que no coincidan, se debe ofrecer alguna posible razón de la discrepancia. Por lo general, la primera explicación que se le ocurre al autor es que hubo algún error en el experimento, lo cual puede o no ser verdad. Si es verdad, se debe señalar de forma exacta cuál fue la falla y *por qué*.

d) Cualquier falla o defecto en el experimento que limite la utilidad o generalización de las conclusiones obtenidas necesita discutirse. Cuando se reporta una falla también se debe explicar por qué se trata de una falla e indicar cómo puede corregirse. No es recomendable crear una larga lista de críticas, pues esto sólo creará una mala impresión en el lector o revisor.

e) Una manera adecuada para terminar la discusión consiste en sugerir cuál podría ser el siguiente experimento sobre dicho tema. Si se intenta hacer esto, es necesario asegurarse de explicar el experimento con suficiente detalle para que sea significativo, y se requiere explicar la razón que lo hace el siguiente paso lógico. Frases como las de los siguientes ejemplos deben evitarse ya que consumen tiempo y espacio.

Ejemplos

En el siguiente experimento es necesario utilizar recompensas más grandes.
Se debieron utilizar mejores participantes.

Aquí sería necesario explicar por qué el conjunto actual de participantes fue deficiente, y de qué manera contar con nuevos participantes sería diferente.

Referencias

Únicamente las referencias bibliográficas citadas en el cuerpo del documento deben incluirse en la lista de referencias. *Todas las citas deben aparecer en la lista de referencias.* La lista se realiza en orden alfabético respecto al apellido del primer autor. Después del apellido se anotan la primera y segunda iniciales del autor. *Los nombres de las revistas se escriben completos.*

En el caso de artículos de revistas, se citan los números de las páginas de todo el artículo. En el caso de los libros no se cita el número total de páginas. En un libro publicado, tan sólo se citan las páginas que pertenecen a la parte del libro (capítulo) escrito por el o los autores que se citan. A continuación se presentan ejemplos del estilo utilizado para los tipos de referencias más frecuentes. El tipo de publicación se anota en corchetes sólo para ayudar a identificar cada una; esto no se hace en la sección de referencias del documento.

Note que el estilo de escritura de las referencias de la Asociación Psicológica Americana lleva el año de publicación dentro de paréntesis, después de los nombres de los autores. El título del artículo sigue a la fecha. Sólo la primera letra de la primera palabra de cada oración del título del artículo va con mayúscula, y la primera letra de la primera

palabra después de una coma también se escribe con mayúscula. A continuación se enlista el nombre de la revista, el número del volumen de la revista y los números de las páginas del artículo. El nombre de la revista y el número de volumen se subrayan o se escriben en *itálicas*. Las letras iniciales del nombre de la revista se escriben con mayúsculas (mayúsculas iniciales).

En el caso de los libros, el nombre del autor va seguido por la fecha de publicación en paréntesis. Después se incluye el título del libro. Observe que sólo la primera letra de la primera palabra de cada frase en el título del libro se escribe con mayúscula. El título del libro también se subraya o escribe en *itálicas*. Después del título del libro se incluye el número de edición (2a. ed. [no en itálicas]), o el número de volumen (vol. 3 [no en itálicas]). Si no se requiere anotar una edición o volumen en especial, entonces después del título se escribe un punto. Para concluir, se anota la ciudad seguida por dos puntos (:), y después el nombre de la editorial seguido por un punto (.).

Ejemplos

Erlich, O. y Lee, H. B. (1978). Use of regression analysis in reporting test results for accountability. *Perceptual and Motor Skills, 47*, 879-882. [Artículo de revista, dos autores]

Hollenbeck, A. (1978). Problems of reliability in observational research. En G. Sackett, Ed., *Observing behavior: Vol. 2. Data collection and analysis methods* (pp. 79-98). Baltimore: University Park Press. [Capítulo en un volumen editado, donde el volumen también tiene un título especial]

Jeffrey, W. E. (1969). Early stimulation and cognitive development. En J. P. Hill, Ed., Minnesota Symposia on Child Development (vol. 3) (pp. 46-61). Minneapolis: University of Minnesota Press. [Capítulo en un volumen editado]

Kerlinger, F. N. (1986). *Foundations of behavioral research* (3a. ed.). Fort Worth, TX: Harcourt Brace. [Libro]

Stevens, S. S. (Ed.). (1951). Handbook of experimental psychology. Nueva York: John Wiley & Sons. [Libro con un editor como autor]

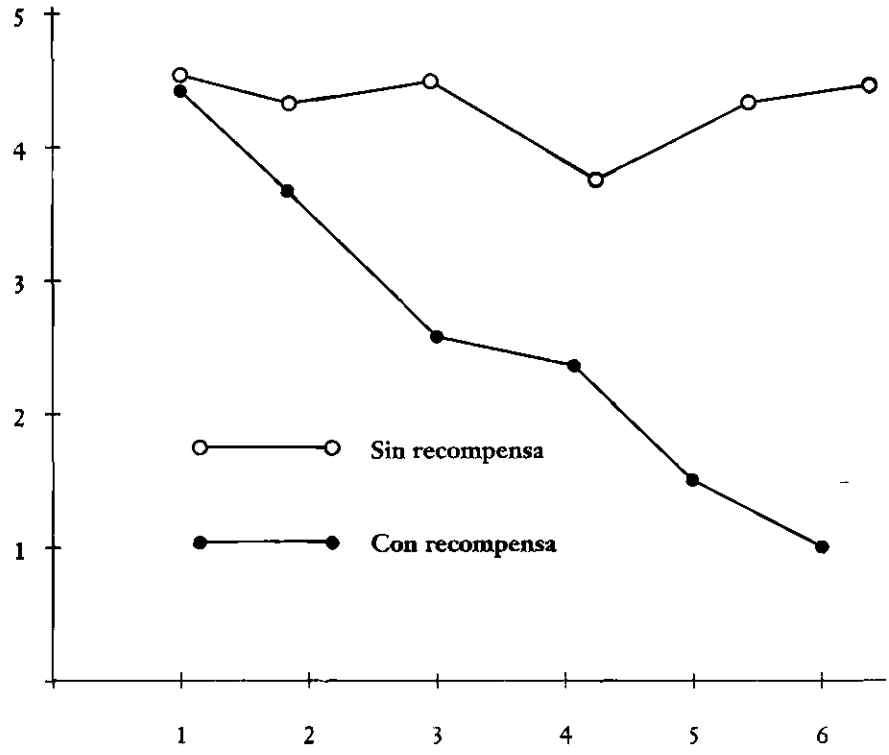
Yi, S. (1977). Some implications of Jeffrey's serial habituation hypotesis: A theoretical basis of resolving one-look versus multiple-look attentional account of discrimination learning. *Journal of General Psychology, 97*, 89-99. [Artículo de revista, un autor]

Preparación de figuras

Las figuras deben contener la información básica necesaria para su comprensión, sin referencia detallada al texto. Lo anterior requiere de la etiquetación cuidadosa de sus partes y un título o pie completo. Cuando se presenta más de una curva en la misma figura, se debe utilizar una leyenda, como en la figura A.1, o poner un nombre directamente a la curva. *El título o pie de la figura aparece debajo de ésta* y consiste de un muy breve resumen de lo que se está graficando. Deben evitarse títulos o pies como "Gráfica de los resultados" o "Una gráfica de..." Sólo se pone mayúscula a la primera palabra del título o pie y se finaliza con un punto. Las figuras se numeran de manera sucesiva con números arábigos. Se debe utilizar una regla para unir los puntos de datos. Hay que evitar dibujar una curva manualmente; de ser posible, debe utilizarse uno de los múltiples programas computacionales para crear gráficas. Por ejemplo, el Excel de Microsoft es capaz de producir gráficas con buena apariencia que son adecuadas para la presentación en artículos. Algunos programas procesadores de textos, como el Word de Microsoft, también son capaces de producir gráficas. En efecto, existen otros programas muy elaborados para la construcción de gráficas.

Ejemplo

FIGURA A.1 *Número promedio de errores en función a los ensayos, bajo condiciones de recompensa y sin recompensa.*



Preparación de tablas

Al igual que con las figuras, las tablas deben contener suficiente información para comprenderse de forma cabal, independientemente del texto. El título debe establecer de forma concisa lo que la tabla contiene. El título debe ser lo más específico posible. Se requiere evitar títulos como: "Tabla de datos", "Tabla de resultados" o "Tabla que muestra..." En general, se deben evitar abreviaturas poco comunes. Si es necesario deben explicarlas en una nota al pie de la tabla.

La tabla se ordena de forma que sea de fácil interpretación para el lector. En caso de ser necesario, debe utilizarse más de una tabla. *El título debe ir centrado encima de la tabla.* Los encabezados, el uso de mayúsculas y otras características importantes de una tabla pueden deducirse a partir del estudio de la tabla A.1 de este documento. Observe que programas de procesadores de textos como el Word de Microsoft poseen la capacidad de generar tablas. Es necesario asegurarse de que los datos incluidos en la tabla estén alineados apropiadamente; es decir, los signos de porcentaje, los decimales y las columnas. Esto facilita al lector el observar y seguir los datos.

Ejemplo

TABLA A.1 *Media de errores en función de los ensayos bajo las condiciones con recompensa y sin recompensa.*

	Media de errores	
Ensayos	*Con recompensa*	*Sin recompensa*
1	4.49	4.51
2	3.80	4.30
3	2.62	4.45
4	2.31	3.87
5	1.46	4.37
6	1.12	4.51

Aunque no existe un límite exacto para el número de figuras y tablas, muchas revistas tienen espacio limitado para los artículos y, por lo tanto, piden que se incluyan únicamente las tablas y figuras más necesarias en el documento real. De ser posible un investigador que posea una gran cantidad de tablas, figuras, hojas de resultados por computadora, etcétera, puede ponerlas a la disposición de los lectores interesados. Existe una organización que acepta dichos materiales y que con una cuota nominal los pone a la disposición de aquellas personas interesadas en ver los datos adicionales. Si el investigador elige utilizar este servicio, debe incluirse una nota de pie o referencia al respecto en el documento real. Dicha nota de pie podría ser como la siguiente.

Ejemplo

[1]Las matrices de correlación en las que se basa este estudio están depositadas en el National Auxiliary Publications Service. Revise el documento NAPS No. ___ con ___ páginas de material suplementario del NAPS, c/o Microfiche Publicatiions, 248 Hempstead Turnpike, West Hempstead, NY 11552. Envíe con anticipación, en dólares estadounidenses únicamente, $ ____ para obtener fotocopias o $ ____ para obtener una microficha. Fuera de Estados Unidos y Canadá añada un franqueo de $ ____ o $ ____ por el franqueo de la microficha.

Uso de abreviaturas

Cuando una palabra o término se utiliza con frecuencia en un reporte, puede abreviarse. Las abreviaturas para el (los) participante(s) y el (los) experimentador(es) son estándares y casi siempre se utilizan:

Ejemplos

sujeto = S sujetos = Ss del sujeto = S's de los sujetos = Ss' experimentador = E experimentadores = Es del experimentador = E's de los experimentadores = Es'

Otras abreviaturas no son estándares y deben definirse la primera vez que se utilicen.

Ejemplo

El aparato era un módulo de prueba visual (MPV). El MPV fue programado para... El instrumento de prueba fueron las escalas de personalidad de Comrey, que de aquí en adelante serán referidas como EPC, que consta de ocho escalas.

Estilo, tiempos, etcétera

La actividad que se describe en el reporte de investigación se llevó a cabo con anterioridad y se debe describir en tiempo pretérito. En raras ocasiones se utilizan pronombres personales, así como los deseos, anhelos, conclusiones, etcétera, del experimentador.

Ejemplos

Pobre:	**El experimentador deseaba encontrar...**
Mejorado:	El propósito del estudio fue...
Pobre:	Nosotros decidimos que el experimento mostraba...
Mejorado:	La conclusión del estudio fue...

Hay que evitar el uso excesivo de expresiones entre paréntesis. Existe la tendencia a referirse a las figuras y a las tablas en paréntesis, lo cual debe evitarse. Por ejemplo, no debe decirse: "La media de errores disminuyó en función a los ensayos (tabla 1)." Sin embargo, en oraciones largas la claridad de la lectura puede mejorarse a través del uso de expresiones entre paréntesis; se requiere tener en mente que se escribe para el lector en general, lo cual significa que los fenómenos deben explicarse. Sin embargo, no se debe adoptar la tarea de enseñar al lector sobre ciertas áreas de comprensión básica. Por ejemplo, es bueno referirse a la teoría de aprendizaje de Skinner sin entrar en los detalles de dicha teoría. No obstante, NO DEBE SUPONERSE QUE EL LECTOR conoce el estudio al que se está haciendo referencia.

Referencias

American Psychological Association (1994). *Publication manual for the American Psychological Association*. Washington, DC: Author.

Gelfand, H. y Walker, C. J. (1990). *Mastering APA style: Student's workbook and training guide*. Washington, DC: American Psychological Association.

Hubbach, S. M. (1995). *Writing research papers across the curriculum* (4a. ed.). Fort Worth, Texas: Harcourt Brace.

Parrott, L. (1994). *How to write psychology papers*. Nueva York: Harper-Collins.

Pyrczak, F. y Bruce, R. R. (1992). *Writing empirical research reports*. Los Ángeles, California: Pyrczak Publishing.

Muestra del reporte de una estudiante

Karen Siegel, quien actualmente es estudiante de doctorado en la Escuela de Psicología Profesional de California, en San Diego, escribió el siguiente reporte cuando era estudiante de licenciatura en la Universidad del Estado de California, Northridge. Éste ilustra de forma clara cómo se escribe un artículo utilizando el estilo de la APA. El reporte se reproduce con el permiso de Karen Siegel.

Forma en que el contacto físico y el género psicológico afectan el comportamiento prosocial

Karen Siegel
Universidad del Estado de California, Northridge

Contacto físico, género psicológico y ayuda

Resumen

El presente estudio predijo que el comportamiento prosocial (de ayuda) entre participantes femeninas y el experimentador se incrementaría con un contacto físico casual e intencional de corta duración. También se investigó el efecto del género psicológico de las participantes (andrógino, femenino o masculino) sobre su conducta de ayuda. 40 participantes voluntarias, que eran estudiantes, completaron un cuestionario (el Inventario del papel femenino de Bem; Bem Sex-Role Inventory), que midió su género psicológico, y después recibieron o no contacto físico, en un diseño entre sujetos de 2 × 3. La variable medida consistió en observar si el sujeto ayudaba a levantar lápices que se caían "accidentalmente". Los resultados mostraron que las participantes que recibieron contacto físico mostraron mayor comportamiento de ayuda que aquellas que no lo recibieron, aunque su género psicológico no tuvo ningún efecto.

Manera en que el contacto físico y el género psicológico afectan el comportamiento prosocial

Existe una gran cantidad de literatura respecto al efecto de diversas formas de contacto físico sobre el comportamiento humano y la salud. Muchos autores consideran que el contacto físico constituye una de las formas más básicas y tempranas de comunicación (Frank, 1957; Montagu, 1971), y que resulta crucial para un desarrollo emocional, social y físico sano (Harlow, 1958). Otra faceta de esta manifestación atávica de comunicación no verbal es su propiedad manipulativa. El contacto físico al servicio del control social (definido por Edinger y Patterson, 1983, como "una respuesta más deliberada y propositiva diseñada para promover un cambio en el comportamiento de la otra persona") ha sido examinado en diversos estudios, empleando variadas modalidades. En un estudio de gran influencia realizado en 1973, Henley (1973) indicó que el contacto físico comunica un mensaje de poder y estatus. Ella reportó una asimetría en los contactos físicos intercambiados entre los sexos (los hombres tocan más a las mujeres en público que a la inversa). Nguyen, Heslin y Nguyen (1975) reportaron que los tipos de contacto físico están asociados a los mismos significados tanto en hombres como en mujeres, aunque los sentimientos difieren y la decodificación precisa de un mensaje táctil depende no sólo de la forma en que se transmite, sino también del lugar donde se aplica.

Un análisis posterior que examinó la generalidad de la asimetría entre los sexos del uso de contacto físico intencional reveló complejidades que impiden una compresión simplista de este aspecto. Ciertos hallazgos de Hall y Veccia (1990) difieren de los de Henley, y sostienen que no está claro qué es lo que los diferentes contactos físicos significan para los sexos, y si la dominancia y el estatus pueden explicar los efectos en el sexo. Major, Schmidlin y Williams (1990) exploraron otras de las muchas variantes del tema de la asimetría. Encontraron que los patrones del género en el contacto físico varían de forma marcada por el ambiente y la edad, lo cual subraya la especificidad situacional de los comportamientos relacionados con el género.

En un contexto menos complejo, se han realizado otros estudios que relacionan el contacto físico (intimidad no verbal) y un incremento en la obediencia a requerimientos hechos (Kleinke, 1976). Willis y Hamm (1980) encontraron que el contacto físico es particularmente importante para lograr la obediencia de personas del mismo sexo. Paulsell y Goldman (1984) examinaron la influencia del contacto físico en diferentes partes del cuerpo (hombro, parte superior e inferior del brazo, y mano) sobre el comportamiento de ayuda. Ellos descubrieron que las mujeres cómplices de los investigadores obtuvieron respuestas de ayuda variables; mientras que los hombres cómplices de los investigadores obtuvieron ayuda con poca variación, independientemente de si los participantes habían o no sido tocados.

Los altos niveles de comportamiento de cuidado y apoyo se dan, discutiblemente, en función de lo que, en la cultura norteamericana, se considera como un comportamiento estereotípico femenino. De acuerdo con Bem (1975), el género psicológico del individuo determina su estilo y rango de comportamiento. Un autoconcepto estrechamente masculino puede inhibir el denominado comportamiento femenino (como ser afectuoso y gentil) y a la inversa; mientras que un autoconcepto andrógino, que incorpora pero no excluye atributos femeninos y masculinos, amplía el rango de comportamientos que pueden elegirse de una situación a otra.

En este experimento se estudió la cualidad manipulativa básica del contacto físico en una interacción limitada. Se probó el efecto de un contacto físico inocuo y casual, dentro del contexto de una interacción verbal, en el comportamiento de ayuda de estudiantes universitarias voluntarias. Además, se evaluó la androginia, feminidad y masculinidad psicológicas de las participantes, por medio del inventario del rol sexual de Bem (1974) (Bem's Sex-Role Inventory).

Contacto físico, género psicológico y ayuda

Se conjeturó que la condición que conduciría a la respuesta de ayuda más consistente sería la del sujeto con contacto físico con un perfil femenino y/o andrógino. De manera inversa, se predijo que aquellos sujetos con un perfil masculino y que no recibieran la situación de contacto físico proporcionarían la menor ayuda de todos.

Método

Se utilizó un diseño entre sujetos de 2×3 combinado, donde las situaciones con contacto físico y sin contacto físico funcionaron como una variable independiente; y el género psicológico de las participantes, como la otra variable independiente. El comportamiento de ayuda registrado consistió en observar si los participantes ayudaban a recoger lápices dejados caer por el experimentador, inmediatamente después de implementar la variable del contacto físico.

Participantes
40 mujeres estudiantes de la Universidad del Estado de California, Northridge, con edades que iban aproximadamente de los 18 a los 28 años, sirvieron como voluntarios. En un intento por eliminar cualquier variable extraña y posiblemente provocadora de confusión, respecto a las diferencias culturales en la actitud hacia el contacto físico y al espacio corporal personal, se incluyeron únicamente participantes criados en Estados Unidos.

Procedimiento
Este experimento empleó una condición con contacto físico y una sin contacto físico. El "contacto físico" consistió en una ligera palmada en la parte superior del brazo (el área que obtuvo el nivel más alto de comportamiento de ayuda en el estudio de Paulsell y Goldman [1984]).

Se aplicó el inventario del rol sexual de Bem a cada uno de los sujetos, en forma individual, después de indicarles que completaran el cuestionario. El experimentador les entregó un lápiz proveniente de una caja de 20, el cual posteriormente sirvió como accesorio. La mayor parte de los participantes completaron el inventario, que consistía en 60 adjetivos y frases impresas en una sola hoja, en 10 minutos o menos. También se les instruyó para que preguntaran por la definición de cualquier frase o palabra que no les fuese familiar. Después de que el sujeto completaba el cuestionario, el experimentador —quien permanecía en el cuarto todo el tiempo— caminaba hacia el sujeto sentado, recogía la forma y le agradecía por su participación, durante lo cual se daba o no el contacto físico en la parte superior del brazo. Inmediatamente después de esto, el experimentador dejaba caer lápices de la caja y se agachaba rápidamente para recogerlos. El sujeto se levantaba a ayudar (y se le otorgaba una puntuación de 1) o no lo hacía (y se le otorgaba una puntuación de 0). Antes de empezar a completar el inventario de Bem se le avisó al sujeto que permaneciera sentado al terminar, para hacer que la ayuda fuera una elección más accesible y como un intento de aplicar de manera más efectiva la variable del contacto físico.

Resultados

La proporción de participantes que ayudaron y recibieron contacto físico (.708) fue significativamente mayor que aquellos que no fueron tocados (.438), tal y como lo determinó una prueba z de una cola para proporciones ($z = 1.71$, $p < .05$). También se analizó la significancia del género psicológico de esta manera, comparando las puntuaciones del comportamiento de ayuda de las participantes andróginas con las femeninas, las andróginas con las masculinas y las masculinas con las femeninas, de manera separada; también comparando la variable de contacto físico en cada género psicológico, sin encontrar diferencias significativas entre las proporciones.

Una chi cuadrada, que comparó los efectos de la variable de contacto físico con el género psicológico de cada sujeto en su comportamiento de ayuda, no resultó significativa [$\chi^2 = .97$, $p > .05$].

Discusión

Tal y como se predijo, los resultados de este estudio indicaron que el comportamiento de ayuda de las participantes hacia un experimentador del mismo sexo se incrementó con el contacto físico casual, de corta duración, en la parte superior del brazo.

Se utilizaron participantes mujeres por dos razones. En primer lugar, no existía un cómplice hombre y era altamente probable que casi todos los participantes hombres ayudaran a una experimentadora mujer. (Paulsell y Goldman [1984] reportaron que el 90 por ciento de los participantes hombres tocados en la parte superior del brazo por cómplices mujeres, ayudaron a recoger objetos dejados caer por dichas cómplices.) En segundo lugar, el uso de participantes mujeres controló cualquier asimetría posible en el estatus relacionado al género percibido (Henley, 1973).

A pesar de que se tuvo cuidado en eliminar cualquier otra explicación para este hallazgo significativo, cualquier investigación futura de este tipo sería mejor si se diseñara como un estudio doble ciego, ya que las expectativas del investigador pueden transmitirse de manera no verbal, como con la dirección de la mirada (Kleinke, 1977) o el tono de voz (Goldman y Fordyce). (El experimentador mantuvo, lo más que pudo, una similitud en la comunicación que involucra tales expresiones en particular.)

La reciente popularidad del tema del comportamiento prosocial ha producido un número de combinaciones de manipulaciones para predecir el comportamiento de ayuda, con resultados significativos. Major, Schmidlin y Williams (1990) estudiaron el impacto de la edad de los participantes, aunado a la situación en que ocurría el contacto físico, sobre los patrones de género del contacto físico intencional. Otro estudio realizado por Hewitt y Feltham (1982) combinó el lugar del contacto físico (seis puntos a partir de la mano y hasta la espalda) con el género del experimentador y del sujeto. Nguyen, Heslin y Nguyen (1975) analizaron las interpretaciones de diferentes tipos de contacto físico y zonas corporales en hombres y en mujeres.

Este estudio se realizó a la luz del enfoque en el paradigma moderno de la condición (preferible) de la androginia psicológica. La androginia psicológica es la habilidad individual, como lo define Bem (1974), de no ser dicotómico en el rol sexual, sino tanto masculino como femenino, asertivo y productivo, instrumental y expresivo, para tener acceso a una gama completa de comportamientos que incluyen aspectos "masculinos" y "femeninos". Aunque no se encontró una relación significativa entre el comportamiento prosocial y el género psicológico en dicho estudio en particular, ésta puede ser aun una variable de interés en experimentos futuros específicamente diseñados para medir comportamientos que se vean afectados por el género psicológico del sujeto. Una variante podría ser emplear un experimentador femenino con participantes masculinos, medidos de la misma forma que en el presente estudio, pero utilizando una variable medible distinta que idealmente no involucre comportamientos de cortesía. Otras combinaciones del género del experimentador y del sujeto merecen ser examinadas, así como combinar el efecto del género psicológico con otras variables. En cualquier caso, estudios futuros están garantizados.

Referencias

Bem, S. L. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology, 42*, 155-162.

Bem, S. L. (1975). Sex role adaptability: One consequence of psychological androgyny. *Journal of Personality and Social Psychology, 31*, 634-643.

Edinger, E.. A. y Patterson, M. L. (1983). Nonverbal involvement and social control. *Psychological Bulletin, 93*, 30-56.

Frank, L. K. (1957). Tactile communication. *Genetic Psychology Monographs, 56*, 219-225.

Goldman, M. y Fordyce, J. (1983). Prosocial behavior as affected by eye contact, touch, and voice expression. *Journal of Social Psychology, 121*, 125-129.

Hall, J. A. y Veccia, E. M. (1990). More "touching" observations; new insights on men women and interpersonal touch. *Journal of Personality and Social Psychology 59*, 1155-1162.

Harlow, H. F. (1958). The nature of love. *American Psychologist, 13*, 673-685.

Henley, N. M. (1973). Status and sex: Some touching observations. *Bulletin of the Psychonomic Society, 2*, 91-93.

Hewitt, J. y Feltham, D. (1982). Differential reaction to touch by mean and women. *Perceptual & Motor Skills, 55*, 1291-1294.

Kleinke, C. L. (1977). Compliance to requests made by gazing and touching experimenters in field settings. *Journal of Experimental Social Psychology, 13*, 218-223.

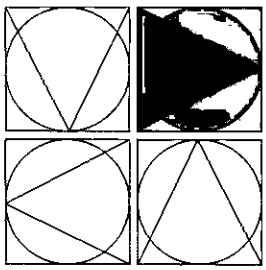
Major, B., Schmidlin, A M. y Williams, L. (1990). Gender patterns in social touch: The impact of setting and age. *Journal of Personality and Social Psychology, 58*, pp. 634-643.

Montagu, A. (1971). *Touching: The human significance of the skin.* Nueva York: Columbia University Press.

Nguyen, T. D., Heslin, R. y Nguyen, M. L. (1975). The meaning of touch: Sex differences. *Journal of Communications, 25*, 92-103.

Paulsell, S. y Goldman, M. (1984). The effect of touching different body areas on prosocial behavior. *The Journal of Social Psychology, 122*, 269-273.

Willis, F. N. y Hamm, H. K. (1980). The use of interpersonal touch in securing compliance. *Journal of Nonverbal Beahavior, 5*, 49-55.



APÉNDICE B[1]

TABLA A *Tabla de números aleatorios*

1	53	95	67	80	79	93	28	69	25	78	13	24	100	62	62	21	11	4	54	44	59	90	78	83	4	97	61	52	75	91
2	62	12	27	41	5	4	19	34	84	78	71	45	73	79	33	57	29	58	75	20	79	78	68	31	25	30	97	31	82	51
3	90	16	47	72	20	60	70	71	2	67	21	65	7	39	58	81	64	11	70	4	79	44	47	7	74	34	55	28	90	19
4	10	59	4	76	80	6	82	20	60	92	33	61	76	83	73	12	84	43	90	71	82	28	21	61	31	92	100	75	22	31
5	32	17	36	64	8	30	80	95	61	33	65	5	39	88	36	44	42	43	5	88	81	13	63	15	47	92	20	62	5	60
6	54	71	27	89	41	53	60	10	2	91	76	95	98	91	64	65	23	57	16	0	90	52	26	90	49	31	68	29	58	10
7	10	60	18	77	34	59	28	99	15	11	70	34	27	78	67	19	97	30	23	60	0	22	11	12	54	50	93	25	69	54
8	42	20	24	36	78	58	82	81	49	91	35	53	30	92	57	19	97	40	58	13	39	42	25	3	97	64	100	55	24	7
9	73	55	87	48	49	97	60	92	27	78	2	55	29	76	99	21	45	72	56	24	16	33	50	84	12	65	4	30	48	56
10	21	56	41	23	58	57	49	49	70	33	6	79	95	3	70	38	26	26	5	89	49	0	68	57	53	91	66	81	53	83
11	9	60	37	99	6	41	69	97	18	44	100	18	46	3	90	57	22	82	15	38	73	97	74	9	35	82	66	34	84	14
12	63	26	41	8	21	38	15	63	38	100	68	89	24	39	19	29	93	97	40	91	70	41	95	83	33	25	33	94	44	39
13	98	72	9	45	69	50	7	86	5	80	0	8	28	96	45	0	0	13	95	24	92	51	11	11	37	91	21	87	89	89
14	87	89	65	22	98	55	86	9	66	43	64	55	80	30	15	99	26	25	71	87	22	39	97	26	50	12	86	22	65	70
15	5	91	68	44	67	2	71	96	15	73	78	3	12	87	53	9	11	12	21	32	57	72	16	35	27	51	91	43	58	61
16	75	93	62	49	95	82	30	81	24	4	11	36	71	96	49	47	65	48	28	8	91	58	40	55	32	7	86	84	95	59
17	76	15	55	38	29	0	8	20	71	42	81	51	44	76	93	42	87	89	38	51	88	65	83	80	66	91	9	68	30	63
18	26	76	93	84	8	40	96	69	84	82	89	5	16	43	34	37	64	39	14	77	95	100	52	99	86	81	65	85	21	9
19	8	35	6	83	76	8	87	81	13	33	14	86	38	23	33	22	58	47	60	36	97	89	20	59	52	9	76	75	52	82

(*continúa*)

[1] Las tablas estadísticas para la Z, t, χ^2 y F se generaron utilizando los algoritmos computacionales encontrados en las siguientes referencias:

Craig, R. J. (1984). Normal family distribution functions: FORTRAN and BASIC programs. *Journal of Quality Technology*, *16*, 232-236.

Kirch, A. (1973). *Introduction to statistics with FORTRAN*. Nueva York: Holt, Rinehart y Winston.

TABLA A *(continuación)*

20	59	73	37	6	26	44	0	24	89	24	78	80	20	8	19	31	32	53	40	32	32	23	57	74	49	17	97	49	71	0
21	87	94	75	45	72	15	39	100	46	99	59	12	22	95	76	18	27	73	88	41	31	99	37	31	24	89	35	14	14	73
22	5	74	8	91	37	5	13	55	13	7	19	24	76	4	25	93	78	9	50	85	98	71	37	53	67	75	9	56	95	71
23	49	82	39	40	51	15	71	53	68	86	50	93	31	22	64	77	46	17	78	25	2	17	69	68	56	44	100	55	80	26
24	2	25	92	97	41	39	98	100	99	67	44	0	99	93	31	69	26	72	56	25	71	42	28	22	96	76	19	63	97	5
25	59	41	49	100	13	0	15	33	82	61	28	59	83	8	17	76	24	58	91	25	3	2	76	87	10	18	23	69	93	27
26	40	13	20	51	81	15	12	45	16	57	47	54	92	60	70	55	98	12	90	27	95	66	23	91	78	86	27	98	16	30
27	80	25	91	36	83	59	19	9	47	61	84	89	98	18	11	56	99	3	26	67	21	24	80	60	44	42	48	77	84	63
28	48	33	7	70	61	95	51	32	89	87	72	6	40	88	52	44	19	96	95	62	12	100	82	5	17	62	65	100	63	9
29	89	5	7	93	48	60	69	97	61	21	87	68	20	4	61	63	75	8	76	92	37	35	40	70	25	86	34	54	53	95
30	97	64	36	36	99	98	23	18	66	28	58	48	34	18	64	71	48	90	63	57	15	14	24	26	65	29	38	85	99	17
31	59	73	71	62	66	34	17	41	32	65	50	73	82	7	20	85	1	65	74	85	23	19	45	61	48	98	84	51	63	70
32	88	75	43	66	66	38	56	31	25	36	26	91	36	100	88	42	74	27	36	40	33	92	18	9	54	51	40	24	82	6
33	34	16	43	38	50	28	34	14	41	2	6	97	56	73	75	17	56	31	100	84	32	25	33	52	26	78	83	44	0	81
34	14	61	81	2	69	73	3	89	79	64	67	80	75	5	66	77	97	30	88	82	52	87	25	63	11	67	93	99	61	39
35	15	39	5	99	29	36	25	40	46	28	34	63	75	18	21	23	13	85	15	43	88	70	92	44	23	73	62	47	60	45
36	68	49	1	55	11	6	63	23	50	33	80	34	82	20	66	48	27	16	86	78	74	89	9	23	66	62	83	28	34	87
37	1	72	18	84	84	86	61	41	22	61	45	36	37	16	20	28	98	36	72	39	67	100	71	8	19	29	0	24	95	26
38	58	73	55	11	9	96	81	84	21	34	50	92	65	91	69	33	23	4	77	93	3	37	95	14	84	27	67	46	61	88
39	91	63	65	63	70	90	57	20	9	13	28	77	72	0	12	30	48	6	28	89	94	6	58	72	73	16	86	19	95	49
40	39	45	31	74	91	85	29	45	98	15	11	50	26	16	36	76	1	40	76	1	88	15	60	27	55	0	83	96	36	53
41	94	12	62	59	14	42	32	75	41	41	0	58	5	78	89	48	35	1	78	70	20	98	38	93	67	35	35	40	38	44
42	3	33	41	22	45	37	65	3	96	27	62	77	16	97	81	78	26	48	94	59	77	82	54	1	63	24	64	31	31	14
43	58	2	83	10	100	50	98	57	32	65	31	87	84	45	0	90	42	78	9	17	21	92	92	47	5	29	6	27	62	72
44	29	73	79	48	66	72	32	1	100	3	2	61	35	0	88	100	45	42	16	18	48	67	36	37	57	12	97	12	95	8
45	55	9	63	66	31	5	8	72	4	85	5	44	4	98	2	79	40	44	96	75	91	59	66	15	41	19	100	33	23	64
46	52	13	44	91	39	85	22	33	4	29	52	6	82	77	25	0	46	100	41	35	46	93	11	9	56	82	97	53	18	86
47	31	52	65	63	88	78	21	35	28	22	91	84	4	30	14	0	97	92	63	87	46	73	55	82	18	76	67	43	76	22
48	44	38	76	99	38	67	60	95	67	68	17	18	46	76	83	5	8	20	87	87	2	42	65	27	16	22	60	18	78	33
49	84	47	44	4	67	22	89	78	44	84	66	15	56	0	90	21	25	88	99	100	32	86	30	50	92	48	55	70	35	20
50	71	50	78	48	65	24	21	24	2	23	65	94	51	82	67	16	35	91	100	35	61	31	75	8	81	58	67	50	28	17
51	42	47	97	81	10	99	40	15	63	77	89	10	32	92	86	32	9	33	79	69	50	7	61	78	15	60	79	47	73	51
52	3	70	75	49	90	92	62	0	47	90	78	63	44	60	13	55	38	64	60	63	92	17	100	2	40	93	83	89	88	20

(continúa)

TABLA A *(continuación)*

53	31	6	46	39	27	93	81	79	100	94	43	39	79	2	18	82	40	30	56	31	81	84	62	41	59	4	46	56	100	58
54	69	27	97	71	52	38	45	35	14	74	40	96	40	88	38	67	44	81	5	12	13	98	21	39	36	74	39	83	77	79
55	2	76	36	72	7	28	55	13	31	78	67	98	50	25	94	39	71	28	0	39	31	69	14	22	50	40	54	12	71	98
56	3	4	20	8	63	33	69	31	69	32	35	18	23	84	69	64	13	43	86	53	10	28	46	41	29	74	46	64	39	4
57	79	55	89	1	25	68	100	58	44	92	73	29	70	47	3	51	37	24	24	29	95	79	80	35	0	9	65	42	99	69
58	99	6	65	35	66	98	66	47	47	22	1	54	94	13	0	31	40	55	69	20	59	12	35	63	52	35	2	56	40	85
59	46	98	1	46	43	86	42	91	63	1	93	84	51	8	79	47	54	85	90	2	19	26	78	95	1	4	72	81	80	60
60	6	14	71	51	7	10	79	41	58	3	27	33	74	67	18	94	4	57	99	37	40	96	68	6	95	55	82	16	36	58
61	92	31	31	40	12	19	74	73	20	94	33	41	40	74	79	42	23	41	29	1	0	13	31	19	63	90	75	17	33	49
62	87	8	68	74	61	66	94	27	71	81	37	82	83	7	8	46	65	63	37	63	88	20	20	75	16	70	26	75	22	48
63	50	48	52	100	68	75	38	65	59	57	78	24	29	52	24	98	78	48	77	64	93	100	50	95	76	94	84	25	67	98
64	67	96	52	88	76	79	16	12	42	33	35	50	54	69	21	57	62	21	84	95	13	66	49	11	48	20	54	51	65	63
65	54	42	22	99	28	90	74	46	26	13	48	45	99	3	38	94	86	53	41	18	35	10	64	79	70	5	55	92	41	92
66	99	51	72	2	75	81	92	71	85	26	77	73	23	14	2	46	7	13	2	40	62	28	72	82	81	51	7	45	9	26
67	35	63	58	46	91	44	56	26	59	56	21	91	19	83	6	61	47	53	10	33	7	97	68	76	44	73	73	0	80	55
68	81	98	63	17	77	45	47	96	25	38	23	26	80	20	47	40	39	14	71	15	60	83	28	56	78	9	27	52	79	68
69	90	47	44	40	40	96	0	62	13	79	39	0	99	57	37	39	2	8	42	58	1	28	1	64	50	28	8	69	70	96
70	29	30	16	54	83	76	50	0	61	100	51	74	78	15	91	61	72	24	44	71	94	59	17	43	50	34	12	14	45	30
71	47	94	70	80	51	26	11	78	34	29	10	55	90	42	4	6	83	72	95	73	24	19	13	98	0	64	44	90	20	13
72	69	14	17	73	79	25	71	14	52	98	77	82	15	25	8	34	38	80	82	97	82	87	98	29	97	69	24	62	100	12
73	54	58	47	9	0	63	6	94	27	3	18	5	36	98	74	36	30	8	87	2	23	76	42	76	87	64	99	5	7	13
74	24	63	57	91	8	58	38	29	72	5	56	71	81	50	67	59	41	9	17	17	85	42	29	80	53	92	6	44	100	18
75	14	24	69	85	97	51	68	80	16	92	59	72	97	23	89	44	16	71	19	83	42	53	54	93	63	19	59	30	80	75
76	86	21	31	59	72	17	77	45	43	29	34	97	67	45	23	88	91	68	12	30	3	41	73	63	76	18	82	8	13	30
77	5	28	80	31	99	77	39	23	69	0	15	49	100	2	22	64	73	92	53	64	7	19	80	64	4	34	30	65	63	11
78	29	71	48	4	87	32	17	90	89	9	99	34	58	8	61	73	98	48	89	90	24	25	98	38	79	45	84	30	49	64
79	90	94	19	80	70	36	2	17	48	63	82	39	85	26	65	27	81	69	83	20	40	25	87	45	88	52	19	33	17	63
80	62	66	48	74	86	6	66	41	15	65	6	41	85	57	84	64	70	39	64	87	62	78	25	71	57	6	98	59	79	34
81	67	54	3	54	23	40	25	95	93	55	59	46	77	55	49	82	26	8	87	54	10	53	29	37	82	5	77	54	4	69
82	75	27	62	15	81	36	22	26	69	42	44	91	55	0	84	48	68	65	5	45	35	11	73	30	16	3	75	56	58	98
83	70	19	7	100	94	53	81	76	73	40	22	58	49	42	96	18	66	89	8	69	17	54	7	86	29	18	86	98	5	56
84	75	7	9	20	58	92	41	42	79	26	91	44	63	87	45	21	23	15	6	72	60	78	88	27	45	80	66	25	37	73
85	55	70	10	23	25	73	91	72	29	47	93	58	21	75	80	52	9	12	36	93	9	58	84	88	90	73	47	49	53	95

(continúa)

Tabla de números aleatorios *(continuación)*

14	94	86	38	11	60	57	16	41	46	20
15	6	62	50	24	11	19	73	14	42	48
16	53	70	54	25	96	38	43	5	2	4
17	28	75	64	90	11	80	94	99	35	54
18	68	57	34	30	29	61	33	49	0	11
19	45	65	89	88	39	93	71	55	29	67
20	73	11	78	58	58	34	20	30	43	40
21	26	59	10	35	75	4	34	38	0	63
22	58	15	70	36	19	49	45	18	36	2
23	87	85	52	76	40	61	50	68	72	7
24	98	44	82	35	0	33	26	68	75	7
25	35	39	8	70	79	48	30	65	65	63
26	79	82	7	23	41	81	8	32	8	8
27	0	30	98	86	100	14	55	86	71	13
28	88	88	48	70	64	81	29	71	62	67
29	45	62	32	83	60	48	0	44	94	22
30	63	8	87	100	28	82	67	65	10	81
31	33	6	49	38	55	78	94	26	4	29
32	79	51	52	9	38	18	13	16	86	42
33	63	29	23	97	64	6	63	74	29	77
34	94	16	38	87	3	25	25	49	22	68
35	32	6	90	100	29	26	31	39	32	93
36	92	99	60	23	79	82	6	62	2	75
37	46	1	2	68	40	8	3	99	19	6
38	65	55	20	58	89	100	74	77	28	30
39	37	58	49	5	51	55	90	22	3	37
40	80	47	63	53	58	95	55	25	67	58
41	2	48	66	86	47	74	48	87	71	21
42	49	71	92	36	55	72	74	13	99	31
43	35	48	56	92	76	75	45	23	91	15
44	77	61	32	6	66	47	66	0	24	26
45	50	83	57	78	38	55	48	97	5	62
46	83	94	8	40	14	39	93	51	42	80

(continúa)

Tabla de números aleatorios *(continuación)*

47	82	1	78	19	94	56	38	8	37	28
48	73	74	13	2	42	64	89	86	72	9
49	54	43	20	13	39	76	59	7	51	19
50	77	32	56	82	56	60	98	80	21	49
51	99	27	39	7	32	7	85	14	22	76
52	1	14	43	75	65	65	63	53	81	57
53	26	51	32	8	24	99	30	36	32	59
54	37	89	4	20	21	91	98	90	37	49
55	25	26	20	61	52	93	90	76	46	19
56	47	55	98	22	69	9	15	34	94	16
57	90	22	16	34	81	44	3	24	96	70
58	2	85	2	58	26	94	48	0	85	70
59	49	67	32	10	28	90	72	25	28	53
60	68	68	69	7	11	31	17	39	82	85
61	13	54	32	26	66	38	1	7	35	16
62	6	1	89	99	21	48	6	9	67	85
63	94	23	75	40	33	86	87	76	24	98
64	33	98	80	13	84	70	85	93	74	22
65	14	63	52	94	56	5	40	55	50	17
66	47	34	47	47	95	45	38	82	85	20
67	84	77	74	27	5	17	57	75	63	2
68	90	48	12	51	55	77	48	10	55	21
69	26	100	6	31	89	0	31	91	5	23
70	79	63	76	72	18	67	87	47	90	93
71	66	81	97	81	11	38	7	37	93	64
72	28	84	86	10	69	25	66	93	21	57
73	33	19	18	37	96	73	95	91	24	24
74	24	31	5	6	37	63	93	42	5	97
75	8	91	48	79	2	40	6	56	57	60
76	78	45	43	77	77	99	98	40	14	82
77	72	20	15	22	30	82	77	51	87	61
78	98	48	25	14	0	12	63	67	12	77
79	60	62	46	12	59	99	5	88	74	89

(continúa)

Tabla de números aleatorios *(continuación)*

80	20	77	87	83	12	74	29	12	16	99
81	7	40	18	32	85	37	73	42	49	49
82	46	93	58	96	29	73	6	71	8	46
83	78	0	78	24	34	73	95	11	44	36
84	7	67	29	27	12	90	60	97	15	94
85	62	28	11	61	0	91	49	32	82	28
86	14	46	52	52	36	21	13	70	24	76
87	26	94	34	57	81	28	49	74	68	50
88	15	11	82	35	77	9	28	11	32	30
89	24	71	92	75	70	60	80	88	21	11
90	18	25	7	100	80	84	97	84	18	53
91	34	91	25	98	77	14	95	100	84	19
92	98	80	72	72	71	66	13	33	24	12
93	22	83	2	33	32	91	78	53	45	63
94	41	39	35	37	66	52	80	1	33	94
95	30	54	73	21	43	68	65	83	26	90
96	65	100	85	12	69	3	72	55	43	5
97	57	69	37	7	62	65	36	9	57	73
98	44	51	38	59	85	91	51	79	14	26
99	39	76	88	46	46	65	72	62	92	67
100	84	60	42	55	48	99	44	66	77	27

	Media	Varianza	Desviación estándar (D. E.)
1	51.8400	895.8144	29.9302
2	46.2000	809.3200	28.4486
3	47.6900	740.6539	27.2150
4	51.8300	872.7611	29.5425
5	53.2100	877.5659	29.6237
6	48.8700	903.9131	30.0651
7	49.6400	778.8704	27.9082
8	51.3700	889.7331	29.8284
9	45.0700	771.7251	27.7799
10	49.2800	872.3016	29.5348
11	48.8700	777.5731	27.8850
12	53.0800	860.2136	29.3294
13	56.5100	773.0099	27.8031
14	47.9900	1110.2299	33.3201

(continúa)

	Media	Varianza	Desviación estándar (D. E.)
15	49.3700	913.6531	30.2267
16	49.0200	714.0396	26.7215
17	45.6800	842.0776	29.0186
18	47.0400	853.3384	29.2120
19	53.5100	977.2499	31.2610
20	52.7400	853.4924	29.2146
21	50.0600	1001.1564	31.6411
22	53.9500	907.7475	30.1288
23	53.6100	737.3779	27.1547
24	49.3100	807.4139	28.4150
25	49.1600	673.9544	25.9606
26	50.2200	855.3316	29.2461
27	58.3600	877.1904	29.6174
28	49.5700	709.7051	26.6403
29	55.4400	868.3664	29.4681
30	49.4300	791.3851	28.1316
31	48.5200	847.9296	29.1192
32	52.9400	802.3564	28.3259
33	46.7900	784.6259	28.0112
34	48.3300	881.4611	29.6894
35	47.2900	759.3059	27.5555
36	35.5100	854.5499	29.2327
37	52.3900	907.8379	30.1303
38	49.9500	851.3275	29.1775
39	46.0000	817.7800	28.5969
40	47.6500	815.1475	28.5508

TABLA B *Tabla de la curva normal (área entre Z = 0 y Z)*

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0001	0.0040	0.0080	0.0120	0.0160	0.0200	0.0240	0.0280	0.0319	0.0359
0.1	0.0399	0.0438	0.0478	0.0518	0.0557	0.0597	0.0636	0.0675	0.0715	0.0754
0.2	0.0793	0.0832	0.0871	0.0910	0.0949	0.0988	0.1026	0.1065	0.1103	0.1141
0.3	0.1180	0.1218	0.1256	0.1293	0.1331	0.1369	0.1406	0.1444	0.1481	0.1518
0.4	0.1555	0.1591	0.1628	0.1664	0.1701	0.1737	0.1773	0.1809	0.1844	0.1880
0.5	0.1915	0.1950	0.1985	0.2020	0.2054	0.2089	0.2123	0.2157	0.2191	0.2224
0.6	0.2258	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2518	0.2549
0.7	0.7581	0.2612	0.2643	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2882	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3314	0.3339	0.3364	0.3389
1.0	0.3413	0.3437	0.3461	0.3484	0.3508	0.3531	0.3554	0.3576	0.3599	0.3621
1.1	0.3643	0.3664	0.3686	0.3707	0.3728	0.3748	0.3769	0.3789	0.3809	0.3829
1.2	0.3850	0.3869	0.3888	0.3907	0.3926	0.3944	0.3962	0.3980	0.3998	0.4015
1.3	0.4032	0.4049	0.4066	0.4083	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4278	0.4292	0.4305	0.4319

(*continúa*)

TABLA B *(continuación)*

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.5	0.4332	0.4344	0.4357	0.4369	0.4382	0.4394	0.4406	0.4417	0.4429	0.4440
1.6	0.4451	0.4462	0.4473	0.4484	0.4494	0.4504	0.4515	0.4524	0.4534	0.4544
1.7	0.4553	0.4563	0.4572	0.4581	0.4590	0.4598	0.4607	0.4615	0.4623	0.4632
1.8	0.4639	0.4647	0.4655	0.4662	0.4670	0.4677	0.4684	0.4691	0.4698	0.4705
1.9	0.4711	0.4718	0.4724	0.4731	0.4737	0.4743	0.4749	0.4754	0.4760	0.4766
2.0	0.4771	0.4776	0.4782	0.4787	0.4792	0.4797	0.4802	0.4806	0.4811	0.4816
2.1	0.4820	0.4824	0.4829	0.4833	0.4837	0.4841	0.4845	0.4849	0.4852	0.4856
2.2	0.4860	0.4863	0.4867	0.4870	0.4873	0.4877	0.4880	0.4883	0.4886	0.4889
2.3	0.4892	0.4895	0.4897	0.4900	0.4903	0.4905	0.4908	0.4910	0.4913	0.4915
2.4	0.4917	0.4919	0.4922	0.4924	0.4926	0.4928	0.4930	0.4932	0.4934	0.4936
2.5	0.4937	0.4939	0.4941	0.4943	0.4944	0.4946	0.4947	0.4949	0.4950	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4958	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4968	0.4969	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4973	0.4974	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979
2.9	0.4980	0.4981	0.4981	0.4982	0.4982	0.4983	0.4983	0.4984	0.4984	0.4985
3.0	0.4985	0.4986	0.4986	0.4987	0.4987	0.4988	0.4988	0.4988	0.4989	0.4989
3.1	0.4990	0.4990	0.4990	0.4991	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992
3.2	0.4993	0.4993	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995
3.3	0.4995	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996
3.4	0.4996	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998

TABLA C *Distribución* t

	Probabilidad					
Grados de libertad	**0.50 0.25**	**0.10 0.05**	**0.05 0.025**	**0.02 0.015**	**0.01 0.00**	de una cola de dos colas
1	1.000	6.34	12.71	31.82	63.66	
2	0.816	2.92	4.30	6.96	9.92	
3	.765	2.35	3.18	4.54	5.84	
4	.741	2.13	2.78	3.75	4.60	
5	.727	2.02	2.57	3.36	4.03	
6	.718	1.94	2.45	3.14	3.71	
7	.711	1.90	2.36	3.00	3.50	
8	.706	1.86	2.31	2.90	3.36	
9	.703	1.83	2.26	2.82	3.25	
10	.700	1.81	2.23	2.76	3.17	
11	.697	1.80	2.20	2.72	3.11	
12	.695	1.78	2.18	2.68	3.06	
13	.694	1.77	2.16	2.65	3.01	
14	.692	1.76	2.14	2.62	2.98	
15	.691	1.75	2.13	2.60	2.95	

(continúa)

TABLA C *(continuación)*

Grados de libertad	0.50 0.25	0.10 0.05	0.05 0.025	0.02 0.015	0.01 0.00	de una cola de dos colas
			Probabilidad			
16	.690	1.75	2.12	2.58	2.92	
17	.689	1.74	2.11	2.57	2.90	
18	.688	1.73	2.10	2.55	2.88	
19	.688	1.73	2.09	2.54	2.86	
20	.687	1.72	2.09	2.53	2.84	
21	.686	1.72	2.08	2.52	2.83	
22	.686	1.72	2.07	2.51	2.82	
23	.685	1.71	2.07	2.50	2.81	
24	.685	1.71	2.06	2.49	2.80	
25	.684	1.71	2.06	2.48	2.79	
30	.683	1.70	2.04	2.46	2.75	
35	.682	1.69	2.03	2.46	2.72	
40	.681	1.68	2.02	2.42	2.71	
60	.678	1.67	2.00	2.39	2.69	
120	.676	1.66	1.98	2.36	2.62	
inf	.674	1.645	1.96	2.33	2.575	

TABLA D *Distribución de la chi cuadrada, cola superior*

gl/α	0.100	0.050	0.025	0.010	0.005	0.001
1	2.71	3.84	5.02	6.63	7.88	10.8
2	4.61	5.99	7.38	9.21	10.6	13.8
3	6.25	7.81	9.35	11.3	12.8	16.3
4	7.78	9.49	11.1	13.3	14.9	18.5
5	9.24	11.1	12.8	15.1	16.7	20.5
6	10.6	12.6	14.4	16.8	18.5	22.5
7	12.0	14.1	16.0	18.5	20.3	24.3
8	13.4	15.5	17.5	20.1	22.0	26.1
9	14.7	16.9	19.0	21.7	23.6	27.9
10	16.0	18.3	20.5	23.2	25.2	29.6
11	17.3	19.7	21.9	24.7	26.8	31.3
12	18.5	21.0	23.3	26.2	28.3	32.9
13	19.8	22.4	24.7	27.7	29.8	34.5
14	21.1	23.7	26.1	29.1	31.3	36.1
15	22.3	25.0	27.5	30.6	32.8	37.7
16	23.5	26.3	28.8	32.0	34.3	39.3
17	24.8	27.6	30.2	33.4	35.7	40.8
18	26.0	28.9	31.5	34.8	37.2	42.3
19	27.2	30.1	32.9	36.2	38.6	43.8
20	28.4	31.4	34.2	37.6	40.0	45.3
21	29.6	32.7	35.5	38.9	41.4	46.8
22	30.8	33.9	36.8	40.3	42.8	48.3

(continúa)

TABLA D *(continuación)*

gl/α	0.100	0.050	0.025	0.010	0.005	0.001
23	32.0	35.2	38.1	41.6	44.2	49.7
24	55.2	36.4	39.4	43.0	45.6	51.2
25	34.4	37.7	40.6	44.3	46.9	52.6
30	40.3	43.8	47.0	50.9	53.7	59.7
35	46.1	49.8	53.2	57.3	60.3	66.6
40	51.8	55.8	59.3	63.7	66.8	73.4
60	74.4	74.4	83.3	88.4	92.0	99.6
80	96.6	101.9	106.6	112.3	116.3	124.8
100	118.5	124.3	129.6	135.8	140.2	149.4

TABLA E *Valores críticos de F (nivel .05 sin negritas, nivel .01 en negritas)**

Grados de libertad (denominador)	*Grados de libertad (numerador)*									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	161.00	200.00	216.00	225.00	230.00	234.00	237.00	239.00	241.00	242.00
	4 052.0	**4 999.0**	**5 403.0**	**5 625.0**	**5 764.0**	**5 859.0**	**5 928.0**	**5 981.0**	**6 022.0**	**6 056.0**
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38	19.39
	98.49	**99.00**	**99.17**	**99.25**	**99.30**	**99.33**	**99.36**	**99.37**	**99.39**	**99.40**
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78
	34.12	**30.82**	**29.46**	**28.71**	**28.24**	**27.91**	**27.67**	**27.49**	**27.34**	**27.23**
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
	21.20	**18.00**	**16.69**	**15.98**	**15.52**	**15.21**	**14.91**	**14.80**	**14.66**	**14.54**
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74
	16.26	**13.27**	**12.06**	**11.39**	**10.97**	**10.67**	**10.45**	**10.29**	**10.15**	**10.05**
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
	33.74	**10.92**	**9.78**	**9.15**	**8.75**	**8.47**	**8.26**	**8.10**	**7.98**	**7.87**
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63
	12.25	**9.55**	**8.45**	**7.85**	**7.46**	**7.19**	**7.00**	**6.84**	**6.71**	**6.62**
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34
	11.26	**8.65**	**7.59**	**7.01**	**6.63**	**6.37**	**6.19**	**6.03**	**5.91**	**5.82**
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13
	10.56	**8.02**	**6.99**	**6.42**	**6.06**	**5.80**	**5.62**	**5.47**	**5.35**	**5.26**
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97
	10.04	**7.56**	**6.55**	**5.99**	**5.64**	**5.39**	**5.21**	**5.06**	**4.95**	**4.85**
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.86
	9.65	**7.20**	**6.22**	**5.67**	**5.32**	**5.07**	**4.88**	**4.74**	**4.63**	**4.54**
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80	2.76
	9.33	**6.93**	**5.95**	**5.41**	**5.06**	**4.82**	**4.65**	**4.50**	**4.39**	**4.30**
13	4.67	3.80	3.41	3.18	3.02	2.92	2.84	2.77	2.72	2.67
	9.07	**6.70**	**5.74**	**5.20**	**4.86**	**4.62**	**4.44**	**4.30**	**4.19**	**4.10**
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60
	8.86	**6.51**	**5.56**	**5.03**	**4.69**	**4.46**	**4.28**	**4.14**	**4.03**	**3.94**
15	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59	2.55
	8.68	**6.36**	**5.42**	**4.89**	**4.56**	**4.32**	**4.14**	**4.00**	**3.89**	**3.80**

(continúa)

TABLA E *(continuación)*

	Grados de libertad (numerador)									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
16	4.49	3.63	3.24	3.01	2.85	**2.74**	**2.66**	2.59	2.54	2.49
	8.53	**6.23**	**5.29**	**4.77**	**4.44**	**4.20**	**4.03**	**3.89**	**3.78**	**3.69**
17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45
	8.40	**6.11**	**5.18**	**4.67**	**4.34**	**4.10**	**3.93**	**3.79**	**3.68**	**3.59**
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
	8.28	**6.01**	**5.09**	**4.58**	**4.25**	**4.01**	**3.85**	**3.71**	**3.60**	**3.51**
19	4.38	3.52	3.13	2.90	2.74	2.63	2.55	**2.48**	2.43	2.38
	8.18	**5.93**	**5.01**	**4.50**	**4.17**	**3.94**	**3.77**	**3.63**	**3.52**	**3.43**
20	4.35	3.49	3.10	2.87	2.72	2.60	2.52	2.45	2.40	2.35
	8.10	**5.85**	**4.94**	**4.43**	**4.10**	**3.87**	**3.71**	**3.56**	**3.45**	**3.37**
22	4.30	3.44	3.05	2.82	2.66	2.55	2.47	2.40	2.35	2.30
	7.94	**5.72**	**4.82**	**4.31**	**3.99**	**3.76**	**3.59**	**3.45**	**3.35**	**3.26**
23	**4.28**	3.42	3.03	2.80	2.64	2.53	2.45	2.38	232.00	2.28
	7.88	**5.66**	**4.76**	**4.26**	**3.94**	**3.71**	**3.54**	**3.41**	**3.30**	**3.21**
25	4.24	3.38	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24
	7.77	**5.57**	**4.61**	**4.18**	**3.86**	**3.63**	**3.46**	**3.32**	**3.21**	**3.13**
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.77	2.22
	7.72	**5.53**	**4.64**	**4.14**	**3.82**	**3.59**	**3.42**	**3.29**	**3.27**	**3.09**
28	4.20	3.34	2.95	2.72	2.56	2.44	2.36	2.29	2.94	2.29
	7.64	**5.45**	**4.57**	**4.07**	**3.76**	**3.53**	**3.36**	**3.23**	**3.12**	**3.03**
29	4.18	3.33	2.93	2.70	2.54	2.43	2.35	2.28	2.22	2.18
	7.60	**5.42**	**4.54**	**4.04**	**3.73**	**3.50**	**3.33**	**3.20**	**3.08**	**3.00**
30	4.27	3.32	2.92	2.69	2.53	2.42	2.34	2.27	2.21	2.16
	7.56	**5.39**	**4.51**	**4.02**	**3.70**	**3.47**	**3.30**	**3.17**	**3.06**	**2.98**
34	4.13	3.28	2.88	2.65	2.49	2.38	2.30	2.23	2.27	2.12
	7.44	**5.29**	**4.42**	**3.93**	**3.61**	**3.38**	**3.21**	**3.08**	**2.97**	**2.89**
38	4.20	3.25	2.85	2.62	2.46	2.35	2.26	2.29	2.14	2.09
	7.35	**5.21**	**4.34**	**3.86**	**3.54**	**3.32**	**3.15**	**3.02**	**2.91**	**2.82**
40	4.08	3.23	2.84	2.62	2.45	2.34	2.25	2.28	2.22	2.07
	7.31	**5.18**	**4.31**	**3.83**	**3.51**	**3.29**	**3.12**	**2.99**	**2.88**	**2.80**
46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.14	2.09	2.04
	7.21	**5.10**	**4.24**	**3.76**	**3.44**	**3.22**	**3.05**	**2.92**	**2.82**	**2.73**
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.02
	7.17	**5.06**	**4.20**	**3.72**	**3.41**	**3.11**	**3.02**	**2.83**	**2.78**	**2.70**
60	**4.00**	**3.15**	**2.76**	2.52	2.37	2.25	2.57	2.10	2.04	1.99
	7.08	**4.98**	**4.13**	**3.65**	**3.34**	**3.12**	**2.95**	**2.82**	**2.72**	**2.63**
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.01	1.97
	7.01	**4.92**	**4.08**	**3.60**	**3.29**	**3.07**	**2.91**	**2.77**	**2.67**	**2.59**
80	3.96	3.11	2.72	2.48	2.33	2.21	2.12	2.05	1.99	1.95
	6.96	**4.88**	**4.04**	**3.56**	**3.25**	**3.04**	**2.87**	**2.74**	**2.64**	**2.55**
100	3.94	3.09	2.70	2.46	2.30	2.19	2.10	2.03	5.97	5.92
	6.90	**4.82**	**3.98**	**3.51**	**3.20**	**2.99**	**2.82**	**2.69**	**2.59**	**2.51**