14. Fundamentos teóricos de metodología estadística

Miguel Ángel Escotet

¿La estadística, para qué?

Este capítulo tiene como objetivo ofrecer una visión general de la estadística aplicada a la psicología, especialmente lo relativo al análisis multivariado y los diseños estadísticos de investigación. Deliberadamente se eliminan los cálculos de las diferentes técnicas, dado que actualmente los mismos pueden ser realizados por los diferentes programas estadísticos adaptados a los microordenadores o microcomputadoras y se escapan a los objetivos de esta obra. Por tanto, el mayor énfasis se hará en la teoría y racionalidad de las diferentes medidas como preámbulo al estudio sistemático de la estadística como disciplina científica.

La técnica estadística es esencialmente derivada de la misma estructura que otras formas científicas en donde se utilizan conjuntamente procesos inductivos y deductivos. Es decir, mediante la observación de fenómenos significativos y a partir de un número de experimentos, se llega por un proceso de inducción a la formulación de una teoría que permita relacionar los resultados a un esquema hipotético. A partir de esta teoría, y mediante un proceso deductivo, se podría predecir los resultados de los subsiguientes experimentos y, en esta forma, verificar o negar el planteamiento inicial de las predicciones.

Sin embargo, la estadística ha sido mal interpretada en sus posibilidades de aplicación. Para muchos y por muchos años, la estadística ha sido un conjunto de hechos numéricos y representaciones generadas de una descripción e inferencia de los datos. Todavía, una buena parte de los textos de estadística, dividen a ésta en dos grandes áreas. La estadística descriptiva y la estadística inferencial o muestral. Un examen cuidadoso de los nuevos enfoques estadísticos nos llevan a redescubrir tres áreas y aplicaciones, que son: *a*) La descripción de las observaciones; *b*) la extracción de inferencias sobre hipótesis científicas derivadas de las observaciones y *c*) el diseño de estudios o experimentos. Es quizá esta última área, la más importante, aun cuando su estudio se derive de las otras dos.

Pero quizá la pregunta que más nos hacemos cuando estamos en los años iniciales del estudio de la psicología es: ¿Para qué y por qué la estadística en la formación del psicólogo? Es aquí donde podríamos incluir uno de los hechos que caracterizan el comportamiento humano: la variabilidad o varianza. No podemos decir que existen dos seres humanos exactamente iguales y por consiguiente sus conductas difieren entre sí. Sin embargo, como señalaba Skinner, la aparición de la variabilidad es el resultado de la falta

de controles adecuados. Para el conductismo ortodoxo, quizá la estadística puede ser desestimada, pero aun en el método del caso histórico de Skinner se genera una serie estadística, a través del registro acumulativo.

Una ley de comportamiento no solamente se deriva de los aspectos o comportamientos comunes, sino de las diferencias. En química podríamos expresar que un conjunto de elementos podría pertenecer a un mismo tronco, pero que la combinación con otros elementos de la naturaleza, generarían derivados que, sin dejar de tener propiedades comunes, poseerían también propiedades específicas. La interacción del hombre y su ambiente, genera propiedades o leyes comportamentales comunes a la especie, y al mismo tiempo, principios específicos a la cultura. En este caso, la estadística es esencial como técnica para describir comportamientos generales y determinar, con cierto grado de confidencia, cuándo una hipótesis particular se mantiene, a pesar de los grados de desviación que existan. Aun cuando pueda parecer una paradoja, la razón de ser de la estadística, radica en las diferencias de los objetos examinados, ya que, asumir que todas las conductas son iguales o similares, determinaría obviamente la eliminación de la estadística como técnica inferencial y experimental.

A este nivel, podríamos decir como Murdock que «los datos de la vida social y la cultura son susceptibles al tratamiento científico como lo son también los hechos de las ciencias físicas y biológicas». Parece verse claramente que los elementos del comportamiento social, en sus permutaciones y combinaciones, se ajustan a las leyes naturales por si solos con una exactitud escasamente menos impresionante, que la que caracteriza las permutaciones y combinaciones de los átomos en la química y de los genes en la biología. El problema, si bien ha sido descubrir las similitudes y diferencias que caracterizan las propiedades de un comportamiento, no por ello podríamos concluir que la combinación y permutación en la conducta humana no existen, sino más bien que el bajo desarrollo de la medición en ciencias humanas es la verdadera causa de la falta de datos exactos. La respuesta se observa claramente en la misma historia de la ciencia.

El análisis en ciencias de la conducta, específicamente en lo concerniente a la investigación, el reto de todo investigador no solamente consiste en describir el fenómeno, sino en inferir que lo que se ha descrito también se puede observar universalmente. Este proceso denominado razonamiento inductivo, conlleva el riesgo de generalización más allá de los límites lógicos establecidos; pero en este caso, la inducción no depende tanto del proceso en si mismo, sino del que aplica dicho proceso. He ahí que el diseño experimental, como tercera etapa de la extensión estadística, deba ser el primer proceso de investigación, a fin de legitimar los límites de las conclusiones.

POSIBILIDADES Y LÍMITES DE LA ESTADÍSTICA

El límite de la estadística, en este caso, es que solamente una parte de lo que se observa es susceptible de generalización; el resto se atribuye al momento específico o al sujeto, es decir, a la especificidad de la muestra seleccionada. Si la observación es generalizable, el resultado se dice que es estadísticamente significativo; si los datos son el resultado de una fluctuación de la muestra, el resultado no es significativo. Este proceso denominado «resolución del problema estadístico» es quizá el más crucial y el que conlleva un límite decisivo hasta ahora no resuelto en la metodología estadística. El primer paso de este proceso es el trasladar la hipótesis científica verbalmente expresada a una hipótesis estadística. Por ejemplo, si deseamos demostrar que estudiantes con razonamiento abstracto

alto son también altamente capaces para la comprensión de la matemática, la hipótesis estadística podría ser que no existe correlación entre dos características. Recuérdese aquí que, en estadística muestral, las hipótesis se rechazan, no se aceptan y que por tanto la expresión de las mismas deben ser hechas en forma contraria a lo que se pretende demostrar. Esto es debido a que el estado de conocimiento acerca de la conducta humana no ha llegado a la precisión de especificar el valor anticipado, por ejemplo, de la correlación entre esas características. El día en que esto sea posible, la hipótesis nula no será necesaria usarla, pero mientras tanto, la hipótesis estadística será planteada como «no relación» o «no diferencia».

El segundo paso se refiere a la elaboración de observaciones, mediante la selección de una de las posibles muestras de un universo y el cálculo de la combinación particular de las observaciones. Hecho esto, se trata de localizar la prueba estadística observada proveniente de la distribución de frecuencia; es decir, la comparación entre el resultado obtenido y el resultado probable esperado. Esto constituirá la tercera etapa del proceso. La última, y en donde mayores problemas existen, está relacionada con la toma de decisión de aceptar o rechazar la hipótesis estadística. En principio, parecería muy sencillo indicar que si el resultado de nuestra prueba estadística proviene de una distribución de frecuencia matemática, la hipótesis sería aceptada y si el valor encontrado fuere descubierto en forma muy casual en la distribución, sería rechazada. Pero realmente, y debido a que la mayoría de las distribuciones muestrales son asintóticas en relación con la línea de base, es decir, nunca tocan los ejes, siempre existe la posibilidad de cometer el error de rechazar una hipótesis, siendo ésta verdadera. La posibilidad entonces sería de aceptar la hipótesis, pero en este caso el error que conllevaría tal decisión sería el de aceptar una hipótesis falsa. Por este motivo, en la toma de decisión de aceptar o rechazar una hipótesis existen dos tipos de errores: *a)* el rechazo de una hipótesis verdadera o *Error Tipo I*; y *b)* la aceptación de una hipótesis falsa o *Error Tipo II*. Estos dos tipos de errores ponen al investigador en un dilema. Si toma la decisión de aceptar la hipótesis se incrementa la posibilidad de cometer el Error Tipo II y si por el contrario decide rechazarla se aumenta el Error Tipo I.

Este dilema no ha sido resuelto y es una de la mayores áreas de investigación de la estadística. Hasta el momento tres procedimientos han sido sugeridos para ayudar a la toma de decisiones en este sentido. El primero de ellos, y el más antiguo, ha sido el de seleccionar un punto de tres desviaciones estándar por encima de la media de la distribución muestreada, es decir, el procedimiento de la Razón Crítica. Si esta Razón es por lo menos tres veces mayor, la hipótesis se rechaza y si es menor se acepta. Este procedimiento fue útil en la utilización de grandes muestras, pues reduce significativamente el Error Tipo I, pero en muestras pequeñas tales como el ji-cuadrado, «t» o F o en muestras que no se aproximan a la curva normal, la *Razón Crítica* varía de una distribución a la otra y por tanto el Error Tipo I no se mantiene constante. El procedimiento que siguió a éste, fue el de establecer el «nivel de Significación» tanto para muestras pequeñas como grandes. Generalmente se rechazaban las hipótesis al uno por ciento (0.01) o cinco por ciento (0.05) de error, pero estos niveles arbitrarios, si bien protegen al investigador contra el Error Tipo I, ignoran completamente el Error Tipo II, ya que al reducir el nivel de significación se incrementa el riesgo de aceptar una hipótesis falsa.

El procedimiento más reciente trata de considerar los dos Tipos de Errores; es decir: *a)* calcular la posibilidad de cometer ambos tipos de errores; y *b)* establecer un juicio de valor sobre que tan serio es cada Tipo. En cierto modo, este procedimiento involucra a la teoría de la decisión, además de los procedimientos clásicos estadísticos. Sin embargo, a través de este procedimiento quizá podemos llegar a conocer ambos errores o más bien su probabilidad, pero difícilmente los reduciríamos al mismo tiempo. La única forma de

reducirlos es, en el momento actual, mediante el incremento del tamaño de la muestra, ya que a mayor población menor desviación estándar de las dos distribuciones.

En resumen, podemos decir que el denominado nivel de significación corresponde a la probabilidad de cometer el Error Tipo I, mientras que la probabilidad de tener Error Tipo II se refiere a la «potencia» de la prueba estadística y que la única forma de reducir ambos errores serían manteniendo el Tipo I constante y aumentando la potencia o número de observaciones de la muestra. El problema radica, en cuánto debería ser el aumento para que el Error II pudiera ser tolerado. Esta es una limitación intrínseca de la estadística y cuya solución está en el futuro.

DEFINICIÓN DE LA ESTADÍSTICA

La estadística es la técnica que computa y enumera los hechos y los individuos susceptibles de enumerarse o de medirse; coordina y clasifica los datos obtenidos con el fin de determinar sus causas, consecuencias y tendencias. Los fenómenos estadísticos se clasifican en *típicos*, cuando iguales circunstancias e idénticas causas producen efectos o resultados iguales, y en *atípicos* cuando idénticas causas producen resultados diferentes.

La estadística no puede alimentarse por si misma, necesita de la buena interpretación de quien la aplica. Como primer elemento para un buen análisis estadístico, se debe tener en cuenta, como criterio orientador, que aun cuando se están empleando cifras que son exactas como símbolos matemáticos, conceptualmente no lo son. Segundo: aunque tengamos prejuicios, no debemos hacer prevalecer nuestro criterio ni acomodar las cifras para llegar a un resultado que satisfaga nuestros deseos. Tercero: para que la estadística sea de utilidad necesita ser una técnica de causalidad que investigue las causas del fenómeno, porque siempre que se presenta una situación patológica se debe buscar la causa, que a veces puede ser múltiple. Cuarto: es prudente comparar datos homogéneos obtenidos análogamente. Quinto: debemos basarnos siempre en hechos concretos, positivos y tangibles y no en simples hipótesis o suposiciones. Y sexto: debemos tener normas escrupulosas para deducir la mayor exactitud posible de los datos.

Medidas de tendencia central y de variabilidad

Las representaciones gráficas nos pueden mostrar de una vez, toda una serie estadística o distribución de frecuencias. Pero en ocasiones, se desea un valor numérico que representa todo el colectivo o muestra que se estudia; este número recibe el nombre de centro o media de la distribución, porque a su alrededor se agrupan todos los demás. La media aritmética, la mediana y otras, reciben el nombre de números o medidas estadísticas de tendencia central o de concentración. Por otra parte, a través del cálculo de las medidas de tendencia central, tenemos una idea del conjunto; pero la información que nos dan es insuficiente; por ejemplo: dos estudiantes han obtenido las siguientes calificaciones mensuales en psicología general durante el año académico:

Estudiante A: 16 - 4 - 18 - 0 - 2 - 14 - 16 - 10.
Estudiante B: 10 -10 - 8 - 10 - 10 -12 -10 -10.

Los dos estudiantes, *A* y *B*, obtienen 10 de promedio y, sin embargo, las calificaciones

que han obtenido ambos son muy distintas. El estudiante *A* se separa mucho de 10, mientras que el alumno *B* ha conseguido una mayor regularidad. Otro ejemplo podría ser el siguiente: si las edades de los pacientes en una psicoterapia de grupo son, respectivamente, 38, 41, 40, 41, 39, 39 y 42 años, es admisible decir que el grupo tienen 40 años de promedio, mientras que si las edades de los miembros de una familia son 13, 15, 18, 40, 45, 67 y 82 años, no tiene ningún sentido decir que la edad general es 40 años, aunque 40 sea el promedio de las edades. Vemos, pues, la necesidad de introducir un método de apreciar la propiedad con que los números estadísticos definidos, caracterizan a la serie. Para ello se utilizan las llamadas «medidas de variabilidad o dispersión», de las cuales presentamos brevemente la desviación semi-intercuartilar, la desviación típica o estándar y la varianza.

MEDIA ARITMÉTICA

Se denomina media aritmética de *n* números, al cociente de dividir entre n la suma de esos números. La media es la medida de concentración o de tendencia central más estable y que posee propiedades matemáticas, que no tienen ni el modo ni la mediana. La media aritmética tiene un valor intermedio entre el menor y el mayor de los números considerados y se representa por M o X.

En efecto, sean los números

$$a_1, a_2, a_3, \ldots a_n,$$

entre los cuales supondremos que a_1 es el menor y a_n el mayor.

De acuerdo con la hipótesis, se verificará:

$$\text{una}_1 < a_1 + a_2 + a_3 + \ldots + a_n < \text{una}_n.$$

Y de aquí,

$$a_1 < \frac{a_1 + a_2 + a_3 + \ldots + a_n}{n} < a_n$$

como se quería demostrar.

Como los números, generalmente, son expresión de diferentes datos diremos que media aritmética simple, *de un grupo o serie de datos, es su suma dividida entre el número de ellos.*

MEDIANA

Si, por ejemplo, se han tomado varias muestras de un mineral y obtenemos, por tanto el porcentaje en peso, o sea determinados valores, y los ordenamos siguiendo el criterio más lógico, en nuestro caso obtendremos una sucesión creciente como ésta:

14.9 - 17.8 - 17.9 - 18.6 - 18.6 -18.8 - 19.6 - 20.2 - 21.2 - 21.3 - 21.6 - 21.9 - 22.9 - 23.2 - 23.5 - 23.8 - 24.3 - 25.3 - 26.0 - 26.1.

Se define la mediana (Mdn) como el elemento que ocupa el lugar central en la sucesión. Así, en nuestro caso, la mediana sería:

$$\text{Mdn} = \frac{21.3 + 21.6}{2} = 21.45$$

Dicho de otra manera, *mediana es la puntuación por encima de la cual se encuentra la mitad de las demás puntuaciones y por debajo, la otra mitad.* Otras medidas de tendencia central que sólo se enumeran son: el modo, la media geométrica, la media armónica, la media móvil y las medias cuadrática, cúbica y bicuadrática.

DESVIACIÓN SEMIINTERCUARTILAR

La desviación semintercuartilar tiene su base en la mediana, a través de los cuartiles. Cuartil es un punto sobre una escala que divide el número de observaciones, dentro de dos grupos con proporciones conocidas en cada grupo. Por ejemplo, hay tres cuartiles: Q_1, Q_2 y Q_3; ellos dividen un grupo de observaciones dentro de cuatro partes. Q_1 es el punto sobre la escala numerada que comprende el primer cuarto del total de observaciones; la mitad de las observaciones están comparadas hasta el punto Q_2 y el 75% de las observaciones hasta Q_3.

La distancia entre el primer y tercer cuartiles de un grupo de puntuaciones ($Q_3 - Q_1$) se llama *amplitud intercuarticular.* Por tanto, *desviación semintercuartilar* (Q) es igual a la mitad de la distancia entre el primero y tercer cuartil.

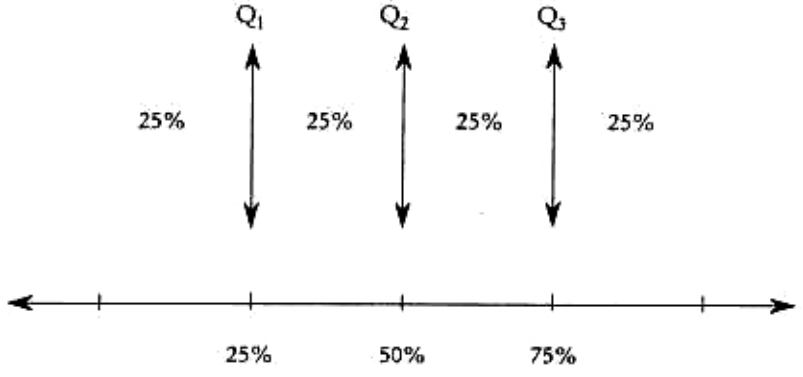

Figura 14.1. Porcentajes de casos entre cuartiles.

DESVIACIÓN TÍPICA O ESTÁNDAR Y VARIANZA

Lo más sencillo para caracterizar una serie parece que debiera ser hallar las desviaciones x de todos los elementos de la misma respecto de la media y calcular el promedio de las desviaciones, tomando ese promedio como medida de la dispersión de la serie; pero este razonamiento resulta ilusorio, ya que está demostrado que la suma de las desviaciones de los elementos de una serie estadística con respecto a la media, es nula.

Pero el promedio de los cuadrados de las desviaciones con respecto a la media, como

cuadrado de un número al que llamaremos *desviación típica*, *desviación estándar* o *desviación cuadrática media*, que no será nunca nula, por no poder serlo una suma de cuadrados, más que en el caso de ser todos nulos, o sea, iguales los elementos de la serie.

Podemos enunciar, en general, que la desviación cuadrática (así llamada por llevar las desviaciones en segunda potencia) es mínima, cuando se calcula con respecto a la media, siendo ésta una de las razones que hacen tomar gran importancia a dicho valor. La desviación típica o estándar es la medida de variabilidad más exacta, más importante y la más utilizada. Se representa mediante los símbolos *DS*, *DT*, s, *S*, σ.

Cuando la desviación estándar es pequeña, quiere decir que los términos se concentran alrededor del promedio; cuando es grande, significa que se encuentran esparcidos ampliamente. En esta forma, por ejemplo, cuando un profesor conoce la media y la desviación estándar de una clase en un examen sabe también cual es el nivel general de aprovechamiento, así como la amplitud de la extensión. Esto le da una base para interpretar la magnitud de una calificación particular en relación con la actuación del resto de la clase.

Imaginemos por un momento, que a un sujeto se le efectúan ocho mediciones de actitudes durante un año y que necesitamos comparar sus resultados con los de otros sujetos sometidos a pruebas similares. Realmente existen diferentes pruebas estadísticas para este tipo de diseños, como veremos más adelante. La medida que está involucrada en la mayoría de esas pruebas de hipótesis es la desviación estándar o su cuadrado, la varianza.

La varianza es por tanto, la media de las desviaciones típicas de las medidas alrededor de sus medias. En términos de cálculo es el cociente resultante de dividir la suma de los productos de las frecuencias por los cuadrados de las desviaciones correspondientes sobre el número de las puntuaciones, o lo que es lo mismo, el cuadrado de la desviación estándar. La misma es un valor numérico basado en la variación de los datos o puntuaciones dentro de una muestra. Si existe una gran variación de los datos dentro de esa muestra, la varianza será proporcionalmente grande. El análisis de varianza que se estudiará más adelante es precisamente un estudio de la variación entre y dentro de los grupos, de sus interacciones y del conjunto total.

Conceptos generales de muestreo y curva normal

Muestra es el conjunto finito que separamos de un colectivo, entendiéndose por colectivo, población o universo al conjunto del cual se han extraído los números o atributos que forman la serie estadística. En otras palabras, colectivo sería la totalidad de valores posibles de una característica particular de un grupo especificado de objetos al cual se llama universo. Fundamentalmente, el propósito de la estadística es el de encontrar formas lo más simples posibles para informar y describir acerca de los aspectos cuantitativos de una serie de datos o atributos. Ahora bien, al hacer un análisis estadístico de los datos, generalmente usamos muestras del colectivo a cuya totalidad generalizamos los resultados obtenidos. Hacer un análisis de toda la población o universo, en muchos casos sería realmente imposible o requeriría un extraordinario esfuerzo y gastos de energías, tiempo, dinero, etc. Es por eso que al realizar una investigación estadística debemos utilizar muestras que sean representativas del colectivo o universo para que los resultados se acerquen lo más posible al total del universo; pero el hecho de tratar con muestras y no con el total de la población da lugar a uno de los principales fundamentos de la inferencia estadística que es la fidedignidad de las muestras escogidas como representantes del colectivo o total de la población. La validez de los resultados obtenidos dependerán, en parte, de la habilidad y corrección en el

estudio de la muestra. Llegados a este punto, podríamos definir como muestra estadística, a *una parte de la población seleccionada de acuerdo con una regla o plan.*

CLASES DE MUESTRAS

Podemos agrupar las muestras en dos grandes clases: *a*) muestras de probabilidad; y *b*) muestras aleatorias. Las *muestras de probabilidad* son aquellas elegidas de acuerdo con un mecanismo casual, para saber si cada elemento del colectivo o universo tiene una probabilidad conocida de pertenecer a la muestra. Dentro de una *muestra aleatoria*, el elemento o cada miembro de la población tiene las mismas probabilidades de ser elegido o pertenecer a la muestra.

La muestra aleatoria es la más preferida, ya que una buena muestra es aquella a partir de la cual pueden hacerse generalizaciones respecto al colectivo, mientras que una mala muestra no permite tales generalizaciones. Para generalizar de la muestra al colectivo, necesitamos deducir, a partir de cualquiera de las suposiciones respecto a la población, cuándo la muestra observada está dentro del rango de variación del muestreo que puede ocurrir para dicha población, bajo el método dado del muestreo. Tales deducciones pueden hacerse aplicando la ley de las probabilidades matemáticas.

Existen diferentes maneras de escoger muestras al azar. En muchos casos podemos extraer muestras de una población especificada mediante una función continua de densidad de probabilidad, o también de una población infinita mediante una función discreta de densidad de probabilidad. Por ejemplo, si queremos averiguar cuántos pacientes en un hospital psiquiátrico, durante los últimos 10 años de la institución han tenido otras enfermedades psicosomáticas, y el colectivo total de enfermos recluidos es de 2.000, podremos escoger una muestra de 200, tomando los historiales de los pacientes en los últimos 10 años, a intervalos regulares de la lista alfabética, o historiales múltiplos de 10, es decir al azar, lo cual concluye que no existe ninguna relación entre el sistema escogido y las informaciones de los historiales de los pacientes. Sin embargo, si escogiéramos 200 casos del mismo colectivo desde el primer historial hasta el número 200, lo más probable es que ésta no sería una muestra aleatoria, pues puede ocurrir, y esto sucede con mucha frecuencia, por ejemplo, que 160 ó 170 de los casos escogidos no hayan tenido otras enfermedades psicosomáticas, pero si seguimos obteniendo otra muestra de 200, puede ser que en ésta suceda todo lo contrario, es decir que 160 ó 170 pacientes han tenido tales enfermedades. Es por esto que debemos tener mucho cuidado cuando tratemos de escoger una muestra aleatoria para que sea representativa del universo que se quiere investigar.

Si un universo está dividido en estratos con respecto a una característica, nosotros no debemos utilizar el procedimiento anterior que denominaríamos *muestra aleatoria simple*; sino que la muestra debe ser *estratificada o de Poisson* (denominada así por ser Poisson, matemático francés, su autor). Por ejemplo, si queremos investigar acerca de los niños de 10 años que van al colegio o que dejan de ir, dentro de una región determinada, tenemos que pensar que a esos niños podríamos dividirlos en diferentes categorías o estratos: matriculados en colegio o no matriculados y estos estratos, a su vez, en diferentes substratos o subcategorías, como por ejemplo, varones o niñas, clase media, alta o baja, etc.; es decir, que dentro de cada estrato deberíamos tomar una muestra al azar, y si se cumplen todos los requisitos de este tipo de muestra, cada estrato debe formar un subuniverso infinito. Este se considera así cuando la serie definida de condi-

ciones actúe teóricamente para producir acontecimientos sin límite. Por ejemplo, podríamos decir que los niños que no asistían a tomar clases era debido fundamentalmente a que procedían de clase muy baja, con padres separados, con un número excesivo de hermanos, etc. Ahora bien, si tomamos una muestra aleatoria simple de uno de los estratos que componen el universo que se desea investigar y se concluye por ella lo que será el universo total, esa muestra recibe el nombre de muestra de Lexis (nombre del matemático alemán). El error de muestra estratificada o de Poisson es menor que el de una muestra aleatoria simple y el error de la *muestra de Lexis* es mayor. Por otra parte, la muestra estratificada es la más representativa en estudios de índole psicológica, sociológica, política o económica. Por último, otro tipo de muestra es la llamada muestra sistemática o controlada, derivada de la de Poisson. Ésta consiste en dividir la población o colectivo en varias categorías o estratos y tomar de cada uno de ellos, al azar, determinado número de sujetos dentro de un plan preconcebido.

Al mismo tiempo debemos anotar que los métodos de análisis no siempre son iguales en cada tipo de muestreo y de variables utilizadas, por lo que se debe tener mucho cuidado en el uso de los métodos más adecuados, pues fallas al respecto pueden conducir a interpretaciones falsas. Muchas veces aparecen algunas diferencias entre los resultados, cuando repetimos la investigación; esto constituye los *errores muestrales*. Por ejemplo, supongamos que hemos realizado un estudio estadístico para averiguar el grado de habilidad mecánica en una muestra de la población y que posteriormente hemos repetido el mismo estudio en otros grupos de la población. Los resultados obtenidos entre los dos grupos son diferentes, en el sentido de que los datos estadísticos obtenidos (media, desviación estándar, coeficientes de correlación, etc.) han variado; esto constituye la llamada *variabilidad de las muestras*, y podremos preguntarnos ¿hasta qué punto nuestros resultados han sido válidos? Estas posibles diferencias son los errores muestrales. Es decir, el hecho de no poder trabajar con todo el universo nos induce a tener que contentarnos con estudiar muestras de él. Esto, como es natural, nos puede llevar a cometer errores muestrales que podemos detectar y, una vez descubiertos, llegar a conclusiones más reales.

Naturalmente, los resultados obtenidos por las muestras son en realidad estimativos del valor verdadero que podría lograrse si hubiera posibilidad de estudiar el universo; pero, como decíamos con anterioridad, es prácticamente imposible en la mayoría de los casos. De ahí que tengamos que utilizar las muestras y estar sujetos a los errores que éstas pueden conllevar. Las constantes estadísticas, como la media, desviación típica estándar, etc., están sujetas a variaciones alrededor del valor verdadero. La teoría general del muestreo se basa en parte, en el procedimiento para detectar o medir los errores, en el sentido del conocimiento del error estándar. Éste lo podemos definir como la desviación estándar de la distribución de los errores del muestreo o también como la desviación estándar de una distribución muestral de un estadístico o población.

PUNTUACIONES ESTÁNDAR Y LA CURVA NORMAL

Las ciencias sociales y de la conducta utilizan generalmente puntuaciones normalizadas para expresar la magnitud de criterio externo que pretenden medir. Una puntuación estándar o normalizada se define como el resultado que se obtiene al dividir una desviación con respecto a la media por la desviación estándar. Se representa con el símbolo Z, y su fórmula es:

$$Z = \frac{X - \bar{X}}{s}$$

X = Puntuación proveniente de una distribución normal.
$\bar{X}$ = Media aritmética correspondiente a la distribución de donde se extrajo la puntuación X.
s = Desviación estándar correspondiente a la distribución de donde se extrajo la puntuación X.

El uso de las puntuaciones normalizadas es particularmente importante para obtener comparaciones de observaciones encontradas por diferentes procedimientos. Esto es posible, debido al hecho de que las puntuaciones normalizadas utilizan como unidad de medida la desviación estándar. La puntuación Z, por tanto, representa el número de desviaciones estándar que una puntuación X muestra en relación a la medida.

Efectuar comparaciones con puntuaciones brutas no nos permite interpretar una puntuación dada con las demás puntuaciones de la distribución. Supongamos que tres estudiantes han tenido en un examen de biología las puntuaciones brutas de 8, 11 y 15, ¿qué significan estas puntuaciones en relación con las obtenidas por el resto de estudiantes en dicho examen?; ¿qué significan dichas puntuaciones entre sí?; ¿qué relación tiene el 8 con el 11? ¿o el 8 con el 15? Este tipo de preguntas pueden ser contestadas convirtiendo las puntuaciones obtenidas en el examen de biología a puntuaciones normalizadas. Dicho procedimiento será ampliado más adelante al referirnos a la escala Z. Por otra parte, debido a que las puntuaciones normalizadas tienen unidades de medida iguales y su amplitud es la misma en una u otra distribución, se utilizan como técnica indispensable para la interpretación de los resultados de pruebas psicológicas o pedagógicas.

Las puntuaciones normalizadas adquieren mayor significado cuando comprendemos su relación con la distribución o curva normal. Ésta, denominada también curva de Gauss (figura 14.2) tiene las siguientes propiedades:

1. La curva es simétrica. La media aritmética, mediana y modo coinciden en la mitad de la curva.
2. La curva es asintótica en relación al eje de la abscisa. Esto indica que las colas de la curva nunca llegan a tocar el eje horizontal y se extienden desde el infinito negativo, hasta el infinito positivo.
3. La ordenada máxima de la curva se ubica en la media, donde la unidad de la curva normal es igual a 0.3989 y z = 0.
4. A partir de los puntos donde se ubican ±1 desviaciones estándar (encima o debajo de la media) la curva cambia en relación al eje de las abscisas de convexa a cóncava.
5. Entre ±1 desviación estándar cubren el 68.26 por ciento del área de la curva.

La mayor ventaja de transformar puntuaciones brutas a puntuaciones normalizadas, es que con las primeras tendríamos un número infinito de distribuciones normales con diferentes medias y desviaciones estándar, mientras que con puntuaciones normalizadas podemos relacionar todas las distribuciones normales a una distribución de frecuencia relativa. En esta forma, cuando la curva normal es utilizada como referencia, a través de los datos normalizados, recibe el nombre de distribución normal estándar. Esta distribución tiene los parámetros $\mu = 0$ y $s_z = 1$. Las principales puntuaciones estándar expresadas en escalas son: la Z, T, percentilar, Stanine, Wechsler, CEEB, ACGT, GRE, Sten, etc. En el fondo todas ellas se basan en la Escala Z y pueden observar su relación con la curva normal en la figura 14.2.

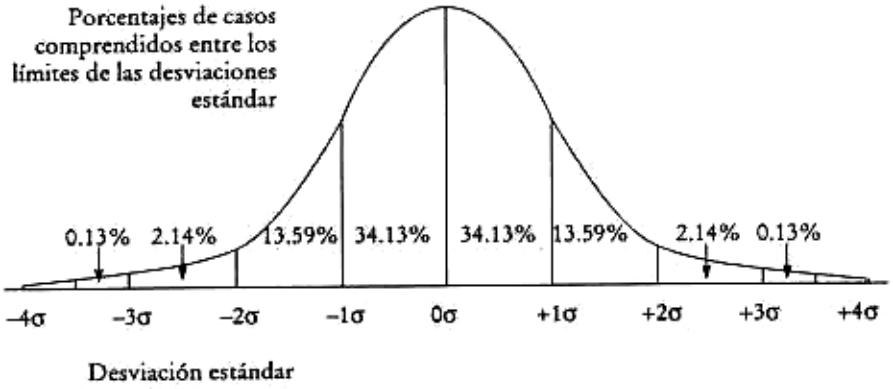

Figura 14.2. Curva de la probabilidad normal o de Gauss.

La escala Z tiene un conjunto de puntuaciones estándar que abarca generalmente cuatro unidades a cada lado de la media, la mitad positiva y la otra mitad negativa. Adopta como unidad la desviación típica o estándar obtenida de las puntuaciones brutas o directas. Debido a que una puntuación normalizada tiene como características que la desviación estándar de una distribución no se altera por la sustracción de una constante y que la variabilidad de un grupo de puntuaciones determina la interpretación de la posición relativa, la distribución de puntuaciones en la escala Z tiene una media 0 y una desviación estándar de sx. La formula fue dada previamente. A través de las puntuaciones o valores Z es posible resolver diferentes problemas de probabilidad y características específicas de un dato que al estandarizarse puede compararse con otros datos normalizados. Tres son los principales tipos de problemas que pueden resolverse en probabilidades con la aplicación de unidades Z:

1. dado un dato X es posible obtener la probabilidad de escoger aleatoriamente un valor que esté por encima o por debajo de dicha puntuación X, siempre y cuando conozcamos los valores de la media y de la desviación estándar;
2. dados dos valores X se puede hallar la probabilidad de escoger aleatoriamente un valor que esté comprendido entre los dos anteriores;
3. se puede obtener un valor X con respecto al cual existe una probabilidad dada de que un valor seleccionado aleatoriamente esté por debajo de él.

De estos problemas básicos se derivan una buena parte de las aplicaciones de los valores normalizados Z. La transformación de puntuaciones directas o valores brutos en unidades Z permiten comparar casos entre si, grupos de casos, muestras e incluso poblaciones.

Estrategias estadísticas en diseño experimental

VARIABLES

La experimentación envuelve la comparación de diferentes grupos o tratamientos entre si. Esto determina la creación convencional de dos grupos como mínimo, uno denominado *Grupo Experimental* y el otro *Grupo Control.* El grupo o grupos experimentales son

expuestos a la influencia de los tratamientos (T ó X) o factores considerados, mientras que al grupo control no se le expone a la influencia de dichos factores. La estrategia experimental está en determinar las diferencias o similitudes, cambios, etc., que aparecen en el grupo experimental al compararse con el de control. Dentro de esta estrategia, el experimentador puede o no manipular ciertos elementos. Estos elementos son las variables. Podemos definir *variable* como una característica que toma diferentes valores en las distintas unidades observadas. En todo aquello que cambia, apareciendo como cantidad, cualidad, atributo, característica, etc.

Si la experimentación envuelve comparación entre grupos: experimental contra control. Clase A contra Clase B, antes contra después, sexo masculino contra sexo femenino, etc., la primera variable que se ha de considerar es aquella que representa la dimensión básica para clasificar los grupos. La variable que define el cambio de grupo a grupo es la variable *independiente*, también denominada *experimental*. Esta variable experimental es la que manipula el investigador, generando cambios sistemáticos para que el efecto de este cambio se pueda observar. También se puede considerar la variable independiente como aquella que clasifica las unidades observables. La variable que es observada con el fin de ver qué le sucede como producto de la manipulación, es denominada variable *dependiente u opinática*, ya que representa el criterio por el cual determinamos si la variable experimental tenía algún efecto. En otras palabras, éstas son las variables que se espera que cambien y que reflejan los efectos en la investigación.

Supongamos que planeamos un experimento para estudiar el efecto del refuerzo positivo sobre la ejecución de una unidad de psicología en estudiantes universitarios. En este experimento, la variable manipulada por el experimentador sería el refuerzo positivo para observar los efectos de esta manipulación sobre la ejecución de psicología. Por tanto, la variable experimental sería el refuerzo positivo, y la ejecución en la unidad de psicología, la variable dependiente. La variable experimental, sin embargo, puede tomar dos formas: la primera, denominada de *tratamiento* que viene siendo la misma variable independiente que el investigador manipula y la segunda, variable *orgánica u organicista*, que es aquella que caracteriza la forma en la cual un grupo particular de organismos cambia. Es por tanto, una variable experimental que representa una clasificación que existe. Edad, sexo, raza, peso, estatura, etc., son ejemplos claros de variables orgánicas. Si deseásemos determinar que los niños y niñas difieren en cuanto a la interpretación del concepto de respeto, la variable experimental orgánica sería el «sexo», ya que es la característica que diferencia a los grupos que se desean comparar y «respeto» sería la variable dependiente como consecuencia de la división del grupo. Es decir, que en este caso, el experimentador no manipuló la variable sino que utilizó dos tipos de grupos que ya existían. Quizá, podríamos decir que la manipulación fue la decisión del experimentador de utilizar dichos grupos y no otros.

Si regresamos al ejemplo del refuerzo positivo sobre la ejecución en el estudio de una unidad de psicología, podría suceder que al imponer dicho refuerzo se generase ansiedad. Este comportamiento vendría a constituir otra variable que se denomina interventora y que aparecen en el proceso del experimento o mediante la utilización de un diseño inadecuado. También podría suceder que aparezcan variables extrañas al experimento en si. Estas se caracterizan por confundir los defectos o resultados y las más importantes son las señaladas en las fuentes de invalidez interna y externa, ampliamente divulgadas por Campbell y Stanley. Estas fuentes son:

1. *Validez interna* de un experimento: Determina el control adecuado sobre las variables externas, selección, control, medición, análisis y procedimientos del experimento. Sus

fuentes de invalidez serían la historia, maduración, comprobación o proceso de aplicación de pruebas, instrumentación, regresión estadística, selección, mortalidad experimental de los sujetos bajo estudio, interacción de variables, contaminación experimental del investigador, inestabilidad de las medidas o errores I y II, y reactividad del sujeto al experimento.

2. *Validez externa* de un experimento: Se refiere a la extensión y aprovechamiento de las generalizaciones de los resultados del experimento a la población de donde se han extraído las muestras. La fuentes de invalidez externa son la propia reactividad o efecto de Hawthorne, los efectos de interacción entre el tratamiento y el proceso de aplicación del test, la interacción entre la selección de la muestra y el tratamiento experimental, las posibles interferencias cuando existen múltiples tratamientos en un mismo experimento, la repetición irrelevante de tratamientos y medidas, y la propia artificialidad del experimento.

Finalmente, existen dos tipos de variables de medición numérica utilizadas en las escalas nominal, ordinal, de intervalo y de proporción. Estas son: a) la variable discreta que es la que toma valores determinados en el rango de medición y la variable continua que puede tomar cualquier valor; la estatura, por ejemplo, es una variable continua por cuanto toma cualquier valor en metros, centímetros, milímetros, etc.; y b) el número de errores en una prueba objetiva es una variable discreta, ya que solamente un número de errores puede aparecer y las medidas podrían integrarse sin valores intermedios.

Los ejemplos señalados permiten identificar variables únicamente en dichos ejemplos en particular, ya que una característica puede ser en un ejemplo, una variable independiente y en otro, una dependiente. En el caso del refuerzo positivo a que nos referíamos anteriormente, el tiempo empleado en la ejecución puede ser una variable independiente o experimental, ya que se desea determinar su efecto sobre el nivel de ejecución, que sería la variable dependiente. Sin embargo, si se desea observar el efecto de los refuerzos positivos sobre el tiempo empleado en la ejecución, el tiempo pasa a constituirse en una variable dependiente y la manipulación de los refuerzos en variable experimental.

ESTRATEGIAS ESTADÍSTICAS EXPERIMENTALES

Los diseños experimentales tienen como objeto controlar los variados factores que pueden influenciar el experimento y, posiblemente, afectar los resultados en los cuales se basan las conclusiones. Estas no provienen de la prueba de las hipótesis, sino de la estructura del experimento y de la naturaleza de los controles expuestos. El control de las variables externas puede ser aumentado cambiando o eliminando la variable. Sin embargo, las estrategias más importantes que se deben utilizar son: repetición, aleatorización y control experimental.

Repetición o replicabilidad significa repetir algún tratamiento de más de una unidad experimental para obtener un estimado del error experimental al cual están sujetas las comparaciones. En estudios simples, la intravariación es tan pequeña, comparada con la variación de tratamientos intermedios, que el efecto es una fluctuación real y no de tipo casual. En cierto modo, el concepto estadístico de repetición viene a ubicarse entre los términos de duplicación y repetición. Cuando asignamos aleatoriamente sujetos a grupos o tratamientos a grupos estamos en un proceso de *aleatorización*. Este proceso permite suponer que cada miembro de la población tiene la oportunidad de pertenecer a la muestra de dicha población y permite a su vez, que las variables extrañas no introduzcan tendencias que distorsionen los efectos de los

tratamientos. Por este motivo, la aleatorización o azarización es la técnica más importante y efectiva en la igualación de grupos y en el control de las variables externas.

Sin embargo, la igualdad de las condiciones experimentales representadas por los diferentes grupos, siempre debe reconocerse como aproximada, independientemente del rigor y cuidado con que se haya tomado la muestra. Si el tamaño de la muestra es suficientemente amplio, la aleatorización logra el control adecuado por cuanto los errores se compensan. Y si aún las condiciones no se compensan en una muestra particular, el investigador, al menos, se asegura de que la desigualdad no es el resultado de estimados tendenciosos, entendiendo por situación no tendenciosa, cuando en todas las muestras posibles, el promedio de los estimados no se desvía del valor de la población. El proceso de muestreo es parte de los pre-requisitos para entender la estadística y el mismo puede ser estudiado en las referencias que aparecen en nuestra bibliografía.

El *control experimental* presupone realizar tratamientos bajo comparación, tan similares como sea posible respecto a condiciones, que no sean factores sobre los cuales se busque información. El control experimental incluye los procedimientos de apareamiento, bloqueo y control estadístico. El primero de ellos, implica aparear los casos mediante la selección de pares de sujetos con idénticas características y asignando un sujeto de cada par al grupo de control, y el otro, al grupo experimental. Sin embargo, el apareamiento ha sido ampliamente discutido por numerosos científicos de la estadística, sobre todo, cuando se utiliza como sustituto de la aleatorización. En este caso, es recomendable asignar los sujetos mediante el apareamiento y ubicarlos aleatoriamente, uno al grupo experimental y el otro al de control. Este uso de una variable de clasificación recibe el nombre de bloqueo.

El *control estadístico* se refiere a la posibilidad de hacer ajustes después del experimento. La técnica más importante es el análisis de covarianza, que veremos más adelante, y que está derivado del análisis de varianza (ANOVA), mediante la utilización de las medidas de las puntuaciones de un pretest como covariados. El análisis de covarianza es una técnica paramétrica que requiere como medición una escala de intervalo pero que, a diferencia de ANOVA, la proporción *F* no prueba las diferencias observadas entre las medias, sino que observa esas diferencias como objeto del ajuste entre las medias sobre la base del covariado. Por tanto, podemos definir un covariado como la medida utilizada en el análisis de covarianza para ajustar las puntuaciones de la variable dependiente. Diversos estudios han concluido, sin embargo, que el análisis de covarianza no es un procedimiento recomendable en todas las ocasiones, y proponen a cambio la utilización de la correlación cuando sea necesario. Veamos entonces, los conceptos de los procedimientos más importantes en estadística y diseño experimental, como son el análisis de varianza (ANOVA) y covarianza (ANCOVA), la correlación, regresión y medidas de asociación. Posteriormente se hará un breve repaso a los diseños multivariados y a las medidas no paramétricas.

Análisis de varianza y de covarianza

El análisis de varianza se debe a R.A. Fisher y universalmente se conoce como ANOVA que son las primeras letras de análisis de varianza en inglés (Analysis of Variance). Algunos autores en español utilizan las siglas de AVAR, pero nosotros preferimos usar la denominación inglesa que es la más conocida. El primer tipo de ANOVA es *unidireccional* o de una vía y representa la forma de manejar las observaciones del diseño completamente aleatorio que estudiaremos más adelante. Por otra parte, el ANOVA es la generalización de una prueba *t* para dos grupos independientes, pero mientras la *t* es apropiada

para dos niveles solamente, el análisis de varianza se utiliza en dos o más niveles. Por tanto, la prueba *t* ha dejado de ser relevante, ya que el análisis unidireccional de varianza es apropiado donde quiera que la prueba *t* lo es. Sin embargo, se hará referencia a esta prueba cuando se vea las medidas de asociación En general, podemos decir que ANOVA estudia la significación de determinada clasificación de una variable, las variables que intervienen o son causa significativa de un efecto específico, y las variables que al actuar conjuntamente son causa del efecto.

ANÁLISIS DE VARIANZA DE UNA VÍA

Supongamos que tomamos cuatro teorías de aprendizaje y diseñamos en función de ellas, cuatro métodos de instrucción de la estadística para probarlas en un grupo de estudiantes. El diseño más simple sería dividir los estudiantes al azar en cuatro grupos iguales y probar cada método (tratamiento) en uno de los grupos. Posteriormente, se administraría una prueba objetiva común de aprovechamiento en estadística y se obtendrían las puntuaciones (variable dependiente) de los distintos grupos comparados.

Obviamente, nadie puede esperar que todos los alumnos que han sido estudiados por un método particular obtengan puntuaciones idénticas. Más bien, se esperaría una especie de distribución de frecuencias de puntuaciones que pueden tener un promedio de $\bar{X}_1$ y una desviación normal de s normal de s_{x1}. Similarmente, cada uno de los otros grupos de tratamiento también producirían distribuciones de frecuencias, cada uno con su propio promedio y desviación normal. En general, entre menos diferencia haya en la eficacia de los métodos, mayor será el traslapo de las distribuciones de frecuencias resultantes, y entre menor sea el traslapo, mayor será la diferencia entre los distintos tratamientos. Es importante observar que la diferencia del traslapo depende tanto de las diferencias entre las medidas (entre la variación) como de la variabilidad dentro de los distintos grupos (dentro de la variación). Básicamente, la información que indica al investigador si su tratamiento ha sido efectivo es la variación entre las medias. Por tanto, el índice que nos interesa es la varianza de medias. Además, es necesario preguntarse cómo sería la varianza de medias si el tratamiento es efectivo comparado con lo que sería dicha varianza si el tratamiento no fuera efectivo.

El problema lógico en el análisis de varianza es determinar qué cantidad de la variación total observada es atribuible al tratamiento dado en el experimento. Ya que éste es un trabajo muy complejo, el investigador pregunta, más simplemente, ¿existe evidencia de que el tratamiento tenga algún efecto? Para contestar esta pregunta, se debe tener en cuenta que el efecto del tratamiento se manifestará como variación entre las medias de los grupos de tratamiento o de la variable independiente. Además, el investigador observa tres aspectos:

a) cuanto más grande sea el efecto del tratamiento, mayor será la variación de medias;
b) aunque no haya ningún efecto, habrá alguna variación entre las medidas observadas;
c) la cantidad de variación entre medias cuando el tratamiento no es efectivo es igual a la que ocurriría si se tomaran distintas muestras aleatorias de la misma población.

Por tanto, la proporción de la varianza observada entre los promedios de la variación muestral estimada en medias, nos indicaría si el tratamiento ha sido efectivo. La presentación convencional (tabla 14.1) de un análisis de varianza incluye los términos de fuente (de variación), la suma de cuadrados, los grados de libertad, la media de los cuadrados y la proporción F. Existen por tanto, dos formas de variación. La primera, denominada *entre-grupos o entre-tratamientos* que se refiera a la variación de un grupo de medias en rela-

ción a la media general o total. La segunda variación se llama *intragrupos o dentro de los grupos* y se refiere a la variabilidad de la media o promedio de las puntuaciones dentro de cada grupo, constituyendo así el término de error. La variación de todas las puntuaciones juntas se llama *varianza total.*

El método de Fisher interpreta los índices de varianza por medio de la *z*, habiendo elaborado dos tablas de significación, para los niveles del 1% y 5%. Ahora bien, para calcular la *z* es necesario determinar la diferencia entre el logaritmo neperiano de la mayor varianza y el logaritmo neperiano de la menor.

Sin embargo, posteriormente, se desarrolló el denominado método *F* de Snedecor, que elimina el trabajo con logaritmos por medio del índice *F*, que se representa mediante el cociente de la varianza entregrupos y la varianza intragrupos consideradas. Este índice se contrasta con los niveles de significación que se encuentran en cualquier libro básico de estadística.

Tabla 14.1. *Presentación convencional del análisis de varianza.*

Fuente	Suma de cuadrados	gl	MC	F
Entre tratamientos	$SC_{entre} = \frac{\sum_j \left(\sum_i X_{ij}\right)^2}{n} - \frac{\left(\sum_j \sum_i X_{ij}\right)^2}{nC}$	C − 1	$\frac{SC_{entre}}{C-1} = MC_{entre}$	$\frac{MC_{entre}}{MC_{dentro}}$
Dentro de tratamientos	$SC_{dentro} = \sum_j \left(\sum_i X_{ij}^2\right) - \frac{\sum_j \left(\sum_i X_{ij}\right)^2}{n}$	Cn − C	$\frac{SC_{dentro}}{Cn-C} = MC_{dentro}$	
Total	$SC_t = \sum_j \left(\sum_i X_{ij}^2\right) - \frac{\left(\sum_j \sum_i X_{ij}\right)^2}{nC}$	Cn − 1		

Por otra parte, se hace recomendable descartar aleatoriamente los datos para obtener frecuencias iguales, siempre que tengan muestras razonablemente grandes. La razón de esto es que el análisis de varianza es muy potente (lo que quiere decir que se pueden violar varias suposiciones de validez sin que tenga ningún efecto importante en los resultados) si cada grupo de tratamiento tiene números iguales. De la misma manera, si las muestras del investigador fueran tan pequeñas que sacando algunas para obtener en cada grupo números iguales le afectara seriamente con relación a los grados de libertad, entonces probablemente se usaría alguna técnica estadística no paramétrica en vez del clásico análisis de varianza.

Sin embargo, el análisis de varianza no es tan potente con relación a todas las suposiciones posibles, aun con números iguales, por lo que es apropiado conocer cuáles son las suposiciones del análisis de varianza y qué hacer con ellas. La primera suposición debe

mantenerse pase lo que pase. Esta es la suposición de *aleatorización*: o sea, que los miembros de los distintos grupos de tratamiento, y los tratamiento en si, se asignen aleatoriamente a los grupos. Esta suposición es absolutamente necesaria para dicha clase de análisis, ya que si no se mantuviere, entonces: *a)* la variación calculada dentro de cada grupo de tratamiento no es un estimado de la varianza de la población total bajo la hipótesis nula; y *b)* cualquier diferencia observada entre los promedios de grupos puede ser el resultado de algo que no son los efectos del tratamiento. Si esta suposición de aleatorización no se puede hacer, entonces debe evitarse el análisis regular de varianza y utilizar algún análisis, tal como el de covarianza, muy útil para «grupos intactos».

La segunda suposición es la de *homogeneidad de varianza* para los distintos grupos. Es decir, se supone que en la población todas las intravarianzas son iguales. Así, el análisis de varianza nuevamente es potente, o sea, las suposiciones pueden violarse, sin cambio en el resultado, respecto a esta suposición. La excepción ocurre cuando: *a)* los promedios de los grupos están correlacionados con las varianzas; y *b)* los tamaños de las muestras varían de un grupo a otro. Por tanto, el procedimiento a seguir podría ser éste:

1.° Si se tienen grupos de tamaños iguales para cada nivel, no es necesario preocuparse por suposiciones de igual varianza.
2.° Si no se pueden lograr grupos iguales, descartando el azar en algunos casos, se debe elaborar una tabla de los promedios obtenidos contra las varianzas logradas en los grupos. Si no existe relación, puede ignorar la suposición, pero si existe una relación tendrá que hacerse la prueba Bartlett y otras pruebas para la homogeneidad de las varianzas que explicamos posteriormente. Si no se rechaza la hipótesis nula (que las varianzas son iguales), se puede continuar con el análisis de varianza, ya sea que la evidencia sugiera que las varianzas de la población intra grupos no son iguales, o ya sea que exista alguna clase de transformación para lograr datos apropiados a fin de utilizarlos en ANOVA o finalmente, se utilizaría una técnica no paramétrica.

Una tercera suposición es que la *distribución* de la población sea *normal*. Esta puede ignorarse excepto bajo condiciones más extremas (un sesgo muy obvio de la curva y muestras muy pequeñas), lo cual puede ocurrir si la variable dependiente es una medida con un efecto máximo o mínimo y si se ha empleado algún grupo externo de una población. Además, si el investigador encuentra una distribución muy sesgada, se dispone de transformaciones que corregirán la forma de la distribución o puede volverse a una de las pruebas no paramétricas, tal como el análisis de varianza unidireccional de Kruskal-Wallis para grupos independientes o el análisis de varianza de dos vías de Friedman para muestras relacionadas.

Finalmente, existe un nuevo principio o suposición, que si bien es indispensable en el análisis de covarianza, también es importante para ANOVA. Este se llama *homogeneidad de regresión*. Por ejemplo, en tres tratamientos es posible elaborar una ecuación de predicción para y en términos de x dentro de cada uno de los tres grupos de tratamiento. La suposición es, entonces, que dichas tres ecuaciones tienen la misma inclinación pendiente. Existen cuatro modelos principales de análisis de varianza que detallaremos en resumen:

— *Modelo I:* Este modelo se supone cuando el investigador se interesa únicamente en los niveles *a* del factor *a*, en los niveles *b* del factor *b* y en los niveles *c* del factor *c*, presentes en el experimento.

— *Modelo II:* Este modelo se supone cuando el investigador se interesa en una población de niveles del factor *a* de la cual únicamente está presente en el experimento una muestra al azar (los niveles *a*); también cuando el investigador se interesa en una población de niveles del factor *b* de la cual únicamente está presente en el experimento una muestra al azar (los niveles *b*) y por último, cuando se interesa en una población de niveles del factor *c* en la cual están únicamente presentes al azar los niveles *c*.
— *Modelo IIIa:* Este se supone cuando el investigador se interesa en *a*) únicamente los niveles *a* del factor *a* que están presentes en el experimento; *b*) únicamente los niveles *b* del factor *b* en el experimento, y por último *c*) una población de niveles del factor *c* de la cual únicamente está presente en el experimento una muestra al azar (los niveles *c*).
— *Modelo IIIb:* Este modelo se supone cuando el investigador está interesado en *a*) únicamente en los factores *a* del factor *a* que están presentes en el experimento; *b*) una población de niveles del factor *b* de la cual únicamente está presente en él una muestra al azar (los niveles *b* y *c*) una población de niveles del factor *c* de la cual únicamente está presente en el experimento una muestra al azar (los niveles *c*).

En resumen, el modelo *I* se refiere a los efectos fijos, el modelo *II* a los efectos aleatorios y el modelo *IIIa*, a efectos fijos y aleatorios y el modelo *IIIb* a los efectos fijos y al azar.

ANÁLISIS DE VARIANZA MULTIDIRECCIONAL

Ahora que hemos conocido el tipo de análisis que se utiliza cuando sólo hay una manera de clasificar los grupos que se comparan, examinemos la situación en la cual una segunda variable independiente se va a considerar en el experimento. Este otro factor puede ser una variable interventora que deseamos controlar o puede ser una segunda variable de tratamiento cuyos efectos queremos estudiar. En el primer caso, nos referiríamos al estudio como un diseño de *bloque aleatorio* y, en el último caso, lo denominaríamos *diseño factorial.*

En ambos casos, el tipo de análisis que se hace se puede llamar análisis de varianza de dos vías, clasificación doble o multidireccional. La única diferencia práctica entre un bloque aleatorio y un diseño de dos factores consiste en la manera de interpretar el análisis. Este último punto se examinará después de haber estudiado las bases lógicas para el ANOVA multidireccional.

Por supuesto, esto no tiene límite en teoría o computación respecto a cuántas variables diferentes se pueden analizar de una vez; existen las limitaciones prácticas de obtener suficientes sujetos en tantos grupos —especialmente si tenemos muchos niveles diferentes aún con pocas variables— y de hacer una interpretación comprensiva de los resultados de las interacciones, aun cuando tuviéramos pocos niveles para cada variable. Por tanto, raramente los experimentadores usan más de tres variables experimentales a un mismo tiempo. De esta forma, la discusión para los conceptos estadísticos introductorios de este libro de psicología general, se limitará únicamente a dos variables experimentales, ya que esto ilustrará los principios incluidos.

Como ejemplo de un problema de clasificación doble, supongamos que estamos interesados en determinar cuál de los libros de texto puede ser el más efectivo en un curso de Introducción a la Psicología. Primero, definiríamos «efectivo» como una puntuación alta en el examen final que se da en el curso y usaríamos dicha puntuación como la variable dependiente o de criterio. Por tanto, una de las variables experimentales es el texto.

Supongamos también que esta variable independiente (en este caso, tratamiento) la distribuimos en cuatro clasificaciones o categorías de tratamientos en la siguiente forma: *a)* texto actual; *b)* un nuevo texto de enseñanza programada; *c)* un texto para instrucción tutorial; y *d)* un texto nuevo y diferente del actual. Por otra parte, podría ser que la efectividad de un texto fuese debida a una segunda variable, por ejemplo, el tipo de conocimientos previos en psicología que tienen los estudiantes. En esta forma, antes de que se pueda dar una respuesta satisfactoria con relación al libro de texto, se debe clasificar a los grupos en categorías, en esta segunda variable. Supongamos que se utilizan tres: *a)* los que tomaron un curso previo de psicología; *b)* los que tomaron un curso formal en psicología elemental diferentes al anterior; y *c)* los que estudiaron el material elemental de psicología por su cuenta y pasaron un examen por el conocimiento requerido.

Entre tanto, se puede establecer un ANOVA de doble clasificación representada por una tabla de clasificación cruzada, indicando casilla o celda, columna, hilera y promedios generales. Esto nos permite aseverar que cada puntuación de un sujeto, es una función de cinco aspectos: *a*) el promedio general; *b*) el texto que se utilizó en el curso; *c*) el tipo de conocimiento previo en psicología que ha tenido; *d*) el texto por la interacción del conocimiento previo, y *e*) la propia singularidad del sujeto. Sin entrar en análisis matemático, sino más bien razonando por analogía del caso unidireccional, ANOVA se puede trasponer el promedio general y luego dividir las sumas totales de los cuadrados. Así, diríamos (tabla 14.2) que las sumas totales de los cuadrados son iguales a las sumas de cuadrados «intra», más la suma de cuadrados entre columnas, más la suma de cuadrados entre hileras, más la suma de cuadrados para la interacción (ajustado entre celdas).

TABLA 14.2. *Caso del análisis de varianza multivariado.*

	Texto (i)				
Conocimiento previo (k)	*Actual*	*Programado*	*Tutorial*	*Nuevo*	*Medias de las hileras*
Curso previo	$\overline{X}_{\cdot 11}$	$\overline{X}_{\cdot 21}$	$\overline{X}_{\cdot 31}$	$\overline{X}_{\cdot 41}$	$\overline{X}_{\cdot\cdot 1}$
Otro curso formal	$\overline{X}_{\cdot 12}$	$\overline{X}_{\cdot 22}$	$\overline{X}_{\cdot 32}$	$\overline{X}_{\cdot 42}$	$\overline{X}_{\cdot\cdot 2}$
Autoestudio y examen	$\overline{X}_{\cdot 13}$	$\overline{X}_{\cdot 23}$	$\overline{X}_{\cdot 33}$	$\overline{X}_{\cdot 43}$	$\overline{X}_{\cdot\cdot 3}$
Medias de las columnas	$\overline{X}_{\cdot 1\cdot}$	$\overline{X}_{\cdot 2\cdot}$	$\overline{X}_{\cdot 3\cdot}$	$\overline{X}_{\cdot 4\cdot}$	Media General $\overline{X} = \overline{X}_{\cdot\cdot\cdot}$

Para obtener el ANOVA multidireccional dividiríamos la suma de los distintos componentes de cuadrados por su grados correspondientes de libertad para calcular los cuadrados promedios. Finalmente, tomando las proporciones apropiadas para estos cuadrados promedios, obtenemos tres pruebas de significación. La primera prueba de «efectos principales» de significación comparará los promedios de hileras y contestará a nuestra pregunta del ejemplo así: ¿La actuación de una persona en el examen final de Introducción a la Psicología depende de su tipo de conocimiento previo en psicología sin tomar en cuenta el texto utili-

zado? La segunda prueba de «efectos principales»comparará los promedios de columnas y contestará la pregunta en nuestro ejemplo: ¿La actuación de uno en el examen final depende del texto que se use sin tomar en cuenta el conocimiento previo en psicología? La tercera prueba —la de interacción— comparará los promedios de celdas ajustados y contestará la pregunta: ¿El tipo de efecto que el texto tiene en la actuación del examen final depende del conocimiento previo que una persona tenga en psicología?

La proporción adecuada para emplear en la prueba de efectos principales, dependerá de si el efecto particular bajo estudio es un efecto variable *fijo* o un efecto *aleatorio* variable. Por efecto fijo queremos decir que hemos incluido, en el experimento, toda la población de niveles de esta variable en que nos interesamos (es decir, sobre la cual sacaremos deducciones). Este es el caso usual. Por ejemplo, en la ilustración de la clase, los únicos textos que interesa estudiar eran los incluidos en el experimento. Además, el único antecedente que consideraba, eran los tres grupos incluidos en el experimento. Cuando ambas (o todas) las variables experimentales se consideran fijas, tenemos lo que llamamos un modelo de efectos fijos. Este es el modelo clásico que generalmente se describe en los textos.

Por efectos aleatorios, queremos decir que, en nuestro experimento, sólo tenemos un ejemplo aleatorio de todos los posibles niveles de la variable experimental, en la cual estamos realmente interesados. Por ejemplo, algunas veces queremos comparar varios métodos de enseñanza, pero creemos que el método puede depender del profesor particular que lo usa. Por tanto, podemos establecer un experimento factorial con el método como un factor y el profesor como otro factor. El método es un efecto *fijo*, pero el profesor es un efecto *aleatorio* porque los profesores incluidos en el estudio son una muestra de todos los profesores que pueden estar usando los distintos métodos. En este último ejemplo, tenemos lo que llamamos *modelo mixto* debido a que una (algunas) variable era fija y la otra (o por lo menos una más) variable era aleatoria. Si ambas (o todas) variables fueran aleatorias, entonces tendríamos lo que conocemos como modelo de *efectos aleatorios*. Un ejemplo de este último sería un experimento de dos factores, siendo uno de ellos la cantidad de tareas y, el otro, los niveles de grados; estudio en donde incluiríamos solamente una muestra de todos los niveles de grados de los cuales queremos hacer una deducción y únicamente una muestra de todos los niveles de cantidades de tareas de los cuales queremos, igualmente, hacer una deducción.

Ahora se ha indicado que la proporción particular de la media de los cuadrados (MC) que ha de utilizarse para probar la hipótesis en estudio, depende de qué modelo tenemos y si la hipótesis que nos interesa es una variable de efecto fijo o efecto aleatorio. La idea básica es que el valor F siempre se construye para que sea la proporción de la media del cuadrado esperado cuando la hipótesis no es verdadera para una media del cuadrado esperado, siendo la hipótesis nula verdadera. Para complementar este concepto, siempre aplicamos la regla general de seleccionar un denominador tal par F que cuando la hipótesis nula es verdadera, la proporción F esperada será igual a uno.

CONCEPTUALIZACIÓN DEL ANÁLISIS DE COVARIANZA

La próxima aplicación especial al estudiar diferencias de grupo, a través de un análisis de varianza, incluye una técnica para controlar la variable molesta o interventora *estadísticamente* en vez de *experimentalmente*, como se hace en diseños experimentales utilizando el bloqueo. El tipo de análisis para lograr este propósito se denomina análisis de covarianza, que generalmente se identifica con las siglas de ANCOVA, COVAR y ANACOVA.

Para entender la lógica del análisis de covarianza es conveniente revisar el modelo

básico del análisis de varianza y observar posteriormente cómo es modificado si no se puede aplicar un control experimental y nos vemos precisados a recurrir a un control estadístico. En la ANOVA de una dirección se empieza con:

$$Y_{ij} = \mu + \tau_j + e_{ij}$$

y se compara la variación entre las medias de los métodos de tratamiento con la variación dentro de los individuos en un grupo de tratamiento, para constatar si en realidad se obtuvo un efecto del tratamiento. En la ANOVA de dos direcciones se nota que puede haber una segunda variable independiente incluida, y cuyo efecto puede expresarse añadiendo algunos términos al modelo básico, para así obtener:

$$Y_{ijk} = \mu + \tau_j + \tau_k + \tau_{jk} + e_{ijk}$$

donde que t representa el efecto de esta nueva variable sola, y τ_{jk} representa el efecto de las dos variables en combinación, fuera de los efectos de cada una por separado, es decir, el efecto de *interacción*. Finalmente, en diseño experimental se observa que si la variable representada por tk era una nueva variable *experimental* tendríamos un diseño factorial; pero si τ_k fuera una variable molesta o interventora que se necesita controlar, entonces nuestro diseño sería del tipo de *bloqueo*. Por lo que, mientras que el experimentador pudo haber clasificado los sujetos antes del experimento y asignado sujetos aleatoriamente en cada grupo de clasificación a los diferentes niveles de la variable experimental (grupos de tratamiento), pudo también haber reducido tanto el término error e_{ijk} y estudiar la importancia de la interacción entre las variables experimentales y de control.

Algunas veces, sin embargo, al intentar hacer cualquier investigación en ambientes educativos o en cualquier otro ambiente, al investigador no le es realmente posible disponer a su antojo de sus alumnos-sujetos como desearía. Debe, más bien tomar todos los grupos *intactos* tal y como sean. Y a veces, estos grupos intactos difieren marcadamente en alguna característica importante. Es decir, una característica altamente relacionada a la variable de criterio. En esta situación el investigador puede disponer de un cierto control sobre la variable molesta importante, si dicha variable puede ser medida y si es tratada como covariante de análisis de covarianza. Por ejemplo, en un intento por ver si tres formas diferentes de presentar la biología para el aprendizaje, producen diferentes niveles en cuanto a progreso, es factible que nos veamos restringidos a emplear clases de biología ya organizadas. Además, si las clases difieren en relación a la inteligencia promedio o con respecto a su conocimiento pasado de la biología, cualquier diferencia de post-tratamiento pudiera ser atribuida a dichas variables incontrolables, más que al tratamiento en si. El análisis de covarianza está diseñado para que aporte una forma de control en situaciones como ésta; y cuando el control es obtenido por análisis de covarianza decimos que tenemos un *control estadístico* y no experimental.

El enfoque de covarianza produce control de dos formas diferentes. Primero, reduce el efecto «dentro», o sea, el error de varianza, al atribuir partes de la «exclusividad» de la actuación del individuo en la variable de criterio a la así llamada *covariante* o variable de control. El análisis de covarianza aporta también una manera para el ajuste de las diferencias observadas en la variable de criterio, sobre los efectos de las diferencias iniciales sobre la variable de control, o sea la covariante. En general, el control experimental (o sea, bloqueo y asignación aleatoria entre niveles) es más preferido que el control estadístico. Existen varias razones para sustentar esta preferencia. Una de las razones por las que se prefiere el control experimental sobre el estadístico es debido a que la reducción del error

es *mayor* con bloqueo que con control de covariante tal como lo indica la siguiente comparación de error inicial, error con bloqueo, y error con control estadístico.

Al examinar los errores de los diferentes tipos de diseños deben notarse dos cosas. Primeramente, el tamaño de la reducción en el término de error depende en gran parte de la magnitud de la correlación entre el criterio y la covariante (o variable de bloqueo). Entonces, por ejemplo, la inteligencia es covariante importante cuando el criterio es logro intelectual mientras que la altura es improbable que lo sea. En segundo lugar, si el experimento tiene grados de libertad grandes en el término de error (es decir, muchas personas en cada grupo de tratamiento) el control estadístico será casi igual al control experimental.

Una segunda razón para preferir el control experimental más que el control estadístico es que es posible obtener más información por bloqueo que por ajuste, a través de medios de covariante. En bloqueo, los *efectos de interacción* entre el tratamiento y la variable de control (por ejemplo, entre el tratamiento y los niveles de inteligencia) puede estudiarse, mientras que con la covarianza el investigador sólo corrige las diferencias de tratamiento ocasionadas por los efectos de no poder controlar una variable. Por tanto, cuando sea factible, se debe preferir el control experimental más que el control estadístico.

Finalmente, ANCOVA es considerado menos deseable que el bloqueo porque requiere un número de suposiciones que son, a veces, difíciles de llenar y que, a la vez, pueden tener un efecto serio en la interpretación de los resultados. No existe ninguna razón, desde luego, para no utilizar tanto el bloqueo como covarianza en una investigación específica. En realidad, es una estrategia de investigación útil, la de colocar en orden todas las variables extrínsecas independientes de acuerdo a su importancia con respecto a su correlación con el criterio y posteriormente, emplear bloqueo con las más importantes, control estadístico con las últimas en la lista, y aleatorización con cantidades grandes para controlar las que no puedan ser manejadas por ninguna de las dos técnicas anteriores.

Estas son las bases para llevar a cabo un ANCOVA. Primeramente tenemos las mismas suposiciones que son necesarias para un análisis de varianza:

a) Absolutamente esencial, en el análisis de varianza, era la suposición de asignación aleatoria a grupos. En el análisis de covarianza esto se convierte en asignaciones aleatorias con respecto a todas las características, a excepción de la covariante. No poder satisfacer esta suposición es muy serio, y podría anular los resultados con respecto a las diferencias iniciales en las características, diferentes a la covariante. Estas podrían, si están casualmente ligadas a la variable de criterio, haber ocasionado una diferencia en las puntuaciones de criterio y por ello haber sido denominados erróneamente efectos de tratamiento. La única forma de asegurarse de que esta suposición ha sido satisfecha se hace mediante un examen cuidadoso de las condiciones que llevaron a las asignaciones de los grupos intactos que se utilizan.

b) Una segunda suposición es la de distribución normal de la variable de criterio en la población. Al igual que el análisis de varianza, la prueba F es sólida en lo que respecta a esta suposición si n iguales aparecen en los grupos de tratamiento (nótese, sin embargo, que la oportunidad de obtener n iguales en el análisis de covarianza no es tan grande como en el análisis de varianza).

c) La tercera suposición es la de homogeneidad de la varianza. También, al igual que el análisis de varianza, esta prueba es sólida si las n son iguales.

En segundo término, además de las suposiciones que se encuentran en el ANOVA, existen otras suposiciones acerca de la covariante misma. La más importante es la de que

la covariante no es afectada por el tratamiento u otra variable experimental. Esto puede satisfacerse obteniendo la medida de la covariante antes de la administración del tratamiento (ésta es la forma más segura) o utilizando una covariante que, lógicamente, no sea afectada por el tratamiento (por ejemplo, la edad). La otra suposición acerca de la covariante misma es que se mide sin error. Esta es importante porque si la fidedignidad de la covariante es baja, entonces el investigador perderá precisión y la ventaja obtenida sobre el ANOVA será pequeña. De hecho, si la fidedignidad de la covariante es tan baja como para resultar en una apreciación inadecuada de la línea de regresión, entonces las cosas podrían resultar peores en vez de mejores al tratar de lograr un control. Estas son:

a) las pendientes de las regresiones de la clase «dentro» son homogéneas (o sea, la pendiente de la regresión Y sobre X es la misma para todos los grupos de tratamiento);
b) la regresión es lineal incluyendo: *1)* regresión de grupo «entre», *2)* regresión de grupo «dentro», *3)* regresión total de grupo.

Finalmente, hay dos puntos adicionales que deben recordarse. El primero se refiere a que el análisis de covarianza puede ser extendido más allá de esta discusión elemental e incluir covariantes múltiple en él, ANCOVAs de doble dirección o superiores y ANCOVAs de medidas repetidas. Segundo, que todas las pruebas a priori, las pruebas post hoc y las interpretaciones de la importancia práctica, contrastadas con la significación estadística, de las diferencias de los grupos resultantes, debe hacerse basándonos en el examen de las medias ajustadas y no en simples diferencias.

Las pruebas de Scheffé y Tukey, denominadas «post hoc», son utilizadas de manera subsiguiente a la F global o a la global y residual, siempre y cuando este índice haya sido significativo. Es decir, supongamos en cualquier análisis de varianza o análisis multivariado de varianza con más de dos niveles, que el investigador obtiene una F significativa. ¿Qué quiere decir esto? Quiere decir que existe una diferencia entre los grupos. La F global no le indica al investigador si todos los grupos son diferentes entre si o si sólo ciertos grupos difieren entre si. Por tanto, un análisis más a fondo se hace necesario para comparar o contrastar cada grupo; estos grupos representan los diferentes métodos o niveles de la variable experimental. Entonces, el enfoque particular que se le dé al análisis que contrasta grupos específicos depende de si el investigador compara pocos grupos a fin de llegar a cuestiones hipotéticas muy específicas, formuladas antes del inicio del experimento, o si el investigador está llevando a cabo el proceso de indagación de los datos para ver qué más se puede obtener. Por tanto, estas pruebas representan un análisis posterior a las pruebas F globales que ya han sido realizadas.

Probablemente el mejor de todos los procedimientos sea la técnica desarrollada por Scheffé. El enfoque de Scheffé puede utilizarse igualmente con grupos de tamaños iguales o desiguales y para todo tipo de comparaciones, además de ser bastante sólido en lo referente a las suposiciones del análisis de varianza, es decir, de la normalidad y de la homogeneidad de las varianzas. Al aplicar el método de Scheffé en vez de probar cada diferencia que ha de ser indagada, primero determina el tamaño de la diferencia que tendría que ocurrir si la diferencia fuera significativa. A continuación se compara esta diferencia con las diferencias observadas; si la diferencia observada es igual o mayor que la diferencia permitida, entonces es significativa, si no, no es significativa. Tukey ha llamado a esta diferencia que se requiere para la significación como «concesión», ya que ésta le indica al investigador cuál es el margen de diferencia admitido por el azar. Tukey también desarrolló una técnica general para la indagación de datos denominada «T protegida». Dicha técnica es algo mejor que el enfoque de Scheffé bajo las siguientes restricciones: *a*) las «n»

son iguales para todos los grupos; *b*) es posible llenar los requisitos de la suposición de homogeneidad de la varianza; *c*) el investigador sólo está interesado en todos los pares posibles de comparación de grupo.

Correlación y regresión

CORRELACIÓN

La correlación es una medida que estudia los cambios sucesivos de dos variables. No se trata de una medida de relación causalista, sino de una relación asociativa. El método de correlación es quizá el más importante en la investigación psicológica y educativa. Por ejemplo, se ha comprobado que a medida que aumenta el urbanismo, aumentan las enfermedades psíquicas: esto es correlación entre una cosa y otra. Como veremos a continuación, se puede medir la magnitud de la correlación y clasificar su resultado.

La correlación se puede clasificar en *positiva*, si al aumentar una variable tiende a aumentar la otra, y *negativa*, si al aumentar una de las variables tiende a disminuir la otra. Las correlaciones son altas y positivas; bajas y positivas; cercanas a cero; iguales a cero; altas y negativas, y bajas y negativas. El *coeficiente de correlación* es un valor numérico que indica si hay concomitancia entre dos variables, y cuál es el grado de concomitancia o relatividad. Se representa por la letra r (tabla 14.3).

TABLA 14.3. *Valores de los coeficientes de correlación.*

r = 1.00	Correlación grande, perfecta y positiva
r = 0.90 a 0.99	Correlación muy alta
r = 0.70 a 0.89	Correlación alta
r = 0.40 a 0.69	Correlación moderada
r = 0.20 a 0.39	Correlación baja
r = 0.01 a 0.19	Correlación muy baja
r = 0.00	Correlación nula
r = −1.00	Correlación grande, perfecta y negativa

El coeficiente de correlación nunca puede ser mayor de 1, ni menor de −1. Su valor está comprendido, por tanto, de 0 a ±1. Por otra parte, mientras que muchos autores entremezclan los problemas de evaluar el grado de relación y de hacer predicciones, es útil, para propósitos conceptuales, el distinguir entre estas dos tareas y discutirlas por separado. La primera pregunta expone: ¿hasta qué punto están relacionadas las variables? La segunda pregunta dice así: dado que existen interrelaciones entre variables, ¿cómo se puede utilizar la información acerca de una o más variables para *predecir* lo que probablemente ocurriría con respecto a otra variable? Indudablemente que la correlación Pearson de producto-momento sería la respuesta a la primera, y la regresión lineal la respuesta a la última. No obstante, a menudo surgen circunstancias en las que estas simples técnicas no son apropiadas y, entonces, se hace necesario el uso de otros índices.

La base del cálculo del coeficiente de correlación se debe a Pearson y es el más sencillo de todos. El coeficiente de la correlación Pearson de producto-momento es una medida del grado de relación que está limitado a situaciones en que:

a) solamente dos variables están involucradas;
b) ambas variables se miden en el intervalo o en niveles superiores;
c) ambas variables son continuas; y
d) la relación entre las dos variables es lineal.

Pero los conceptos para la evaluación del grado de relación o correlación pueden haberse extendido a variables que estuvieran a diferentes niveles de medición y a situaciones en las que una o más de las variables no fuera continua. Una descripción sumaria de las medidas de asociación y relación más utilizadas se presenta en la tabla 14.4.

REGRESIÓN Y PREDICCIÓN

En esta parte se estudian las bases de la regresión estadística y se amplían los conceptos y medidas de correlación y la evaluación del grado de relación hasta llevarlos a situaciones en que otra correlación, que no sea ninguna de las de Pearson de producto-momento, sea apropiada al discutir las circunstancias en las que la relación entre variables no es lineal (*coeficiente eta*), cuando la intención es la de asegurar la similaridad intra o dentro de grupos definibles (*correlación intraclase*), donde ambos grupos son categorías sin ordenar (*coeficiente de contingencia e índice de Cramer*) y donde más de dos variables se incluyen (coeficiente de concordancia, correlación parcial, y correlación múltiple). Pero previamente definamos la regresión.

La regresión estadística se debe a Galton. Él efectuó una serie de trabajos, entre los que se destacan los relacionados con la estatura y la herencia. Galton encontró que lo hijos de padres altos, tienden a tener una estatura más baja que la de los padres, y viceversa. Conclusión: Se tiende a que todos tengan estatura normal, es decir, regresión a la media. Hoy la regresión se emplea en el sentido de conocer una variable, a partir de otra variable. Ejemplo: Si la estatura es x ¿cuál será el peso y? Por tanto, regresión es estimar valores de una variable, conocidos los valores de la otra variable. Las variables son dos: la *predictor* y la *predicando*. La primera se llama independiente y la otra, la que se busca, dependiente. Para hacer una regresión se utilizan las coordenadas, con el fin de ver qué tipo de regresión es: rectilínea, circular, elipsoide, hiperbólica o parabólica. A través de la intercepción de los resultados de x e y se construye la *nube de puntos*. La geometría analítica nos dice cuáles son las ecuaciones de cada línea o curva resultante (por ejemplo, la ecuación de la recta es: $y = a + bx$). Por medio de la representación en el eje de coordenadas obtenemos la línea de ajuste. De ahí se deriva el *coeficiente de regresión lineal*, que es aquel que expresa el número de unidades en que varía el valor más probable de y por cada unidad de variación de x. En el caso de representaciones de curvas la esencia del proceso es el mismo, aún cuando las características sean distintas.

Para entender una correlación curvilínea, también llamada coeficiente eta o proporción de correlación, se necesita recordar que la idea de regresión es bastante general y que servirá para cualquier tipo de relación entre las variables X e Y. Aunque la experiencia anterior con correlación y regresión haya estado limitada al caso lineal, se puede enfatizar que la idea básica de regresión es la de fijar la relación de un conjunto de distribuciones condicionales de Y tomadas en valores específicos de X. Por tanto, el análisis general de regresión sugiere que simplemente debemos conectar los medios de una serie de distribuciones condicionales. Si consideramos lo que sucedería si se intentara acomodar una línea recta en un diagrama de dispersión que hubiese sido representado por una regresión curvilínea, se podría indicar que

TABLA 14.4. *Medidas de asociación y relación.*

Tipo de variaciones involucradas	*Restricciones, suposiciones, o comentarios*	*Medición de la asociación que se va a utilizar*
2 Variables continuas	Relación lineal Escalas de intervalo o de proporción	Correlación Pearson de producto-momento
2 Variables continuas	Escalas ordinales	Correlación de orden de rango o tau de Kendall (véase medidas no-paramétricas)
2 Variables continuas, dicotomizadas artificialmente	Distribución bivariante normal de las dos	Correlación tetracórica
1 Variable continua 1 Variable, o continua o un conjunto discreto de categorías	Relación no-lineal	Proporción de correlación (coeficiente eta)
1 Variable continua 1 Variable es un conjunto discreto de categorías	La intención es la de asegurar el grado de similitud intra-grupos Intervalo o escalas de proporción	Correlación intraclase
1 Variable continua 1 Variable continua dicotomizada artificialmente	Intervalo o escalas de de proporción	Correlación biserial
1 Variable continua 1 auténtica dicotomía	Intervalo o escalas de proporción	Correlación punto biserial
2 auténticas dicotomías	Escala nominal u ordinal (Véase menú principal en Distribuciones el punto 3 de frecuencias bivariantes)	Coeficiente Fi
2 Conjuntos de categorías sin ordenar	Escala nominal	Coeficiente de contingencia
1 Conjunto de categorías sin ordenar 1 o más variables de cualquier tipo	La intención es la de determinar el grado de similitud entre los grupos en base a varias mediciones	Mahalonobis D2 (de la función discriminante lineal) o R biserial múltiple
3 o más variables continuas	La intención es la de encontrar el grado de relación entre dos y con los efectos de las demás constantes mantenidas	Correlación parcial
3 o más variables continuas	La intención es la de determinar la pronosticabilidad de una variable en base a otras. Relaciones lineales	Correlación múltiple (de regresión múltiple)
3 o más variables continuas	La intención es la de determinar la cantidad global de acuerdo Escala ordinal	Coeficiente de Kendall de concordancia (Véase medidas no-paramétricas)
3 o más variables continuas	Calcula el promedio de las inter-correlaciones	Correlación intraclase clase entre pares

una línea recta no se acomoda muy bien en una relación curvilínea y debe ser aparente que las sumas de los cuadrados de las desviaciones verticales (que se minimizan al calcular a y b en una ecuación lineal de regresión) serían demasiado grandes. Por lo que es de esperarse que el coeficiente de la correlación Pearson de producto-momento basado en un análisis de regresión lineal al ser aplicado a una relación curvilínea sería bastante pequeño y subestimaría considerablemente el tamaño de la relación indicada por el diagrama de dispersión (gráfico de puntos en que se dispersan o agrupan todas las relaciones x e y).

El índice que proporciona una mejor estimación del grado de asociación bajo estas circunstancias es la proporción de correlación *eta*. La lógica de la eta es bastante clara. De un diagrama disperso dado (que muestre, desde luego, una relación curvilínea) el investigador simplemente divide las variables X en conjuntos de columnas. Luego, busca la media Y (la Y) para cada columna X. La variación total de las observaciones individuales de la medida global de criterio Y, puede considerarse como la suma de dos componentes: la variación de la observación individual de la media por columna, y la variación de la media por columna de la media global. Otra medida bivariante especial del grado de relación a ser considerado es la correlación intraclase. Este índice es similar a eta en que también busca responder a la cuestión de qué proporción de la varianza de una variable está relacionada a otra variable.

Por otra parte, supongamos que vamos a considerar la tarea de evaluar el grado de relación entre status marital y el lugar en que trabajan los profesores. En este caso tenemos una situación general R c C ji-cuadrado, es decir, tenemos *clasificación cruzada* y el número de personas en cada categoría. El índice descriptivo del grado de asociación bajo estas circunstancias se denomina *coeficiente de contingencia* y que puede variar entre 0 y 1, pero no puede alcanzar un valor máximo, a menos que el número de hileras y el número de columnas sea infinito. Entonces, aunque el coeficiente de contingencia, ha sido utilizado tradicionalmente, recientemente se ha desarrollado otro índice que sobrepasa este problema. Este nuevo índice es una extensión del coeficiente *Fi*, pero a menudo se le denomina prueba de Cramer. En muchas situaciones nos vemos enfrentados con no solamente dos variables, entre las que se ha de determinar el grado de relación, sino con muchas. Por tanto, surge el problema de evaluar las relaciones en un caso de variantes múltiples. Cuando el investigador tiene tres o más variables pueden hacerse varias preguntas tomando como base a los datos.

La primera pregunta es: ¿hasta qué extremo puede una serie de variables predecir otra de manera efectiva? Cuando la variable por predecir es un conjunto de categorías sin ordenar, la respuesta estadística a esta pregunta es la D^2 de Mahalanobis; cuando la variable por predecir es continua, la respuesta estadística la da el *coeficiente de correlación múltiple*. La segunda pregunta importante es: ¿cuál es la relación entre las dos variables mientras se mantienen los efectos de otras constantes? Esa pregunta se responde por medio de los coeficientes de correlaciones parciales. La tercera pregunta es: ¿hasta qué extremo hay acuerdo global entre varias medidas? Cuando las variables utilizadas han sido medidas en una escala de intervalo o de proporción, la solución a esto es la correlación intraclase; cuando se han obtenido medidas de rango, la solución correcta es el *coeficiente de concordancia*.

Primero, examinemos la D2 de Mahalanobis o la así llamada distancia generalizada. Esencialmente, este índice nos muestra hasta que extremo se pueden separar los grupos (es decir, distinguirlos entre sí) por medio de un número de medidas diferentes. Entonces, se estima lo que ofrece la distancia entre dos grupos que puede ser logrado con un conjunto de medidas dado. Por ejemplo, supongamos que se desea saber hasta dónde es posible distinguir entre educadores y psicólogos en base a varios resultados de pruebas: memoria, razonamiento, habilidad cuantitativa, interés en las personas; etc. Para resolver este proble-

ma, el investigador podría tomar un grupo de psicólogos y un grupo de educadores, administrarles todas las pruebas y luego elaborar un tipo de puntuación compuesto:

$$C = b_1X_1 + b_2X_2 + b_3X_3 + \ldots = \sum_{i=1}^{n} b_iX_i$$

donde X_1 X_2 X_3 etc., representan las diferentes medidas (de las cuales hay m) y b_1 b_2 b_3 etc., son ponderaciones. El investigador determinará las ponderaciones para maximizar la suma de cuadrados de los grupos «entre» relativo a la suma de cuadrados de los grupos «dentro». Esto último tiene el efecto de establecer a b_i; de manera que la distribución de los puntajes compuestos de los dos grupos se traslapen lo menos posible. Por lo ya dicho, el investigador espera obtener un puntaje compuesto que produciría una distribución bimodal.

En el caso de los dos grupos la D^2 de Mahalanobis es equivalente al cuadrado del coeficiente de correlación múltiple (R^2), que se describe más adelante, que se obtendría si los puntajes de criterio son valores dados de *0* para todas las personas que son miembros de un grupo y *1* para todas las materias que pertenecen al otro grupo. Si el investigador tiene solamente dos grupos o está evaluando la diferencia entre solo dos de los grupos, puede utilizar correlación múltiple en vez de la D^2 de Mahalanobis. Pero si el investigador tiene más de dos grupos, entonces puede obtener un cálculo global del extremo hasta el cual se puede distinguir entre todos los grupos por medio de la distancia generalizada de Mahalanobis.

De una manera muy similar cuando la variable por predecir es continua, en vez de ser una serie de categorías sin ordenar como en el caso de la función discriminante y de la D^2 de Mahalanobis, el investigador establece una variable predictiva compuesta

$$Y = a + b_1X_1 + b_2X_2 \ldots + b_mX_m$$

y puede calcular las constantes a, b_1, b_2, b_3, etc. de manera de minimizar el error cuadrado de predicción. R es interpretada de la misma forma que r, por lo que R^2 representa la reducción relativa en cuanto a error de predicción y cuando la información en las variables de predicción m se utilizan y pueden ser interpretadas como porcentaje, las correlaciones múltiples también pueden ser expresadas en términos de variantes.

Por otra parte, conviene señalar los puntos importantes que se deben tomar en cuenta en relación a la interpretación del coeficiente de correlación múltiple. El coeficiente de correlación múltiple calculado, tomando como base un conjunto dado de datos, representa la correlación entre valores reales y predichos (es decir, r_{yy} en el conjunto de observaciones dado). Debe ser obvio que si tomamos un segundo ejemplo, un conjunto en algo diferente de b_i sería obtenido. Por tanto, si el investigador desea utilizar las ponderaciones de regresión para predecir una nueva muestra, como ocurre frecuentemente (por ejemplo, el funcionario encargado de las admisiones en una universidad puede utilizar a la clase de principiantes del año en curso para hacer predicciones para la clase de principiantes del año entrante) habrá una pérdida o disminución del coeficiente de correlación múltiple logrado. Mientras tengamos más variables mayor será la disminución. Por lo que, para calcular el grado verdaderamente accesible de predicción por medio del coeficiente de correlación múltiple, los investigadores deben llevar a cabo el proceso llamado de validación cruzada o cross-validación. Esto incluye la división de la muestra original de materias en dos partes: un «análisis» y una muestra de «validación». Posteriormente se derivará los valores bi en el grupo de «análisis» y se calculará el coeficiente de regresión múltiple como r_{yy} en el grupo de «validación».

La próxima pregunta, de las que pueden formularse cuando existen más de dos variables, interroga acerca del punto hasta el cual dos características pueden estar relacionadas cuando los efectos de otras se han mantenido constantes. Debe recordarse que la razón principal por la que no se puede utilizar la correlación entre dos variables para inferir causación se debe a que, a veces, la relación entre dos variables puede ser el resultado de que una tercera variable actúe sobre las otras dos. Por ejemplo, se podría preguntar hasta qué extremo están relacionadas la velocidad de lectura y la comprensión de la misma, si el factor de la inteligencia se mantiene constante, o se podría desear conocer dónde el número de horas de estudio y de progreso en estadística están relacionadas, si mantenemos constantes la habilidad de estudio y el conocimiento previo de las matemáticas. La respuesta a este tipo de pregunta la da el así llamado *coeficiente parcial de correlación.* Cuando mantenemos constante a una variable y aseguramos la correlación entre otras dos, tenemos lo que se denomina parciales de primer orden. Cuando deseamos mantener constantes a dos variables, mientras observamos la correlación entre otras dos variables tenemos coeficientes parciales de segundo orden.

La última pregunta que utiliza tres o más variables, trata de determinar hasta qué punto están de acuerdo en forma global varias variables. Ya se ha examinado la correlación *intraclase* que es aplicable cuando las observaciones han sido registradas en base a la proporción o escala de intervalo. Esta correlación ofrece un cálculo del promedio de relaciones de pares, es decir, las correlaciones entre pares de variables. Es aconsejable notar que muchas veces se ha utilizado la mediana de todas las correlaciones de pares como una alternativa de la correlación intraclase. No obstante, obtener las correlaciones de pares implica considerable trabajo, algo innecesario si todo lo que se desea es un cálculo global del punto hasta el cual hay un acuerdo entre las variables de un conjunto. En esta misma situación, cuando tenemos rangos, la prueba a utilizar es el coeficiente de concordancia de Kendall, el cual describiremos al final de este trabajo, relacionado con las medidas o procedimientos no-paramétricos.

MEDIDAS DE ASOCIACIÓN: LA PRUEBA *t*

Las medidas de asociación tienen una base parecida a las de correlación pero no pueden considerarse tales. Son pruebas de medición de hipótesis basadas en el cálculo de probabilidades, por las cuales podemos predecir los límites probables en que nuestros resultados pueden experimentar fluctuaciones a causa de los posibles errores que lleve implícita la experimentación. Varias son estas pruebas. El chi o ji cuadrado, el coeficiente de contingencia o el coeficiente fi y la prueba *t*. Las tres primeras las estudiaremos en el apartado de pruebas no-paramétricas, dado que lo mismo actúan como paramétricas o no-paramétricas. Sin embargo, la prueba *t* sólo puede ser utilizada en supuestos paramétricos y cuando un análisis de varianza no es aconsejable.

La prueba *t* es un método conocido como *t de Student*, por ser éste el seudónimo de su descubridor de nombre W.S. Gosset. El estableció un procedimiento para comprobar el nivel de significación de las medias, la diferencia entre ellas, de muestras pequeñas. La proporción de *t* es definida de la misma manera que la de z, es decir, la desviación o diferencia entre medias dividida por la desviación estándar o su error. De ahí, que la prueba *t* sea de gran utilidad en psicología para la comprobación de hipótesis cuando nuestra muestra no tiene mas de 30 datos o elementos.

Los cálculos del valor de t varían en función de la dependencia e independencia de los grupos, de la homegeneidad de la varianza, de la igualdad o desigualdad del tamaño de las

muestras y de la correlación entre ellas. Supongamos que deseamos comprobar si la diferencia entre las medias de dos poblaciones es significativa estadísticamente. Se asume previamente que los grupos son independientes y extraídos aleatoriamente de poblaciones normales. El Grupo A tiene como media = 40, su varianza es = 8 y el número de sujetos es de 28. El grupo B tiene una media = 30, una varianza = 10 y un número total de sujetos de N = 18. El cálculo de estos datos sobre la base de una muestra independiente nos daría como resultado una $t = 11.3$. Este valor debe ser contrastado en función de los grados de libertad (gl) de nuestras muestras. Es decir, $gl = N_a+N_b - 1$; o sea, gl = 44. Los grados de libertad se buscan en una tabla de probabilidades de la prueba t que puede encontrarse en cualquier libro de estadística elemental, y se observa que para gl = 44 corresponde un valor de 2.69 con 1% de error. Dado que nuestro valor t es muy superior al valor *t* para 44 grados de libertad, podemos concluir que las diferencias encontradas entre las medias de los grupos A y B son significativas al 1%.

Pero todas estas medidas, como podrá haberse inferido, tienen su base lógica en el objetivo, tipo y diseño del experimento. Cada medida de análisis de varianza, correlación, regresión, predicción, asociación o las técnicas no paramétricas, requieren de un diseño estadístico experimental previo. Veamos muy brevemente algunos de esos diseños generales y especiales.

Diseños generales experimentales

Los diferentes diseños experimentales estadísticos se pueden concentrar en cuatro diseños generales, algunos especiales y en una serie de principios que optimizan dichos diseños. Los cuatro diseños generales son: el diseño completamente aleatorio, el diseño de medidas repetidas, el diseño de bloques aleatorios y el diseño factorial.

DISEÑO COMPLETAMENTE ALEATORIO

El diseño más básico de todos es el completamente aleatorio. Este es el plan experimental para el cual el análisis de varianza unidireccional, es adecuado. En este sentido, el diseño completamente aleatorio (DCA) (CRD) se usa cuando el experimentador analiza solamente una variable experimental y no controla, especificamente, ninguna otra variable independiente. Sin embargo, como el diseño es aleatorio, se logra cierto control siempre que la variación de un sujeto a otro sea pequeña y el tamaño de la muestra sea suficientemente grande.

Por otra parte, el DCA es muy útil cuando el experimentador supone encontrar grandes efectos. Es decir, está en una situación donde las diferencias experimentales son suficientemente grandes para superar con facilidad las diferencias individuales de sujeto a sujeto. Esto ocurriría con más probabilidad si: *a)* el experimentador posee suficiente control sobre los sujetos para reducir diferencias de un sujeto al próximo en el comportamiento observado; o *b)* el experimentador genera y utiliza potentes tratamientos poco comunes. Si el experimentador puede producir las dos últimas condiciones, como frecuentemente es posible en estudios con ratas en condicionamiento operante, quizá no necesite hacer ningún análisis de varianza, ya que los efectos serán completamente obvios sin dicho análisis. Este diseño determina que cada nivel de tratamiento es aplicado a un grupo aleatorio de sujetos experimentales. Es decir, los sujetos son asignados a grupos en forma aleatoria y, posteriormente, los tratamientos también son asignados en forma aleatoria a los grupos, por lo que se establece un diseño completamente aleatorizado.

DISEÑO DE MEDIDAS REPETIDAS

El análisis especial de cada caso de varianza que ocurre cuando los datos que tenemos como información de criterio representan medidas tomadas en varias ocasiones diferentes es el punto fundamental del tema. Las medidas repetidas ocurren más a menudo cuando existe una puntuación de método de pre-tratamiento y una puntuación de método de post—tratamiento del mismo individuo.

Pero también, pueden desearse un pre-test, un post-test y un post-test retrasado. De igual manera, hay algunas situaciones en las que una serie de métodos de tratamiento le son administrados al mismo sujeto (o sea, cada sujeto recibe todos los tratamientos), y algunas situaciones en las que el nivel de actuación es evaluado después de aplicar cada una de las series de ensayos de aprendizaje. En el caso de estas últimas situaciones lo más apropiado es uno de los análisis de varianza (ANOVA) para diseños de medidas repetidas o tendencias. El más sencillo de los análisis de varianza, en cuanto a tendencias, es el conocido generalmente como ensayos con una condición estandard, experimentos de un solo factor con medidas repetidas sobre los mismos elementos o simplemente como ANOVA para medidas repetidas.

Por otra parte, se podría buscar una respuesta a la pregunta del efecto de algún tratamiento sobre tiempo con un diseño completamente aleatorio, en el cual, sujetos *diferentes* aparecieron en cada grupo de ensayo diferente o grupo de tiempo. Esto, sin embargo, es ineficaz ya que el tratamiento puede afectar a personas diferentes de diferente manera, e igualmente, los diversos grupos pueden ser muy diferentes entre si. También este sistema podría utilizarse en una situación en que las medidas repetidas representen diferentes ensayos de aprendizaje; el diseño completamente aleatorio requería que para el segundo grupo se descartase información acerca de su actuación en el primer ensayo, para el tercer grupo se descartase la información obtenida en los dos primeros ensayos, etc., y se continuaría en esta misma forma hasta terminar toda la serie.

Para mejorar esto, se puede estratificar a los grupos y utilizar un diseño de bloque aleatorio. Pero para que este último diseño sea eficiente, sería necesario clasificar a las personas en grupos que sean homogéneos con respecto a su reacción al método de tratamiento sobre tiempo (o con respecto a las curvas del aprendizaje o a los récords de frecuencia acumulativa) y obviamente podría resultar bastante difícil. Pero ya que la mayoría de los sujetos serán observados de cualquier manera, en varias ocasiones, resulta más eficiente permitir que cada sujeto sea su propio control; es decir, utilizar un diseño de medidas repetidas de manera de que cada sujeto sea un «bloque». *Si procedemos de esta manera habremos alcanzado la mejor clasificación posible, pues se puede aparear a cada sujeto consigo mismo.* A la vez se vencería también el problema que surge, si el tratamiento afecta a diferentes sujetos de manera diferente. Cuando los sujetos en el grupo «antes» (pre) son diferentes a los del grupo «después» (post), la diferencia observada de la actuación de los dos grupos se atribuye, en parte, a que los sujetos de los dos grupos eran diferentes y no a los efectos del método de tratamiento. Por ello, si los sujetos en los grupos «antes» y «después» son los mismos, la razón de ser de una de las fuentes principales de diferencias inaplicables queda explicada, especialmente en referencia a la búsqueda de los efectos de los métodos de tratamiento.

Siguiendo los procedimientos usuales de ANOVA para dividir las sumas de los cuadrados en componentes asignables, puede observarse que la suma total de cuadrados puede fraccionarse en componentes de sujetos de «entre» (entre filas) y sujetos de «intra» (dentro de las filas). Por otra parte, al investigador no le interesa verdaderamente el compo-

nente de sujetos «entre», ya que espera que los sujetos sean diferentes entre si. No obstante, si le interesan los cambios en el componente de los sujetos «dentro» sobre tiempo. O sea, que aceptará variación de las personas «entre» como un fenómeno real y, como consecuencia, lo extraerá como componente pero no lo someterá a más examen. Además, ya que el mismo sujeto ha sido medido en varias ocasiones diferentes (o expuesto a varios métodos de tratamientos diferentes), una parte de la varianza del sujeto «dentro» se atribuye a este hecho (es decir, hay algunos cambios reales que no están sujetos al azar, sobre veces, que reflejan el método de tratamiento o los efectos de tiempo) y el resto de la varianza de las personas «entre» se asigna a factores residuales o desconocidos que también producen un cambio, sobre las veces. Por este motivo, el análisis que le sigue, busca separar estas dos cosas, dividiendo la varianza sujetos dentro, en dos subcomponentes: un componente de métodos de tratamiento entre (o tiempo-veces) y un componente residual.

DISEÑO DE BLOQUES ALEATORIOS

Si recordamos que una variable independiente interventora puede producir diferencias significativas en la variable dependiente y, por tanto, ocasionar cambios que pueden confundirse con el efecto del tratamiento, se hace necesario su control. La manera más fácil de lograr un control de esta variable independiente es aplicar una técnica análoga al muestreo estratificado. Es decir, el investigador puede dividir los sujetos en grupos homogéneos respecto a dicha variable, y entonces asignar, aleatoriamente, los sujetos a los distintos niveles de tratamiento de estos grupos y bloques homogéneos. Por ejemplo, en las Ciencias de Comportamiento cuando utilizamos sujetos humanos, el nivel intelectual (NI) puede ser, a menudo, una variable importante para controlar. De este modo, el experimentador quizá pudiera utilizar un diseño que se le conoce generalmente como de bloque aleatorio (DBA). Se le llama también completo porque, por ejemplo, algunos sujetos de cada nivel intelectual aparecen en cada nivel de tratamiento; y se le define como de bloque aleatorio porque lo sujetos se asignan aleatoriamente a los distintos niveles de tratamiento dentro de cada bloque.

Para evitar sesgos en este diseño, es importante que los sujetos dentro de una categoría en la variable de clasificación (es decir, dentro de un bloque) se asignen aleatoriamente a grupos y que los grupos sean asignados aleatoriamente a los tratamientos. A primera vista, puede parecer innecesario asignar tratamientos aleatoriamente si ya el investigador ha asignado los grupos aleatoriamente. Sin embargo, debemos reconocer que el primer paso produce interacciones aleatorias dentro de grupos, por ejemplo, para que todos los hombres y todas la mujeres no aparezcan en un tratamiento particular; pero que el segundo paso se realiza para asegurar al experimentador que sujetos o combinaciones de sujetos particulares no se seleccionen para tratamientos específicos debido a algunos sesgos no intencionales que pueda tener el investigador. Por supuesto, si para empezar, los grupos de tratamientos se numeran uno, dos, tres, cuatro, etc., no es necesario sacar dos series de números aleatorios para realizar las dos gestiones anteriores. Se puede hacer simplemente, asignando sujetos aleatoriamente a cada uno de los grupos arbitrariamente descritos. Este diseño de bloque aleatorio podría llegar a constituirse en un diseño de grupos apareados. Esto ocurriría a medida en que el bloque se hace más y más preciso; es decir, a medida en que ocurre la clasificación de personas en grupos más o menos homogéneos, el experimentador logra una situación en la cual el número de sujetos en cada grupo es igual al número de niveles de tratamiento. El último paso para obtener control por medio de este mismo procedimiento,

aparece cuando se utiliza un solo sujeto como bloque, sirviendo de este modo, como su propio control, en una situación donde se aplican con éxito los diferentes tratamientos, conduciéndonos así, a un diseño de *medidas repetidas*, tal como presentamos previamente.

DISEÑO FACTORIAL

El último de los cuatro diseños básicos —el diseño factorial— (DF) (FD) simplemente extiende el concepto del diseño completamente aleatorio para incluir dos o más variables experimentales, en la misma forma que el diseño de bloque aleatorio extiende el diseño completamente aleatorio para incluir una variable independiente. De este modo, tal como en el caso del DBA, el diseño factorial es la estructura experimental apropiada para un análisis de varianza doble o superior. Por otra parte, ya que en el diseño factorial todas las variables son de tratamientos, el investigador examinará todos los efectos principales y todas las interacciones. Esto está en contraste con el diseño de bloque aleatorio, donde los principales efectos de tratamiento e interacción se examinan frecuentemente, pero el efecto principal de la variable clasificación puede ser o no examinado. Así, el diseño factorial es un plan experimental en el cual el investigador obtiene la información más completa posible. Contrarresta esta ventaja el hecho de que el diseño resulta difícil de usar —es decir, requiere un gran número de grupos aleatoriamente asignados— si el número de variables (o sea, el número de factores) o si el número de niveles de los factores resulta grande.

Por otra parte, si todos los factores de un diseño factorial tienen el mismo número de niveles, la notación utilizada para describirlo tiene la forma de lf cuando

l = número de niveles
f = número de factores

Por ejemplo, 2^4 describe un estudio en el cual hay cuatro factores experimentales, cada uno con dos niveles, e implicaría que el investigador tendría que encontrar $2^4 = 16$ grupos de sujetos diferentes para dar a cada grupo de sujetos una combinación de tratamientos distintos. Por otra parte, si los diferentes factores tienen distintos números de niveles, entonces la notación descriptiva se expresa en forma de multiplicación. Por ejemplo, un diseño factorial $2 \times 4 \times 4 \times 3$ inferiría una situación en la cual un factor tendría dos niveles, dos de los factores tendrían cuatro niveles, y una de las variables experimentales, tres.

En la descripción de diseños factoriales es una práctica común utilizar letras en bastardilla, mayúsculas cerca del comienzo del alfabeto para señalar un factor particular, y usar letras minúsculas con pequeñas notas escritas abajo para indicar los diferentes niveles de este factor. Por ejemplo, en el estudio arriba mencionado que incluye cuatro factores con diferentes números de niveles, un experimentador puede dar a estos factores las letras *A*, *B*, *C* y *D*, y entonces describir los niveles respectivos de estos factores como los valores: a_1, a_2; b_1, b_2, b_3, b_4; c_1, c_2, c_3, c_4; d_1, d_2, d_3.

Otro ejemplo de diseño factorial sería de $3 \times 2 \times 4$ en donde un factor tendría tres niveles (A), el siguiente (B) dos niveles y (C) cuatro niveles, lo que implicaría 24 grupos de sujetos diferentes (G). Supongamos un experimento en el cual la variable de tratamiento A tiene tres niveles, la B dos niveles y la C cuatro niveles. Si las tres variables de tratamiento A, B y C tienen p, q, y r niveles, respectivamente, el experimento requiere un total de pqr grupos de sujetos experimentales. El uso de subscriptos o números

debajo de G (Grupos) indican: el primero el nivel de tratamiento de la variable C; el segundo, el nivel de tratamiento de la variable B; y el tercer número, el nivel de tratamiento de la variable A.

Los diseños factoriales generan a su vez otros tipos de diseño. Los más importantes son:

a) *El diseño factorial: Dos factores.* Este diseño viene a ser una extensión del diseño completamente aleatorio, pero con la diferencia de que se puede contrastar un conjunto de variables con otro. Se denomina también factorial 2 × 2, aun cuando puede ser que se esté interesado en comparar entre un conjunto de 2 variables con otro conjunto de tres variables, lo que nos daría un diseño factorial 2 × 3 con dos factores, o tres variables con otras tres. Lo que sería un factorial 3 × 3, y así sucesivamente. Por otra parte, se deben tener en cuenta los tres modelos de efectos, el fijo, el aleatorio y el mixto. Esto es decisivo para cualquier diseño. Recordemos que fijo implica que los niveles de los factores A y B son fijados por el experimentador; aleatorio que los niveles de los factores A y B son seleccionados aleatoriamente de todos los posibles niveles existentes; y mixto cuando A es fijo y B aleatorio o viceversa.

b) *Diseño factorial: Tres factores.* Este diseño que utiliza un análisis de varianza multidireccional o de clasificación múltiple, es una extensión del anterior pero en donde se determina la combinación de los efectos de tres variables; así se podría construir diseños factoriales 2 × 2 × 2, 2 × 3 × 2, 3 × 2 × 3, 2 × 3 × 3, etc., de acuerdo al número de niveles que se establezcan en cada factor.

c) *Diseño mixto de dos factores.* Este diseño está comprendido en el modelo de efectos mixtos a que nos referíamos en el análisis de varianza. Realmente, este tipo de modelo no es usado frecuentemente en investigación psicológica y educativa, sin embargo, nos permite comparar las diferencias de actuación total de los sujetos en todos los grupos experimentales y al mismo tiempo el estudio de los cambios en la actuación de dichos sujetos. Este diseño es una combinación del diseño de tratamiento por sujetos y el completamente aleatorio y su uso, al contrario de la mayoría de los modelos mixtos, es ampliamente utilizado en psicología y ciencias sociales.

d) *Diseño mixto de tres factores:* El diseño mixto de tres factores puede tomar dos modalidades distintas. La primera con medidas repetidas sobre un factor, y la segunda con medidas repetidas sobre dos factores. Los dos tienen las ventajas y desventajas comunes pero mientras el de un factor es una combinación de un diseño factorial y de un tratamiento por sujetos, el de dos factores combina el diseño completamente aleatorio con el de tratamiento por tratamiento por sujetos. En el primer diseño mixto de tres factores, los grupos experimentales y la asignación de sujetos a los mismos se hace exactamente igual que en un diseño factorial. Sin embargo, los efectos de un tratamiento o factor combinados con el otro factor no se miden una sola vez, sino en varias oportunidades; de ahí su nombre de medidas repetidas sobre un factor. Las consideraciones para aplicar el diseño son las mismas que para el diseño anterior. El segundo tipo, denominado diseño mixto de tres factores en medidas repetidas sobre dos factores consiste en que en vez de que un solo grupo reciba todos los tratamientos son dos o más grupos los que se evalúan. Se asemeja por tanto al anterior, pero como aumentan los tratamientos y grupos, existe el riesgo de la interacción de efectos entre cada ensayo y tratamiento. Pero por otra parte, el número de sujetos para realizar el experimento es inferior.

Las estadísticas, o pruebas no paramétricas

Las pruebas no-paramétricas se consideran de distribución libre, por cuanto no plantean suposiciones con relación a la distribución de las puntuaciones en la población, mientras que la prueba F asume puntuaciones distribuidas normalmente. Esta posición es mantenida por muchos especialistas en estadística, sin embargo, otros infiere que las técnicas paramétricas estarían limitadas a las escalas de intervalo y de proporción debido a la transformación no lineal, hecho que es permitido con las escalas ordinales y nominales. Este argumento sería rechazado si se tiene en cuenta que la consideración más importante para utilizar pruebas paramétricas no son las escalas de datos, sino que los datos se distribuyan conforme a las suposiciones matemáticas, como independencia, normalidad, etc.

Estos argumentos y muchos otros, nos llevan a la conclusión que todavía no existe un acuerdo generalizado al respecto. Por ello, el investigador antes de utilizar pruebas no paramétricas, debe considerar si existe la posibilidad de usar pruebas de mayor potencia como la F, t, etc. Esto debe considerarse en todo momento, ya que los métodos de distribución libre tienen muy baja potencia para detectar diferencias significativas. En este sentido, las desventajas más importantes de las pruebas no-paramétricas en relación a las paramétricas son:

a) Los estimados de parámetros son imposibles de obtener con pruebas no-paramétricas, mientras que un ANOVA sí.
b) Con el ANOVA se puede obtener un estimado de la fidelidad del experimento, mientras que con las pruebas no-paramétricas esto es imposible.
c) Aun en la medida en que las suposiciones exigidas para las pruebas paramétricas se cumpliesen en la aplicación de pruebas no-paramétricas, éstas serían menos potentes.
d) Casi no existen pruebas no-paramétricas para el estudio de los efectos de interacción.

Sin embargo, por otra parte, las pruebas paramétricas usualmente no son apropiadas cuando las variables del experimento son medidas con escalas ordinales, y en la mayoría de los casos la prueba no-paramétrica debe ser la seleccionada.

Las ventajas que generalmente se atribuyen a las pruebas no-paramétricas son:

a) Simplicidad y velocidad de aplicación.
b) Menor número de sujetos en la muestra, ya que en estos casos, las suposiciones para las pruebas paramétricas son susceptibles de violación.
c) Los estados de probabilidades son exactos cuando se obtienen de la mayoría de las pruebas no-paramétricas, sin que importe la forma de la distribución de la población de donde se obtuvo la muestra.

Resumiendo, se podría decir, que si las escalas de datos son en intervalos o proporciones, siempre se debe utilizar las pruebas paramétricas. Si las escalas son nominales u ordinales se puede utilizar una prueba no paramétrica, siempre que no exista otra paramétrica en el caso específico bajo estudio. Finalmente, si la muestra es muy pequeña, la única alternativa válida es el uso de una prueba no paramétrica.

Las pruebas no paramétricas más importantes de acuerdo a la clasificación de sus variables y al número de muestras, es tal como sigue:

- Una muestra
  - Prueba de Rachas
  - Chi-cuadrado (Escala nominal)
- Dos muestras independientes
  - Chi-cuadrado (Escala nominal)
  - Prueba U de Mann-Whitney (Escala ordinal)
- Dos muestras dependientes
  - Prueba de rangos señalados de Wilcoxon (Escala ordinal)
  - Prueba Q de Cochran (Escala nominal o dicotomía ordinal)
  - Prueba Bidireccional de Friedman (Escala ordinal)
- Más de dos muestras independientes
  - Chi-cuadrado (Escala nominal u ordinal)
- Más de dos muestras dependientes
  - Prueba A de Cochran (Escala nominal o dicotomía ordinal)
  - ANOVA Bidireccional de Friedman (Escala ordinal)
  - Correlación por rangos
  - Ro de Spearman (2 variables)
  - Tau de Kendall (2 variables y otra variable constante)
  - Correlación parcial de Kendall (K variables)
  - Coeficiente de concordancia de Kendall (K variables)

Veamos algunas de estas pruebas de forma muy sucinta. La obra de Siegel, *Diseño experimental no paramétrico,* es un clásico que incluye todos los procedimientos.

PRUEBA Q, DE COCHRAN

La prueba Q de Cochran se utiliza para las estrategias de diseño que determinan la observación de un mismo elemento o sujeto en diferentes condiciones, tres o más, siempre que las observaciones puedan dicotomizarse. Ejemplos como votaciones (si-no), calificaciones (aprobado-suspenso), actitudes (en favor-en contra), etc., que configuran variables dicotómicas y cuyo procedimiento puede repetirse varias veces a través de diferentes enfoques, la prueba Q es muy útil para determinar los efectos diferenciales de los distintos enfoques. De la misma forma que se utiliza un sólo sujeto en distintos procesos es posible tener diferentes sujetos que estén apareados en grupos. La prueba Q es muy potente para grupos correlacionados, siempre y cuando las observaciones o puntuaciones no estén expresados en una escala de proporción o intervalo, pues en este caso, el análisis de varianza es el más adecuado.

PRUEBAS CHI-CUADRADO O JI-CUADRADO

Cuando realizamos un muestreo y aplicamos a éste los métodos estadísticos, nos proponemos deducir de la muestra, conclusiones del universo de donde ella proviene. Por tanto, al efectuar este procedimiento estamos estableciendo una hipótesis. La muestra que hemos escogido corresponde al total del universo. Puede bien la prueba de chi o ji-cuadrada (c^2), así como las otras técnicas, probarnos esta hipótesis. Cuando se calcula el coeficiente de contingencia, nos basamos para calcular éste, en el chi-cuadrado, que nos

permite establecer una concordancia entre valores empíricos y teóricos y llega hasta la comparación de distribuciones enteras. En otras palabras, el chi-cuadrado permite la comparación global del grupo de frecuencias teóricas calculadas a partir de la hipótesis que se quiere demostrar. Si el chi-cuadrado demuestra que la disparidad entre la frecuencia que se ha obtenido y la que teóricamente se debía obtener es demasiado grande, debe ser atribuida al azar y de esta manera nuestra hipótesis debemos considerarla falsa.

Es conveniente aclarar que la prueba de chi-cuadrado no se debe emplear cuando el número total de observaciones sea inferior a 50, cuando la frecuencia teórica es menor de 5 y cuando el total de las frecuencia empíricas no es igual al total de las frecuencias teóricas. Para la comprobación de muestras pequeñas, en el sentido de comprobar la significación de sus medias, tenemos la prueba t de Student.

ANÁLISIS DE VARIANZA BIDIRECCIONAL POR RANGOS

El análisis de varianza bidireccional o de dos vías por rangos es una prueba elaborada por Friedman para ser aplicada a K muestras correlacionadas. Si existen razones suficientes para pensar que el ANOVA paramétrico no puede cumplir con las suposiciones de homogeneidad, normalidad, etc., esta prueba es la indicada. Los datos son un conjunto de K observaciones para una muestra de sujetos que han sido sometidos a diferentes condiciones experimentales o tratamientos. Esta es una prueba similar a la Q de Cochran, pero requiere que las distintas observaciones en un mismo sujeto estén en condiciones de ser ordenadas en rangos. Si la hipótesis nula es verdadera, la suma de los rangos de cada una de las diferentes condiciones sobre las que el sujeto es observado, deben ser iguales.

PRUEBA *U*, DE MANN-WHITNEY

Esta prueba se utiliza para evaluar la diferencia entre dos distribuciones de población. Las dos muestras deben ser aleatorias y seleccionadas independientemente. La hipótesis nula es que las poblaciones de donde se obtuvieron las muestras son iguales. Sin embargo, esto no es equivalente a decir que las medias de la población son diferentes, ya que es posible que dos distribuciones tengan diferente forma pero sus medias sean iguales. Por ello, cuando la forma de las distribuciones es igual, la prueba compara las tendencias centrales de los grupos; mientras que si son diferentes, los resultados deben interpretarse en términos de las diferencias entre las distribuciones en general.

PRUEBA DE WILCOXON

La prueba de Wilcoxon, denominada de Rangos Señalados y Pares Igualados se utiliza con dos muestras correlacionadas cuyos datos, a partir de la diferencia entre ellas, se convierten en rangos absolutos. Esta prueba necesita dos clases de ordenación. En primer término se ordenan las diferencias entre cada par de puntuaciones igualadas y, en segundo lugar, se ordenan las diferencias entre los pares. Una mayor diferencia entre los pares igualados recibirá obviamente mayor peso que si fuese más pequeña. La Prueba de Wilcoxon es tan potente como la prueba «t» de Student, para muestras dependientes, si se cumplen los requisitos sobre diferencias entre pares de observaciones.