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.

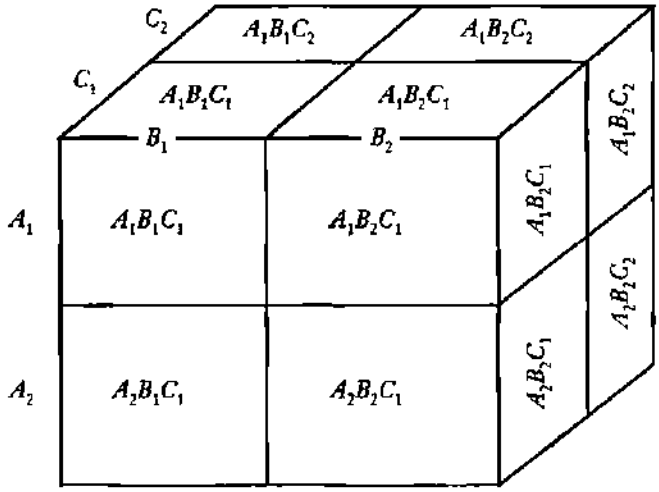
Parecen no existir reglas aceptadas de forma general respecto a cómo construir tabulaciones cruzadas. Se sabe, sin embargo, que son particiones cruzadas y que deben seguir las reglas de la partición o categorización discutidas anteriormente. Esas reglas eran: 1) las categorías se establecen de acuerdo 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_1B_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.

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(1 - 25.1616)^2}{25.1616} + \frac{(52 - 27.8384)^2}{27.8384} + \frac{(46 - 21.8384)^2}{21.8384} + \frac{(0 - 24.1616)^2}{24.1616}$$

$$= 23.2013 + 20.9704 + 26.7319 + 24.1616 = 95.0653$$

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:

$$V = \sqrt{\frac{95.0653}{99(1)}} = \sqrt{.9602} = .9799 = .98$$

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	Afroamericanos (%)	Raza	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	

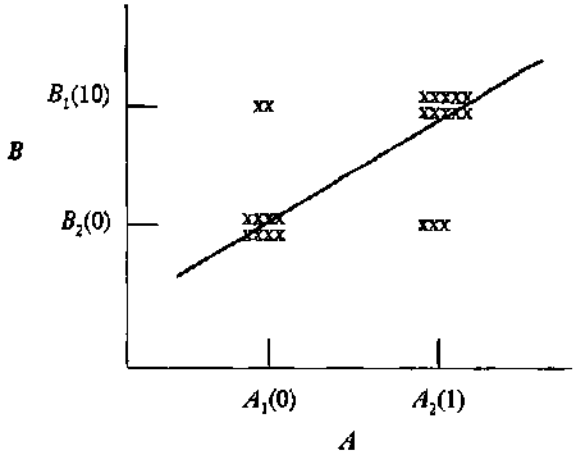
B_1 se les asignan ceros y a A_2 y B_2 se les asignan unos. Si un país tiene bajo CE y alto DP, entonces es un $A_2 B_1$; consecuentemente el par ordenado asignado para éste es (1, 0). Los primeros 10 países de la tabla 10.16 pertenecen a la categoría $A_2 B_1$, por lo tanto se les asignan (1, 0). De 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

$$\text{Probabilidad } (\textit{éxito} \mid \textit{bajo}) = \frac{.3}{.7} = .43$$

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

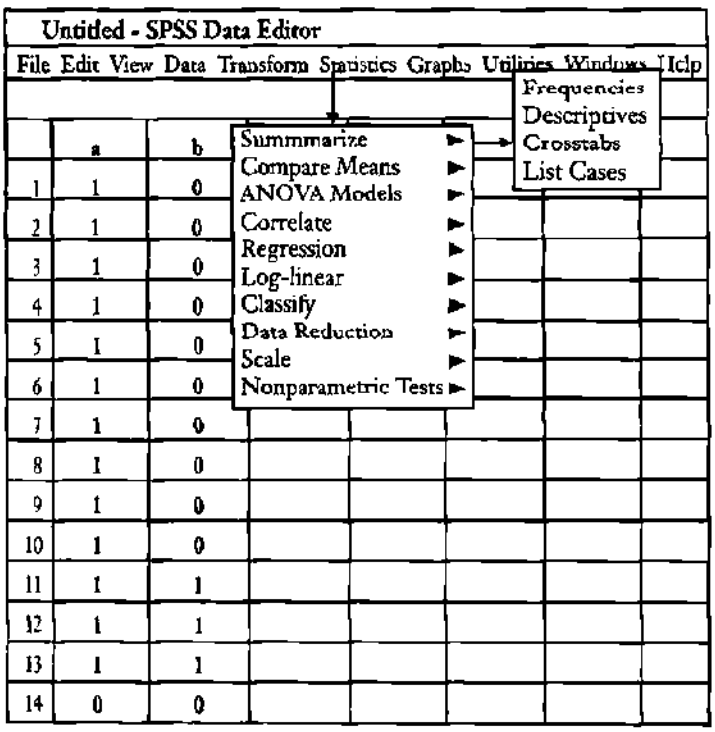
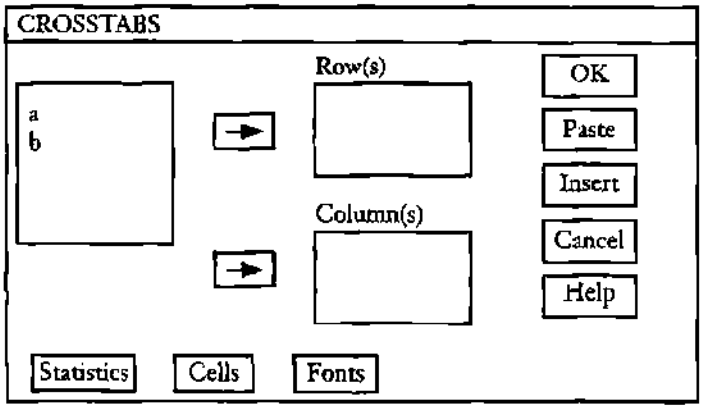


FIGURA 10.6



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

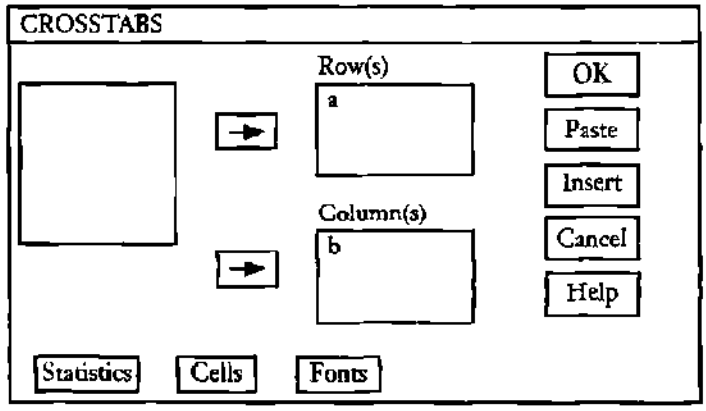
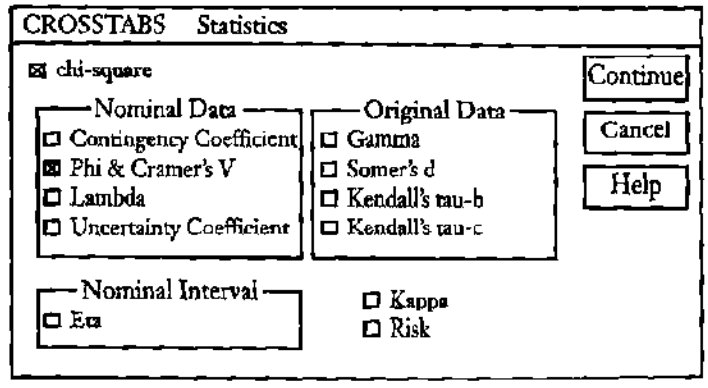


FIGURA 10.8



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

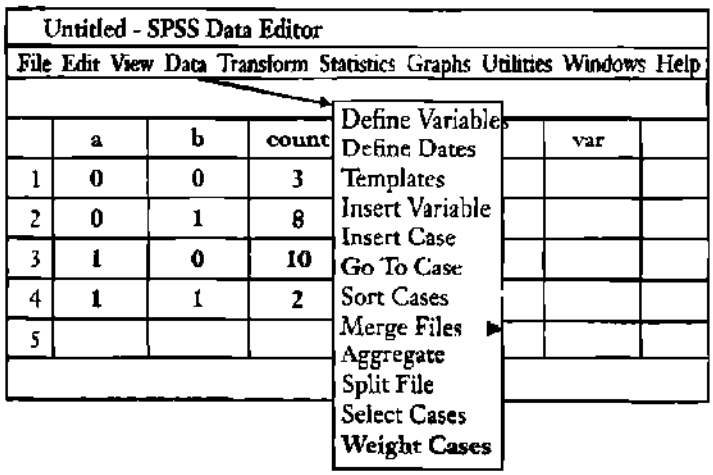
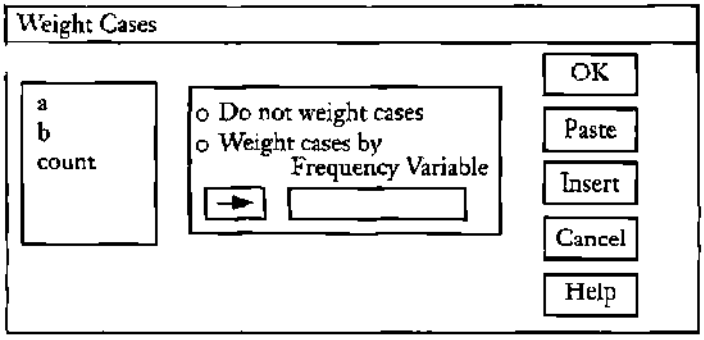


FIGURA 10.12



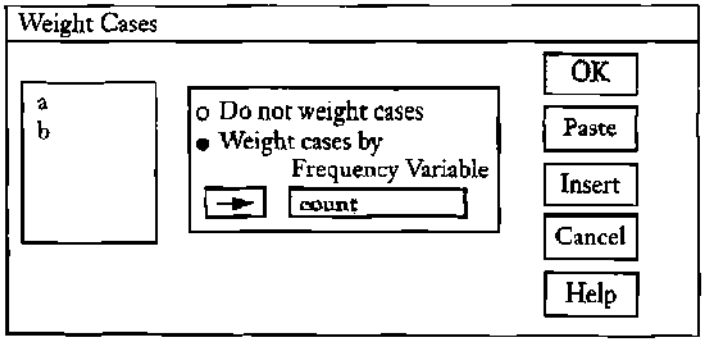
caja de la extrema izquierda están las tres variables: "a", "b" y "count". Primero, hacer clic en el botón etiquetado "Weight cases by" (figura 10.13). Entonces, con el ratón se resalta la variable "count" en la caja de la extrema izquierda. Haciendo clic en el botón de la flecha derecha del panel central, 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.

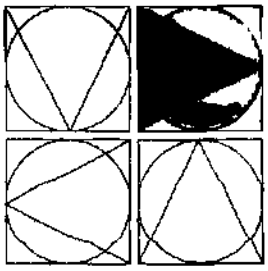
6. Los siguientes datos fueron recolectados por Glick, DeMorest y Hotze (1988) en su estudio acerca de la pertenencia al grupo, el espacio personal y la solicitud de un pequeño favor. Este estudio fue descrito brevemente en un capítulo anterior. Los investigadores querían determinar si la similitud de 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	$w(X) = p$	$(X - M)^2$	$(1/2)^2$
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

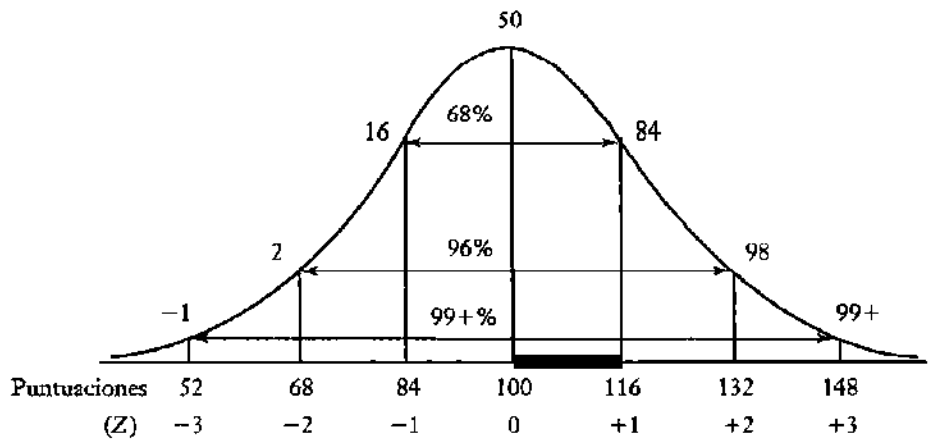
entre el éxito y las calificaciones del GRE no tienen valores muy altos (Morrison y Morrison, 1995). Se considera que las puntuaciones del GRE se distribuyen normalmente; sin embargo, esto no sucede con las personas que han sido admitidas en una escuela de posgrado *de alto nivel, donde las calificaciones 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:

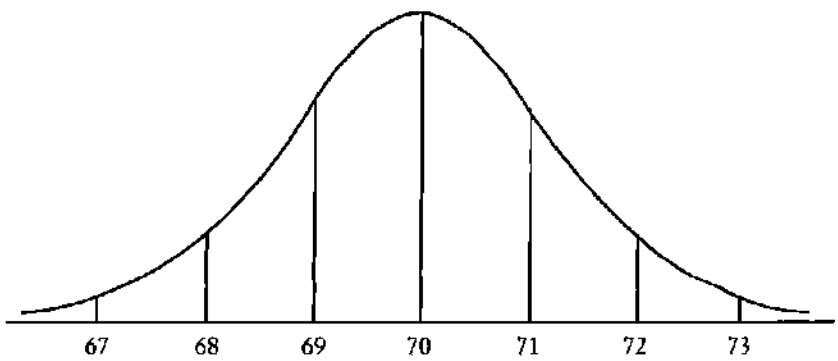
$$EE_M = \frac{DE}{\sqrt{n}} \tag{11.10}$$

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.

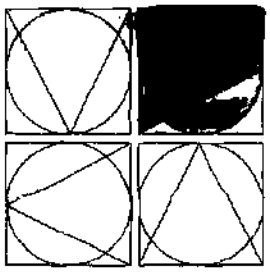
RESUMEN DEL CAPÍTULO

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

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 aleatorios y muestras de números aleatorios. 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

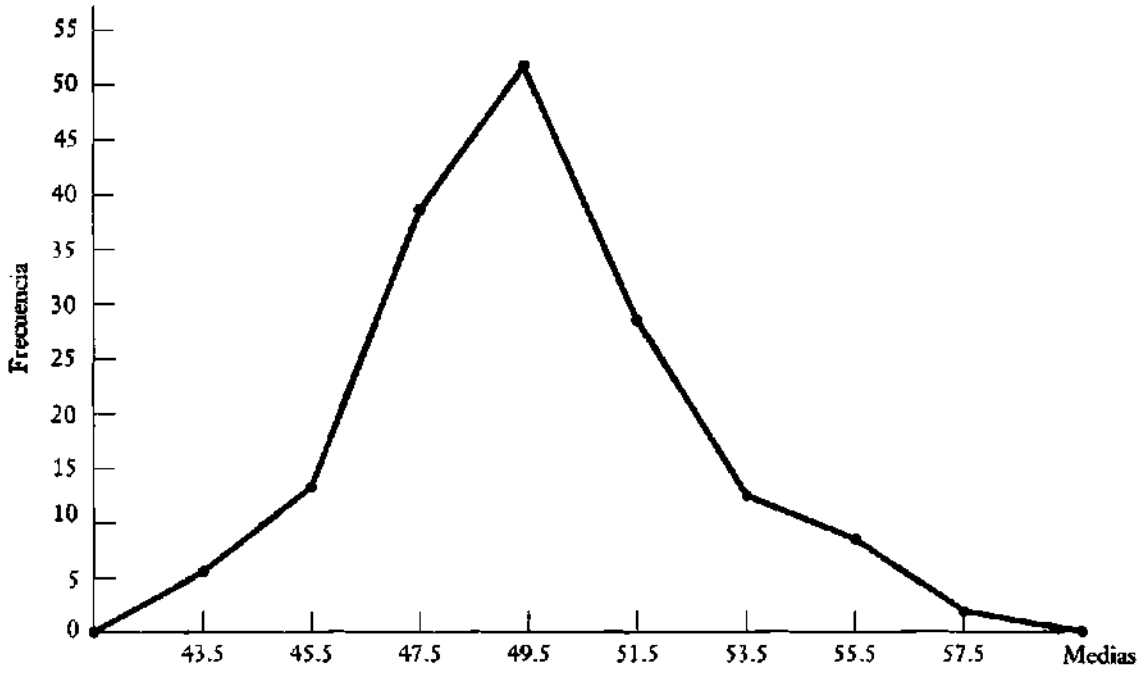
Antes de estudiar el uso real del error estándar de la media, se requiere conocer un poco acerca de una generalización extremadamente importante sobre las medias: *si las muestras son extraídas aleatoriamente de una población, las medias de las muestras tenderán a distribuirse normalmente*. Mientras 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:

$$EE_{M_A} = \frac{DE_A}{\sqrt{n_A}} = \frac{8}{\sqrt{100}} = .8, \qquad EE_{M_B} = \frac{DE_B}{\sqrt{n_B}} = \frac{9}{100} = .9$$

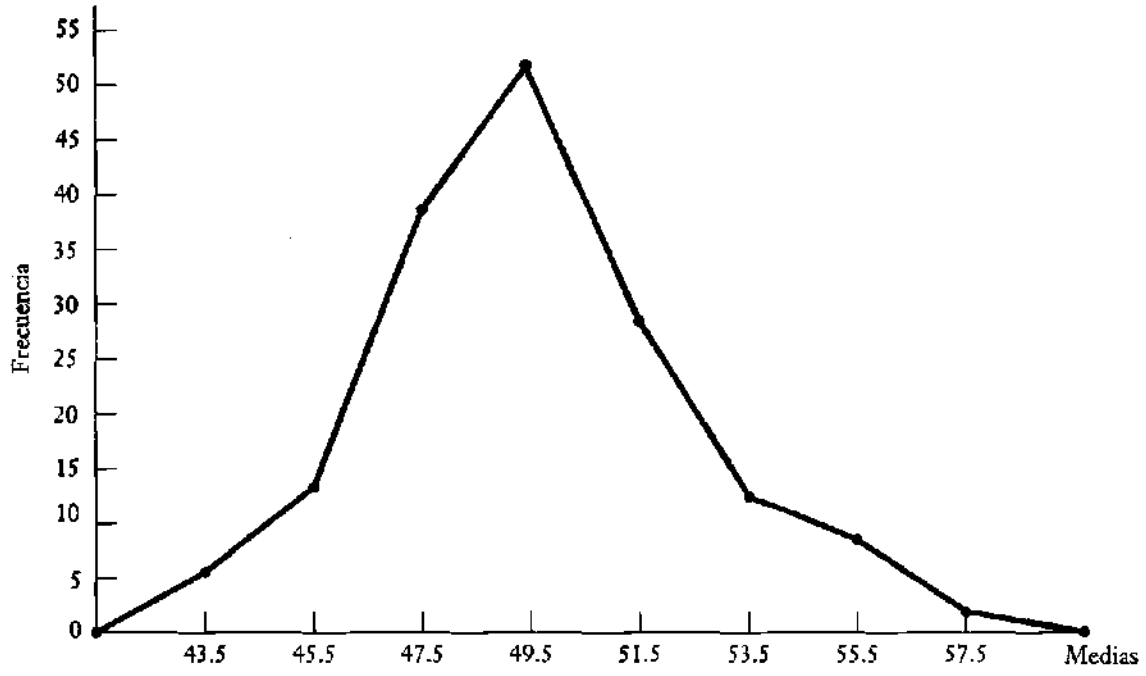
Mediante la ecuación 12.2 se calcula el error estándar de las diferencias entre las medias:

$$EE_{M_A - M_B} = \sqrt{EE^2_{M_A} + EE^2_{M_B}} = \sqrt{(.8)^2 + (.9)^2} = \sqrt{.64 + .81} = \sqrt{1.45} = 1.20$$

¿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*:

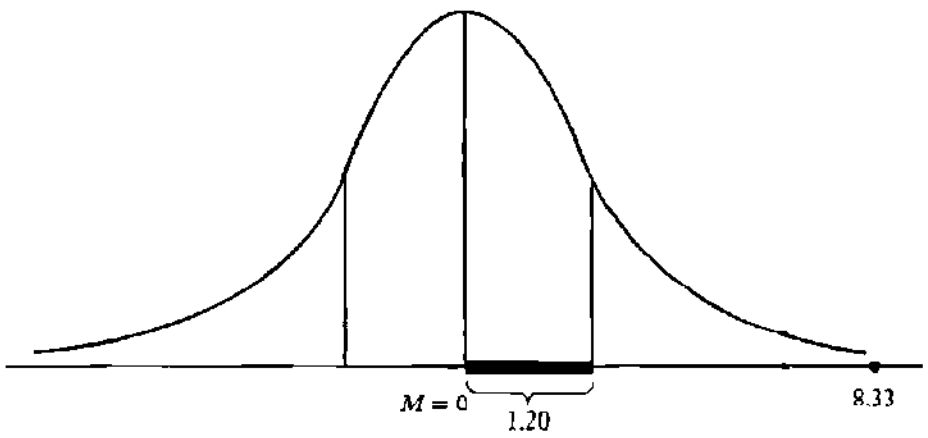
$$t = \frac{M_A - M_B}{EE_{M_A - M_B}} = \frac{110 - 100}{1.20} = \frac{10}{1.20} = 8.33$$

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. Ésta 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

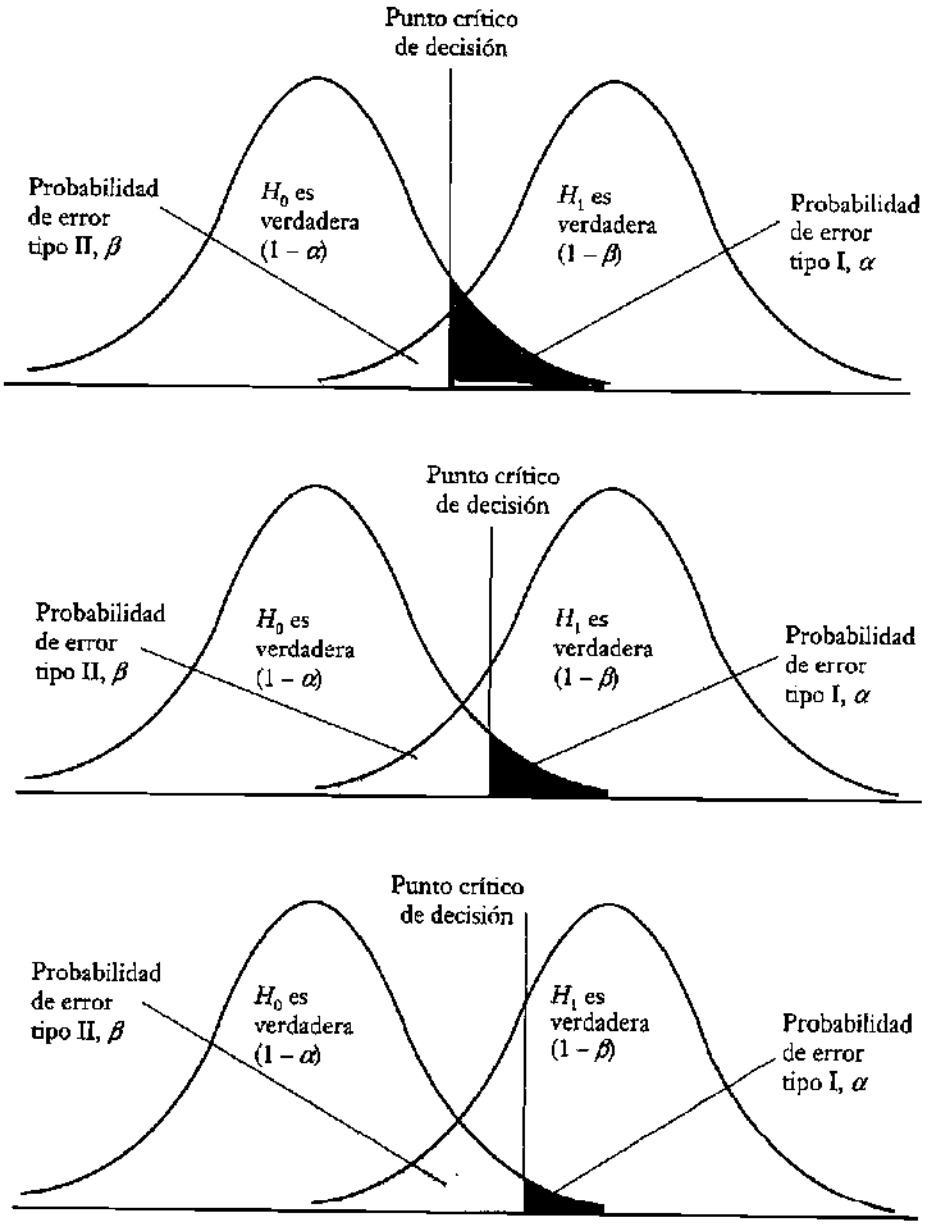

Punto crítico de decisión
Probabilidad de error tipo II, β
H_0 es verdadera (1 − α)
H_1 es verdadera (1 − β)
Probabilidad de error tipo I, α
Punto crítico de decisión
Probabilidad de error tipo II, β
H_0 es verdadera (1 − α)
H_1 es verdadera (1 − β)
Probabilidad de error tipo I, α
Punto crítico de decisión
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:

$$n = \frac{1.96^2(0.6^2)}{.2^2} = \frac{3.842(.36)}{.04} = \frac{1.383}{.04} = 34.6 \approx 35$$

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:

$$n' = \frac{70}{1 + 70/1\,000} = 65.421 = 66, \text{ o } 33 \text{ en cada grupo}$$

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).

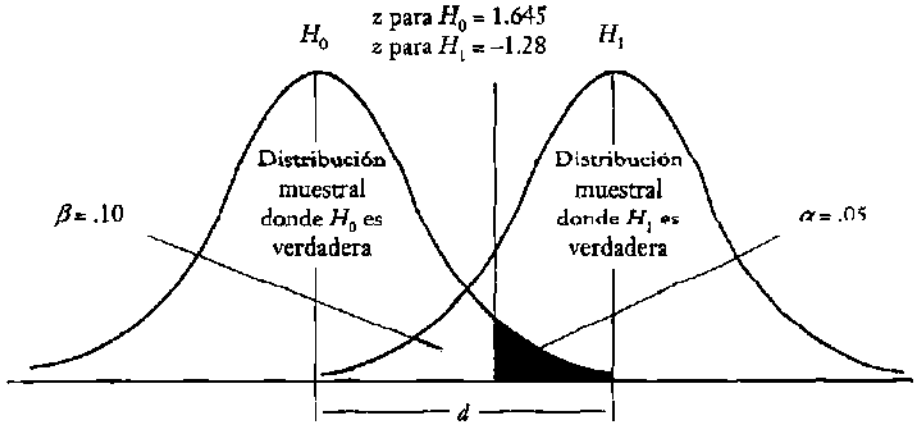
Para demostrar cómo funciona esta fórmula, es necesario referirse a la figura 12.3, que muestra la relación entre α y β. El tamaño de la muestra puede determinarse especificando tanto α como β, junto con la desviación estándar. Suponiendo que la desviación estándar se midió con precisión, el número de 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:

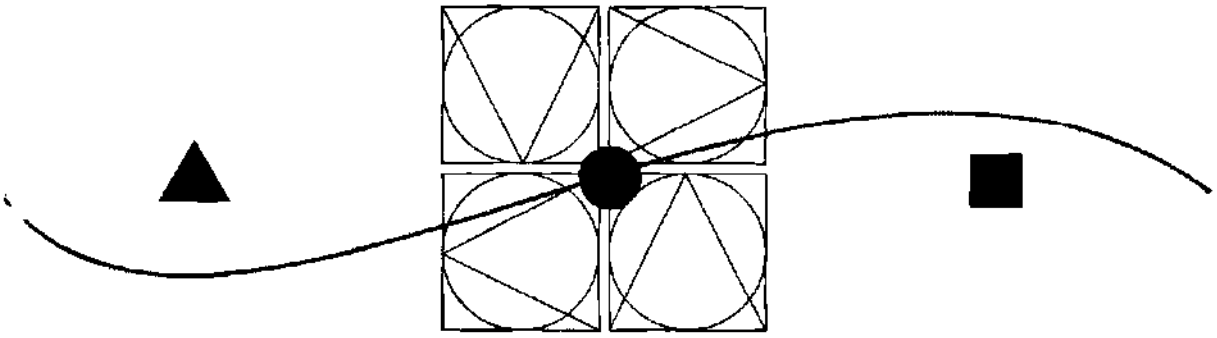
	Significado de palabras		*Creatividad verbal*	
	Tradicional	*Abierto*	*Tradicional*	*Abierto*
N	50	50	50	50
M	4.84	4.35	135.38	129.60
DE	1.19	.78	23.5	19.2

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

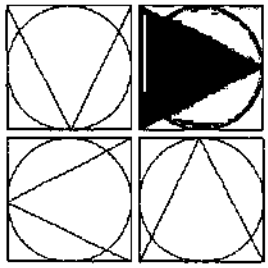


Capítulo 13
Análisis de varianza: fundamentos

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 lineal (*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_e 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

$$V_e = \frac{\Sigma x_k^2}{k-1} = \frac{.50}{2-1} = .50$$

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 esras 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

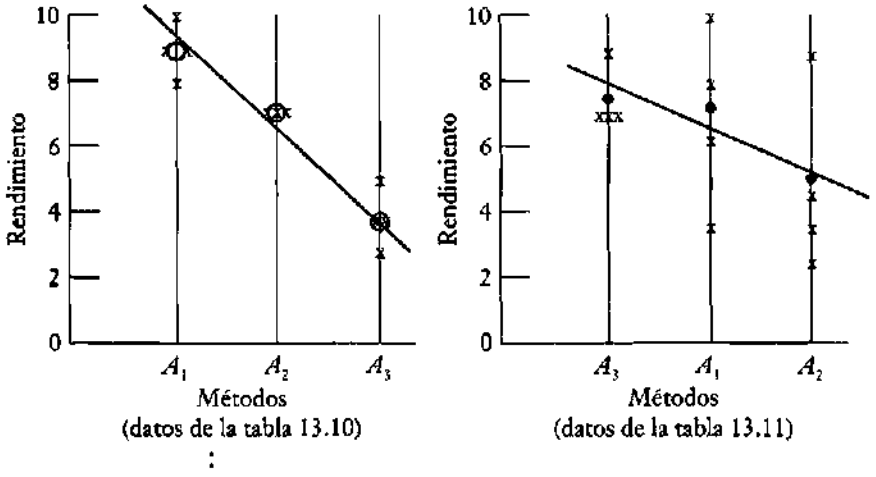
las medias, cuanto más lejos están una de la otra y mayor será la relación, siempre y cuando los demás aspectos se mantengan constantes.

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:

$$\omega^2 = \frac{SC_t - (k-1)CM_d}{SC_t + CM_d} \tag{13.5}$$

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:

$$H_1\text{: } C_1 = \frac{M_1 + M_2}{2} > \frac{M_3 + M_4}{2}$$
$$H_2\text{: } C_2 = M_1 > M_2$$

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

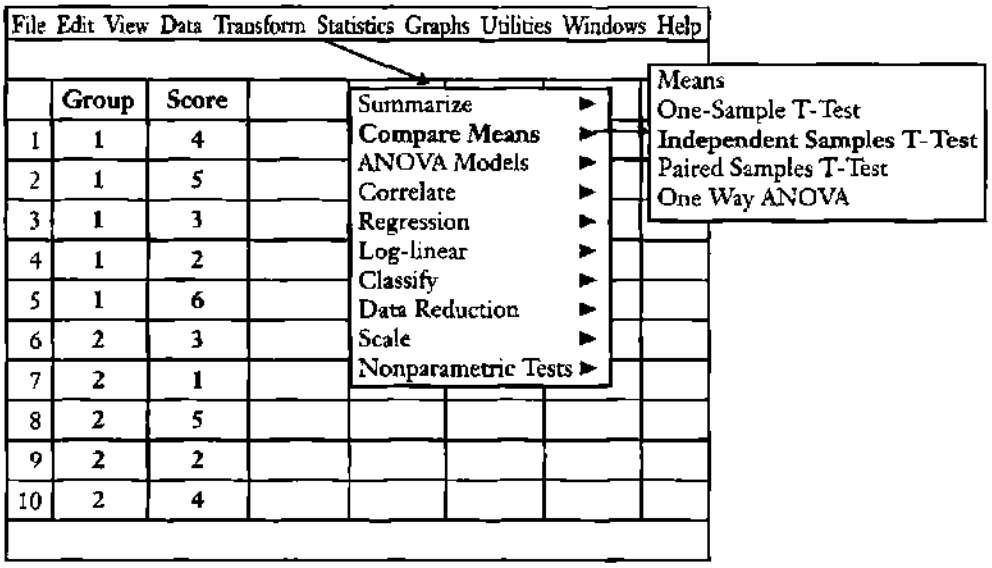
En este capítulo se examinó la razón t que se utilizó para analizar la diferencia entre dos medias y el análisis de varianza de un factor que puede usarse para analizar la diferencia entre dos o más medias grupales. Técnicamente nos referimos a los grupos como niveles de la variable independiente, y 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

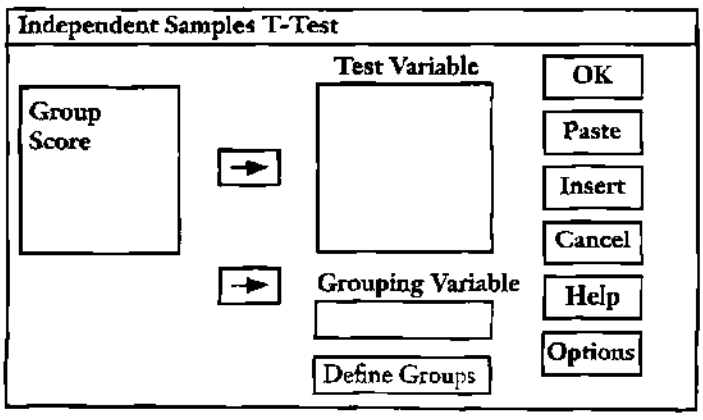
	Group	Score					
1	1	4					
2	1	5					
3	1	3					
4	1	2					
5	1	6					
6	2	3					
7	2	1					
8	2	5					
9	2	2					
10	2	4					

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*

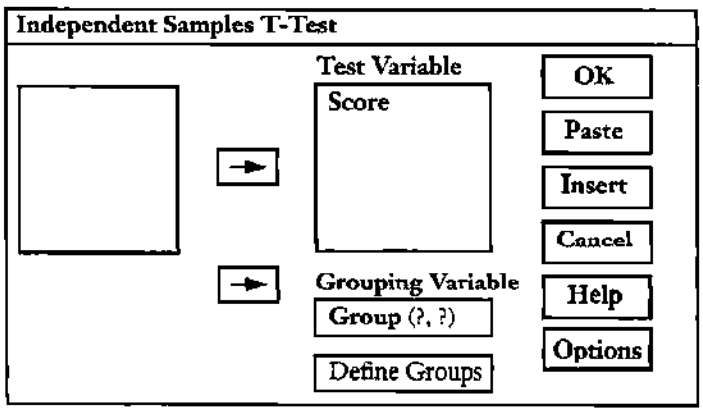
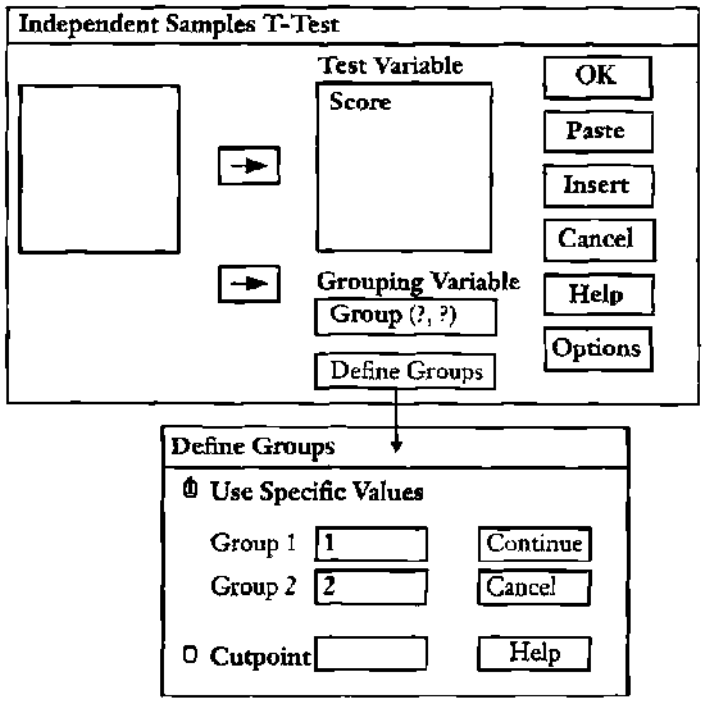


FIGURA 13.6 *Pantalla utilizada para definir niveles de la variable independiente*

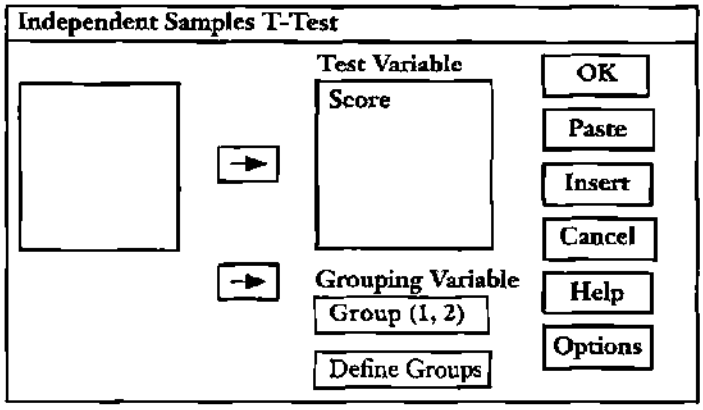


Después de hacerlo aparece otra pantalla que permite definir los niveles de la variable Group. La figura 13.6 muestra dicha pantalla. Observe que se anota "1" para el Grupo 1 y "2" para el grupo 2. Para regresar a la pantalla previa se hace clic en el botón "Continue". Sin embargo, los dos signos de 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*

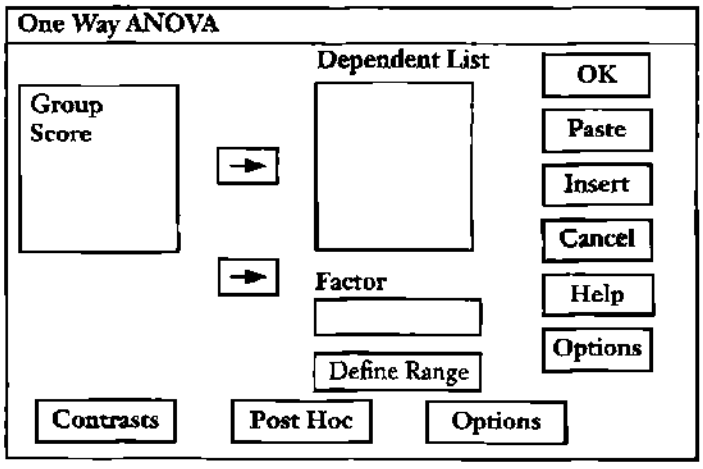


FIGURA 13.9b *Pantalla para especificar las variables dependiente e independiente*

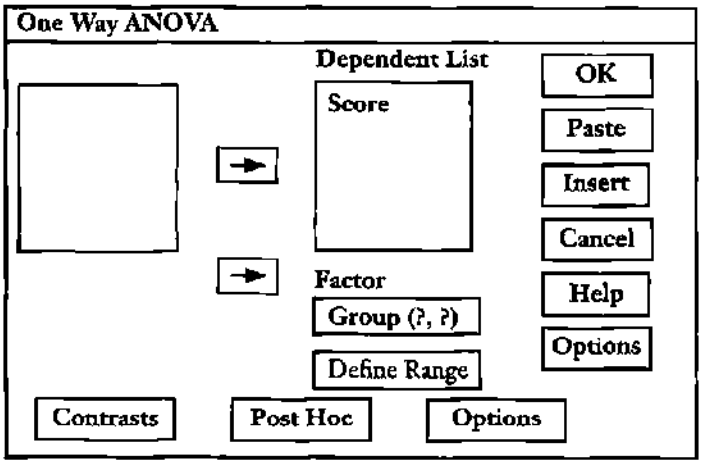


FIGURA 13.10 *Pantalla para definir el rango de valores para la variable independiente*

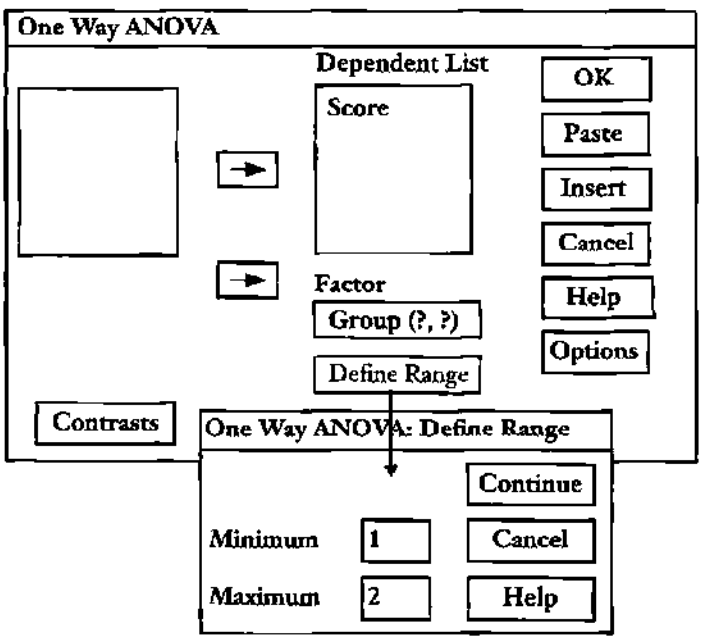
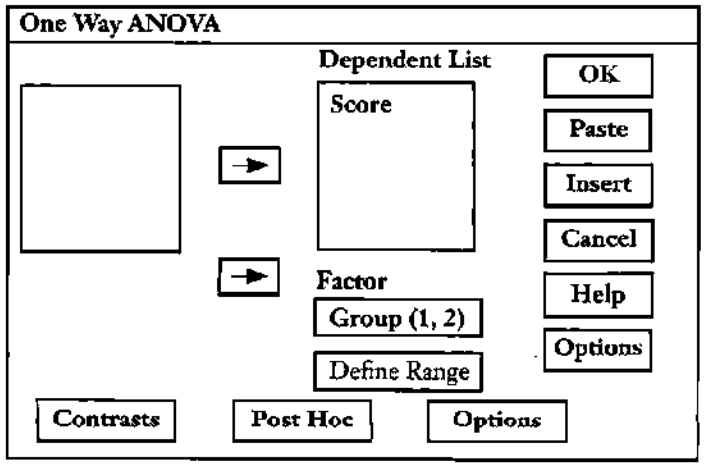


FIGURA 13.11 *Pantalla del ANOVA de un factor después de definir el rango*



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, ΣX_j y súmelas para obtener la ΣX_t. Calcule la N total a partir de las n de los grupos.

$$\Sigma X_t = \Sigma [M_j n_j] = (5)(6) + (5)(3) = 45; N = 5 + 5 = 10$$

2. Término de corrección: $(\Sigma X_t)^2/N = 45^2/10 = 202.50$ (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:

$$SC_e = \Sigma [n_j M j^2] - C$$
$$SC_e = [(6^2)(5) + (3^2)(5)] - C = 225.00 - 202.50 = 22.50$$

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 Behavior, 13*, 449-464.

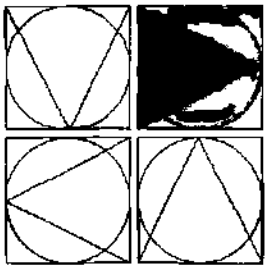
Kirsch, I., Mobayed, C. P., Council, J. R. y Kenny, D. A. (1992). Expert judgments of hypnosis from subjective state reports. *Journal of Abnormal Psychology, 101*, 675-662.

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.

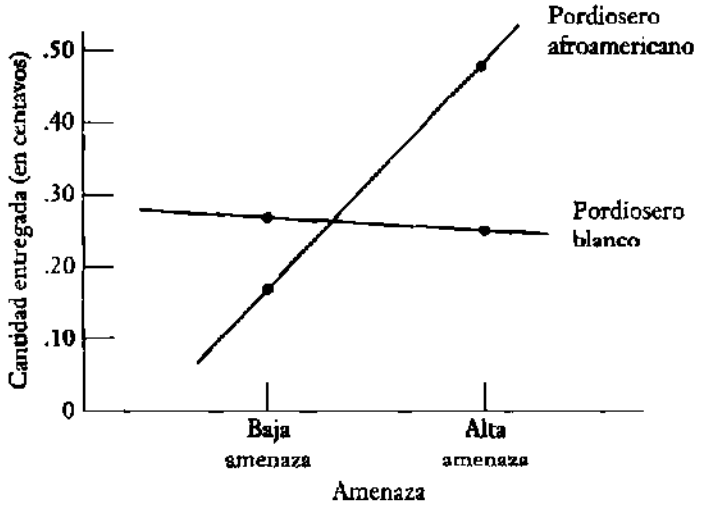
La segunda variable experimental, "raza", se manipuló de la siguiente manera: después de completar la manipulación de la variable *amenaza*, a los sujetos se les pagó con monedas de 25 centavos y se les informó que podían irse. A la salida del laboratorio, un cómplice afroamericano se dirigió a la mitad 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:

$$y = a_o + A + B + AB + e \qquad (14.1)$$

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:

$$y = a_o + A + B + C + AB + AC + BC + ABC + e \qquad (14.2)$$

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
− (entre métodos + 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*

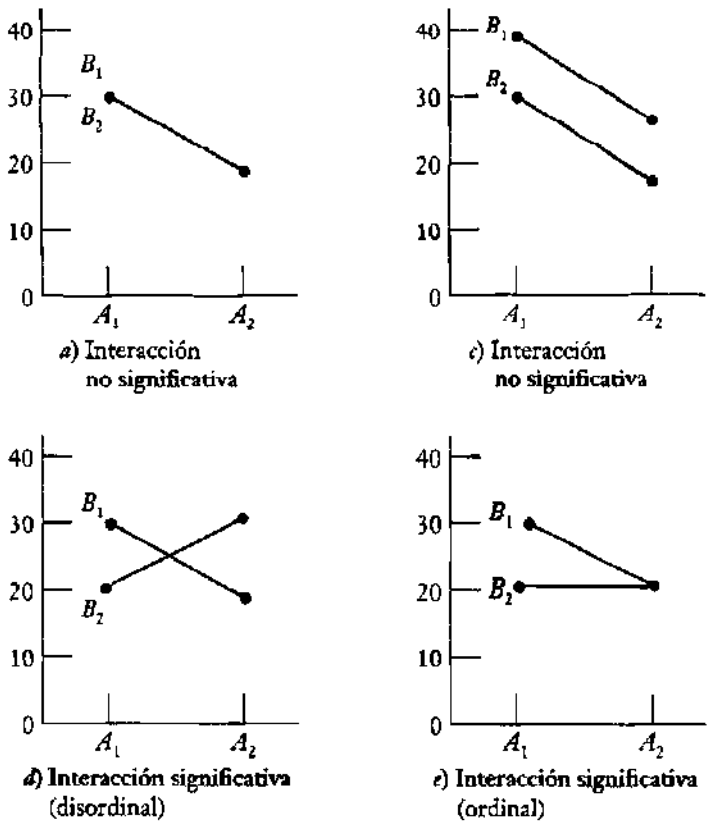
	A_1	A_2		A_1	A_2		A_1	A_2	
B_1	30	20	25	30	30	30	30	20	25
B_2	30	20	25	20	20	20	40	30	35
	30	20		25	25		35	25	
a) A significativa; B no significativa; interacción no significativa				*b*) A no significativa; B significatitiva; interacción no significativa			*c*) A significativa; B significativa; interacción 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 (1984) 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_a, la varianza de la clase social y reintroduciendo V_d:

$$V_t = V_e + V_a + V_d \tag{14.6}$$

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_w, 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 × 2 × 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]

	Habilidad			
Género	**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

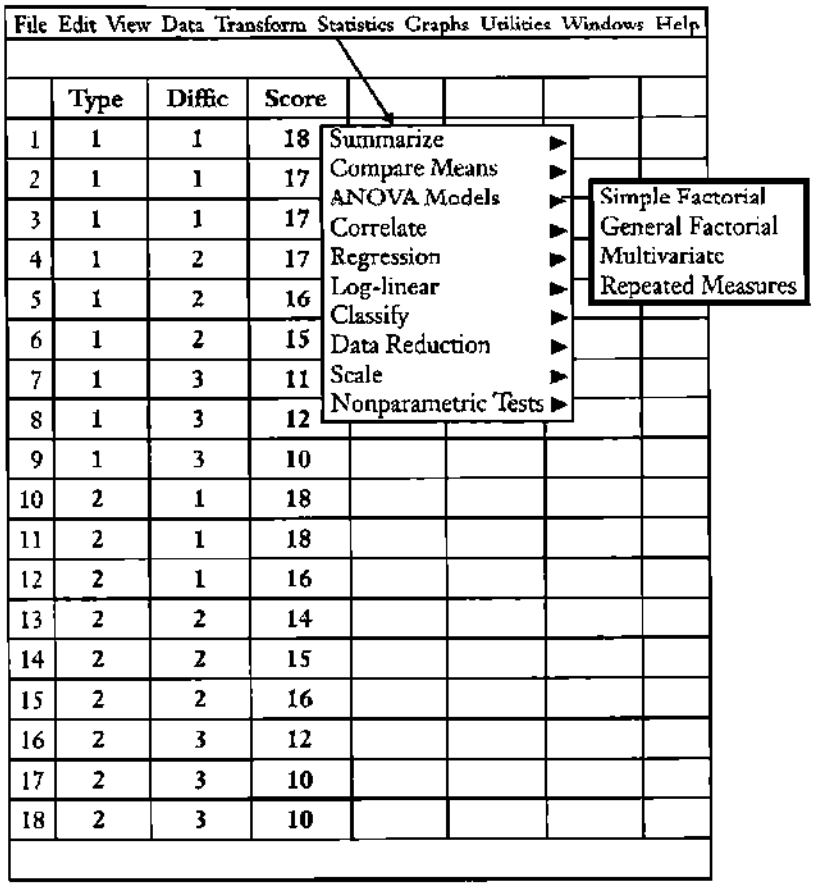
resulta muy similar. La tabla 14.18 presenta datos ficticios en la tradicional forma de tabla. La figura 14.4 muestra cómo se reestructuraron tales datos en el formato de la hoja de cálculo del SPSS. Es muy importante que el lector sepa cómo moverse de la presentación de los datos en la tabla 14.18 a 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



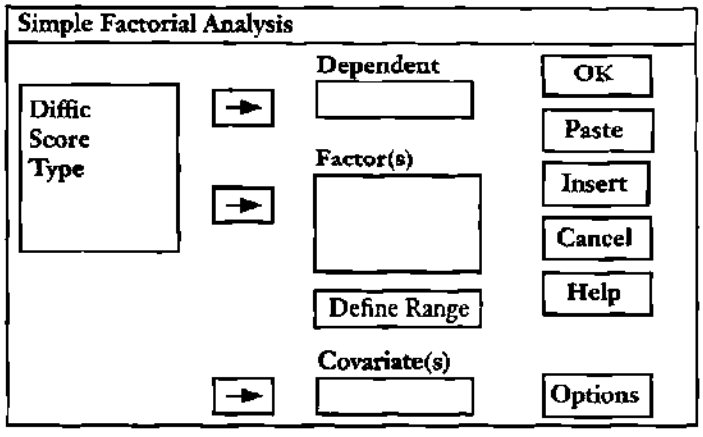
	Type	Diffic	Score
1	1	1	18
2	1	1	17
3	1	1	17
4	1	2	17
5	1	2	16
6	1	2	15
7	1	3	11
8	1	3	12
9	1	3	10
10	2	1	18
11	2	1	18
12	2	1	16
13	2	2	14
14	2	2	15
15	2	2	16
16	2	3	12
17	2	3	10
18	2	3	10

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

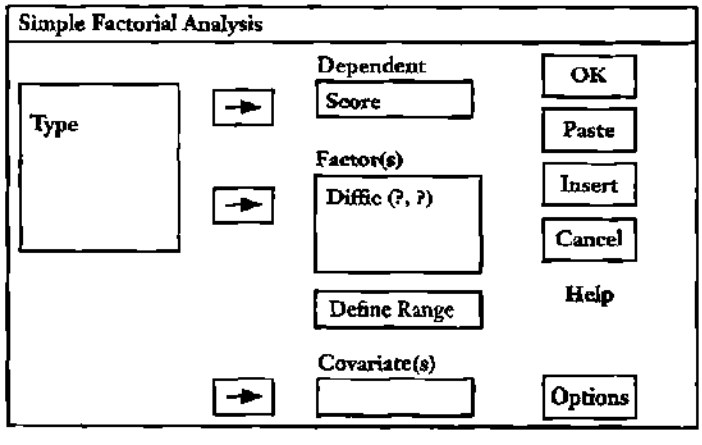
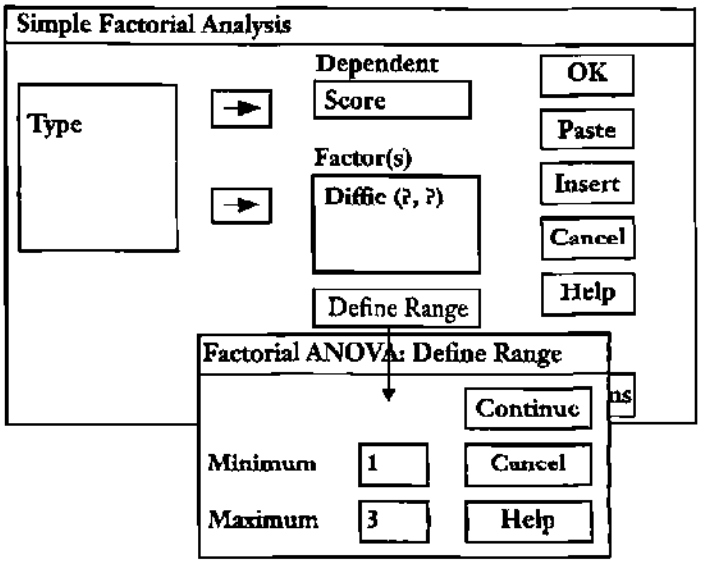


FIGURA 14.8



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

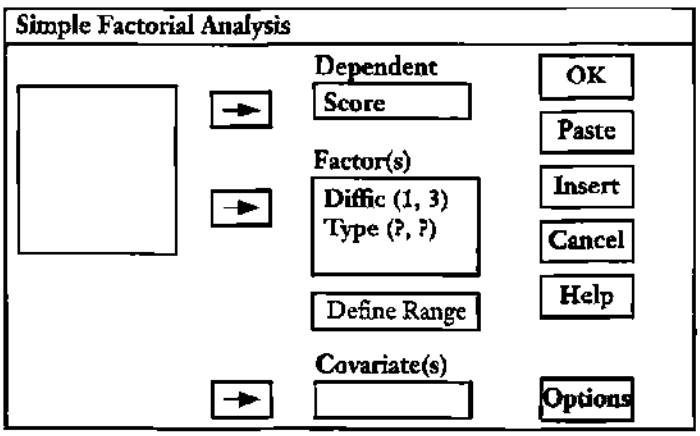
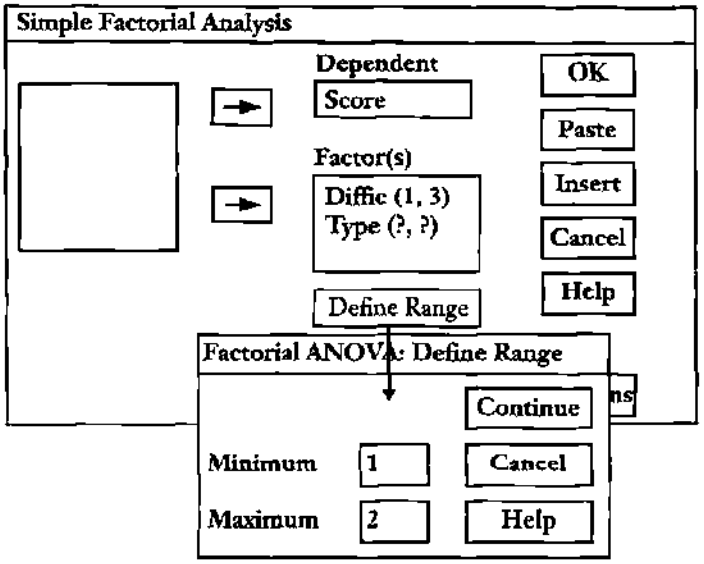
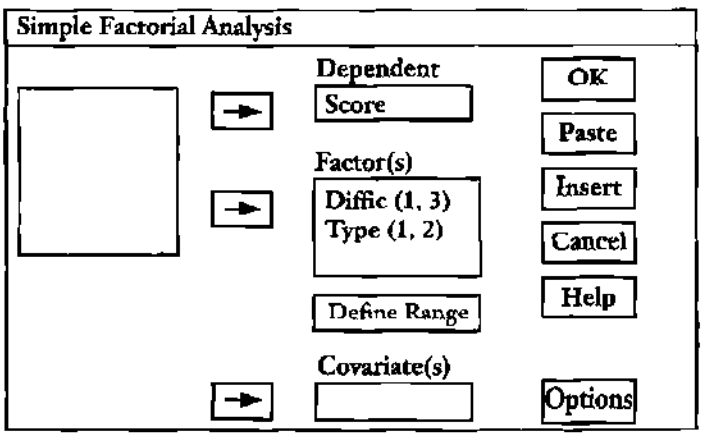


FIGURA 14.10



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)

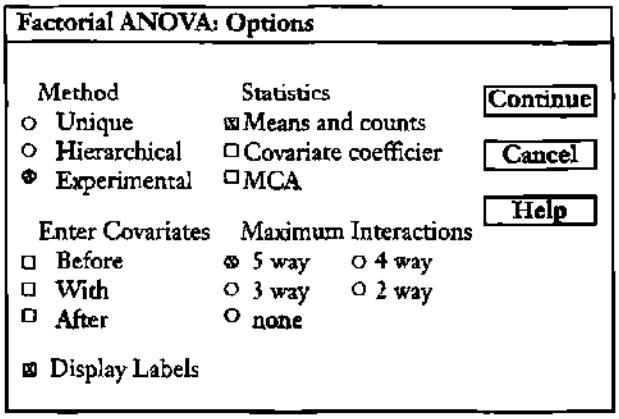
ANALYSIS OF VARIANCE

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.

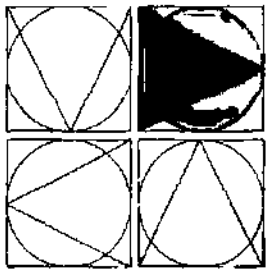
4. A partir de los números aleatorios del Apéndice a, obtenga 40 números, del 0 al 9, en grupos de 10. Considere a los cuatro grupos como A_1B_1, A_1B_2, A_2B_1 y A_2B_2.
   a) Realice un análisis factorial de varianza como se explicó en el capítulo. ¿Cómo deben ser las razones F de A, B y $A \times B$ (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
	$M_c = 57.80$					$M_t = 57.80$	

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:

$$sc_t = sc_e + sc_d \qquad (15.1)$$

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_e, 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:

$$sc_t = sc_c + sc_r + sc_{res} \qquad (15.2)$$

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	$\Sigma X_t^2 = 145$
			$M_t = 3.5$			$M_t = 3.5$

$$C = \frac{(35)^2}{10} = 122.50$$

$$Total = 145 - 122.50 = 22.50$$

$$Entre\ columnas\ C = \left[\frac{15^2 + 20^2}{5}\right] - 122.50 = 2.50$$

$$Entre\ renglones\ R = \left[\frac{6^2 + 4^2 + \ldots + 8^2}{2}\right] - 122.50$$

$$= 132.50 - 122.50 = 10$$

$$C = \frac{(35)^2}{10} = 122.50$$

$$Total = 145 - 122.50 = 22.50$$

$$Entre\ columnas\ C = \left[\frac{15^2 + 20^2}{5}\right] - 122.50 = 2.50$$

$$Entre\ renglones\ R = \left[\frac{3^2 + 6^2 + \ldots + 9^2}{2}\right] - 122.50$$

$$= 141.50 - 122.50 = 19$$

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 1		
Variable independiente 2	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 1		
Variable independiente 2	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
B_1	S_1 S_2 S_3 Grupo 1	S_1 S_2 S_3 Grupo 1	S_1 S_2 S_3 Grupo 1
B_2	S_4 S_5 S_6 Grupo 2	S_4 S_5 S_6 Grupo 2	S_4 S_5 S_6 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.

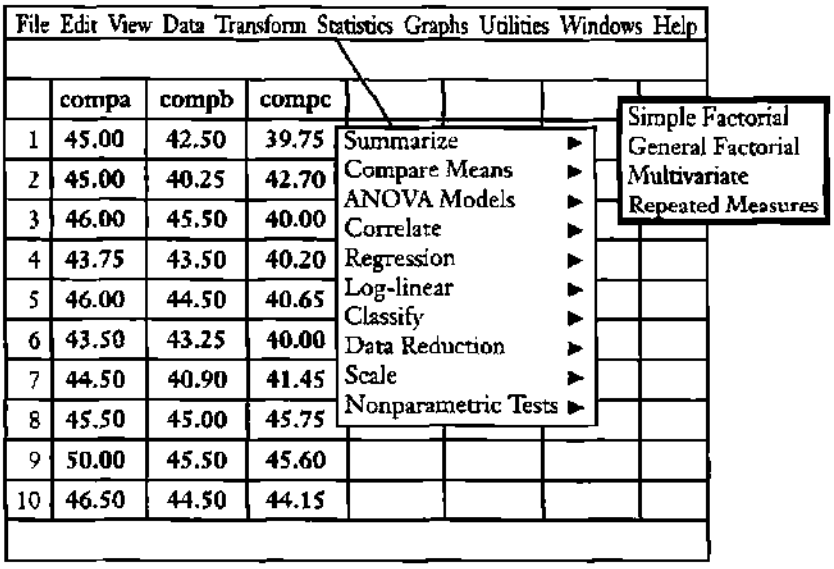
En un estudio de Reinmuth y Barnes (1975) no se utilizó el ANOVA de bloque aleatorizado para analizar sus datos. Sin embargo, los datos que recolectaron en su estudio sobre las propuestas de tres compañías extractoras de petróleo se ajusta al diseño de bloques aleatorizados. El estudio constituía en 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

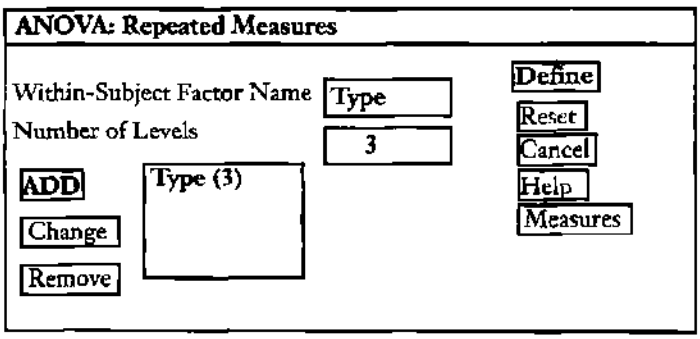
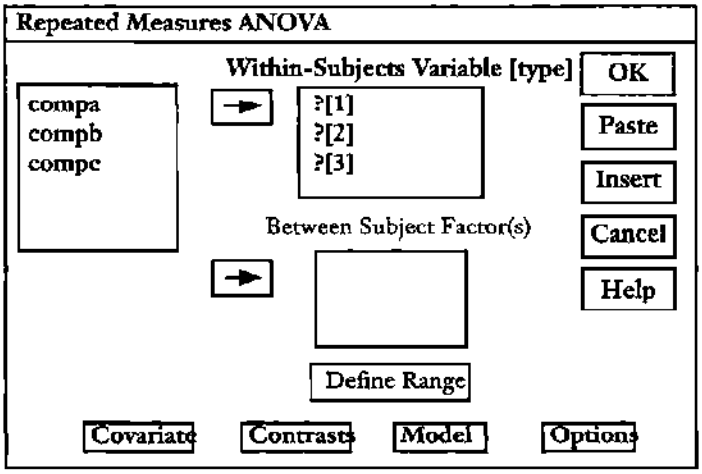


FIGURA 15.3



(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

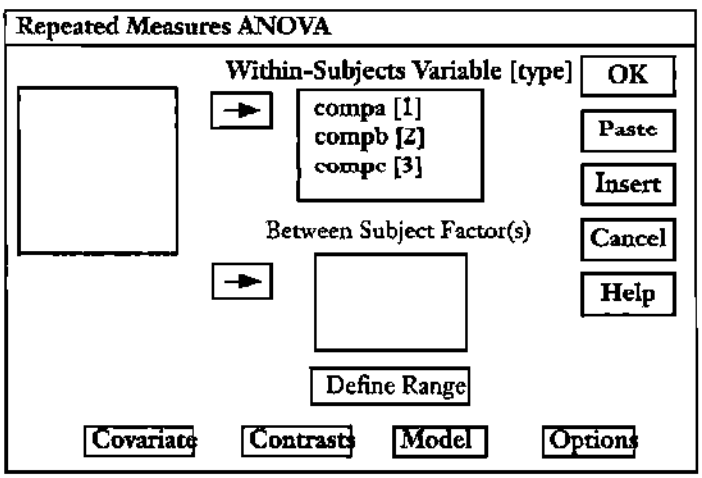


FIGURA 15.5

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 DiCara)*[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.

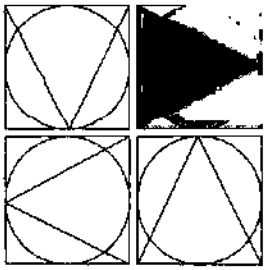
de orina; parte de los datos están reproducidos en la tabla 15.15. Los datos contenidos en II a la derecha son los incrementos en la secreción de orina de siete ratas seleccionadas aleatoriamente de un grupo de 14 ratas, antes y después del "entrenamiento", el cual consistió en un condicionamiento instrumental: 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

$$H = \frac{12}{N(N+1)} \sum \frac{R_j^2}{n_j} - 3(N+1) \qquad (16.1)$$

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. Él 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_e. (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:

$$EE_{MR(}=\frac{\text{observación mayor} - \text{observación menor}}{N}$$

Una prueba t de la diferencia entre dos medias puede realizarse mediante la siguiente fórmula:

$$t_r = \frac{M_1 - M_2}{\frac{1}{2}(R_1 - R_2)}$$

donde t_r 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*

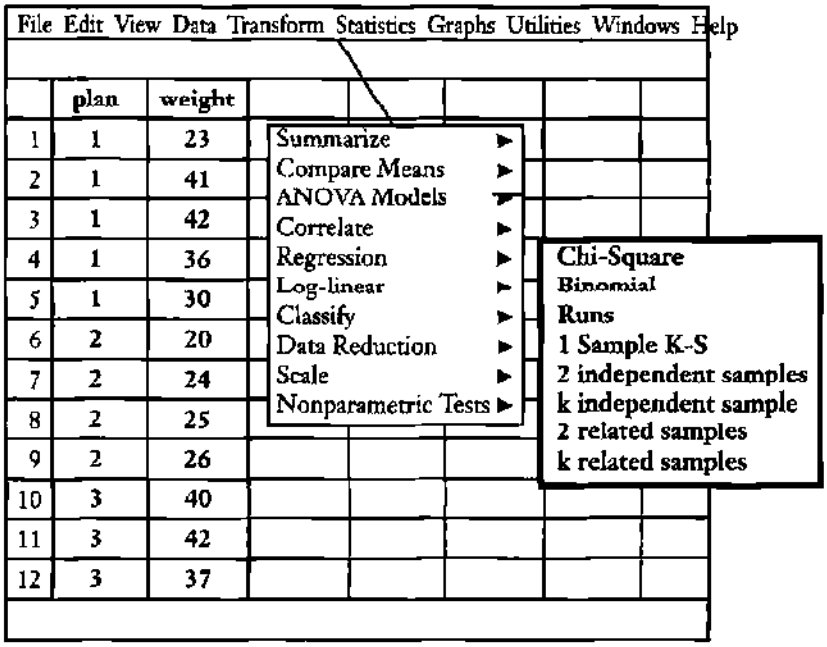


FIGURA 16.2 *Panel del SPSS para la especificación de las variables*

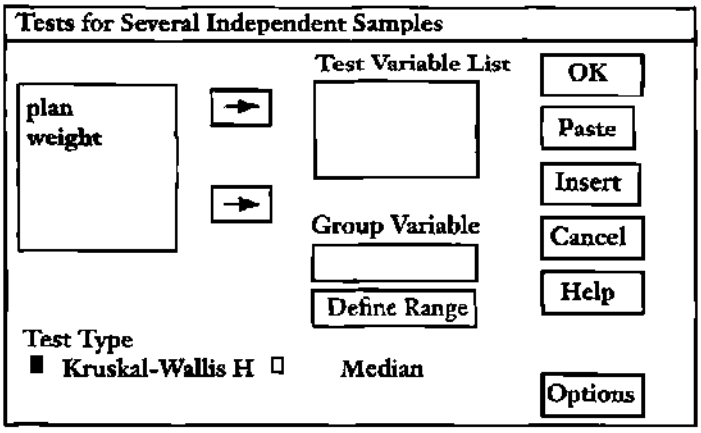
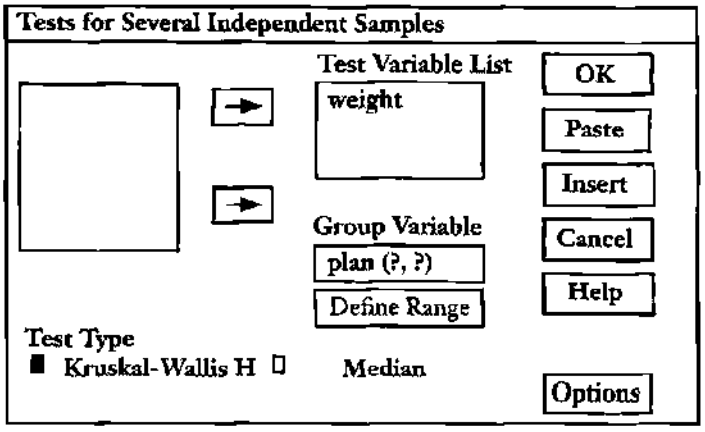


FIGURA 16.3 *Desplazamiento de las variables "peso" y "plan" dentro de los cuadros apropiados*



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*

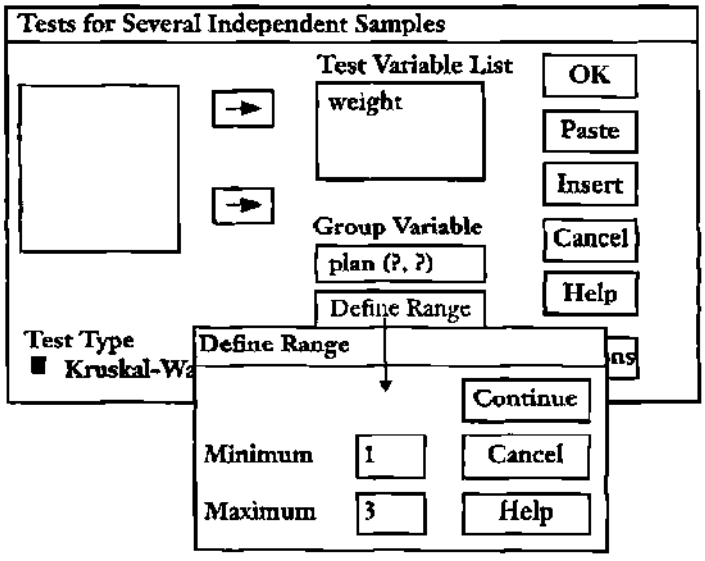
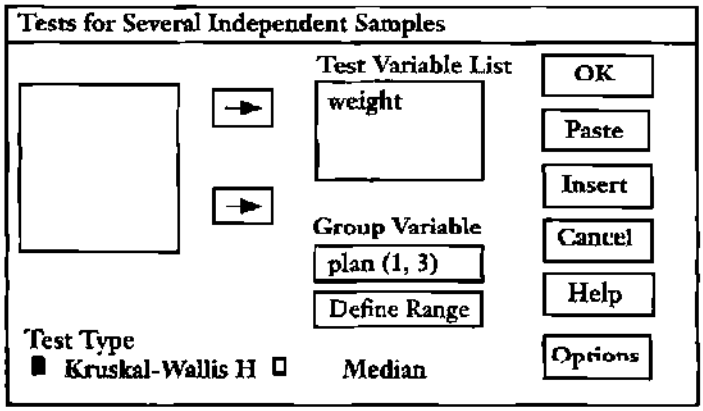


FIGURA 16.5 *Pantalla del SPSS antes de ejecutar el análisis*



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

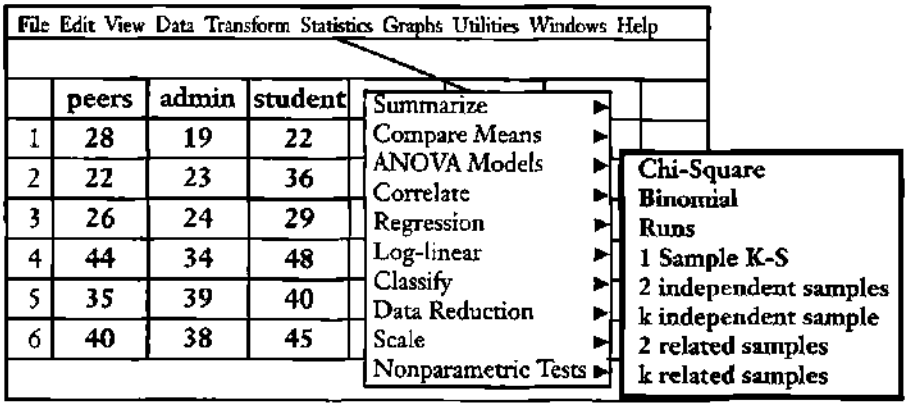
Mean Rank Cases

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*

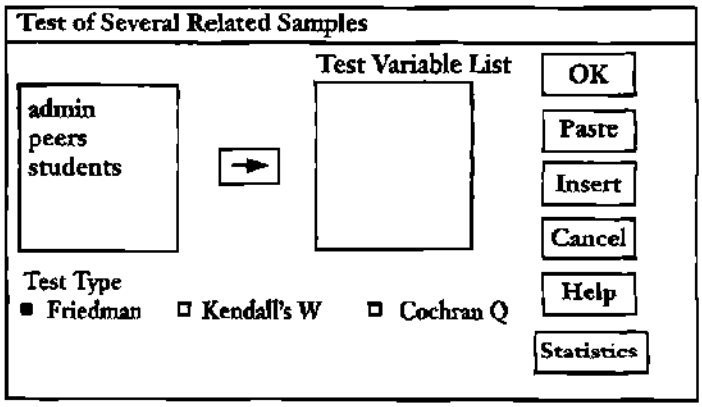
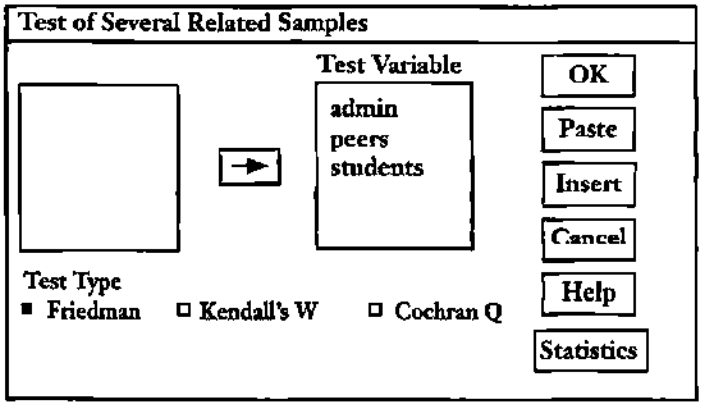


FIGURA 16.9 *Pantalla resultante del SPSS, previa al análisis*



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 $\eta^2 = sc_t/sc_t$.) 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!)

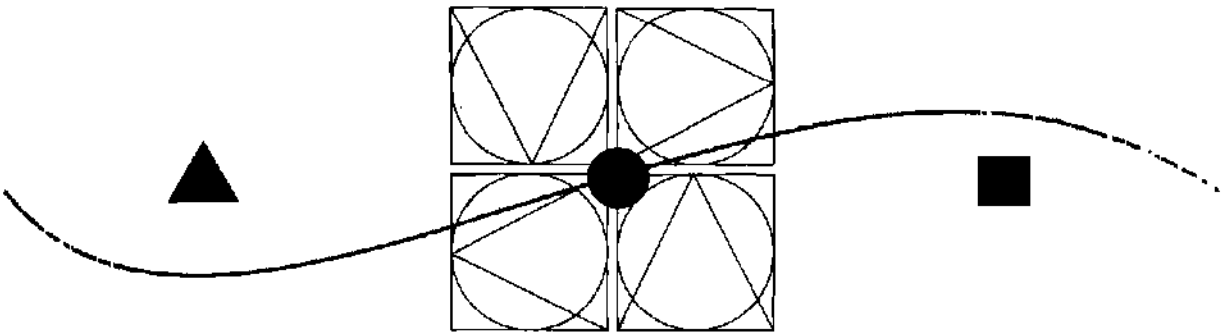
b) Invente dos casos de muestras de 20 cada una, que contenga diferentes números de series y utilice su prueba para evaluar la significancia del número de series en las muestras.

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

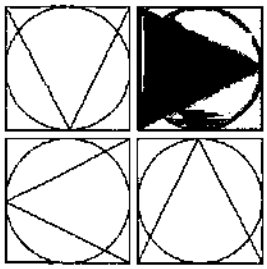
Diseños inadecuados y criterios para el diseño

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.

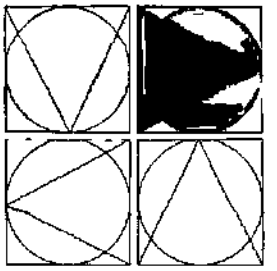
Brody, R. G. y Bowman, L. (1998). Accounting and psychology students' perceptions of whistle blowing. *College Student Journal, 32*, 162-166. (¿Debe el currículum universitario incluir ética?)

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.

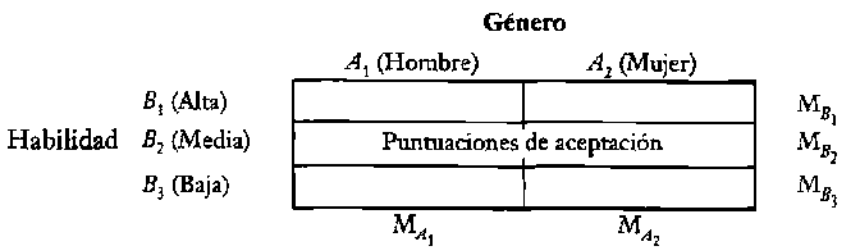
Un diseño más fuerte

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