

HUBERT M. BLALOCK

ESTADÍSTICA SOCIAL

$\frac{c}{f}e$

 BIBLIOTECA 21

CONSULTA EN SALA





HUBERT M. BLALOCK, JR.

ESTADÍSTICA SOCIAL



BIBLIOTECA
UNIVERSIDAD
EMPRESARIAL
SIGLO VEINTIUNO



FONDO DE CULTURA ECONOMICA
MÉXICO

Primera edición en inglés	1960
Primera edición en español,	1966
Segunda edición en inglés,	1972
Segunda edición en español, de la segunda en inglés,	1978
Quinta reimpresión,	1994

A
ANN, SUSIE
y KATIE

Título original:

Social Statistics

© 1960, 1972, Mc Graw-Hill, Inc. Nueva York

D. R. © 1966, FONDO DE CULTURA ECONÓMICA

D. R. © 1986, FONDO DE CULTURA ECONÓMICA, S. A. DE C. V.
Carretera Picacho-Ajusco 227; 14200 México, D. F.

ISBN 968-16-0135-1

Impreso en México

PREFACIO

ESTE libro ha sido escrito fundamentalmente para aquellos estudiantes de sociología, tanto los ya titulados como los que aún no lo están, que se propongan dedicarse a la investigación social.

Durante los doce años transcurridos desde que apareció la primera edición, tanto el nivel de preparación como la complejidad de la estadística aplicada han experimentado una mejoría considerable, no sólo en el campo de la sociología, sino en los de la ciencia política, la antropología, la geografía y el trabajo social. A pesar de ello, una abrumadora mayoría de los estudiantes, o de los que ejercen en estos terrenos, carecen de la necesaria base matemática que les permita obtener plena ventaja de la bibliografía técnica sobre estadística, matemática y econometría, cuyos materiales aumentan rápidamente. Teniendo en mente tales datos básicos ha sido escrito este texto, tratando de evitar, hasta donde es posible, las derivaciones matemáticas, bastando una rápida revisión de ciertos principios algebraicos, listados en el Apéndice 1, para que el estudiante medio obtenga una preparación suficiente. Aun cuando no resulte necesario en un primer curso de estadística poner de relieve dichas derivaciones matemáticas, el autor está convencido de la necesidad de entender perfectamente ciertas ideas básicas y fundamentales sobre las que se asientan los principios de la deducción estadística. Tal cosa resulta indispensable si ha de lograrse algo más que un conocimiento limitado a "recetas" estadísticas. Hemos puesto por ello especial interés en la lógica que fundamenta la deducción estadística, incluyendo asimismo un capítulo relativo a la probabilidad, en tanto se ha prestado menor atención a materias más o menos rutinarias estudiadas en textos elementales.

Uno de los problemas más difíciles en la enseñanza de la estadística aplicada es el de lograr incitar a los estudiantes, de manera que éstos superen su temor a las matemáticas y aprendan a aplicar la estadística en sus propios campos de interés. Por esta última razón el autor no ha intentado cubrir una gama amplia de aplicaciones, eligiendo por el contrario ejemplos de interés fundamental para los sociólogos. También ha escogido otros ejemplos de campos fronterizos de la sociología, tomados de terrenos tales como la psicología social, el trabajo social y la conducta política. En la mayoría de los casos cada nuevo tema ha sido ilustrado con un solo ejemplo, por suponer que muchos estudiantes perderían el hilo básico del razonamiento si se utilizaran muchos de ellos para ilustrar un punto. Se proporcionan, sin embargo, ejemplos adicionales bajo la forma de ejercicios que aparecen al final de cada capítulo. En general ha tratado

el autor de lograr un equilibrio razonable entre la conveniencia de establecer los principios básicos en forma tan clara y concisa como resulte posible, y la necesidad de repetir algunas de las ideas más dificultosas cada vez que examina un nuevo tema. Hasta donde es posible, las ideas nuevas han sido presentadas gradualmente y —lo que es igualmente importante— se han hecho esfuerzos para que cada nuevo tema quede relacionado con los que han precedido. Al hacerlo así se tuvo presente la meta fundamental, consistente en brindar una apreciación de las semejanzas básicas que fundamentan muchas de las pruebas (*tests*) y mediciones más comúnmente usadas.

Casi todas las sugerencias que he recibido de personas deseadas de ayudar a mejorar la primera edición se referían más a la conveniencia de aumentar que a la de reducir el libro, dando a entender que muchos de los temas tratados originalmente deberían ser examinados más técnicamente. En mi opinión, tanto los sociólogos como los estudiosos de ciencias políticas, en particular, precisan verse más expuestos, tanto a una bibliografía de mayor nivel técnico sobre diseños experimentales, como a procedimientos para el uso de ecuaciones simultáneas relacionadas con la investigación no experimental. Resultó claro que si se agregaban estos materiales al texto original, éste perdería su atractivo como introducción adecuada para los estudiantes que aspiran a la maestría en ciencias sociales. Se resolvió, pues, que los diseños experimentales, el análisis de factores y los métodos de ecuaciones simultáneas, así como otros temas más avanzados, serían tratados en un texto separado, a cargo de dos de mis colegas: Lewis F. Carter y Krishnan Namboodiri.

Se incluye en el texto cierto número de secciones, párrafos y ejercicios que, o bien son conceptualmente difíciles o presuponen que el estudiante posee cierta familiaridad con temas cubiertos en cursos acerca de los métodos de investigación. Estas partes del texto han sido señaladas con asteriscos (*), y pueden ser leídas sin detenerse en ellas, o bien eliminadas del todo. Los instructores que utilicen el texto en cursos de un semestre podrán indicar a los estudiantes la conveniencia de omitir dichos materiales.

Se ha llevado a cabo una leve revaluación en los aspectos técnicos, empero sin cambiar la estructura básica del texto. Hay ciertas modificaciones en la sección relativa a la estadística descriptiva, sección a la que el autor ha añadido discusiones sobre supuestos y conceptos básicos, confiando así en aclarar la relación entre los modelos estadísticos y el mundo real con el que el científico social ha de tratar. Además de estos cambios, la presente edición contiene discusiones acerca de diversos procedimientos, *tests* y mediciones que han venido siendo usados cada vez más durante la década de 1960-1969.

El capítulo IX, sobre la probabilidad, ha sido ampliado mediante el examen de permutas, diagramas en árbol, Teorema de Bayes y cálculos relacionados con las probabilidades condicionales y la noción de valores previstos. También al capítulo X, que incluye una explicación de la distribución binomial, ha agregado el autor una breve discusión de dicha distribución binomial, la distribución hipergeométrica y la distribución de Poisson. Estas agregaciones facilitarán la transición a otros textos, orientados específicamente a la estadística no paramétrica.

El autor ha aumentado asimismo el espacio dedicado a técnicas no paramétricas, incluyendo el *test* de Friedman para análisis recíprocos de fluctuaciones con orden; gamma y d_{ym} , como medidas de asociación ordinal; un *test* para la interacción relativa a la diferencia de diferencias de proporciones, y la normalización en el caso de procedimientos con escala nominal. El lector encontrará además una discusión ampliada acerca de las propiedades de las varias medidas ordinales y técnicas de divisibilidad con escalas ordinales.

De manera análoga ha sido ampliado el comentario de los acercamientos paramétricos para incluir en él la discusión de los supuestos para el modelo lineal general, cubriendo además el acercamiento de la variable simulada como perspectiva alterna en el análisis de la covariación. También se han agregado los principios esenciales básicos de la teoría que fundamenta el uso de combinaciones lineales, aplicándola a la discusión del error normal de la media, la diferencia de medias, la diferencia de las diferencias de proporciones y el empleo de comparaciones ortogonales en el caso de muestras múltiples.

En un esfuerzo para ayudar al lector en la apreciación del cuadro global, el autor ha incluido una tabla resumen de *tests* y mediciones, tabla que aparece en el interior de la tapa, así como amplios sumarios al final de los capítulos II, XIV y XX.

Son muchas las personas que han colaborado en esta revisión, pero el autor desearía dar gracias de manera especial a Richard G. Ames, Erica Borden y Louis Goodman por sus comentarios en relación con la totalidad del manuscrito.

Por su ayuda en la preparación de la primera edición, deseo dar gracias de manera destacada a los estudiantes y colegas de la Universidad de Michigan, quienes leyeron varios borradores del libro y sugirieron mejoras. A Richard T. LaPierre, Sanford Dornbusch, Robert Ellis, Santo Camilleri y Theodore Anderson mi aprecio por leer y criticar el manuscrito original. Deseo igualmente agradecer por la corrección de pruebas, el mecanografiado y la revisión de los cálculos, la ayuda prestada por Ann Blalock, Diane Etzel, Ann Laux y Doris Slesinger. Gracias igualmente a Daniel O. Price, merecedor de amplio reconocimiento por haber estimulado mi interés en la estadística.

Quedo en deuda con el profesor Sir Ronald A. Fisher, de Cambridge, con el doctor Frank Yates, de Rothamsted y con los señores Oliver and Boyd, Ltd., de Edimburgo, por su autorización para reimprimir los cuadros III, IV y V de su libro *Cuadros estadísticos para investigaciones biológicas y agrícolas*. También estoy reconocido a los distintos editores y autores, mencionados en los lugares adecuados, quienes amablemente dieron permiso para el uso de varios cuadros y formas de computación.

HUBERT M. BLALOCK, JR.

PRIMERA PARTE

INTRODUCCIÓN

I. INTRODUCCIÓN: OBJETIVOS Y LÍMITES DE LA ESTADÍSTICA

EL CAMPO de la estadística tiene múltiples aplicaciones, como lo demuestra el hecho de que se den cursos de ella en materias tan dispares como son la odontología y la sociología, la administración de negocios y la zoología, la salud pública y la enseñanza. A pesar de ello, existen todavía muchas concepciones erróneas acerca de la naturaleza de esta disciplina en estado de rápido desarrollo. La idea que de la estadística se hace el lego, puede diferir mucho de la que tiene de ella el estadígrafo profesional. A veces se supone que el estadígrafo es una persona que manipula números para demostrar su punto de vista. Por otra parte, en cambio, algunos estudiantes de sociología o de otras ciencias sociales, propenden a admirarle como a alguien que, con la ayuda de su calculadora, puede convertir casi cualquier estudio en "científico". Debido posiblemente al respeto que muchas personas sienten por todo aquello que en alguna forma se relaciona con las matemáticas, a muchos estudiantes les resulta difícil inscribirse en un curso de estadística sin cierta aprehensión. Pese a que les infunda temor la perspectiva de trabajar con números, es posible también que esperen demasiado de una disciplina que parece tan formidable. Antes, pues, de entrar demasiado rápidamente en materia, con lo que corremos el riesgo de perder la perspectiva, empecemos por preguntarnos qué es exactamente la estadística y qué es aquello que puede y aquello que no puede hacer.

Tal vez resulte más fácil empezar indicando aquello que la estadística no es. En primer lugar, la estadística no es en modo alguno un método con el que uno pueda probar casi todo aquello que desea probar. Veremos, antes bien, que los estadígrafos ponen especial empeño en establecer las reglas del juego de tal manera que las interpretaciones no vayan más allá de los límites de los datos. Sin embargo, no hay nada en los métodos estadísticos en sí mismos que sea capaz de evitar que el individuo superficial o intelectualmente poco escrupuloso saque sus propias conclusiones, a pesar de los datos, y uno de los aspectos más importantes de un curso de introducción a la estadística consiste precisamente en poner a los estudiantes en guardia contra los posibles abusos de esta herramienta.

La estadística no es sencillamente una colección de hechos. Si lo fuera, no valdría mucho la pena estudiarla. Ni constituye tampoco un sustitutivo del pensamiento abstracto teórico o del examen minucioso de los casos excepcionales. En algunos de los libros de texto más antiguos solían encontrarse prolijas discu-

siones acerca de los méritos del estudio casuístico frente al método estadístico. Ahora, en cambio, admítase claramente que los métodos estadísticos no se "oponen" en modo alguno al análisis cualitativo de los casos particulares, sino que ambos métodos se complementan. Y ni siquiera es exacto que la estadística sólo sea aplicable en presencia de un gran número de casos, o que no pueda emplearse en los estudios de exploración. Finalmente, la estadística no es tampoco un sustituto de la medida, o de la preparación cuidadosa de una cédula de investigación o de otros instrumentos para la recolección de datos. Se insistirá con mayor detalle en este último aspecto al final del presente capítulo y en el siguiente.

Y ahora, habiendo indicado lo que la estadística no es, ¿podemos acaso afirmar decididamente aquello que es? Infortunadamente, los estadígrafos mismos parecen discrepar algo entre sí en cuanto a la extensión de aquello que deba comprenderse bajo el apelativo general de "estadística". Adoptando un punto de vista pragmático, podemos decir por nuestra parte que la estadística comprende dos funciones muy vastas, y que nada de aquello que no cumple dichas dos funciones forma parte de ella. La primera es la de la descripción, el resumen de la información de tal modo que se pueda emplear mejor. Y la segunda es la de la inducción, consistente en formular generalizaciones a propósito de una determinada población sobre la base de una muestra extraída de la misma. Estas dos funciones se examinarán a su tiempo.

1.1. Funciones de la estadística

La estadística descriptiva. En la investigación social, una persona se encontrará a menudo en la situación de disponer de tantos datos, que le resulte difícil absorber la información entera. Puede haber reunido 200 cuestionarios y preguntarse con todo, embarrasamente, "¿qué hago con todo ello?" Con tanta información habría de resultar excesivamente difícil, excepto tal vez para las mentes extraordinariamente fotográficas, captar intuitivamente lo que los datos contienen. En una forma u otra, pues, la información ha de reducirse hasta un punto en que pueda verse claramente lo que hay en ella: ha de resumirse. Con el empleo de medidas de cálculo, tales como porcentajes, promedios, desviaciones estándar y coeficientes de correlación, resulta posible reducir los datos a proporciones manuales. Al resumir los datos sustituyendo grandes cantidades por unas pocas medidas, cierta información ha de perderse necesariamente y, lo que es más grave, es posible obtener resultados engañosos, a menos que se los interprete con mucha precaución. De ahí que convenga indicar claramente las limitaciones de toda medida resumida.

La estadística descriptiva es muy útil en aquellos casos en que

el investigador necesita manejar relaciones mutuas entre más de dos variables. Supongamos, por ejemplo, que resulte preciso emplear ocho o diez variables como ayuda para explicar las tasas de delincuencia, y supongamos por otra parte que aquellas variables explicativas o *independientes* están altamente relacionadas entre sí. Si se desea aislar el efecto ocasionado por una o dos de tales variables, limitándonos a las consecuencias de las demás, ¿cómo habría que proceder? ¿Qué género de supuestos resultarían necesarios? Situaciones de este grado de complejidad se plantean en una rama de la estadística conocida con el nombre de *análisis multivariado*. En los capítulos xv, xvi, xix y xx examinaremos algunos problemas relativamente sencillos de análisis multivariado, reservando otros casos más complejos para un segundo volumen.

La estadística inductiva. La estadística resultaría una materia muy fácil si la atención pudiera limitarse a las medidas descriptivas. Tal vez una función mucho más importante de la estadística, y en todo caso la que retendrá la mayor parte de nuestra atención en este texto, es la de la inducción, consistente en inferir propiedades de una población sobre la base de una muestra con resultados conocidos. La inducción estadística, como se la acostumbra llamar, implica un razonamiento mucho más complejo que el de la estadística descriptiva, pero, si se la comprende y utiliza bien, se convierte en un instrumento muy importante para el desarrollo de una disciplina científica. La estadística inductiva se basa directamente en la teoría de la probabilidad, que es una rama de las matemáticas. Tenemos, pues, así, una disciplina puramente deductiva que proporciona una base racional para el razonamiento inductivo. Que el autor sepa, no existe otra base racional alguna para la inducción. Este punto general se examinará con mayor detalle en el capítulo VIII.

Existen algunas razones de orden práctico en cuya virtud resulta a veces necesario tratar de generalizar sobre la base de una información limitada. La más obvia de ellas es la del factor tiempo-costo. Sería absolutamente impracticable, y no digamos ya prohibitivamente costoso, preguntar a cada elector cómo se propone votar, con objeto de predecir en esta forma el resultado de una votación nacional. Ni puede el investigador corriente permitirse visitar a todos y cada uno de los residentes de una gran ciudad para estudiar sus prejuicios, la movilidad social o cualquier otro fenómeno por el estilo. Lo primero que hace, en efecto, es decidir la naturaleza exacta del grupo que se propone generalizar ("la población"). Puede escoger a todos los ciudadanos en edad de votar, o todos los varones blancos de dieciocho años cumplidos, que viven en los límites de la ciudad de Detroit. En tal caso suele por lo regular extraer una muestra consistente

en una proporción relativamente pequeña de las personas en cuestión, pero interesándose ante todo no en esa muestra particular, sino en la población más numerosa de la que ha sido extraída. Puede encontrar, por ejemplo, que, en esa muestra particular de 200 varones blancos, existe una relación negativa entre la educación y el prejuicio. Aun admitiendo que en otro conjunto de 200 individuos muestreados el resultado pudo haber sido totalmente distinto, propenderá sin embargo a establecer ciertas inferencias acerca del carácter de la relación en el caso de haberse estudiado la población entera de los varones blancos adultos en Detroit.

Otra razón que lleva a generalizar sobre la base de una información limitada es la de que puede ser imposible utilizar a toda la población, porque ésta sea infinita o difícil de definir. Al replicar un experimento en las ciencias naturales o sociales, el objetivo parece ser siempre cierta clase de generalización de la que se espera que se verificará "en circunstancias similares". O bien un especialista en ciencias sociales puede haber reunido datos de todos los casos de que dispone. Puede haberse servido, por ejemplo, como unidades de análisis, en un estudio sobre la migración interior, de todos los 50 estados [de los Estados Unidos], deseando sin embargo generalizar acerca de la migración en condiciones "semejantes". En cada uno de dichos casos, la situación requiere el recurso a la estadística inductiva.

Llegados a este punto, alguien pondrá tal vez una pregunta por el estilo de ésta: "si la estadística es tan importante, ¿cómo es que ciencias como la física y la química, por ejemplo, hayan podido progresar tanto sin el empleo extenso de las técnicas estadísticas? ¿Difieren acaso éstas en algo?" Es obvio que sí lo hacen. Algunas de las ciencias naturales se han desarrollado, sin duda, por espacio de siglos sin el empleo de la estadística inductiva. Pero esto parece ser ante todo cuestión de suerte o, para reconocer el mérito de los esfuerzos de los científicos, se da un control relativamente satisfactorio de los elementos perturbadores del medio. En efecto, tal como se pondrá de manifiesto en capítulos ulteriores, en la medida en que imperan condiciones de laboratorio escrupulosamente controladas, la necesidad práctica de las técnicas estadísticas es menor. En este sentido, la estadística es el sustitutivo, para el indigente, de los experimentos complicados de laboratorio en los que se han tenido en cuenta todas las variables relevantes importantes. Hay que subrayar, con todo, que muchos de los mismos principios estadísticos se aplican a los experimentos de laboratorio en materia de física, a los experimentos algo menos precisos en materia de agricultura y a las investigaciones sociales. Así, por ejemplo, si un experimento en física se ha replicado 37 veces con los mismos resultados, es perfectamente concebible, sin embargo, que ensa-

yos subsiguientes den resultados distintos. Por consiguiente, el científico ha de generalizar sobre la base de un número limitado de experimentos, y las inferencias que establece son en esencia estadísticas por su carácter. En forma análoga, el problema del error de medición puede concebirse también en términos de estadística. En efecto, por muy preciso que sea el instrumento de medición, el científico nunca obtiene exactamente el mismo resultado con cada replicación. Puede atribuir dichas diferencias ya sea a error de medición o a efectos perturbadores de algunas variables incontroladas. La estadística se hace especialmente necesaria cuando de una replicación a otra las diferencias son tales, que ni se las puede ignorar ni atribuir a error de medición. Por lo tanto, fundamentalmente, la inferencia estadística puntualiza todas las generalizaciones científicas, aunque la necesidad de una preparación estadística y el empleo de técnicas estadísticas complicadas varíe considerablemente de un campo de actividad a otro.

1.2. El lugar de la estadística en el proceso de la investigación

La importancia de la estadística en el proceso de la investigación se exagera en ocasiones debido al destacado lugar que ocupa en los planes de estudios de graduación. La estadística misma no comprende problemas de medición, tales como la elaboración de índices o la puntuación de las preguntas de un cuestionario. Comprende, antes bien, una manipulación de cifras, partiendo del supuesto que se han cumplido determinados requisitos en el proceso de medición. De hecho, las consideraciones estadísticas sólo se introducen en la fase de análisis del proceso de investigación una vez que se han reunido todos los datos, al principio de la misma, cuando se proyectan los planes iniciales del análisis y cuando se ha de extraer una muestra.

Mientras que la indicación que acaba de hacerse en el sentido de que la estadística sólo entra en las fases técnicamente correctas del análisis y del muestreo del proceso de investigación, podría con todo inducir a error, a menos que fuera preciso. No significa ciertamente que el científico en materia social pueda planear y llevar a cabo su investigación entera sin conocimiento alguno de estadística, y ponerla luego en manos del estadígrafo diciéndole: "He aquí, mi labor está terminada: ahora, analícela usted." Si así lo hiciera, los resultados probablemente serán poco satisfactorios, cuando no inútiles por completo. Es obvio, en efecto, que los problemas que habrán de encontrarse en el análisis han de anticiparse en cada etapa del proceso de investigación, y en este sentido las consideraciones estadísticas hallan aplicación a todo lo largo del mismo. Un análisis estadístico, por muy elaborado que sea, raramente o nunca llegará a compensar

las fallas de un proyecto mal concebido o de un instrumento de recolección de datos deficiente. Este último punto merece un comentario especial. Significa, en efecto, que la estadística puede ciertamente constituir un auxiliar valioso de un acertado discernir juicioso, pero nunca, en cambio, un sustituto del mismo. Desde el punto de vista del sociólogo no es más que un instrumento.

Dicho lo anterior, agregaré que la estadística resulta en los exámenes exploratorios una herramienta mucho más flexible de lo que podría imaginarse. Buena parte de la investigación social se basa en ideas teóricas sumamente tentativas, las que no constituyen una guía precisa en función de las interrelaciones que cabe esperar, de las variables que han de ser controladas en el análisis, o incluso de las prioridades y secuencias a que han de sujetarse las etapas del análisis. Con frecuencia se sorprenden los estudiantes ante la complejidad que adquiere el análisis de datos, tan pronto como se introducen en el cuadro hasta una media docena de variables. Es especialmente en estos casos cuando un conocimiento de la teoría estadística de diseños experimentales, o de la técnica de la estimación mediante ecuaciones simultáneas pasa a ser un instrumento valioso, mediante el cual pueden ser clarificadas algunas relaciones de una gran complejidad. Los métodos verbales o intuitivos resultan absolutamente inadecuados. En un texto general, tal como el presente, sólo pueden abordarse temas de diseño experimental y análisis multivariado, pero es importante tener en cuenta que hay numerosas materias mucho más avanzadas, las que han mostrado su valía incluso en aquellas investigaciones exploratorias cuyo propósito consiste en determinar la importancia relativa de numerosos factores, al objeto de reducir de manera sistemática el margen de alternativas, creando hipótesis más precisas para su uso en investigaciones ulteriores.

I.3. Advertencia

En presencia de un número o de una ecuación matemática, algunos estudiantes experimentan un temor que va desde una ligera aprehensión hasta la inhibición mental completa. Si el lector es de éstos, deberá tratar especialmente de deponer toda idea por el estilo acerca de que "la estadística es algo que ya sé que nunca llegaré a entender". En efecto, el grado de matemáticas requerido en este texto es tal, que los cursos de álgebra de la escuela secundaria, añadidos a las pocas operaciones algebraicas elementales que se exponen en el Apéndice 1, constituyen una preparación suficiente. Hay que recordar, con todo, que los textos de matemáticas y estadística no se leen como una novela. Por lo regular, en efecto, la materia se presenta en forma muy conden-

sada. De ahí, pues, que se requieran una lectura atenta y una disposición de espíritu activa, y no simplemente pasiva, frente al material presentado. Esta es la razón de que no se pueda prescindir de un trabajo cotidiano y de la resolución de los problemas prácticos incluidos al final de cada capítulo.

BIBLIOGRAFÍA

1. Downie, N. M. y R. W. Heath: *Basic Statistical Methods*, 2ª ed. Harper and Row, Publishers, Incorporated, Nueva York, 1965, caps. 1 y 2.
2. Hagoood, M. J. y D. O. Price: *Statistics for Sociologists*, Henry Holt and Company, Inc., Nueva York, 1952, caps. 1 y 2.
3. Hammond, K. R., y J. E. Householder: *Introduction to the Statistical Method*, Alfred A. Knopf, Inc., Nueva York, 1962, cap. 1.
4. Hays, W. L.: *Statistics*, Holt, Rinehart and Winston, Inc., Nueva York, 1963, pp. 1-12.
5. Tippett, L. H. C.: *Statistics*, 2ª ed., Oxford University Press, Nueva York, 1956.
6. Walker, H. M.: *Mathematics Essential for Elementary Statistics*, Henry Holt and Co., Inc., Nueva York, 1951.
7. Wallis, W. A. y H. V. Roberts: *Statistics: A New Approach*, The Free Press of Glencoe, Ill., Chicago, 1956, caps. 1-3.

II. TEORÍA, MEDICIÓN Y MATEMÁTICAS

ESTE capítulo tiene por objeto esbozar en líneas generales las relaciones existentes entre las proposiciones teóricas, las hipótesis empíricas, la medida y los modelos matemáticos. Muchos de los problemas tratados en este capítulo no suelen examinarse en conexión con los cursos de estadística, lo que se debe en parte a la tendencia poco afortunada consistente en dividir la materia en cursos con las apelaciones de "teoría", "métodos de investigación" y "estadística". Esto ocasiona que las relaciones internas entre dichas materias resulten a veces oscurecidas. Con objeto de situar a la estadística en la debida perspectiva, conviene prestar atención a las relaciones entre las proposiciones teóricas y las hipótesis de investigación por una parte, y entre estas últimas y los modelos matemáticos por la otra.

Se suele decir con frecuencia que el objeto de la investigación está en verificar hipótesis desarrolladas teóricamente y que los métodos estadísticos capacitan para efectuar dichas pruebas. Hay que tener presente, sin embargo, que los procesos implicados en pasar de la teoría a las hipótesis reales de investigación y de éstas a los enunciados de probabilidad del tipo empleado en la inferencia estadística no son en modo alguno directos. En efecto, en ambos casos hay que tomar decisiones, las cuales pueden dar lugar a un grado considerable de controversia. Examinemos primero el carácter de las decisiones que se requieren para desarrollar, a partir de proposiciones teóricas, hipótesis verificables.

*II.1. Teoría e hipótesis: definiciones operativas

En el instante en que empezamos a diseñar un proyecto de investigación enderezado a verificar una proposición que puede aparecer en un trabajo teórico, resulta evidente que hay que hacer varias cosas antes de poder proceder a la prueba. Tomemos a título de ejemplo concreto la siguiente proposición: "Cuanto más elevada es la condición social de una persona, tanto menores son sus prejuicios en relación con los negros." Supongamos que la "condición social" se haya definido como la posición que la persona ocupa en relación con otras en la jerarquía social, y los "prejuicios" como tendencia latente a la discriminación de

* El asterisco que precede una sección, párrafo o ejercicio indica que la materia que contiene o es de comprensión difícil o trata de conceptos con los que probablemente no están familiarizados los estudiantes que sólo disponen de una preparación limitada en materia de metodología de investigación. El estudiante principiante puede perfectamente omitir dichos pasajes o leerlos superficialmente. Por su parte, el asterisco que precede al título de una sección indica que la sección entera puede omitirse si se quiere.

una minoría o como actitud negativa basada en juicios preferidos. Aun si se prefiere sustituir por otras las definiciones de esos dos conceptos, se descubrirá sin duda alguna que, cualesquiera que sean las definiciones escogidas, resulta imposible servirse directamente de ellas para decidir cuál sea exactamente la condición de Jones, pongamos por caso, o el grado de sus prejuicios.

La razón de ello reside en que la mayoría de las proposiciones son más bien teóricas que operativas. En la definición teórica, en efecto, un concepto se define en términos de otros conceptos que se dan por comprendidos. En el modelo ideal del sistema completamente deductivo, se tomarían ciertos conceptos sin definir (primarios), y todos los demás se definirían en términos de aquéllos. En la geometría euclidiana, por ejemplo, los conceptos de *punto* y *recta* pueden tomarse sin definir, pudiendo luego definirse las nociones de *ángulo*, *triángulo* o *rectángulo* en función de aquellos términos. Pese a que la elección de conceptos no definidos es hasta cierto punto arbitraria, el hecho de que tengan que existir siempre algunos conceptos primeros o primarios resulta de la necesidad inherente de definir los conceptos teóricos en términos unos de otros.

Por otra parte, las definiciones operativas son definiciones que enuncian efectivamente los procedimientos empleados en la medición ([8], pp. 58 a 65). La definición operativa de "longitud", por ejemplo, indicará exactamente cómo deba medirse el largo de un cuerpo. El ejemplo de una definición operativa del prejuicio implicará una prueba como la de la escala de la distancia social de Bogardus o, tal vez, una relación de conceptos anti-negros en una lista de 24 puntos, juntamente con instrucciones detalladas para recoger los datos, valorar los puntos, etcétera. Como quiera que toda medición implica como requisito mínimo alguna clasificación, la definición operativa puede considerarse como un conjunto detallado de instrucciones que permiten clasificar a los individuos en forma inequívoca. De este modo, la noción del grado de confianza o garantía de seguridad queda integrada en dicho concepto de la definición operativa. La definición ha de ser lo bastante precisa para que todas las personas que se sirvan del procedimiento lleguen a los mismos resultados. Lo que las definiciones teóricas del prejuicio y de la condición social consignadas más arriba no lo permitirán, por supuesto, directamente.

Sostenemos, pues, que en toda ciencia se utilizan dos tipos distintos de definiciones. Diversas maneras alternativas de enfocar la relación entre la teoría y la investigación conducen esencialmente a la misma conclusión. Northrop designa lo que hemos llamado definiciones teóricas como "conceptos por postulación", y las definiciones operativas como "conceptos por intuición" [9].

Por nuestra parte nos hemos servido de una terminología que parece implicar que hay dos maneras distintas de definir un "mismo" concepto, en tanto que Northrop prefiere referirse a dos tipos distintos de conceptos. Otros, todavía prefieren pensar en términos de índices, más que de definiciones operativas. El concepto de *índice* implica por lo regular que el procedimiento empleado da sólo un indicador imperfecto de alguna variable señalada que no es medible directamente. De acuerdo con este punto de vista, pues, hay dos cosas: una variable señalada y un indicador de esta variable. Pero, independientemente del punto de vista que cada cual prefiera, es indispensable comprender el carácter del nexo entre las dos clases de definiciones, de conceptos o de variables. Podemos preguntar si existe o no un método puramente *lógico* de juntar las dos clases de definiciones. Otra forma de plantear la cuestión consistiría en preguntar si existe o no modo lógico alguno de decidir si una definición operativa determinada (o un índice) mide "realmente" el concepto o la variable teóricamente definidos. La respuesta a ambas cuestiones parece ser negativa.

Northrop sostiene esencialmente que no hay manera alguna de asociar las dos clases de conceptos o definiciones, excepto por vía de convención o de común acuerdo. La gente en general está simplemente de acuerdo en que debería emplearse una determinada definición operativa como medida de un determinado concepto, si las operaciones parecen razonables sobre la base de la definición teórica. Puede presumirse que, si varias definiciones operativas son posibles, se escogerán aquellas que parecen más apropiadas y al mismo tiempo más seguras. El carácter de "apropiado" ha de juzgarse inevitablemente sobre la base de la comprensión que uno tiene de la definición teórica. Se emplea a veces el término de *validez aparente* para designar el carácter apropiado de un índice o de una definición operacional ([11], p. 165). Idealmente, según lo señala Bridgman, las operaciones y las definiciones teóricas habrían de asociarse sobre la base de uno a uno ([2], pp. 23 ss). O en otros términos, si cambiamos la operación, deberíamos servirnos de otro concepto. Sin embargo, semejante ideal es tal vez irreal en el estado actual de desarrollo de las ciencias sociales. Su aplicación conduciría sin duda alguna ya sea a una rigidez capaz de ahogar todo nuevo progreso metodológico o a una proliferación de conceptos teóricos [1].

¿Qué puede hacerse, pues? Podemos admitir la posibilidad de tener asociado un número de diversas operaciones o de índices a cada concepto teórico. Pero en tal caso podemos encontrarnos con una dificultad común: dichos procedimientos pueden dar resultados distintos. Uno de los procedimientos empleados para medir el prejuicio puede llevar a resultados que indiquen que

nuestra "hipótesis" ha sido confirmada. En tanto que, en otro caso, otro procedimiento puede conducir a la conclusión opuesta. En cierto sentido, así es como se opera el progreso, a condición de que no conduzca a una disputa interminable acerca de cuál procedimiento mide "realmente" el prejuicio (cuya esencia se supone comprendida). Con objeto de prevenir confusiones importa darse cuenta de que *la prueba efectiva se hace en términos de los conceptos tal como se los ha definido operacionalmente. Por lo tanto, las proposiciones que comportan conceptos definidos teóricamente no son verificables directamente.* Así, pues, si se dan dos definiciones operativas distintas del prejuicio se verificarán dos hipótesis distintas.

Se ha admitido que puede resultar deseable tener asociada más de una operación con cualquier concepto teórico dado, y se ha señalado que semejantes operaciones pueden conducir a resultados diversos. Estamos ahora en condiciones de proporcionar un criterio eficaz, pragmático, para una definición teórica empíricamente satisfactoria del concepto. Supongamos que tenemos un concepto definido teóricamente y varias definiciones operativas susceptibles de asociarse con dicha definición teórica. Sobre la base de esta última definición, la mayoría de los científicos estarán probablemente de acuerdo en que algunas de las operaciones deberían eliminarse por cuanto no se aplican a lo que está contenido en la definición teórica. Pueden decidir, por ejemplo, que las preguntas relativas a las tendencias delictivas o los gustos musicales no deberían emplearse para medir el prejuicio. Pero puede haber varias operaciones que ocupen más o menos el mismo lugar en la opinión de dichos jueces. En otros términos: sobre la base de la definición teórica, los expertos pueden no estar en condiciones de ponerse de acuerdo acerca de que un determinado procedimiento operativo debiera escogerse con preferencia a otros. Podemos decir en tal caso que, *en la medida en que dichos diversos procedimientos dan resultados diferentes* (en igualdad de circunstancias), *la definición teórica es deficiente*, en el sentido de que necesita probablemente revisión o aclaración. Por ejemplo: el concepto *prejuicio* puede acaso haberse definido de tal modo que resulte demasiado vago. Tal vez se considerará necesario distinguir entre varias clases o dimensiones del prejuicio, asociando operaciones distintas a cada una de ellas. En una forma por el estilo de ésta —que se la reconozca explícitamente o no—, el proceso de investigación puede utilizarse para ayudar a aclarar los conceptos teóricos.

Parece, pues, haber en esta forma, dos lenguajes distintos, relacionados por una especie de diccionario al que se ha llegado por consenso, que permite asociar los conceptos de uno de ellos con los del otro. Los científicos piensan en el lenguaje teórico y realizan sus experimentos en el lenguaje operativo. No es ne-

cesario asociar operaciones con todos los conceptos del lenguaje teórico. Sin embargo, importa percatarse de que los conceptos que no han sido definidos operativamente no deberían por lo regular aparecer en enunciados que pretenden constituir hipótesis comprobadas. En efecto, si esto ocurre, las cuestiones planteadas por las "hipótesis" carecerán por lo regular de sentido desde el punto de vista operativo y conducirán probablemente a un debate interminable.

II.2. El nivel de medición: escalas nominales, ordinales y de intervalo

Acabamos de ver que el proceso consistente en pasar de los conceptos definidos teóricamente a los definidos operativamente no es en modo alguno directo. En efecto, al asociar un tipo de concepto con el otro, han de tomarse ciertas decisiones. Y en forma análoga, el proceso enderezado a seleccionar el modelo matemático o estadístico apropiado para emplearlo en una técnica de investigación determinada o en un procedimiento operativo comporta asimismo cierto número de decisiones importantes. Podría acaso pensarse que, una vez que un fenómeno ha sido medido, la elección de un sistema matemático sería cosa de simple rutina. Esto depende de lo que se entiende por *medida*. Si empleamos el término para referirnos únicamente a aquellos tipos de medición usualmente empleados en una ciencia como la física (v.gr. la medición de la longitud, del tiempo o la masa), entonces la elección de un sistema matemático no constituye prácticamente problema. Pero si ampliamos el concepto de la medición para incluir en él ciertos procedimientos menos precisos de empleo corriente en las ciencias sociales como se hará en este texto, entonces el problema se hace más complejo. Podemos, pues, distinguir entre distintos niveles de medición, y habremos de encontrar diversos modelos estadísticos apropiados a cada uno de ellos.¹

Escalas nominales. La operación básica y a la vez más sencilla en toda ciencia es la de la clasificación. Al clasificar tratamos de separar elementos desde el punto de vista de determinadas características, decidiendo acerca de cuáles son más semejantes y cuáles más distintos. Nuestro propósito consiste en agruparlos por categorías que sean lo más homogéneas posible en comparación con las diferencias entre las categorías. Si la clasificación es útil, se verá que las categorías son también homogéneas con respecto a otras variables [10]. Así por ejemplo, agrupamos unas personas de acuerdo con sus respectivas religiones (metodistas, presbiterianos, católicos, etc.) y vemos si la religión guarda al-

guna relación con el prejuicio o el conservadurismo político. Podríamos acaso hallar que los presbiterianos tienden a ser más conservadores que los católicos, siendo las puntuaciones de aquéllos relativamente altas en comparación con éstos. Si se hubieran seleccionado los individuos según el color del pelo, criterio de clasificación perfectamente adecuado, probablemente no se habrían encontrado diferencias significativas entre las clases en relación con otras variables estudiadas. En otros términos: las diferencias *entre* las clases de color del pelo habrían sido ligeras en comparación con las diferencias *dentro* de cada categoría.

Así, pues, la clasificación es fundamental para toda ciencia. Todos los demás niveles de medición, cualquiera que sea su precisión, comprenden básicamente la clasificación como operación mínima. Podemos, pues, considerar la clasificación como el nivel más bajo de medición, en el sentido más amplio del término. Damos a las categorías nombres arbitrarios, a manera de etiquetas convenientes, sin formular supuesto alguno acerca de las relaciones entre aquéllas. Así por ejemplo, colocamos a los presbiterianos y a los católicos en categorías distintas pero no suponemos que los unos sean "mayores que" o "mejores" que los otros. A condición de que las categorías sean exhaustivas (que comprendan todos los casos) y no se superpongan o se excluyan mutuamente (que ningún caso figure en más de una categoría), tenemos las condiciones mínimas necesarias para la aplicación de los métodos estadísticos. Se ha utilizado la expresión de *escala nominal* con referencia a ese nivel, el más simple de todos, de medición. Desde el punto de vista formal, las escalas nominales poseen las propiedades de simetría y transitividad. Por simetría entendemos que una relación que sea verdad entre *A* y *B* lo es también entre *B* y *A*. En tanto que por transitividad entendemos que si $A = B$ y $B = C$, entonces $A = C$. Resumido, esto significa simplemente que si *A* está en la misma clase que *B*, o en una clase distinta, *B* está en la misma clase que *A*, o en una clase distinta, respectivamente, y que si *A* y *B* están en una misma clase y *B* y *C* también en una misma clase entonces *A* y *C* han de estar también en la misma clase.

Habría que señalar que los números pueden asociarse arbitrariamente con cada categoría, pero esto no autoriza en ningún modo el empleo de las operaciones aritméticas usuales con dichos números. La función de los números, en este caso, es exactamente la misma que la de nombres, esto es la de designar las categorías. Es obvio que no tendría sentido alguno adicionar cifras de seguridad social y números de cuartos en un hotel. Pese a que nunca caeremos en la tentación de efectuar una operación tan ridícula como ésta se dan casos, sin embargo, en la investigación científica social, en que el absurdo no resulta en modo alguno tan obvio. Así pues, pese a que los valores numéricos

¹ Para exámenes más detallados de estos distintos niveles de medición véanse [5], [7], [12] y [13].

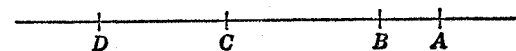
puedan atribuirse arbitrariamente a diversas categorías el empleo de ciertas operaciones matemáticas de las más corrientes (suma, resta, multiplicación y división) requiere, con todo, la ejecución de ciertas operaciones *metodológicas* en el procedimiento de clasificación. Tendremos en breve ocasión de ver cuál deba ser el carácter de dichas operaciones.

Escala ordinal. Resulta a menudo posible ordenar las categorías según el grado en que poseen una característica determinada, sin que por ello estemos en condiciones de decir cuántas poseen. Así pues, nos representamos un simple continuo a cuyo largo pueden ordenarse los individuos. Tal vez podamos colocar a los individuos de tal modo que nunca estén dos en el mismo lugar del continuo. Sin embargo, por lo regular existirá cierto número de conexiones. En tal caso no estamos en condiciones de distinguir entre determinados individuos, y los hemos agrupado juntos en una misma categoría. Pero estamos en condiciones, con todo, es decir que todos esos individuos tienen puntuaciones superiores a las de otros individuos determinados. Podemos, por ejemplo, clasificar familias conforme a su respectivo estado socioeconómico, en "superior", "media superior", "media inferior" e "inferior". Podríamos incluso limitarnos a dos categorías, la "superior" y la "inferior".

El tipo de medición que estamos examinando se sitúa manifestamente a un nivel algo superior al que empleamos para obtener una escala nominal, ya que con ella podemos no sólo agrupar a los individuos en categorías separadas, sino ordenar además estas categorías unas con respecto a las otras. Designamos este nivel de medición como "escala ordinal". Además de poseer las propiedades simétricas de la escala nominal, la escala ordinal es asimétrica en el sentido de que algunas relaciones especiales pueden ser verdad entre A y B y no serlo, en cambio, entre B y A . Así por ejemplo, la relación "mayor que" ($>$) es asimétrica, por cuanto si $A > B$, no puede ser cierto que $B > A$. La transitividad, en cambio, sigue subsistiendo, ya que si $A > B$ y $B > C$, entonces $A > C$. Son estas propiedades, por supuesto, las que nos permiten colocar A , B , C ... a lo largo de un mismo continuo.

Importa advertir que el nivel ordinal de medición no nos proporciona información alguna acerca de la *magnitud* de las diferencias entre los elementos. En efecto sabemos solamente que A es mayor que B , pero no sabemos cuánto mayor sea. Como tampoco podemos decir que la diferencia entre A y B sea menor que la que hay entre C y D .² Por consiguiente, no podemos adicionar o restar distancias sino en un sentido muy restringido. Así por ejemplo, si tuviéramos las siguientes relaciones

² Se ha empleado el término *métricamente ordenado* para designar escalas en las que es posible ordenar la magnitud de las *diferencias* entre elementos. Véase [7].



podemos decir que la distancia

$$\overline{AD} = \overline{AB} + \overline{BC} + \overline{CD}$$

pero no podemos tratar de comparar entre sí las distancias \overline{AB} y \overline{CD} . En otros términos, cuando transformamos relaciones de orden en operaciones matemáticas, no podemos por lo regular servirnos de las operaciones usuales de suma, resta, multiplicación y división. En cambio, sí podemos emplear las operaciones "mayor que" y "menor que", siempre que esto presente alguna utilidad.

Escala de intervalo y de proporción. En el sentido restringido de la palabra, el término *medida* puede emplearse para designar casos en los que no sólo estamos en condiciones de ordenar objetos según el grado en que poseen una característica determinada, sino que podemos indicar asimismo la distancia exacta entre ellos. Si esto es posible, podemos obtener lo que se designa como una *escala de intervalo*. No resulta difícil comprender que el nivel de medición de escala de intervalo requiere el establecimiento de algún tipo de unidad física de medición que pueda considerarse por todos como una norma común y sea repetible, esto es, que pueda aplicarse indefinidamente con los mismos resultados. El largo se mide en términos de pie o metros, el tiempo en segundos, la temperatura en grados Fahrenheit o centígrados, el peso en libras o gramos, y el ingreso en dólares, etcétera. Por otra parte, no existen unidades semejantes de inteligencia, autoridad o prestigio, en las que todos los sociólogos puedan ponerse de acuerdo y de las que se pueda suponer que permanecen constantes de una situación a otra. Dada una unidad de medida, resulta posible decir que la diferencia entre dos marcas es de veinte unidades, o que una diferencia es dos veces mayor que otra. Esto significa que es posible adicionar o restar marcas en forma análoga a como añadimos pesas a una balanza o quitamos 6 pulgadas de una tabla con una sierra ([3], pp. 296 a 298). Y en forma semejante podemos adicionar los ingresos de marido y mujer, en tanto que carece de sentido adicionar sus cuotas de inteligencia (IQ).

Si además es posible situar en la escala un punto cero absoluto o no arbitrario, entonces tenemos un nivel de medición algo mayor, que suele designarse como *escala de razón* (*ratio scale*). En tal caso estamos en condiciones de comparar marcas sirviéndonos de sus proporciones. Podemos, por ejemplo, decir que una marca es dos veces más alta que otra. Si el punto cero fuera

arbitrario, como es el caso en las escalas de grados centígrados y de Fahrenheit, aquello no sería legítimo. Así, por ejemplo, no decimos que 70° Fahrenheit son el doble de 35° centígrados si bien podemos decir que la *diferencia* entre dichas temperaturas es la misma que entre 105° y 70° Fahrenheit. Sin embargo, prácticamente en todos los casos que conoce el autor, esta distinción entre escala de intervalo y escala de proporción es puramente académica, ya que es extremadamente difícil encontrar una escala legítima de intervalo que no sea al propio tiempo una escala de proporción. Esto se debe al hecho de que, una vez establecida la magnitud de la unidad, es casi siempre posible concebir cero unidades, pese a que nunca podamos hallar un cuerpo que no posea largo o masa, u obtener una temperatura de cero absoluto. Así pues, prácticamente en todos los casos en que se dispone de una unidad, será legítimo emplear todas las operaciones corrientes de la aritmética, incluidas las raíces cuadradas, las potencias y los logaritmos.

* Suscítanse algunas importantes cuestiones acerca de la legitimidad de servirse de escalas de intervalo en el caso de cierto número de variables sociológicas y sociopsicológicas. Infortunadamente, no resulta posible discutir estas cuestiones en detalle en un texto general como el presente, pese a lo cual vamos a mencionar brevemente alguna de ellas. Se sostiene en ocasiones que una variable como, por ejemplo, el ingreso no constituye en realidad, si se calcula en dólares, una escala de intervalo ya que una diferencia de \$ 1 000 posee un significado psicológico distinto según que se dé entre ingresos de \$ 2 000 y \$ 3 000 o entre ingresos de \$ 30 000 y \$ 31 000. Al parecer, este argumento confunde la cuestión. Porque lo que aquí se dice efectivamente es que el ingreso calculado en dólares y el "ingreso psicológico" (a suponer que se lo pueda medir en términos de alguna unidad) no se relacionan directamente o en forma lineal. Y esto es una cuestión de hecho que carece de importancia en relación con la cuestión de saber si existe o no una unidad legítima de medida.

* Al llevar a cabo la enumeración de actos de conducta, de personas, de ocupaciones, o de grupos de diversa índole, se obtendrán muchas escalas de razón. La proporción de actos criminales, por ejemplo, se obtiene contando el número registrado de tales actos y comparándolos con la base de población. La mayor parte de nuestros datos censales de ciudades, estados o regiones, se obtienen contando varias clases de gente y dividiendo las cifras así obtenidas por la base de la población: por ciento urbano, porcentaje de la fuerza de trabajo en situación de desempleo, tamaño promedio de las familias, porcentaje de no blancos, etcétera. La complejidad de la división del trabajo puede ser medida en función del número de ocupaciones diferentes, o bien puede obtenerse un índice de la complejidad organizativa contan-

do el número de oficinas sucursales. En ocasiones surgen discusiones acerca de si tales mediciones constituyen realmente "escalas de razón" (ver Coleman [4] para un excelente planteamiento de este problema). Si se toma el punto de vista estrictamente operativo, según el cual la medida utilizada constituye la definición de la variable de interés, cabrá poca duda de que se habrá así obtenido una legítima escala de razón, ya que han sido contadas unidades precisas, y tales unidades han sido tomadas como equivalentes (y por tanto intercambiables). De esta manera, si añadimos a una determinada población 1 000 negros y le restamos 1 000 blancos, haremos la suposición fundamental de que, en orden a la medida usada, no hay diferencia, bien sean unos u otros los negros o los blancos implicados. Por otra parte el punto cero está bien definido. La afirmación de que el porcentaje de no blancos en una ciudad es igual a cero no ofrece ambigüedad.

* En cuantas ocasiones surge una discrepancia acerca de la adecuación de las medidas enumeradas, o de si éstas legitiman la adopción de mediciones relativas al nivel proporcional, experimento una vehemente sospecha de que el problema básico tiene un carácter totalmente distinto, a saber: el de la relación entre la medida utilizada y la construcción teórica que intenta medirse. Por ejemplo: la proporción de desempleados puede ser utilizada como indicación del mal funcionamiento de la economía; un porcentaje minoritario, como indicador de una amenaza planteada por la minoría, o un porcentaje urbano como indicador de la influencia ejercida por los valores urbanos. En tales casos nunca podrá la estadística resolver *per se* una controversia, resultando por ello necesario soslayar los problemas básicos, suponiendo, por el contrario, que sólo nos interesa la variable que intentamos medir.

* Puede suscitarse otra cuestión a propósito de si es posible o no conseguir una escala de intervalo en materia de medición de la actitud. Se han efectuado varios intentos enderezados a conseguir dicho fin. En el método Thurstone de los intervalos de apariencia igual, se pide a los jueces que agrupen objetos en montones situados a distancias iguales a lo largo del continuo de la actitud ([11], pp. 359 a 365). Se discurre esencialmente diciendo que, si se da un alto grado de consenso entre los jueces, puede emplearse legítimamente una escala de intervalo. Este procedimiento, así se sostiene, es esencialmente el mismo que se emplea para obtener escalas de intervalo en otras disciplinas. Este argumento parece legítimo, a condición que se dé efectivamente un alto grado de consenso entre los jueces y a condición que éstos dispongan de un gran número de montones en los que se puedan clasificar los objetos. Así, por ejemplo, si se vieran obligados a clasificar los objetos en uno de tres o cuatro montones, podría-

mos contar con un grado elevado de consenso, debido simplemente a la tosquedad del instrumento de medición. Habría, en efecto, tal margen de variabilidad dentro de cada montón, que difícilmente podría sostenerse que los objetos de los diversos montones estaban a igual distancia unos de otros. Pero aun admitiendo una concordancia perfecta y la máxima libertad en el agrupamiento de los objetos en montones, aun así sigue el método de Thurstone presentando dificultades por lo que se refiere al concepto de la unidad de referencia. Se hace necesario postular que es la existencia de dicha unidad la que hace posible el acuerdo entre los jueces. Puede afirmarse razonablemente que, en este punto del desarrollo de la medición de la actitud, la mayoría de las técnicas dan unas aproximaciones muy mediocres de las escalas de intervalo. Probablemente de muchas de ellas no debiera siquiera considerarse que proporcionan escalas ordinales legítimas. Las consecuencias de ello por lo que se refiere al análisis estadístico se irán haciendo más claras a medida que vayamos avanzando.

II.3. Medición y estadística

Hemos visto que existen diversos niveles de medición, con sus propiedades peculiares cada uno. Debe observarse que estos distintos niveles forman una escala cumulativa ellos mismos. En efecto, la escala ordinal posee todas las propiedades de la escala nominal además de la ordinal. A su vez, la escala de intervalo posee todas las propiedades de las escalas nominal y ordinal y, además, una unidad de medida, en tanto que la escala de proporción presenta el nivel más elevado, ya que posee no sólo una unidad de medida, sino, además, un cero absoluto. El carácter cumulativo de estas escalas significa que, al analizar nuestros datos estamos siempre autorizados a descender uno o más grados en el nivel de medición. En efecto, si tenemos una escala de intervalo, tenemos al propio tiempo una escala ordinal, y podemos servirnos de esta circunstancia en nuestros análisis estadísticos. Esto resultará a veces necesario, cuando no dispongamos de técnicas estadísticas o éstas sean en algún modo deficientes en cuanto a manipular la variable como escala de intervalo. Sin embargo, al proceder así perdemos información. Así, por ejemplo, si sabemos que Jones tiene un ingreso de \$ 11 000 y Smith uno de \$ 6 000 y sólo nos servimos del hecho de que Jones cuenta con el mayor de los dos ingresos en cuestión, entonces desperdiciamos la información relativa a que la diferencia de los ingresos es de \$ 5 000. Por lo tanto, en la mayoría de los casos resultará ventajoso servirnos del nivel de medición más alto que podamos legítimamente adoptar.

¿Y qué puede decirse del proceso inverso consistente en subir

la escala de medición, pasando, por ejemplo, de la ordinal a la de intervalo? Estamos a veces tentados de hacerlo, ya que estaríamos en condiciones de servirnos de técnicas estadísticas más potentes. Es incluso posible que lo hagamos sin darnos cuenta en absoluto de lo que ha sucedido exactamente. Importa percatarse de que no hay nada en los procedimientos estadísticos o matemáticos de los que nos servimos en última instancia que nos permita verificar la legitimidad de nuestros métodos de investigación. *El empleo de un determinado modelo matemático supone que se ha alcanzado cierto nivel de medición.* La responsabilidad en cuanto a decidir si sus procedimientos operativos permiten o no el empleo de determinadas operaciones matemáticas recae exclusivamente sobre el investigador. Este ha de decidirse en primer lugar por el nivel de medición adecuado, y esto decidirá a su vez acerca del sistema matemático apropiado. En otros términos: un determinado modelo matemático puede asociarse a cierto nivel de medición conforme a las consideraciones examinadas en la sección precedente. Así, por ejemplo, las operaciones aritméticas corrientes sólo pueden emplearse por lo regular con las escalas de intervalo y de razón.

* Nos enfrentamos aquí una vez más con el problema de tener que traducir de un lenguaje a otro. El lenguaje operativo comporta determinadas operaciones físicas, tales como el empleo de una unidad de medida. El lenguaje matemático, a su vez, implica un sistema totalmente abstracto de símbolos y operaciones matemáticas, y es útil no sólo porque es preciso y está altamente desarrollado, sino debido también a que su carácter abstracto permite la aplicación a una gran variedad de problemas empíricos. Las matemáticas se sirven del razonamiento deductivo por el que se pasa de un conjunto de definiciones, supuestos y reglas de operación a un conjunto de conclusiones mediante un razonamiento puramente lógico. En sí mismas, las matemáticas nada nos dicen acerca de la realidad, ya que todas las conclusiones están contenidas en las definiciones, los supuestos y las reglas originales, no habiéndose determinado empíricamente. Así, pues, si han de ser de alguna utilidad para el científico las conclusiones matemáticas han de traducirse inversamente a los lenguajes operativo y teórico [5].

Sostenemos, pues, que no es legítimo servirse de un sistema matemático que comporta las operaciones de sumar o restar, si esto no está legitimado por el método de medición. Aunque el sentido de este hecho sólo nos resultará plenamente claro cuando empecemos a servirnos de las diversas escalas de medida, estamos diciendo en realidad que no podemos remontar legítimamente en la jerarquía de medición, a menos que el proceso mismo de la medición haya sido mejorado. Lo que ninguna manipulación matemática puede hacer. ¿Cómo decidimos, pues, cuál nivel de

medición es el legítimo? Infortunadamente, el problema no es tan sencillo como podría suponerse. Unos pocos ejemplos bastarán para dar una idea de la complejidad del mismo.

* Para ilustrar uno de estos problemas es necesario distinguir las escalas ordinales y de intervalo de la *escala parcialmente ordenada* que resulta de la combinación de dos o más escalas ordinales (o de intervalo) en un solo índice. Ocurre con frecuencia, en sociología y en las otras ciencias sociales, que aquello que por lo pronto parece ser una simple escala ordinal (o de intervalo) es en realidad una combinación de varias escalas ordinales (o de intervalo), con el resultado de que no puede hacerse una clasificación inequívoca de individuos sin adoptar previamente ciertas otras decisiones. Tomemos, por ejemplo, el caso de la condición socioeconómica. Por lo regular, solemos determinar la condición de una persona examinando cierto número de criterios distintos, tales como su ingreso, ocupación, educación, antecedentes familiares o la zona de residencia. Si *A* se clasifica mejor que *B* según todos y cada uno de dichos criterios, entonces *A* puede obviamente clasificarse como más alto que *B* por lo que se refiere a la condición general. Pero, ¿qué ocurre si *A* tiene un ingreso superior al de *B*, y éste, en cambio, un nombre de familia más prominente? ¿Cuál de los dos ocupa en este caso el mejor rango social? Tenemos aquí varias alternativas. La primera consiste en dejar de lado la noción de condición general y pensar en términos de dimensiones separadas de la misma, cada una de las cuales pueda acaso admitir un nivel ordinal de medición. Terminamos así no con una, sino con varias escalas ordinales, y la cuestión empírica está en saber hasta qué punto las distintas dimensiones puedan relacionarse entre sí. Por supuesto, si existe una relación perfecta entre todas las dimensiones, la cuestión se convierte en puramente académica, ya que *A*, si es superior a *B* en cada una de las dimensiones, lo será también en todas ellas. En la práctica, sin duda, esto no ocurre nunca.

* Nuestra segunda alternativa consiste en tratar de "forzar" la aplicación de una escala ordinal a los datos, adoptando algunas decisiones acerca del peso relativo de cada dimensión y de las equivalencias que ello implica. Así, por ejemplo, si podemos admitir que un año suplementario de instrucción equivale a \$1 338.49 de ingreso suplementario, podemos traducir las unidades educativas en unidades de ingreso, llegando así a una escala unidimensional. Obviamente, el problema de traducir los antecedentes familiares o el área de residencia es más complicado todavía. El método de medición que aquí estamos examinando comporta un tipo de construcción de índice. Baste decir que semejante construcción de índice comporta usualmente algunas decisiones arbitrarias a propósito de los pesos relativos que haya que atribuir. Si el sistema de ponderación se deja justificar, entonces puede

emplearse una escala ordinal; en caso contrario, subsiste la duda acerca de si los individuos pueden o no clasificarse legítimamente en relación con el rango.

* Uno de los métodos comúnmente empleados para obtener una escala ordinal consiste en servirse de uno o más jueces para clasificar a los individuos conforme a un criterio como, por ejemplo, el del poder o del prestigio. Supongamos, para simplificar, que no hay más que un juez y que se le ha impuesto la tarea de clasificar a los individuos según su "posición social" en la localidad. Suponiendo que la persona coopere, el método empleado nos garantiza la obtención de una escala ordinal *independientemente* de cómo los individuos se comparen realmente a los ojos del juez. Es posible que, si se hubiera empleado otro método, no se habría obtenido escala ordinal alguna. Si se hubiera utilizado una técnica de comparaciones apareadas, en la que se pronunciaran juicios entre cada combinación por pares, el juez podría haber tasado a Smith más alto que a Brown, a éste más alto que a Jones, pero a este último más alto que al primero, Smith; violando en esta forma la propiedad de transitividad de las escalas ordinales. El investigador ha de proceder ahora a una elección. Puede llegar a la conclusión de que existe una escala parcialmente ordenada de una clase u otra. O puede considerar que el juez es inconsecuente o comete "error". Como lo señala Coombs, este problema relativo a lo que haya que designar como error de medición es un dilema básico con el que se encuentra el sociólogo ([7], pp. 485 a 488). En términos generales, éste puede adoptar un alto nivel de medición y considerar las desviaciones del tipo que se acaba de señalar como errores de medición, o puede descender a un nivel más bajo de ésta.

* Puede ilustrarse el mismo dilema en el caso de la escala de Guttman. En el tipo perfecto de ésta, las preguntas tienen una propiedad cumulativa que justifica la adopción de una escala ordinal [14]. Las preguntas pueden ordenarse de modo que se vaya pasando de un límite inferior a un límite superior, de tal modo que el tipo exacto de respuesta de un individuo pueda reproducirse a partir de su puntuación total. Así, por ejemplo, si se tienen cinco problemas aritméticos que vayan del más fácil al más difícil, la persona que resuelva el más difícil estará también lógicamente en condiciones de resolver los más fáciles. Si resuelve correctamente tres de los problemas, éstos serán los tres más fáciles, fallando en los otros dos. En una escala perfecta de distancia social, las preguntas relativas al prejuicio pueden disponerse conforme al grado de la intimidad de contacto con la minoría considerada. Una persona que esté dispuesta a casarse con un negro, estará dispuesta, por supuesto, a vivir en la misma calle que uno de ellos; si lo acepta como vecino, no tendrá inconveniente en sentarse a su lado en el autobús. Así, pues, pode-

mos ver en la escala perfecta de Guttman que la persona que conteste afirmativamente cuatro preguntas habrá contestado exactamente las mismas que una persona con tres afirmaciones, *más una*. Si la escala sólo estuviera parcialmente ordenada, podría decirse que en ciertos aspectos A tiene más prejuicios que B, y en otros aspectos menos, ya que los dos individuos han aceptado combinaciones distintas de preguntas.

* Sin embargo, en la práctica raramente alcanzamos una escala perfecta de Guttman, si es que la alcanzamos alguna vez. En efecto, hay siempre algunas personas cuyo tipo de respuestas se desvía del tipo ideal. ¿Son éstas acaso inconsecuentes porque aceptan a un negro como vecino pero se niegan a sentarse a su lado en el autobús? Tal vez. Pero, por otra parte, tal vez no sea así. A menos que el investigador esté dispuesto a *suponer* que dispone de una escala ordinal legítima, no puede sostener que el individuo considerado cometa error. Y si el número de errores aumenta, empezamos a sospechar de nuestra escala. Por otra parte, siempre estamos dispuestos a tolerar cierto número relativamente pequeño de errores. Es este principio el que se halla a la base de la decisión relativa a aceptar la escala de Guttman como escala ordinal, si el número de errores, medido por el coeficiente de reproductibilidad, es muy pequeño. Sin embargo, conviene percatarse del hecho de que la decisión es hasta cierto punto arbitraria, y de que en última instancia nos enfrentaremos con el problema de decidir a qué debemos llamar error.

* Estos ejemplos deberían bastar para indicar que no siempre es cosa fácil decidir cuál tipo de escala pueda emplearse legítimamente. Desde un punto de vista ideal, habría que servirse de una técnica de reunión de datos que permita los niveles más bajos de medición, si éstos son los únicos que los datos admiten, antes que recurrir a técnicas que adapten violentamente la escala a los datos. Así, pues, el método de las comparaciones apareadas sólo dará una escala ordinal si el juez está efectivamente en condiciones de clasificar a los individuos. Por otra parte, si se le invita a colocarlos en un orden preciso de clasificación, habrá de hacerlo, así crea o no que esto puede lograrse legítimamente. Habiéndose servido de este último método de reunión de datos y no estando en condiciones de demostrar empíricamente que los individuos pueden ordenarse sin violentar los datos, habrá de *suponer* la existencia de un solo continuo.

Con objeto de insistir en el hecho de que toda técnica estadística considerada presupone siempre un nivel específico de medición, nos acostumbraremos a indicar siempre el nivel de medición requerido por cada procedimiento. Al elegir entre procedimientos alternativos una de las preguntas más importantes a formular es ésta: "¿Es legítimo aceptar el nivel de medición que una determinada técnica requiere?" Si no lo es, tal vez deba

encontrarse un procedimiento alternativo. Si la única consideración fuese el nivel de medición, se simplificaría el problema de la elección entre procedimientos alternativos.

Encontramos con frecuencia, sin embargo, que ciertos procedimientos que no tienen grandes exigencias en cuanto a la medida, y que por tal razón parecen preferibles, resultan menos satisfactorios en relación con otras características deseables. Así se ve uno enfrentado con decisiones difíciles, en las que está implícita la necesidad de sopesar la seriedad relativa de las diversas clases de presunciones violadas. En tales casos puede resultar deseable analizar nuestros datos mediante la aplicación de diferentes métodos, observando si las conclusiones así obtenidas difieren entre sí en forma considerable.

En este punto puede ocurrir que nuestro examen de estos diferentes niveles de medición y de los problemas de elección entre pruebas y mediciones alternativas, no nos ilustre gran cosa. Uno de los peligros de la estadística "de recetario" consiste en una excesiva simplificación de los criterios y los problemas implícitos en la adopción de decisiones relativas al análisis de datos. Es imposible exagerar la importancia que tiene, al utilizar cualquier técnica estadística, el tener presente las presunciones implícitas que el procedimiento requiere. En el curso del presente examen, una de las primeras preguntas a formularse es la relativa al nivel de medición que puede legítimamente aceptarse.

II.4. Organización del libro

La organización de los restantes capítulos viene determinada por ciertas consideraciones, la primera de las cuales consiste en presentar ante todo las ideas más simples, pasando gradualmente a las de mayor complejidad. Como cada una de las secciones presupone el conocimiento de materiales que previamente han sido tratados, resulta conveniente seguir esta organización, pasando por alto solamente los párrafos o secciones precedidos de asterisco. El capítulo xiv, por excepción, puede ser saltado en su totalidad, o bien englobado con las pruebas y procedimientos "no paramétricos" contenidos en los capítulos xvi y xviii. El capítulo xxi, acerca del muestreo, puede ser leído en relación con el capítulo ix relativo a la probabilidad, aun cuando el capítulo sobre muestreo contiene varias secciones que sólo podrán entenderse cuando hayan sido leídos los capítulos xi, xiii y xvi. Lo fundamental del capítulo xvii podrá ser asimilado sin haber previamente abordado el capítulo xvi sobre análisis de diferencias. Se recomienda, en general, que se estudien los distintos temas en el orden en que se presentan.

Los instrumentos estadísticos no son fácilmente agrupables bajo uno o dos apartados, y por tal razón los títulos que ostentan

las principales divisiones del libro son sólo parcialmente adecuados, limitándose por el contrario a centrar la atención primaria. La Segunda Parte se limita al tema de la estadística descriptiva, en tanto que en las partes Tercera y Cuarta el principal, aunque no el único, foco de atención, se refiere a la inducción, a la prueba de hipótesis y a la estimación de parámetros de población basada en datos de muestreo. En las partes Segunda y Tercera nos limitaremos casi por entero a los procedimientos que traen implícita una sola variable por vez, en tanto que en la Cuarta Parte pasamos a problemas más difíciles tales como el manejo simultáneo de dos o más variables.

Entrelazada en estas distinciones entre descripción e inducción, así como entre las estadísticas univariadas y bivariadas o multivariadas, se observa un tercer principio organizativo, a saber: el relativo a los niveles de medición para cada una de las variables. Muchos de los títulos de los capítulos señalan este nivel de medición, pero tal vez el mejor método para lograr una perspectiva resumida del contenido consista en acudir al cuadro de pruebas y mediciones que aparece en las guardas. En su primera columna aparecen los procedimientos a usar con variables simples. Vemos allí que en el capítulo III nos ocuparemos de las mediciones muy simples (porcentajes, proporciones y razones), utilizados tanto con las dicotomías como con las escalas generales nominales con más de dos categorías. Las pruebas de hipótesis relativas a escalas nominales simples serán estudiadas en los capítulos X, XI y XII. Las medidas (mediana, desviación cuadril) adecuadas para ser utilizadas con una escala ordinal, sencillas, serán examinadas brevemente en los capítulos V y VI, en tanto que en el capítulo X se discutirá una muy sencilla prueba (la binomial), aplicable a datos ordinales. Dedicaremos algo más de nuestra atención a las escalas de intervalos y razones, las que examinaremos en los capítulos IV a VII, relativos a procedimientos descriptivos univariados, y de nuevo en los capítulos XI y XII de la Tercera Parte, relativos a la estadística inductiva.

Comenzando con el capítulo XIII volveremos nuestra atención a las relaciones entre dos o más variables, lo que desde luego supone que habremos de ocuparnos del nivel de medición de la variable segunda (y adicional), tanto como de la primera. Las columnas 2 a 5 del cuadro ofrecen varias combinaciones en relación con el nivel de medición de las dos variables. Por ejemplo: la casilla superior de la columna 2 se refiere a aquellas situaciones en que se dan dos dicotomías relacionadas entre sí (por ejemplo, sexo contra preferencias políticas). En la segunda casilla de la columna 2 se tiene en cuenta la posibilidad de que la primera escala nominal tenga más de dos categorías (por ejemplo: protestantes, católicos y judíos). En la tercera casilla una de las variables es una dicotomía (por ejemplo, sexo) en tanto

que la segunda es una escala ordinal, y así sucesivamente. Hay una sola casilla sin ocupar, a saber: aquella en que una variable se mide en el nivel ordinal y la segunda en el nivel de intervalo o de razón. Aun cuando tales situaciones pueden por supuesto ser resueltas, carecemos de instrumentos realmente satisfactorios que no requieran una pérdida de información al reducir el nivel de medición en cualquiera de las dos variables. No hay necesidad alguna de llenar las casillas situadas sobre las que ocupan la diagonal del cuadro, ya que las cubren aquellas situadas bajo dicha diagonal.

Resulta prematuro en estos momentos poner a discusión cada una de las posibilidades relacionadas en el cuadro. El punto más importante a señalar es el de que el nivel de medición afectado constituye una de las consideraciones más importantes al llevar a cabo una elección entre diversos procedimientos. La elección es relativa, aunque no enteramente, simple, en tanto uno se encuentre limitado al uso de sólo dos variables. Resulta mucho más difícil en el caso del análisis multivariado, en el que con frecuencia puede uno trabajar con cinco o incluso hasta quince o veinte variables al mismo tiempo, y donde resalta como sumamente improbable que todas ellas sean medidas al mismo nivel, y donde a menudo es poco deseable utilizar demasiados métodos de prueba y medición. En los capítulos XV, XVI, XIX y XX se examinan estos problemas de análisis multivariado. En ciertos lugares, particularmente al final de los capítulos XIV y XX, se encontrarán, en forma resumida, algunas de las consideraciones relativas a la selección entre procedimientos alternativos.

Como se observará, no todas las combinaciones posibles son manejadas en este texto con el mismo grado de minuciosidad. Ocurre así, no sólo por limitaciones de espacio y por la necesidad de detenerse en el examen de las ideas fundamentales, sino porque la teoría estadística se encuentra mucho más avanzada en ciertos aspectos. Se ha trabajado mucho más, en particular, en el sector de la llamada "estadística paramétrica", relativa a las escalas de intervalo y de razón, que en el de los procedimientos ordinales, por lo cual nuestros instrumentos para el uso de las escalas de intervalo y de tiempo están mucho más desarrollados, especialmente en el caso del análisis multivariado. La diferencia entre las escalas de intervalo y razón no ha sido tampoco explotada en la teoría estadística, por lo menos hasta el nivel que a nosotros nos interesaría. La razón básica estriba en que los modelos estadísticos con los que generalmente trabajamos están basados en una ecuación lineal general que es aditiva, en lugar de abarcar razones o variables. Por tal motivo, y para todo objetivo práctico, no es necesario tener presente tal distinción según se avanza en la lectura. Será empero necesario acudir periódicamente al cuadro de las guardas.

GLOSARIO

El lector hará bien en acostumbrarse a explicar en sus propios términos el significado de los conceptos importantes. Los nuevos conceptos introducidos en este capítulo son:

- la escala de intervalo,
- la escala nominal,
- * la definición operativa,
- la escala ordinal y
- la escala de razón.

BIBLIOGRAFÍA

1. Blalock, H. M.: "The Measurement Problem: A Gap between the Languages of Theory and Research"; en H. M. Blalock y Ann B. Blalock (E.) *Methodology in Social Research*, McGraw-Hill Book Company, Nueva York, 1968, cap. 1.
2. Bridgman, P. W.: *The Logic of Modern Physics*, The Macmillan Co., Nueva York, 1938, pp. 1-39.
3. Cohen, M. R. y E. Nagel: *An Introduction to Logic and Scientific Method*, Harcourt, Brace and Company, Inc., Nueva York, 1937, caps. 12 y 15.
4. Coleman, James S.: *Introduction to Mathematical Sociology*, The Free Press, Nueva York, 1964, cap. 2.
5. Coombs, C. H., H. Raiffa y R. M. Thrall: "Some Views on Mathematical Models and Measurement Theory", *Psychological Review*, vol. 61, pp. 132-144, marzo de 1954.
6. Coombs, C. H.: *A Theory of Data*, John Wiley & Sons, Inc., Nueva York, 1964.
7. Coombs, C. H.: "Theory and Methods of Social Measurement", en L. Festinger y D. Katz (ed.), *Research Methods in the Behavioral Sciences*, The Dryden Press, Inc., Nueva York, 1953, pp. 471-535.
8. Lundberg, G. A.: *Foundations of Sociology*, The Macmillan Company, Nueva York, 1939, caps. 1-2.
9. Northrop, F. S. C.: *The Logic of the Sciences and the Humanities*, The Macmillan Company, Nueva York, 1947, caps. 5-7.
10. Radcliffe-Brown, A. R.: *A Natural Science of Society*, The Free Press of Glencoe, Ill., Nueva York, 1957, pp. 28-42.
11. Selltiz, C., M. Jahoda, M. Deutsch y S. W. Cook: *Research Methods in Social Relations*, Henry Holt and Company, Inc., Nueva York, 1959, caps. 5 y 10.
12. Senders, V. L.: *Measurement and Statistics*, Oxford University Press, Nueva York, 1958, cap. 2.
13. Stevens, S. S.: "Mathematics, Measurement, and Psychophysics", en S. S. Stevens (ed.), *Handbook of Experimental Psychology*, John Wiley & Sons, Inc., Nueva York, 1951, pp. 1-49.
14. Stouffer, S. A., et al.: *Measurement and Prediction*, Princeton University Press, Princeton, N. J., 1950, caps. 1 y 3.
15. Weiss, R. S.: *Statistics in Social Research*, John Wiley & Sons, Inc., Nueva York, 1968, cap. 2.

SEGUNDA PARTE

ESTADÍSTICA DESCRIPTIVA UNIVARIADA

III. ESCALAS NOMINALES: PROPORCIONES, PORCENTAJES Y RAZONES

Es MUCHO más sencillo resumir los datos que comportan escalas nominales que en el caso en que se emplean escalas de intervalo. La operación aritmética básica es, en el primer supuesto, la de contar el número de los casos al interior de cada categoría y de anotar sus tamaños relativos. Un grupo determinado puede constar de 36 varones y 24 mujeres, o de 25 protestantes, 20 católicos y 15 judíos. Sin embargo, para poder establecer comparaciones con otros grupos, hay que tener en cuenta el número de casos en cada uno de los grupos considerados. Las medidas que se examinan en el presente capítulo permiten establecer comparaciones entre diversos grupos, mediante normalización esencialmente en relación con el tamaño. Sin duda alguna, dos de las medidas en cuestión, la de las proporciones y la de los porcentajes, son ya conocidas de todos.

III.1. *Proporciones*

Con objeto de poder servirnos de las proporciones, hemos de presumir que el método de clasificación ha sido tal que las categorías son mutuamente exclusivas y exhaustivas. En otros términos: cada individuo ha sido puesto en una categoría y en una sola. Con fines de simplificación, tomemos una escala nominal que conste de cuatro categorías, con N_1 , N_2 , N_3 y N_4 casos respectivamente. Supongamos que el número total de los casos sea N . La proporción de casos en cualquier categoría dada está definida como el número en la categoría dividido entre el número total de casos. Por lo tanto, la proporción de individuos de la primera categoría se halla dada por la cantidad N_1/N , y las proporciones de las demás categorías son respectivamente de N_2/N , N_3/N y N_4/N . Es obvio que el valor de una proporción no puede ser mayor que la unidad. En efecto, como quiera que

$$N_1 + N_2 + N_3 + N_4 = N$$

tenemos que

$$\frac{N_1}{N} + \frac{N_2}{N} + \frac{N_3}{N} + \frac{N_4}{N} = \frac{N}{N} = 1$$

Así, pues, si adicionamos las proporciones de los casos en todas las categorías (mutuamente exclusivas), el resultado es la unidad.

Es ésta una propiedad importante de las proporciones que se deja extender fácilmente a cualquier número de categorías.

Ilustremos el empleo de las proporciones con los datos dados en el cuadro III.1.

CUADRO III.1. *Número de delincuentes y de no delincuentes en dos localidades hipotéticas*

Sujetos	Localidad 1	Localidad 2
Delincuentes		
Primer delito	58	68
Reincidentes	43	137
No delincuentes	481	1 081
Total	582	1 286

Resulta más bien difícil decir cuál de las dos localidades cuenta con mayor número de delincuentes, porque son diversamente grandes. En cambio, si expresamos los datos en términos de proporciones, podemos establecer una comparación directa. En efecto, la proporción de primeros delitos es, en la comunidad 1, de 58/582, o .100; la de la localidad 2, en cambio, es de 68/1 286 o sea .053. Las demás proporciones pueden calcularse en forma análoga, resumiendo los resultados en forma de cuadro (cuadro III.2). El cuadro en cuestión nos permite apreciar que los números relativos de delincuentes son muy parecidos en las dos localidades, pero que la segunda de ellas contiene un número considerablemente más bajo de primeros delitos y una proporción más alta de reincidentes.

CUADRO III.2. *Proporciones de delincuentes y de no delincuentes en dos localidades hipotéticas*

Sujetos	Localidad 1	Localidad 2
Delincuentes		
Primer delito	.100	.053
Reincidentes	.074	.107
No delincuentes	.826	.841
Total	1.000	1.001

La suma de las proporciones de la localidad 2 no da exactamente la unidad, debido a los errores de redondeo. En ocasiones es conveniente presentar los datos de tal modo que las sumas

sean exactamente igual a 1.000. Esto puede acaso exigir el ajuste de algunas de las proporciones de las categorías, en cuyo caso modificamos por convención las cifras de las categorías que comprenden el mayor número de casos.¹ El argumento en favor de ese procedimiento está en que un cambio en la última cifra decimal de una proporción mayor es relativamente menos importante que el mismo cambio en una cifra menor. Así, por ejemplo, podría cambiarse la proporción de los no delincuentes de la localidad 2 en .840, de modo que la suma resultante sea igual a la unidad.

El cuadro III.2 comprende proporciones del número total de casos en cada una de las comunidades. Supóngase, sin embargo, que el interés se centraba sobre todo en los delincuentes, y que deseábamos conocer la proporción de los reincidentes *entre los delincuentes*. El número total de delincuentes en las dos localidades es respectivamente de 101 y 205. Por lo tanto, entre los delincuentes, las proporciones de los reincidentes son respectivamente de 43/101, o .426 y 137/205, o .668. A primera vista estas cifras pueden proporcionar una impresión ligeramente diferente de la del primer conjunto de proporciones. Habríamos de guardarnos especialmente de concluir que el segundo espécimen es "más delictivo" que el primero. Por supuesto, este último conjunto de proporciones nada nos dice en absoluto acerca de las cifras relativas de no delincuentes en los dos especímenes considerados. Es obvio que no existe sustitutivo alguno de la lectura atenta de los cuadros. Constituye un buen principio acostumbrarse a determinar siempre las categorías que se hallan comprendidas en el número total de casos que sirve de denominador de la proporción. El lector debiera siempre preguntar: "*¿de qué es esto la proporción?*" Y la respuesta resultará clara del conjunto.

III.2 Porcentajes

Los porcentajes pueden obtenerse de las proporciones multiplicando simplemente por 100. La palabra porcentaje significa *por ciento*. Por lo tanto, al servirnos de los porcentajes normalizamos en relación con el volumen, calculando el número de individuos que habría en una categoría determinada si el total de los casos fuera 100, permaneciendo inalterada la proporción en cada categoría. Y como quiera que las proporciones sumadas dan la unidad, es obvio que los porcentajes sumarán 100, a menos que las categorías no sean mutuamente exclusivas o exhaustivas.

Al reproducir resultados, los porcentajes se emplean con mucha mayor frecuencia que las proporciones. Las cifras del cuadro III.2 habrían podido expresarse lo mismo en términos de

¹ Puede utilizarse exactamente el mismo procedimiento en el caso de porcentajes.

porcentajes. Mejor que servirnos de los mismos datos, tomemos otro cuadro que puede servir para ilustrar otros diversos aspectos. Supongamos que tenemos tres agencias de servicios domésticos con una distribución de casos como la que se indica en el cuadro III.3.

Como es usual, los porcentajes se han dado hasta el primer decimal y se han operado los ajustes de los últimos dígitos, de modo que los totales den exactamente 100. Aquí el número de casos de cada agencia es lo suficientemente grande como para justificar el empleo de porcentajes. Sin embargo, si el número de casos hubiera sido menor, el empleo de aquéllos habría resultado equivoco. En efecto, supóngase que la agencia C había tratado sólo 25 casos en total. Si hubiera habido cuatro madres solteras y siete parejas de novios, los porcentajes en dichas categorías habrían sido respectivamente del 16 y del 28 por ciento. Y como quiera que muchas personas acostumbran mirar sólo los porcentajes y no el número efectivo de casos comprendidos, podría fácilmente obtenerse la impresión de que había muchas más parejas de novios que de madres solteras. Como se verá cuando lleguemos a la estadística inductiva, la diferencia entre cuatro y siete casos puede deberse perfectamente a factores puramente casuales. El empleo de los porcentajes y las proporciones comporta por lo regular una estabilidad mucho mayor de las cifras. Por lo tanto, he aquí dos reglas generales importantes: 1) *indíquese siempre el número de casos juntamente con los porcentajes o las proporciones, y 2) no se calcule nunca un porcentaje, a menos que el número de casos en que está basado se halle a proxi-*

CUADRO III.3. *Distribución de los números y porcentajes de casos tratados por tres agencias hipotéticas de servicios domésticos*

Clase de casos	Agencia A		Agencia B		Agencia C		Total	
	Nº	%	Nº	%	Nº	%	Nº	%
Matrimonios	63	47.3	88	45.5	41	36.6	192	43.8
Divorciados	19	14.3	37	19.2	26	23.2	82	18.7
Novios	27	20.3	20	10.4	15	13.4	62	14.2
Madres solteras	13	9.8	32	16.6	21	18.8	66	15.1
Otros	11	8.3	16	8.3	9	8.0	36	8.2
Total	133	100.0	193	100.0	112	100.0	438	100.0

midad de los 50 o más. Si el número de casos es muy pequeño, será preferible indicar el número efectivo de ellos en cada categoría, sin recurrir a los porcentajes. En el caso anterior, por ejemplo, indicaríamos simplemente que la agencia C había tratado cuatro madres solteras y siete parejas de novios.

Véase ahora la columna del total que indica la distribución en porcentajes de las tres agencias juntas. Esas cifras se han obtenido sumando el número de casos de cada tipo y el número total de casos tratados por las tres agencias juntas. Para el cálculo de los porcentajes totales se utilizó, pues, como base un N de 438. Supóngase, sin embargo, que el número de casos no nos hubiera sido dado en el cuerpo del cuadro, sino que se hubiera presentado como en el cuadro III.4. En tal caso podría darse la tentación de obtener los porcentajes totales tomando directamente la media aritmética de los tres porcentajes de cada hilera. Semejante procedimiento no tendría en cuenta el hecho de que las tres agencias habían tratado números diferentes de casos; sólo se justificaría si los números de éstos fueran efectivamente iguales. El procedimiento correcto consistiría en ponderar cada porcentaje por el número correspondiente de casos. Uno de los medios para hacerlo consistiría en calcular hacia atrás para obtener el número efectivo de casos de cada casilla. Lo que podría efectuarse multiplicando el número total de casos tratados por la agencia por la *proporción* de una categoría determinada. Por ejemplo, $(133)(.473) = 63$.

CUADRO III.4. *Distribución en porcentajes de los casos tratados por tres agencias hipotéticas de servicios domésticos, con los porcentajes dispuestos verticalmente*

Clase de casos	Agencia A ($N = 133$)	Agencia B ($N = 193$)	Agencia C ($N = 112$)
	%	%	%
Matrimonios	47.3	45.5	36.6
Divorciados	14.3	19.2	23.2
Novios	20.3	10.4	13.4
Madres solteras	9.8	16.6	18.8
Otros	8.3	8.3	8.0
Total	100.0	100.0	100.0

Obsérvese que los porcentajes dados en los cuadros III.3 y III.4 tienen por objeto contestar a ciertas preguntas y no otras. Nos permiten examinar cada agencia por separado y ver la distribución de los casos tratados. Permiten además la comparación de las agencias entre sí en relación con los casos tratados. Así, por ejemplo, las agencias B y C trataron relativamente más madres solteras y personas divorciadas de las que trató la agencia A. Supóngase, sin embargo, que nos interesaban ante todo los casos de cierto tipo y el número relativo de ellos tratados por cada agencia. Así, por ejemplo, podría eventualmente interesar-

CUADRO III.5. *Distribución en porcentajes de los casos tratados por tres agencias hipotéticas de servicios domésticos, con los porcentajes calculados horizontalmente*

Clase de casos	Agencia A (N = 133) %	Agencia B (N = 193) %	Agencia C (N = 112) %	Total (N = 438) %
Matrimonios (N = 192)	32.8	45.8	21.4	100.0
Divorciados (N = 82)	23.2	45.1	31.7	100.0
Novios (N = 62)	43.5	32.3	24.2	100.0
Madres solteras (N = 66)	19.7	48.5	31.8	100.0
Otros (N = 36)	— *	— *	— *	— *

* Los porcentajes no se calculan cuando la base es inferior a 50.

nos saber el porcentaje de todos los matrimonios que pasaron por la agencia B. En estas condiciones resultaría más conveniente calcular los porcentajes a través del cuadro. En efecto, podríamos tomar el número total de matrimonios y ver cuáles porcentajes de dicha categoría fueron tratados respectivamente por las agencias A, B y C. Los porcentajes sumarían entonces 100 en el sentido horizontal del cuadro, y no en el vertical, y los resultados se resumirían como en el cuadro III.5.

De modo que los porcentajes pueden calcularse tanto en sentido vertical como en sentido horizontal. Por lo tanto, los cuadros han de examinarse siempre cuidadosamente para ver exactamente cómo se han calculado aquéllos. Para los casos en que la propia teoría nos dicta cuál es la variable que debe ser tomada como causalmente dependiente y cuál ha de ser considerada causalmente primaria o independiente, podrá bastarnos una simple regla empírica. Si tenemos la costumbre de situar la variable independiente en la parte alta del cuadro, y la variable dependiente al lado izquierdo, los porcentajes sumarán 100 hacia *abajo*, y las comparaciones se harán de *izquierda a derecha*. En el ejemplo relativo a la comparación de niveles de delincuencias en dos localidades, cabría normalmente suponer que ciertas características locales pueden tener influencia sobre la delincuencia, más bien que a la inversa.

Cuando computamos los porcentajes para que sumen 100 hacia abajo, lo que en realidad hacemos es normalizar los tamaños de las localidades, ya que reconocemos que los factores que se refieren a sus tamaños relativos, o los muestreos realizados dentro de cada localidad, *no* dependen causalmente de sus niveles de delincuencia. Al computar hacia abajo los porcentajes estamos controlando aquellos factores que afectan al tamaño de los dos muestreos. Este punto quedará más en claro una vez que hayamos considerado el concepto de inclinación de una línea recta

en la que una de las variables figura como dependiente de la otra (ver capítulo XVII).

Resultará que los porcentajes computados en la dirección sugerida pueden ser considerados como casos especiales de dichos declives.

III.3. Razones

La razón de un número A con respecto a otro número B se define como A dividido entre B. La cantidad que precede se pone en el numerador, en tanto que la que sigue forma el denominador. Supóngase que en una elección local se hallan inscritos 365 republicanos, 420 demócratas y 130 independientes en calidad de votantes. En este caso la razón de los republicanos a los demócratas es de 365/420, y la de los republicanos y los demócratas a los independientes es de $(365 + 420)/130$. Obsérvese que, a diferencia de la proporción, la razón puede tomar un valor superior a la unidad. Vemos asimismo que la expresión que precede o que sigue pueden constar, una y otra, de cantidades distintas (v.gr. republicanos y demócratas). Generalmente la razón se reduce a su expresión más simple eliminando en el numerador y el denominador los factores comunes. Así, pues, la razón de los demócratas a los independientes se escribirá como 42/13 o bien, en forma equivalente, como 42:13. En ocasiones es conveniente expresar la razón en términos de un denominador formado por la unidad. Por ejemplo, la razón de los demócratas a los independientes puede escribirse como 3.23 a 1.

Es obvio que las proporciones representan un tipo especial de razón en la que el denominador es el número total de los casos y el numerador una cierta fracción de aquél. Sin embargo, el término de *razón* se emplea por lo regular para referirse a casos en los que A y B representan categorías separadas y distintas. Podríamos, por ejemplo, establecer la razón de los delincuentes a los no delincuentes, o de los matrimonios a los novios. Es evidente que con cuatro o cinco categorías el número de razones posibles susceptible de calcularse es muy grande. En consecuencia, a menos que el interés se centre ante todo en uno o varios pares de categorías, será en general más económico y menos sujeto a confusión por parte del lector servirse de los porcentajes y las proporciones. Obsérvese que, si las categorías sólo son dos, será posible calcular la proporción directamente a partir de la razón y viceversa. Así, por ejemplo, si sabemos que la razón de los varones a las mujeres es de 3:2, entonces en cada cinco personas ha de darse un promedio de tres varones y dos mujeres. La proporción de los varones es, pues, de 3/5, o .6.

Las razones pueden expresarse en términos de cualquier base que resulte conveniente. La base de la razón está indicada por la

magnitud del denominador. Así, por ejemplo, las razones relativas al sexo se indican convencionalmente en términos del número de varones por 100 mujeres. Por lo tanto, una razón de 94 en materia de sexo indicará que el número de los varones es ligeramente inferior al de las mujeres, en tanto que una razón de sexos de 108 significaría una ligera preponderancia de los primeros. Las bases que comportan números grandes, tales como 1 000 o 100 000, se emplean a menudo al calcular *cuotas*, otro tipo de razón, cuando el empleo de las proporciones o los porcentajes conduciría a valores decimales pequeños. Las cuotas de natalidad, por ejemplo, suelen darse en términos del número de nacimientos vivos por 1 000 mujeres en edad de procrear. Las cuotas de asesinatos pueden darse en términos del número de asesinatos por 100 000 habitantes.

Las cuotas de crecimiento constituyen otro tipo corriente de razón. Al calcular una de estas cuotas, tomamos el crecimiento efectivo durante el periodo considerado, dividido entre el volumen al principio del periodo. Así, por ejemplo, si la población de una ciudad aumenta de 50 000 a 65 000 entre 1940 y 1950, la cuota de crecimiento durante el decenio en cuestión será de

$$\frac{65\,000 - 50\,000}{50\,000} = .30$$

o 30 por ciento. En el caso de cuotas de crecimiento, es obvio que los porcentajes se prestan bien más allá del 100 por ciento, en tanto que serán negativos si la ciudad ha experimentado un descenso de población.

GLOSARIO

Porcentaje
Proporción
Tasa
Razón

EJERCICIOS

1. Supóngase que se da el siguiente cuadro que muestra la relación entre la asistencia a la iglesia y el año de clase en una determinada universidad:

Asistencia a la iglesia	Año de clase				Total
	1er. Año	2º Año	Inferior	Superior	
Asistencia regular	83	71	82	59	295
Asistencia irregular	31	44	61	78	214
Total	114	115	143	137	509

- ¿Cuál es el porcentaje de asistencia regular en el conjunto? Respuesta, 57.96 %.
- ¿Cuál es la razón de los estudiantes de primer año a los del año superior?
- Entre los asistentes regulares, ¿cuál es la razón de los años inferiores a los superiores (de los 1º y 2º años a los años inferior y superior)? Respuesta, 1.09 a 1.
- ¿Cuál es la proporción de los asistentes irregulares entre los estudiantes del año superior? ¿La proporción de estudiantes de año superior entre los asistentes irregulares? Respuesta .364; .569.
- ¿Hay relativamente más asistentes irregulares entre los estudiantes de 1º y 2º años que entre los de las clases inferior y superior? Exprésense los resultados en porcentajes.
- Resúmanse los datos en varias proposiciones.

2. Al estudiar la relación entre la productividad industrial y el tipo de líder de los grupos, un psicólogo social obtiene los siguientes datos, que muestran los niveles de productividad agrupados en tres tipos distintos de dirección:

Productividad	Grupos de tipo de líder del grupo			Total
	Democrático	Liberal	Autoritario	
Alta	37	36	13	86
Mediana	26	12	71	109
Baja	24	20	29	73
Total	87	68	113	268

- ¿En qué dirección preferiría el lector calcular los porcentajes? ¿Por qué?
- Calcúlense los porcentajes y resúmanse los datos en forma breve.
- ¿Cuál es la razón de los productores de nivel alto a los de nivel bajo en cada uno de los grupos? En relación con estos datos particulares, ¿resumen las tres razones la situación de modo adecuado? Explíquese.

3. Si la razón de los blancos a los no blancos es de 8/5 en una determinada localidad, ¿cuál es la proporción de los no blancos? Supóngase que la razón de los blancos a los negros fuera de 8/5, ¿podría obtenerse la proporción de negros en la misma forma? ¿Por qué, o por qué no?

4. Si una ciudad tenía una población de 153 468 habitantes en 1940 y de 176 118 en 1950, ¿cuál fue la tasa de crecimiento (expresada en porcentaje) entre 1940 y 1950? Respuesta, 14.76 %.

5. Si en un determinado condado hay 12 160 varones y 11 913 mujeres, ¿cuál es la razón entre los sexos (expresada en términos del número de varones por 100 mujeres)?

BIBLIOGRAFÍA

1. Anderson, T. R. y M. Zelditch: *A Basic Course in Statistics*, 2ª ed., Holt, Rinehart and Winston, Inc., Nueva York, 1968, pp. 24 a 31.
2. Freeman, L. C.: *Elementary Applied Statistics*, John Wiley & Sons, Inc., Nueva York, 1965, cap. 4.
3. Hagood, M. J. y D. O. Price: *Statistics for Sociologists*, Henry Holt and Company, Inc., Nueva York, 1952, cap. 7.
4. Weiss, R. S.: *Statistics in Social Research*, John Wiley & Sons, Inc., Nueva York, 1968, cap. 4.
5. Zeisel, Hans: *Say It with Figures*, 5ª edición, Harper and Row, Publishers, Incorporated, Nueva York, 1968, caps. 1 y 2.

IV. ESCALAS DE INTERVALO: DISTRIBUCIONES DE FRECUENCIA Y REPRESENTACIÓN GRÁFICA

EN EL presente capítulo nos ocuparemos de métodos para el resumen de datos muy parecidos a los del capítulo precedente. Vamos a agrupar las escalas de intervalo en categorías, a ordenar éstas y a servirnos de dichos grupos para dar una visión conjunta de la distribución de los casos. Al proceder en esta forma, podemos reducir la información relativa a un número muy grande de casos a una forma muy simple, que permita al lector representarse en qué forma están distribuidos los casos. Más adelante comprobaremos que agrupando los datos podemos asimismo simplificar considerablemente ciertos cálculos. En los dos capítulos siguientes nos ocuparemos de métodos de resumen de datos en forma más compacta, de modo que puedan ser descritos por varios números expresando medidas que representan formas típicas y grado de homogeneidad.

IV.1. Distribuciones de frecuencia: agrupamiento de los datos

En el capítulo precedente nos hemos encontrado con sólo pocas decisiones importantes, si ha habido alguna, en relación con el resumen de los datos. Esto se debe al hecho de que, presumiblemente, las clases estaban ya determinadas y lo único que había que hacer era contar el número de casos en cada clase y luego normalizar en relación con el número de casos del espécimen conjunto, calculando una proporción, un porcentaje o una razón. En cambio, si los datos de la escala de intervalo han de resumirse del mismo modo, hay que adoptar una decisión inicial en relación con las categorías que se van a utilizar. Ya que por lo regular los datos estarán distribuidos de modo continuo, sin o con pequeñas lagunas, entre cifras contiguas, el esquema de clasificación puede ser muy arbitrario. Será menester decidir cuántas categorías se van a utilizar y en dónde deban establecerse los

39.2 %	11.6 %	36.3 %	26.3 %	37.1 %	15.3 %	27.3 %	23.5 %	13.3 %
28.1	26.3	27.1	35.1	23.0	26.1	31.0	36.3	27.3
22.8	33.4	25.6	21.6	46.8	7.1	16.8	26.9	46.6
44.3	58.1	33.1	13.4	27.8	33.4	22.1	42.7	33.0
36.3	20.7	9.3	26.3	29.9	39.4	5.3	24.3	17.8
18.2	37.1	21.6	17.5	12.3	23.6	37.2	37.1	25.1
27.1	28.8	27.8	33.6	26.5	28.3	26.9	24.8	41.0
33.6	19.3	43.7	28.2	19.9	83.6	47.1	4.8	9.7
39.5	32.3	22.4	15.1	26.3	26.1	29.2	14.3	14.6
21.6	37.9	37.1	24.9	10.0	20.7	11.8	22.9	36.0
46.1	21.5	13.3						

puntos de intersección. Infortunadamente, no existen reglas simples para hacer esto, ya que la decisión depende de los objetivos perseguidos por medio de la clasificación. Sirvámonos, para ilustrar el carácter del problema, de un ejemplo sencillo. Supóngase que los números indicados al final de la página anterior representan el porcentaje de electores elegibles que votan en la elección de un consejo escolar, en 93 colegios electorales de una determinada ciudad.

Los datos brutos presentados en esta forma no sirven prácticamente de nada en cuanto a proporcionar al lector una idea clara de lo que está sucediendo. Y esto es tanto más así cuanto mayor sea el número de los casos. Supóngase que deseáramos comparar dicha localidad con otra en relación con la participación electoral. Una rápida ojeada echada a los datos indica que la mayoría de los distritos tuvieron una participación de 20 a 40 % y que hubo uno con una cifra extremadamente alta.

Resulta sin embargo realmente difícil obtener una idea clara de la distribución total.

Número y magnitud de los intervalos. Con objeto de representarnos dicha distribución total, será útil clasificar las cifras vecinas en una misma categoría. Sin embargo, nos encontramos en seguida con un problema. ¿De cuántos intervalos habremos de servirnos al agrupar los datos? ¿Cuál ha de ser su extensión? Por lo pronto, no tiene objeto emplear intervalos de amplitud o límites peculiares. Así, pues, escogeremos más bien intervalos de amplitud 5, 10 o 20 que uno de amplitud 4.16, pongamos por caso. Y también nuestros puntos terminales, o límites de clase como se los suele llamar, serán por lo regular números redondos, tales como 5.0 o 10.0. Si tenemos duda acerca de los intervalos de los que habremos de servirnos definitivamente es preferible clasificar las cifras sirviéndonos de un número mayor de intervalos relativamente pequeños. La razón de ello es obvia: si nos servimos de intervalos pequeños, siempre podemos agrupar, inmediatamente, los casos en intervalos mayores. En tanto que si empezáramos con un pequeño número de intervalos grandes, no podemos luego subdividirlos, como no sea rehaciendo todos los cálculos. Por lo tanto, nos decidiremos probablemente a clasificar los datos en intervalos de amplitud 5 por ciento, como en el cuadro IV.1.

Y si examinamos ahora las frecuencias en cada categoría, vemos que la imagen que presentan es relativamente angulosa e irregular. Podemos probablemente explicarnos las variaciones entre categorías contiguas en términos de fluctuaciones casuales. Si hubiera habido más casos, habríamos podido contar con una distribución más suavizada. El razonamiento que se halla a la base de este juicio intuitivo se destacará más claramente en capítulos ulteriores. Baste de momento decir que empíricamente

CUADRO IV.1. *Distribución de la frecuencia, con datos agrupados en intervalos de 5 por ciento*

Intervalo	Frecuencia, <i>f</i>	Intervalo	Frecuencia, <i>f</i>
0.0 — 4.9	1	45.0 — 49.9	4
5.0 — 9.9	4	50.0 — 54.9	0
10.0 — 14.9	9	55.0 — 59.9	1
15.0 — 19.9	8	60.0 — 64.9	0
20.0 — 24.9	16	65.0 — 69.9	0
25.0 — 29.9	23	70.0 — 74.9	0
30.0 — 34.9	8	75.0 — 79.9	0
35.0 — 39.9	14	80.0 — 84.9	1
40.0 — 44.9	4		93

siempre parece ocurrir así. Sin embargo, dado nuestro *N* de 93 distritos, lo mejor que podemos hacer para obtener una distribución de aspecto más regular es servirnos de un número menor de intervalos más amplios. Sirviéndonos de intervalos de 10 en 10, obtenemos el cuadro IV.2.

CUADRO IV.2. *Distribución de la frecuencia, con datos agrupados en intervalos de 10 por ciento*

Intervalo	Frecuencia, <i>f</i>
0.0 — 9.9	5
10.0 — 19.9	17
20.0 — 29.9	39
30.0 — 39.9	22
40.0 — 49.9	8
50.0 — 59.9	1
60.0 — 69.9	0
70.0 — 79.9	0
80.0 — 89.9	1
	93

Si hubiéramos empleado intervalos mayores todavía, digamos, por ejemplo, de 20, el cuadro se presentaría como el cuadro IV.3. Aquí empezamos a oscurecer ya la mayor parte de nuestra información inicial. En efecto, sabemos sólo que aproximadamente las dos terceras partes de los casos se sitúan entre 20.0 y 39.9, pero viendo los datos en esta forma, no podemos decir mucho acerca de dónde se sitúe el grueso de los casos al interior de ese

CUADRO IV.3. Distribución de las frecuencias con datos agrupados en intervalos de 20 por ciento

Intervalo	Frecuencia, <i>f</i>
0.0 — 19.9	22
20.0 — 39.9	61
40.0 — 59.9	9
60.0 — 79.9	0
80.0 — 99.9	1
	93

intervalo realmente muy grande. En resumen, hemos de encontrar una forma a modo de servirnos de gran número de intervalos de modo que la visión no resulte demasiado detallada o irregular, ni servirnos de tan pocos que se pierda demasiada información. Y dicho sea de paso, observamos que, al resumir los datos de la escala de intervalo, *se pierde prácticamente siempre algo de información importante*. En tanto que, por otra parte, incluir toda la información conduce a presentar tanto detalle, que la visión resulta más bien oscurecida que aclarada.

Pese a que se han indicado fórmulas matemáticas que pueden servir de guía por lo que se refiere al número de intervalos a utilizar, esas fórmulas dan a menudo la impresión de exactitud, en tanto que la mejor decisión se basará normalmente en el sentido común y en el objeto a que se destine la tabla de frecuencia. Independientemente del número de casos o de la regularidad de la línea, lo más prudente consiste en seguir la regla práctica de que el intervalo no debería ser mayor que la magnitud de diferencia entre valores que pueden ignorarse sin perjuicio. Una diferencia de \$ 5 entre precios de casas, por ejemplo, es insignificante, en tanto que no es así si se trata de los precios de camisas. Por consiguiente, el intervalo deberá comprender los casos cuyos valores puedan considerarse para fines prácticos como semejantes.

Los datos indicados más arriba presentan otro problema. ¿Qué pasa con el único colegio que ostenta una participación del 83.6 por ciento a la vista? Si nos servimos de intervalos de una amplitud de 10, varias clases quedan vacías, con dicho único colegio abandonado, por así decir, a sí mismo. Sin duda, esto es lo que hay que hacer, si es que los datos han de resumirse cuidadosamente. Dicho colegio es efectivamente único. Por otra parte, en determinadas circunstancias puede ser conveniente abreviar la tabla. Si los porcentajes fueran bien más allá de 100 y si hubiera varios extremos que se extendieran por sobre de 10 o más intervalos, nos enfrentaríamos a una decisión más difícil todavía. En

tal caso, en efecto, se presentan varias alternativas. Primero, podemos servirnos de intervalos de *amplitudes diversas*, permitiendo que los intervalos extremos sean mucho más grandes que los otros. Así, por ejemplo, podríamos servirnos de un solo intervalo de 50.0 a 89.9, lo que comprendería las dos marcas mayores. Por supuesto, al proceder en esta forma perdemos información, ya que ahora tenemos una indicación mucho menos precisa de las cifras correspondientes a los dos casos extremos.

En segundo lugar, podríamos servirnos de un *intervalo abierto* para comprender los casos extremos. La última categoría podría leerse en tal caso como "50 por ciento o más". Aquí, sin embargo, perdemos todavía más información que anteriormente, aunque sabemos que en este ejemplo concreto los porcentajes no pueden ir más allá de 100. Pero si los datos se refirieran a ingresos y que el último intervalo fuera de "\$ 20 mil o más", el lector no tendría en absoluto manera alguna de adivinar, sobre la base de la sola tabla, cuáles pudieron haber sido los ingresos más altos. Conviene observar, con todo, que en determinadas circunstancias puede no revestir importancia alguna saber cuáles sean esos ingresos más altos. En tal caso, las simplificaciones introducidas mediante el empleo de intervalos abiertos pueden compensar con ventaja los inconvenientes. Con distribuciones que presentan un número reducido de casos muy extremos, puede no darse alternativa satisfactoria alguna. Si alguien desea, por ejemplo, indicar los ingresos de los ciudadanos más ricos sin desfigurar su tabla, le resultará más fácil hacerlo en el texto de su exposición. Como lo veremos en capítulos sucesivos, no debieran emplearse intervalos abiertos si el objetivo primero de la agrupación de los datos consiste en simplificar los *cálculos* y no en exponer aquéllos de modo significativo.

Límites verdaderos. El lector habrá observado que, al indicar los intervalos, los límites de las clases se han establecido de tal modo que éstas no se entrecorten. De hecho, existe un pequeño vacío entre una y otra. Los límites suelen por lo regular fijarse en esta forma para evitar toda ambigüedad frente al lector. En efecto, si se hubiera fijado como de 10 a 20, de 20 a 30, etcétera, se habría planteado la cuestión de qué hacemos con una marca de 20 exactamente. En realidad, siempre habrá ambigüedad, cualquiera que sea la forma en que se fijen los intervalos, como podemos apreciarlo al preguntarnos ahora qué habrá que hacer con un caso que se sitúe entre 19.9 y 20. Observamos, por supuesto, que no hay tales casos, pero un poco de reflexión nos convencerá de que esto es debido al hecho de que los datos se han *redondeado* a la décima del porcentaje más próximo. Por lo tanto, hemos de contestar a la siguiente cuestión: "¿cuáles casos corresponden en realidad a un intervalo determinado, puesto que los datos se han redondeado?" Vemos inmediatamente que los ver-

daderos límites de las clases no son los mismos que los que se han *fijado*. Si hubiéramos seguido las reglas convencionales del redondeo, un colegio con una participación ligeramente superior a 19.95 se habría redondeado en 20.0, situándolo en el intervalo de 20.0 a 29.9. Y si el porcentaje hubiera quedado por debajo de 19.95, por poco que así fuera, lo habríamos redondeado en 19.9, colocando el colegio en cuestión en la categoría inmediatamente inferior. Por lo tanto, los verdaderos límites efectivamente empleados son los siguientes:

de -0.05 a 9.95
de 9.95 a 19.95
de 19.95 a 29.95
etcétera.

Vemos que, al servirnos de los verdaderos límites, cada intervalo tiene una amplitud exactamente de 10.0 (más bien que de 9.9) y que el límite superior de un intervalo coincide exactamente con el límite inferior del siguiente.¹ Si la marca hubiera sido exactamente de 9.95000, habríamos seguido el procedimiento convencional redondeando hacia arriba, ya que el número dígito que precede al último cinco es impar.² Podemos, pues, asignar a cada caso, de modo inequívoco, su intervalo propio. Obsérvese que si el redondeo se ha operado hacia la cifra *próxima*, como suele ser el caso, el verdadero límite comportará siempre la separación de la diferencia entre los límites fijados de dos intervalos contiguos. Así, por ejemplo, si partimos la diferencia entre 19.9 y 20.0, obtenemos 19.95. La convención consiste en indicar las cifras de tal modo que se exprese el grado de exactitud de la medición, o sea que 10.45 indica una exactitud a dos lugares decimales, 10.450 a tres y 10.4 a uno. Dicho grado de exactitud debe indicarse siempre, de modo que el lector pueda averiguar los límites verdaderos si desea servirse de ellos en sus cálculos. Así, por ejemplo, si se indica que los límites son respectivamente 10.00 a 19.99, sabemos que la medición es exacta hasta dos decimales, que el redondeo se ha operado a la próxima centésima de $\frac{1}{100}$ del 1 por ciento, y que, en consecuencia, los verdaderos límites van de 9.995 a 19.995. Si los límites se hubieran indicado como

¹ Si el límite más bajo es cero y que los valores no pueden ser negativos (como en el caso de los porcentajes), consideramos de todos modos que todos los intervalos son de la misma amplitud, imaginando que el límite inferior del primer intervalo es en realidad -0.05 y que las marcas se han redondeado en 0.00.

² Obsérvese que en el caso de los intervalos de los que nos hemos servido habría una desviación muy ligera, ya que los casos que quedan exactamente entre intervalos se situarán siempre en la categoría superior. En la mayoría de los casos prácticos dicha desviación puede ignorarse.

10 a 19, entonces los verdaderos límites habrían sido, por supuesto, 9.5 a 19.5.

En unos pocos casos, como, por ejemplo, el de la edad en relación con el *último* aniversario, los datos pueden no haberse redondeado en la forma convencional. Sin embargo, si nos preguntamos a cuál intervalo corresponda un caso determinado, la respuesta habría de ser siempre clara. Como quiera que, en efecto, una persona que vaya a cumplir 20 años mañana cuenta hoy 19, es obvio que el intervalo fijado como de 15 a 19 tiene como verdaderos límites los valores 15 y 20. Pese a que pueda parecer que andamos con sutilezas al distinguir entre los límites indicados y los límites verdaderos, veremos, sin embargo, en los capítulos sucesivos que estos últimos han de utilizarse en los cálculos, aunque por lo regular no se indiquen explícitamente al presentarse los datos en forma de distribución de frecuencia.

Datos discretos y continuos. Los datos de los que nos hemos servido son *continuos*, en el sentido de que cualquier valor hubiera podido obtenerse teóricamente para un porcentaje, a condición que la exactitud de medición fuera lo suficientemente precisa y que los intervalos fueran muy grandes. Así, por ejemplo, el valor de 17.4531 por ciento es tan posible como el de 17.0000 por ciento. Algunos otros tipos de datos son *discretos*, ya que no todos los valores son posibles. En efecto, una mujer puede tener exactamente 0, 1, 2 o inclusive 17 niños, pero no puede tener 2.31 niños. El ingreso y el volumen de una ciudad son variables teóricamente discretas, ya que no es posible tener un ingreso de \$ 3 219.5618, o que una ciudad tenga una población de 43 635.7 habitantes. Debido a las limitaciones de todo instrumento de medición y a la necesidad subsiguiente de haber de redondear en un punto u otro, los datos empíricos vienen siempre en forma discreta; pero en muchos casos podemos por lo menos concebir una distribución continua susceptible de alcanzarse con un instrumento de medición perfecto. Como lo veremos en el capítulo relativo a la curva normal, los matemáticos han de desarrollar a menudo distribuciones teóricas que adoptan una variable continua.

En algunos casos, como los del ingreso o número de habitantes de una ciudad, no resulta demasiado difícil concebir los datos como continuos, aunque se trate en realidad de unidades muy pequeñas (centavos, personas) que no se dejan subdividir. Pero, ¿qué ocurre con el número de niños en una familia? Aquí pareceríamos violentar excesivamente los hechos si admitiéramos continuidad. Al presentar los datos en una distribución de frecuencia no se nos ocurrirá, por supuesto, servirnos de intervalos que vayan de 0.5 a 2.4 o de 2.5 a 4.4 niños. Emplearemos sencillamente intervalos como de 0 a 2, de 3 a 4, etcétera, y no habrá ambigüedad alguna por lo que se refiere a los huecos entre aqué-

llos. En algunos cálculos, sin embargo, será necesario, por razones pragmáticas, tratar los casos como continuos y disponer marcas discretas en intervalos pequeños. En efecto, por raro que se nos pueda antojar, podemos necesitar considerar a las madres con un hijo como en un intervalo de 0.5 a 1.5 niños. Para la mayoría de los objetos obtendremos los mismos resultados que obtendríamos manteniendo los datos en forma discreta. Con el fin de adaptarse a los modelos establecidos por los matemáticos, en este y otros casos será necesario hallar un compromiso con la realidad. A condición de que nos demos perfecta cuenta de lo que estamos haciendo, no resultará de ello confusión alguna o sólo muy poca.

IV.2. Distribuciones de frecuencia acumulativa

Para algunos objetos es conveniente presentar los datos en una forma algo distinta. En lugar de indicar el número de casos en cada intervalo, podemos indicar el número de marcas que son menores (o mayores) que un valor determinado. En el caso de los intervalos de los que nos hemos estado sirviendo, no hay, por supuesto, colegios electorales con una participación de votantes inferior a cero, hay cinco con menos del 9.95 por ciento, 22 con menos del 19.95 por ciento, y los 93 juntos tienen una participación inferior al 89.95 por ciento. Así, pues, podemos presentar los datos en forma acumulada, tal como se indica en el cuadro IV.4. Obsérvese que podemos acumular lo mismo hacia arriba que hacia abajo preguntando cuántos casos están *por encima* de un valor determinado. Las frecuencias acumulativas suelen indicarse por lo regular con una F mayúscula, en lugar de la mi-

CUADRO IV.4. Distribución de frecuencia acumulativa

Acumulación hacia arriba			Acumulación hacia abajo		
Número de casos por debajo de	Frecuencia acumulada, F	Por ciento	Número de casos por encima de	Frecuencia acumulada, F	Por ciento
0.0	0	0.0	0.0	93	100.0
9.95	5	5.4	9.95	88	94.6
19.95	22	23.7	19.95	71	76.3
29.95	61	65.6	29.95	32	34.4
39.95	83	89.2	39.95	10	10.8
49.95	91	97.8	49.95	2	2.2
59.95	92	98.9	59.95	1	1.1
69.95	92	98.9	69.95	1	1.1
79.95	92	98.9	79.95	1	1.1
89.95	93	100.0	89.95	0	0.0

núscula. Si queremos, podemos convertir las frecuencias efectivas en porcentajes. Tendremos ocasión de servirnos de las distribuciones acumulativas en el capítulo V al calcular las medias, así como más adelante en el capítulo XIV.

IV.3. Presentación gráfica: histogramas, polígonos de frecuencia y ojivas

Hay personas que sienten reparo en interpretar los cuadros y que captan mejor los materiales presentados en forma gráfica

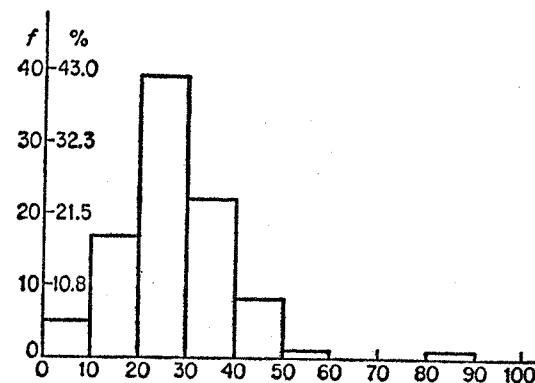


FIG. IV.1. Histograma de intervalos iguales.

o visual. Uno de los modos más sencillos y útiles de presentar los datos de tal manera que las diferencias entre las frecuencias se destaquen fácilmente consiste en servirse de figuras de áreas o alturas proporcionales a las frecuencias en cada categoría. Puede, por ejemplo, utilizarse una barra para representar cada categoría, indicando la altura de la misma su magnitud relativa. Si la escala es nominal, la ordenación efectiva de las barras no reviste importancia. Por lo que se refiere a las escalas ordinales y de intervalo, las barras pueden disponerse en su propio orden, con lo que dan una buena indicación visual de la distribución de la frecuencia. La figura resultante se llama *histograma*. La frecuencia absoluta o la proporción de los casos pueden indicarse a lo largo de la ordenada, como en la figura IV.1.

Hay que observar que si las *alturas* de las barras se toman como proporcionales a las frecuencias en cada intervalo de clase, el cuadro visual puede resultar confuso, a menos que todos los intervalos sean cerrados y de amplitud igual. Supóngase, por ejemplo, que uno de los intervalos centrales hubiera sido de ancho 20 en lugar de 10. Encontraríamos en consecuencia un

mayor número de casos en el intervalo, y el resultado sería como en la figura IV.2. Es obvio que si deseamos obtener un histograma que represente los datos en forma más adecuada, deberíamos dar a la barra la mitad solamente del alto, ya que he-

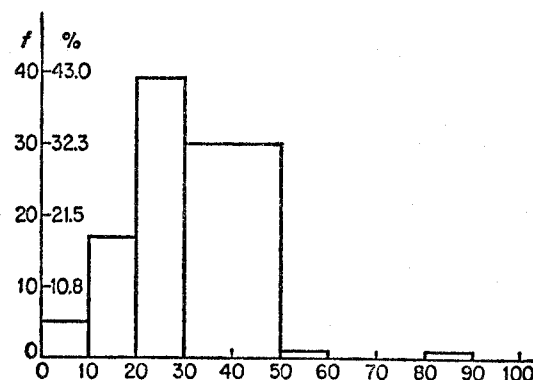


FIG. IV.2. Histograma de intervalos desiguales y alturas proporcionales a las frecuencias.

mos doblado el ancho y, en promedio, hemos incluido un doble número de casos en el intervalo mayor de lo que sería el caso en uno u otro de los dos intervalos de tamaño normal. Esto nos daría un histograma (véase figura IV.3) mucho más semejante al

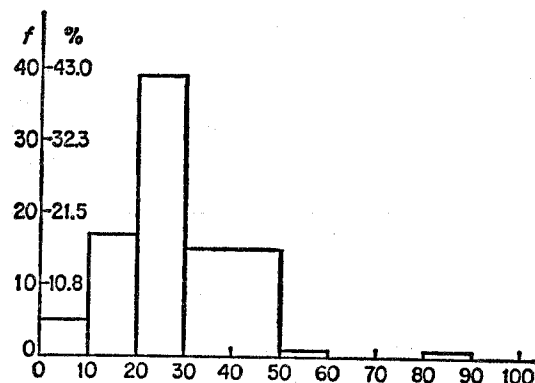


FIG. IV.3. Histograma de intervalos desiguales y áreas proporcionales a las frecuencias.

obtenido inicialmente. Una breve reflexión nos convencerá de que si hemos de pensar en términos de áreas más que en alturas, podremos manipular más fácilmente los datos que comportan intervalos desiguales. En otros términos: dejamos que las

áreas de los rectángulos sean proporcionales al número de los casos. En el caso especial importante en que todos los intervalos sean de ancho igual, las alturas serán también, por supuesto, proporcionales a las frecuencias. Si el ancho de cada rectángulo se toma como unidad y si las alturas se representan como pro-

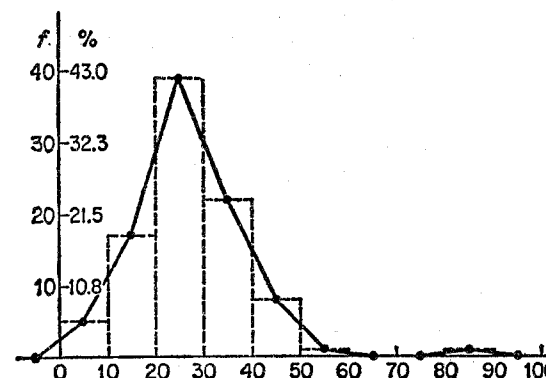


FIG. IV.4. Polígono de frecuencia.

porciones, entonces el área total comprendida en el histograma será la unidad. Así, por ejemplo:

$$1(5/93) + 1(17/93) + 1(39/93) + \dots + 1(1/93) = 1$$

Al estudiar la curva normal en el capítulo VII, veremos que es necesario tratar con áreas, antes que con alturas, y será conveniente tomar el área total bajo el histograma como unidad.

Otro modo muy parecido de presentar gráficamente una distribución de frecuencia es el del *polígono de frecuencia*. Para obtenerlo, unimos simplemente los puntos medios de los lados superiores de cada rectángulo por medio de rectas y borramos luego los rectángulos, como en la figura IV.4. Obsérvese que los puntos extremos del polígono de frecuencia se han colocado sobre la línea base (eje horizontal) en los puntos medios de los intervalos a uno y otro lado de los dos intervalos de los extremos. Normalmente no nos serviríamos de los dos tipos de figuras, pero, superponiendo el polígono de frecuencia sobre el histograma, vemos que el área delimitada por las dos figuras ha de ser igual. Esto es así porque por todo triángulo que queda al interior del polígono de frecuencia, pero exteriormente al histograma, hay un triángulo idéntico debajo del histograma, pero fuera del polígono de frecuencia. Así, pues, podemos también considerar como unidad el área delimitada por dicho polígono. Obsérvese, sin embargo, que no hemos hecho más que conectar

por medio de rectas cierto número de puntos. Los puntos mismos pueden representar el número de casos en cada intervalo, pero hemos de guardarnos de inferir que hay cierto número de casos en cualquier otro punto a lo largo del trazo continuo. Así, por ejemplo, *no* hemos de inferir que hay aproximadamente 28 casos con marcas de 20 exactamente.

Los polígonos de frecuencia pueden emplearse asimismo para representar distribuciones de frecuencia acumulativa. La figura

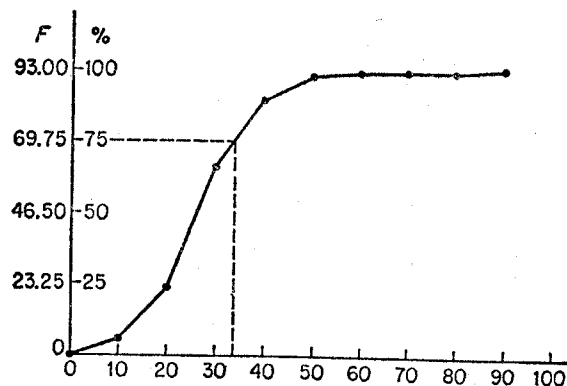


FIG. IV.5. Ojiva que representa una distribución de frecuencia acumulativa.

que en tal caso resulta se designa como *ojiva*. A lo largo de la ordenada o eje Y podemos indicar frecuencias o porcentajes. Colocamos, en cambio, las marcas de la variable de escala de intervalo a lo largo del eje de las X (abscisa), lo mismo que anteriormente, en el bien entendido de que las frecuencias representadas indican el número de casos de *valor inferior* al eje de la X. Por ejemplo, en la figura IV.5 vemos que aproximadamente el 75 % de las marcas son menores que 34. Por lo tanto, las ojivas se pueden usar como un método gráfico de determinar el número de casos por encima o por debajo de un cierto valor. Es obvio que la forma de la ojiva habrá de ser siempre o creciente o decreciente según que se acumule hacia arriba o hacia abajo. La curva será, en cambio, horizontal en los intervalos vacíos. Si la distribución de frecuencia es del tipo de nuestros datos anteriores, con el número mayor de casos en los intervalos que quedan cerca del centro de la distribución, la ojiva tendrá forma de S, con la inclinación más rápida a proximidad de los intervalos que contienen el mayor número de casos.

GLOSARIO

Datos continuos y datos discretos
Distribución acumulativa
Distribución de frecuencia
Polígono de frecuencia
Histograma
Ojiva
Límites verdaderos

EJERCICIOS

1. Supóngase que las cifras a continuación representan los ingresos anuales de un grupo de residentes de una localidad:

\$ 2 760	\$ 3 850	\$ 3 340	\$ 3 890	\$ 2 860
4 340	4 360	4 350	11 740	4 350
5 210	2 140	2 610	3 560	7 310
3 410	3 330	8 190	2 740	3 550
4 570	7 810	4 250	7 110	4 210
9 300	5 340	3 460	10 300	5 490
3 320	2 970	19 310	4 440	2 110
1 790	4 140	2 670	3 370	23 400
4 560	3 000	3 100	5 170	3 760
3 800	1 610	5 130	3 160	4 170
13 460	4 570	1 710	2 800	6 170
5 210	1 940	4 320	3 180	2 350
2 690	2 780	9 830	4 240	8 340

- Constrúyase una distribución de frecuencia y una distribución acumulativa.
- ¿Cuáles son los verdaderos límites?
- Trácese un histograma, un polígono de frecuencia y una ojiva.

2. En un examen de tipos de visita entre amigos íntimos y parientes, 81 interrogados son invitados a indicar el número de los amigos y parientes que visitan por lo menos una vez al mes. Los resultados son los siguientes (las cifras indican el número efectivo de personas regularmente visitadas):

3	5	2	3	3	4	1	8	4
2	4	2	5	3	3	3	0	3
5	6	4	3	2	2	6	3	5
4	14	3	5	6	3	4	2	4
9	4	1	4	2	4	3	5	0
4	3	5	7	3	5	6	2	2
5	4	2	3	6	1	3	16	5
3	11	4	5	19	4	5	2	2
4	3	14	5	2	1	4	3	4

- Constrúyase una distribución de frecuencia y una distribución acumulativa.

- b) Justifíquese lo mejor que se pueda la elección de los intervalos.
c) Trácese un histograma, un polígono de frecuencia y una ojiva.

3. Indíquense los límites verdaderos en cada uno de los siguientes intervalos:

- | | |
|------------------|--------------------------------|
| a) 1 000 — 1 900 | c) 1.000 — 1.999 (Respuesta, |
| 2 000 — 2 900 | 2.000 — 2.999 0.9995 — 1.9995) |
| b) 1 000 — 1 999 | d) .010 — .019 |
| 2 000 — 2 999 | .020 — .029 |

¿Qué se ha supuesto en cada uno de los casos a propósito del método de redondeo?

BIBLIOGRAFÍA

1. Anderson, T. R. y M. Zelditch: *A Basic Course in Statistics*, 2ª ed., Holt, Rinehart and Winston, Inc., Nueva York, 1968, cap. 4.
2. Downie, N. M. y R. W. Heath: *Basic Statistical Methods*, 2ª ed., Harper and Row, Publishers, Incorporated, Nueva York, 1965, cap. 3.
3. Hagood, M. J., y D. O. Price: *Statistics for Sociologists*, Henry Holt and Company, Inc., Nueva York, 1952, caps. 4 y 5.
4. McCollough, C., y L. van Atta: *Introduction to Descriptive Statistics and Correlation*, McGraw-Hill Book Company, Nueva York, 1965, cap. 1.
5. Mueller, J. H., K. Schuessler y H. L. Costner: *Statistical Reasoning in Sociology*, 2ª ed. Houghton Mifflin Company, Boston, 1970, cap. 4.
6. Weiss, R. S.: *Statistics in Social Research*, John Wiley & Sons, Inc., Nueva York, 1968, cap. 5.

V. ESCALAS DE INTERVALO: MEDIDAS DE TENDENCIA CENTRAL

VIMOS que las escalas nominales pueden resumirse fácilmente en términos de porcentajes, proporciones o razones, y que dichas medidas de resumen son fundamentalmente intercambiables. En otros términos: basta un tipo determinado de medida para describir los datos. En el caso de las escalas de intervalo, a su vez, vimos que los datos pueden describirse por medio de una distribución de frecuencia. Podemos servirnos también de tipos distintos de medidas, siendo las más importantes de ellas las de tipismo o de *tendencia central* y las de heterogeneidad o *dispersión*. Veremos que existe en cada caso cierto número de medidas distintas entre las que podemos elegir, cada una de las cuales reúne propiedades, ventajas e inconvenientes algo diferentes. Por lo tanto, el resumen de las escalas de intervalo es algo menos directo que en el caso de las nominales. En el presente capítulo nos ocupamos de las medidas de tipismo, en tanto que en el siguiente examinaremos las de dispersión. Tomados juntos, dichos dos tipos de medidas resultarán normalmente adecuados para la descripción de los datos de escala de intervalo.

La idea que tiene el lego a propósito del término *promedio* propende a ser más bien vaga o ambigua. En efecto, puede no darse cuenta de que existen varias medidas diversas del tipismo y que, en determinadas circunstancias, dichas medidas dan resultados muy distintos. El hecho de que sea posible obtener tales medidas diferentes de tendencia central supone que es necesario comprender las ventajas y los inconvenientes de cada una de ellas. Importa, pues, saber en cuáles circunstancias cada una sea adecuada. ¿Por qué la Oficina del Censo indica ingresos medianos y no ingresos medios? ¿Tendría algún sentido indicar al lego que la familia "media" tiene 2.3 hijos y vive en una casa de 4.8 cuartos? ¿En cuáles circunstancias es de poca importancia la medida que se emplee? Estas son algunas de las numerosas cuestiones que podrían plantearse acerca del tipo de promedio que hemos de calcular.

V.1. La media aritmética

Hay dos medidas importantes de tendencia central empleadas en la investigación sociológica: la media aritmética (designada a continuación simplemente como *media*) y la mediana. La media es con mucho la más común de las dos y se define como la suma de las marcas dividida por el número total de los casos comprendidos. Para indicar la media se utiliza por convención el símbolo

\bar{X} , aunque a veces se emplee también la letra M. Por lo tanto, la fórmula de la media aritmética es la siguiente:

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_N}{N} = \frac{\sum_{i=1}^N X_i}{N} \quad (\text{V.1})$$

en la que X_1 representa la puntuación del primer individuo, X_2 la del segundo, y X_i la del individuo general.¹ Si no existe ambigüedad, podemos prescindir de los subíndices y escribir simplemente

$$\bar{X} = \frac{\sum X}{N}$$

en donde se entiende que todas las cantidades se suman.

La media posee la propiedad algebraica de que la suma de las desviaciones de cada marca con respecto a la media será siempre cero. Simbólicamente esto puede expresarse mediante la ecuación siguiente:

$$\sum_{i=1}^N (X_i - \bar{X}) = 0$$

Este hecho no ha de sorprender en absoluto si tenemos en cuenta la definición de la media. La prueba es sencilla. Como quiera que tenemos una suma de números cada uno de los cuales, es en realidad, una diferencia, podemos descomponer la expresión indicada en la diferencia de dos sumas. En la siguiente forma:

$$\sum_{i=1}^N (X_i - \bar{X}) = \sum_{i=1}^N X_i - \sum_{i=1}^N \bar{X}$$

Pero, como quiera que \bar{X} es una constante, tenemos:

$$\sum_{i=1}^N \bar{X} = N\bar{X} = N \frac{\sum_{i=1}^N X_i}{N} = \sum_{i=1}^N X_i$$

y vemos inmediatamente que la diferencia es cero.

La propiedad mencionada puede utilizarse para simplificar el

¹ Para el examen de la notación de adición véase el Apéndice I.

cálculo de la media. Supóngase, por ejemplo, que hemos de calcular la media de los números 72, 81, 86, 69 y 57. Sumando y dividiendo por cinco obtenemos una $\bar{X} = 73.0$. Si sustraemos ahora esta media de cada una de las cifras y adicionamos los residuos, verificamos que la suma resultante es cero.

X	X - 73	X - 70
72	- 1	2
81	8	11
86	13	16
69	- 4	- 1
57	-16	-13
	<hr/> 0	<hr/> 15

Supóngase, en cambio, que hubiéramos anticipado una media de 70 y la hubiéramos restado de cada una de las cifras en cuestión. Entonces la suma resultante no es cero, sino que observamos que cada una de las nuevas diferencias es mayor en tres unidades (en dirección positiva) que las diferencias originarias. Vemos así que hemos anticipado una media que es demasiado pequeña en tres unidades. Si añadimos ahora un factor de corrección de tres a la media anticipada, obtenemos la media correcta. En la práctica, sin embargo, no compararíamos los dos juegos de diferencias en esta forma, sino que, observando que la suma del segundo grupo de diferencias es de +15 y sabiendo que hay cinco términos, esto indica que en promedio estábamos de 15/5, o sea 3.0 unidades, por debajo de la media verdadera. Y como puede verificarse fácilmente, si hubiéramos anticipado un valor demasiado alto, entonces la suma de las diferencias habría sido negativa, y hubiéramos debido sustraer de la media anticipada para obtener la correcta. Si \bar{X}' representa la media anticipada, podemos establecer una fórmula de la media en términos de la media supuesta y de un factor de corrección:

$$\bar{X} = \bar{X}' + \frac{\sum_{i=1}^N (X_i - \bar{X}')}{N} \quad (\text{V.2})$$

o bien, en palabras:

La media verdadera = a la media supuesta + $\frac{\text{la suma de desviaciones de ésta}}{\text{número de casos}}$.

Con objeto de verificar la corrección de esta fórmula desarrollamos la expresión de la derecha y obtenemos:

$$\begin{aligned}
 \bar{X}' + \frac{\sum_{i=1}^N (X_i - \bar{X}')}{N} &= \bar{X}' + \frac{\sum_{i=1}^N X_i}{N} - \frac{\sum_{i=1}^N \bar{X}'}{N} \\
 &= \bar{X}' + \frac{\sum_{i=1}^N X_i}{N} - \frac{N\bar{X}'}{N} \\
 &= \frac{\sum_{i=1}^N X_i}{N} = \bar{X}
 \end{aligned}$$

Pese a que pueda parecer que nos hayamos tomado mucha molestia calculando \bar{X} por rodeo en esta forma, este método permite sin embargo ahorrarse a menudo una considerable cantidad de trabajo cuando no se dispone de calculadoras de escritorio. El empleo de una medida anticipada permite por lo regular reducir la magnitud de los números que han de adicionarse. En efecto, cuanto más cerca quede la media supuesta de la verdadera, tanto menores serán en magnitud las diferencias resultantes. Este principio nos será particularmente útil cuando emprendamos el cálculo de las medias de datos agrupados.

Otra propiedad de la media puede formularse como sigue: la suma de las desviaciones *cuadradas* de cada cifra con respecto a la media es menor que la suma de las desviaciones cuadradas con respecto a cualquier otro número. O en otros términos:

$$\sum_{i=1}^N (X_i - \bar{X})^2 = \text{mínimo.}$$

* La prueba de esta propiedad es muy sencilla. Consideremos las desviaciones de X_i alrededor de cualquier otro número \bar{X}' que previamente hayamos tratado como media anticipada. Sumando y restando la media real \bar{X} de cada una de dichas expresiones podremos anotar:

$$X_i - \bar{X}' = (X_i - \bar{X}) + (\bar{X} - \bar{X}')$$

Elevando los dos términos al cuadrado obtenemos:

$$(X_i - \bar{X}')^2 = (X_i - \bar{X})^2 + 2(X_i - \bar{X})(\bar{X} - \bar{X}') + (\bar{X} - \bar{X}')^2$$

Resumiendo para todos los casos N obtendremos:

$$\begin{aligned}
 \sum_{i=1}^N (X_i - \bar{X}')^2 &= \sum_{i=1}^N (X_i - \bar{X})^2 \\
 &+ 2(\bar{X} - \bar{X}') \sum_{i=1}^N (X_i - \bar{X}) + \sum_{i=1}^N (\bar{X} - \bar{X}')^2
 \end{aligned}$$

en donde ha sido posible escribir la cantidad $2(\bar{X} - \bar{X}')$ frente al signo de sumar en el segundo término, ya que se trata de una constante. Inmediatamente veremos que todo el segundo término debe ser igual a cero, pues acabamos de mostrar que $\sum_{i=1}^N (X_i - \bar{X}) = 0$. Por otra parte, el último término consta de N términos, todos iguales a $(\bar{X} - \bar{X}')^2$. Tendremos por tanto

$$\sum_{i=1}^N (X_i - \bar{X}')^2 = \sum_{i=1}^N (X_i - \bar{X})^2 + N(\bar{X} - \bar{X}')^2$$

y así se comprueba que la suma de las desviaciones alrededor de \bar{X}' al cuadrado es igual a la suma de las desviaciones alrededor de la media verdadera, al cuadrado, *más* un término al cuadrado que nunca puede ser negativo.

Cuanto más grande sea la diferencia entre \bar{X}' y \bar{X} , tanto mayor será el segundo término situado a la derecha.

Tendremos frecuentes ocasiones para utilizar esta propiedad de los cuadrados de la media, y la cantidad $\sum_{i=1}^N (X_i - \bar{X})^2$ habrá de aparecer en gran parte de lo que sigue, como una medida de la variación total o heterogeneidad.

V.2. La mediana

A menudo necesitamos localizar la posición del caso medio cuando los datos se han ordenado de mayor a menor. O podemos dividir un grupo de estudiantes en porcentajes localizando los individuos que tienen exactamente el 10 por ciento de la clase que queda debajo de ellos, exactamente el 32 por ciento debajo de ellos, etcétera. Las medidas de este tipo se designan a menudo como *medidas de posición*, ya que localizan la posición de algún caso típico (o atípico) en relación con otros individuos. La mediana es tal vez la más importante de estas medidas de posición. Definimos la mediana como un número que posee la propiedad de tener el mismo número de marcas con valores menores que las que hay de valores mayores. La mediana divide habitualmente el total de los datos en dos mitades. Si el número de los casos es impar, la mediana será simplemente la marca del caso

del medio. Si N es par, no habrá caso central y, de hecho, cualquier número entre los valores de los dos casos centrales tendrá la propiedad de dividir las marcas en dos grupos iguales. Así, pues, si N es par, la mediana queda definida ambiguamente. Por convención tomamos entonces como valor único de la mediana la media aritmética de los dos datos centrales.

Si tuviéramos los números 72, 81, 86, 69 y 57, la mediana sería 72 (en tanto que la media es 73). Si hubiera un sexto término, digamos, por ejemplo, 55, las dos marcas centrales serían 69 y 72, y tomaríamos como mediana $(69 + 72)/2$, o sea 70.5. Si se da el caso de que los dos casos centrales tengan la misma marca, la mediana será, por supuesto, este mismo dato. Obsérvese que si N es impar, la mediana será el dato $(N + 1)/2$. Si el número de los datos es par, la mediana se encontrará en el centro entre el dato $N/2$ y el dato $(N + 1)/2$. Así, por ejemplo, si $N = 251$, la mediana será el dato del caso centésimo vigésimo sexto, y si $N = 106$, tomamos un valor medio entre las cifras de los casos quincuagésimo tercero y quincuagésimo cuarto. Estas fórmulas resultarán útiles por lo regular cuando N sea relativamente grande.

Vimos que la media posee las propiedades siguientes:

$$\sum_{i=1}^N (X_i - \bar{X}) = 0$$

y

$$\sum_{i=1}^N (X_i - \bar{X})^2 = \text{mínimo.}$$

La razón de que la primera propiedad se verifique es fundamentalmente que, cuando se sustrae la media de cada uno de los datos, las diferencias resultantes son tales que las marcas negativas se equilibran exactamente con las positivas. Pero supóngase que hubiéramos prescindido por completo de los signos, considerando todas las diferencias como positivas, ¿qué ocurrirá en este caso? Puede demostrarse que si se hubiera restado la mediana de cada una las marcas prescindiendo del signo de las diferencias y sumando los residuos, se obtendría una suma menor que la cifra comparable de cualquier otra medida de tendencia central. En símbolos esto se expresa así:

$$\sum_{i=1}^N |X_i - Md| = \text{mínimo}$$

en donde Md representa la mediana y las barras a ambos lados de la expresión $(X_i - Md)$ indican que hay que tomar el valor positivo (o "absoluto") de cada diferencia. Aunque esta propie-

dad de la mediana posea tal vez algún interés, no parece, sin embargo, tener aplicaciones directas de alguna significación sociológica.

V.3. Cálculo de la media y la mediana de datos agrupados

Método largo para el cálculo de la media. Cuando el número de datos se hace grande y los cálculos se realizan a mano, el computar la media o la mediana puede resultar tedioso. La mayoría de los científicos sociales cuentan con programas de computación que resuelven estos y otros cálculos con facilidad. En general resulta preferible utilizar tales programas cuando así parece conveniente, pues así disminuyen los riesgos de incurrir en errores de computación y redondeo, a la vez que se obtiene una economía considerable en tiempo y dinero. Debe, sin embargo, conocerse el procedimiento para computar varias medidas sin recurrir a tales programas, ya que con frecuencia resulta inconveniente disponer los datos en forma adecuada para su manejo por computadoras rápidas. En tales casos resulta útil agrupar los datos por categorías, computando la media o la mediana, tomando como base las resultantes distribuciones de frecuencias. En ocasiones se trata de datos que nos son dados ya en forma agrupada, pudiendo resultar imposible o inconveniente regresar a los datos originales para proceder a su computación. Un ejemplo de datos en grupos lo constituyen los censos. Por ellos sabremos que hay cierto número de personas con edades de 0 a 4 o de 5 a 9 años, pero desconoceremos la edad exacta de cada individuo.

Como veremos más abajo, el empleo de los datos agrupados puede simplificar nuestra labor considerablemente. Pero, por otra parte, al agruparlos en categorías, perdemos sin poderse evitar información. Podemos saber solamente, por ejemplo, que hay 17 personas con ingresos entre \$ 2 000 y \$ 2 900, pero no sabemos cómo se hallan distribuidas exactamente en el interior de dicho intervalo. Con objeto de calcular la media o la mediana de tales datos agrupados, hemos de proceder a hacer ciertos supuestos simplificadores acerca de la posición de los individuos en el interior de cada categoría. En el caso de la media, trataremos todos los casos como si se hallaran concentrados en los puntos medios de sus intervalos respectivos. Y al calcular la mediana suponemos que aquéllos se hallan esparcidos a distancias iguales en el interior de cada intervalo. Por supuesto, esas simplificaciones llevan aparejada cierta inexactitud. En efecto, no podemos esperar obtener en esta forma exactamente los mismos resultados que nos proporcionarían los datos brutos. Pero, por otra parte, si el número de datos es grande, las distorsiones introducidas serán por lo regular insignificantes y compensarán sobradamente el ahorro de tiempo. Es obvio, por lo demás, que cuanto más an-

gostos sean los intervalos, tanto menos información perderemos y tanto mayor será la exactitud. Así, por ejemplo, si sabemos que hay 17 casos entre \$ 2 000 y \$ 2 900 y 26 casos entre \$ 3 000 y \$ 3 900, podemos obtener resultados más exactos imaginando que los 17 casos se hallan en el punto medio del primer intervalo y los 26 en el punto medio del segundo, que si hubiéramos de situar los 43 casos juntos en el punto medio del intervalo mayor de \$ 2 000 a \$ 3 900. Estas simplificaciones tienen mayores probabilidades de conducir a errores en el caso de intervalos extremos, ya que los datos de dichos intervalos pueden resultar desviados hacia el centro de la distribución total. En esta forma, si hay 17 casos en el intervalo más bajo, la mayoría de ellos pueden encontrarse en la mitad superior del mismo. Sin embargo, si el número de los individuos en dichos intervalos extremos es muy pequeño, como suele suceder, es probable que la distorsión introducida sea insignificante.

De ahí que al calcular la media de datos agrupados tratemos todos los casos como si estuvieran situados en el punto medio de sus intervalos respectivos. Si lo prefiriéramos, podríamos suponerlos esparcidos a distancias iguales en el interior del intervalo, pero, como es fácil verificar, esto conduciría a los mismos resultados, ya que la media de cada intervalo quedaría exactamente en el punto medio del mismo. Como quiera que todos los casos de un intervalo se tratan como si tuvieran el mismo valor, podemos multiplicar el número de casos de cada intervalo por su valor común, en lugar de adicionar los datos separadamente. Así, por ejemplo, si hemos colocado 26 casos a la altura del valor de 3 450, el producto de $26 \times 3\,450$ será igual a la suma de 26 marcas separadas de 3 450 cada una. Y si hacemos esto con todos los intervalos, sumamos los productos y dividimos entre el número total de casos, obtendremos la media aritmética. La fórmula de ésta se convierte en tal caso en:

$$\bar{X} = \frac{\sum_{i=1}^k f_i m_i}{N} = \frac{\sum_{i=1}^k f_i m_i}{\sum f_i} \quad (V.3)$$

en la que f_i = número de casos de la categoría i -ésima con $\sum f_i = N$
 m_i = punto medio de la categoría i -ésima
 k = número de las categorías.

El ejemplo expuesto en el cuadro V.1 aclarará el proceso.

En el cuadro V.1 todos los intervalos son de la misma amplitud. Esto no es esencial, a condición que se empleen puntos medios correctos. Sin embargo, es necesario servirse de intervalos cerrados. Supóngase, en efecto, que el último intervalo hubiera

sido de \$ 7 000 para arriba, ¿Qué punto medio tomaríamos? No poseemos absolutamente base alguna que nos permita juzgar, a menos que nos remontemos a los datos originales. Algunas veces esto resulta posible, ya que las categorías extremas sólo comprenden a menudo relativamente pocos datos. En éstos resulta por lo regular más lógico servirse de la media real de los datos

CUADRO V.1. *Cálculo de la media de datos agrupados por el método largo*

Límites fijados	Límites verdaderos	Puntos medios (m_i)	f_i	$f_i m_i$
\$ 2 000-2 900	\$ 1 950-2 950	\$ 2 450	17	\$ 41 650
3 000-3 900	2 950-3 950	3 450	26	89 700
4 000-4 900	3 950-4 950	4 450	38	169 100
5 000-5 900	4 950-5 950	5 450	51	277 950
6 000-6 900	5 950-6 950	6 450	36	232 200
7 000-7 900	6 950-7 950	7 450	21	156 450
Totales			189	\$ 967 050

$$\bar{X} = \frac{\sum_{i=1}^k f_i m_i}{N} = \frac{967\,050}{189} = \$ 5\,117$$

de la categoría extrema que del punto medio de algún intervalo mayor. En los casos en que no resulta posible remontarse a los datos originales, será necesario adoptar un supuesto razonable en relación con el valor del punto medio. De ahí que sea decididamente más ventajoso para nosotros servirnos de intervalos cerrados siempre que haya de calcularse una media. Según veremos en el capítulo VI, esto se aplica asimismo al cálculo de la desviación estándar, la medida más comúnmente empleada de dispersión.

Método corto para el cálculo de la media. El método arriba indicado comportará por lo regular la multiplicación de números bastante grandes (v.gr., $2\,450 \times 17$), a menos que resulte que los puntos medios son números simples. Con una calculadora moderna dichos productos pueden calcularse y acumularse fácilmente. Pero, si los cálculos han de hacerse a mano, existe un medio mucho más sencillo de calcular la media de datos agrupados. Este método, llamado "corto", parece a primera vista comportar más trabajo que el "largo", pero, una vez dominado, se revela como mucho más sencillo que el otro. Fundamentalmente, el método corto consiste en anticipar una media y servirse en

esta forma de números más pequeños en la multiplicación. Luego se añade, como anteriormente, un factor de corrección a la media supuesta.

Con objeto de simplificar nuestros cálculos, tomemos como media anticipada el punto medio de uno de los intervalos. En el ejemplo arriba tratado podemos ver por inspección que la media será algo inferior a \$ 5 450, punto medio del cuarto intervalo. La ventaja de servirnos de un punto medio como media supuesta es obvia. En efecto, todos los demás datos estarán en tal caso a cierto número de intervalos de distancia de la media supuesta, ya que cada marca se supone hallarse en uno u otro de los puntos medios. Si restamos ahora la media supuesta de cada una de las marcas, obtendremos diferencias de exactamente \$ 1 000, \$ 2 000 o \$ 3 000 en ambas direcciones. Multiplicamos luego esas *diferencias* por las frecuencias apropiadas para obtener el factor de corrección que ha de añadirse a la media anticipada. En otros términos, habrá 17 casos con marcas de exactamente \$ 3 000 menos que aquélla; habrá 26 casos con una diferencia de \$ 2 000, etcétera. Si nos servimos de una columna d_i que represente la diferencia entre las marcas efectivas y la media *anticipada*, podemos modificar la fórmula (V.2) y escribir la fórmula de la media como sigue:

$$\bar{X} = \bar{X}' + \frac{\sum_{i=1}^k f_i d_i}{N} \quad (V.4)$$

donde

$$d_i = X_i - \bar{X}'$$

y podemos disponer nuestros cálculos en un cuadro como en el cuadro V.2. Una vez más, el factor de corrección se obtiene tomando la desviación total con respecto a la media anticipada (aquí -63 000) y después dividiendo entre el número de casos, lo que da la cantidad promedio en que la media anticipada se separa de la verdadera.

En este ejemplo, el factor de corrección ha resultado ser negativo, indicando que la media anticipada era demasiado grande. Hay que observar que si hubiéramos anticipado para la media otro valor cualquiera, habríamos llegado al mismo resultado. Si se elige como media anticipada el punto medio de tercer intervalo (\$ 4 450), el factor de corrección es de \$ 667, el cual, *adicionado* a \$ 4 450 da el resultado correcto. Dicho sea de paso, esto constituye un medio de control muy útil de nuestra labor. Obsérvese que si hubiéramos elegido el punto medio de cualquier otro intervalo, habríamos realizado más trabajo, ya que los números a sumar en la columna $f_i d_i$ habrían sido numéricamente mayores. Y si hubiéramos fallado en servirnos de un punto me-

dio, las desviaciones respecto de la media supuesta habrían comportado números mucho menos simples, con lo que no nos habríamos ahorrado trabajo alguno. Una vez que el proceso se haya comprendido bien, es posible omitir en el cuadro de cálculo la columna de los puntos medios.

El lector habrá sin duda observado que cada una de las desviaciones respecto de la media presunta del ejemplo anterior es un

CUADRO V.2. Cálculo de la media de datos agrupados por el método corto

Límites verdaderos	Puntos medios	f_i	d_i	$f_i d_i$
\$ 1 950-2 950	\$ 2 450	17	\$ - 3 000	\$ - 51 000
2 950-3 950	3 450	26	- 2 000	- 52 000
3 950-4 950	4 450	38	- 1 000	- 38 000
4 950-5 950	5 450	51	0	0
5 950-6 950	6 450	36	1 000	36 000
6 950-7 950	7 450	21	2 000	42 000
Totales		189		\$ - 63 000

$$\begin{aligned} \bar{X} &= \bar{X}' + \frac{\sum_{i=1}^k f_i d_i}{N} \\ &= 5 450 + \frac{- 63 000}{189} = 5 450 - 333 \\ &= \$ 5 117 \end{aligned}$$

múltiplo exacto de 1 000, o sea la magnitud del intervalo utilizado. Esto será siempre así, a condición que todos los intervalos tengan la misma amplitud. Por lo tanto, podemos poner la amplitud del intervalo como factor en cada uno de los productos $f_i d_i$, multiplicando por dicha amplitud una vez terminada la adición. En otros términos: pudimos haber obtenido la suma de - 63 000 de la manera siguiente:

$$- 63 000 = 1 000(- 51 - 52 - 38 + 0 + 36 + 42).$$

En lo que equivale a lo mismo, pudimos haber expresado las desviaciones originales en términos del número de *intervalos* (o "desviaciones graduantes") respecto de la media supuesta. Por lo tanto, determinamos cuántos intervalos dista la media supuesta de la verdadera y, finalmente, transportamos la magnitud del error hacia atrás a las unidades originales, multiplicando este

factor de corrección por la magnitud del intervalo. Designando la desviación en amplitudes de intervalo como d' , podemos revisar nuestro cuadro en la forma indicada en el cuadro V.3.

Si se han empleado intervalos desiguales, habrá que modificar esta segunda fórmula del método breve. A algunas personas les parecerá más fácil remontarse al método anterior, sirviéndose

CUADRO V.3. Cálculo de la media de datos agrupados por el método corto y de las desviaciones graduales

Límites verdaderos	Puntos medios	f_i	d'_i	$f_i d'_i$
\$ 1 950-2 950	\$ 2 450	17	-3	-51
2 950-3 950	3 450	26	-2	-52
3 950-4 950	4 450	38	-1	-38
4 950-5 950	5 450	51	0	0
5 950-6 950	6 450	36	1	36
6 950-7 950	7 450	21	2	42
Totales		189		-63

La fórmula modificada es ahora:

$$\bar{X} = \bar{X}' + \frac{\sum_{i=1}^k f_i d'_i}{N} \quad (V.5)$$

en donde i representa la amplitud de intervalo. Por consiguiente:

$$\bar{X} = 5\,450 + \frac{-63}{189} 1\,000 = 5\,117$$

de d_i en lugar de d'_i y escribiendo las diferencias efectivas en las unidades originales. Y alternativamente, si sólo difieren del resto en cuanto a amplitud uno o dos intervalos, podemos tomar como amplitud i de intervalo la amplitud de la mayoría de los intervalos de clase. Las desviaciones de los puntos medios de los intervalos restantes respecto de la media supuesta pueden en este caso expresarse en forma de fracciones de los intervalos enteros. Así, por ejemplo, si el último intervalo hubiera sido de \$ 6 950 a \$ 8 950, en lugar de \$ 6 950 a \$ 7 950, entonces el punto medio habría sido \$ 7 950 en lugar de \$ 7 450. Por lo tanto, la desviación respecto de la media presunta habría sido de \$ 2 500, o sean 2.5 amplitudes de intervalo. Si el intervalo hubiera ido hasta \$ 9 950, el valor d'_i hubiera sido de 3.0, según se deja comprobar fácilmente.

Cálculo de la mediana. Al calcular la mediana de datos agrupados, trataremos todos los casos al interior de un intervalo dado como si estuvieran distribuidos a distancias iguales en el mismo. Localizamos primero el intervalo que contiene el caso medio, e interpolamos luego para encontrar la posición exacta de la mediana. Al determinar el intervalo que contiene a ésta, es por

CUADRO V.4. Cálculo de la mediana de datos agrupados

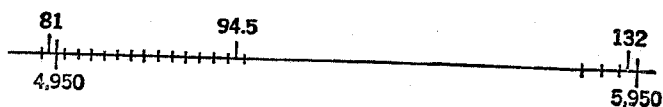
Límites verdaderos	f	F	Nº de casos inferiores a
\$ 1 950-2 950	17	17	\$ 2 950
2 950-3 950	26	43	3 950
3 950-4 950	38	81	4 950
4 950-5 950	51	132	5 950
5 950-6 950	36	168	6 950
6 950-7 950	21	189	7 950
Total	189		

lo regular conveniente obtener la distribución de frecuencia acumulativa. Pese a que no es absolutamente necesario, es preferible acostumbrarse a disponer por escrito la distribución acumulativa completa y a indicar en una columna separada el significado de cada una de las cifras de dicha columna (F). La distribución acumulativa de los datos anteriores se da en el cuadro V.4. A título de control de nuestra adición, observamos que todos los 189 casos han de quedar por debajo de \$ 7 950.

A continuación localizamos el intervalo que contiene el dato medio o el $N/2$ -ésimo. Aquí es $189/2 = 94.5$, de modo que buscamos el intervalo que contenga los casos noagésimo cuarto y noagésimo quinto. Obsérvese que, si los datos no hubieran estado agrupados, habríamos localizado el dato $(N+1)/2$ -ésimo, o sea el noagésimo quinto. La razón de esta inconsecuencia aparente se examinará más abajo. Como quiera que hay 81 casos por debajo de \$ 4 950 y 132 por debajo de \$ 5 950, la mediana ha de quedar en algún lugar del intervalo que va de \$ 4 950 a \$ 5 950. Constituye un buen procedimiento marcar dicho intervalo con un paréntesis, ya que se da a veces la tendencia de leer los datos a partir de la cifra 81, con lo que se obtiene el intervalo incorrecto de \$ 3 950 a \$ 4 950.

Examinemos ahora más de cerca el intervalo que contiene la mediana. Hay en éste 51 casos y, en consecuencia, habremos de dividir el intervalo entero en 51 subintervalos de amplitud \$ 1 000/51, o \$ 19.61 cada uno. Situamos cada uno de los 51 casos en el punto medio de su subintervalo propio. El caso octogésimo

primero quedará así situado en el último subintervalo del intervalo de \$3 950 a \$4 950, y el caso 132-avo será sólo ligeramente inferior al límite superior del intervalo que contiene la mediana. Ahora procedemos simplemente a contar subintervalos hasta llegar a aquélla. Si los datos no estuvieran agrupados, habríamos localizado la marca del caso $(N+1)/2$, o sea el nonagésimo quinto. De acuerdo con nuestra convención, dicho caso se situaría en el *punto medio* del decimocuarto subintervalo o, exactamente, a 13.5 subintervalos del límite inferior del intervalo. Obsérvese que este mismo valor se hubiera obtenido restando 81



de 94.5 o $N/2$. Es porque estamos operando con puntos medios de intervalos pequeños que contamos exactamente $N/2$ intervalos, con objeto de localizar la posición del caso $(N+1)/2$.

El valor de la mediana puede ahora obtenerse multiplicando simplemente el número de subintervalos abarcados por la magnitud de cada uno de ellos y añadiendo el resultado al límite inferior del intervalo. El procedimiento conjunto puede resumirse en la fórmula siguiente:

$$Md = l + \frac{N/2 - F}{f} i \quad (V.6)$$

en la que F = frecuencia acumulativa correspondiente al límite inferior,

f = número de casos del intervalo que contiene la mediana,

l = límite inferior del intervalo que contiene la mediana,

i = amplitud del intervalo que contiene la mediana.

La cantidad i/f representa la magnitud de cada subintervalo, y $N/2 - F$ da la distancia (en subintervalos) entre el límite inferior del intervalo y la mediana. En nuestro problema tenemos, pues:

$$\begin{aligned} Md &= 4\,950 + \frac{94.5 - 81}{51} 1\,000 = 4\,950 + 13.5 \frac{1\,000}{51} \\ &= 4\,950 + 265 = \$5\,215. \end{aligned}$$

Existe un camino alternativo, pero equivalente, de representar el proceso conducente a la obtención de la mediana. En efecto, en lugar de buscar la magnitud de cada subintervalo y multipli-

cando por el número de los subintervalos, podemos discurrir que, como quiera que hay 51 casos en el intervalo entero y que hemos de recorrer 13.5 de estos intervalos más pequeños para llegar a la mediana, hemos de recorrer $13.5/51$ del intervalo entero. Por lo tanto, si multiplicamos la magnitud del intervalo (1 000) por la fracción de la distancia total que hemos de recorrer, obtenemos el resultado deseado llamado *interpolación*. Al utilizar la fórmula es indiferente, por supuesto, cuál de las dos explicaciones nos parezca más satisfactoria. Con objeto de no hacernos demasiado dependientes de la fórmula, es mejor discutir el proceso cada vez, sirviéndonos de aquélla como control, hasta que se haya comprendido a fondo. A título de otro control hay que observar que la mediana pudo haberse asimismo obtenido *restando* cierta cantidad del límite superior u . Como puede demostrarse fácilmente, la fórmula se convierte en tal caso en:

$$Md = u - \frac{F - N/2}{f} i \quad (V.7)$$

en la que F representa ahora la frecuencia acumulativa correspondiente al límite superior del intervalo. Numéricamente esto da:

$$Md = 5\,950 - \frac{132 - 94.5}{51} 1\,000 = \$5\,215.$$

V.4. Comparación de la media y la mediana

Habiendo examinado los métodos de cálculo utilizados en la obtención de la media y la mediana tanto de datos agrupados como no agrupados, tócanos ahora comparar sus propiedades. Saltan a la vista varias diferencias entre las dos medidas. Primero, la media utiliza más información que la mediana, por cuanto al calcular la media nos servimos de la totalidad de las marcas exactas, en tanto que la mediana sólo comporta la marca del caso medio. Volviendo a las marcas 72, 81, 86, 69 y 57, vemos que si la marca más alta hubiera sido 126 en lugar de 86, la mediana habría permanecido inalterada, en tanto que la media habría aumentado considerablemente. Y en forma análoga, si la marca inferior hubiera sido cero, la media habría bajado, permaneciendo la mediana nuevamente inalterada. Por consiguiente, podemos establecer una diferencia muy importante entre ambas medidas, a saber: *La media resulta afectada por cambio de los valores extremos, en tanto que la mediana permanece inalterada, a menos que cambie asimismo el valor del caso medio*. En nuestro ejemplo, mientras 72 siga siendo el tercer caso después del reordenamiento, la mediana permanecerá inalterada.

Esta importante diferencia entre las dos medidas nos permite decidir en la mayoría de los casos cuál de ellas resulta más apropiada. Por lo regular deseamos que nuestra medida se sirva de toda la información disponible. En una forma u otra ponemos intuitivamente más fe en la medida que cumple dicha condición. Pese que al presente no sea posible reforzar dicha fe con un sólido razonamiento estadístico, puede darse, con todo, cierta justificación de la preferencia de la media en las circunstancias corrientes. Resulta, en efecto, que la media es por lo regular una medida más estable que la mediana, en cuanto varía menos de una muestra a otra. Cuando enderecemos nuestra atención a la Estadística inductiva, veremos que por lo regular el investigador tiene más interés en generalizar a propósito de la población que en su muestra particular. Está perfectamente percatado de que si se hubiera tomado otra muestra los resultados no habrían sido exactamente los mismos. Si se hubiera tomado una gran cantidad de muestras del mismo tamaño, habría podido ver simplemente en cuánto las medianas de las muestras diferían entre sí. Lo que aquí decimos es que las medianas de las muestras difieren de uno a otro de ellos más que las medias correspondientes. Pero como quiera que en la práctica sólo extraemos por lo regular una sola muestra, importa saber que la medida que empleamos dará resultados seguros, en cuanto que habrá un mínimo de variabilidad de una muestra a la próxima. Podemos, por consiguiente, establecer la siguiente regla práctica: *en caso de duda, empléese la media con preferencia a la mediana.*

Debido al hecho de que utiliza todos los datos, en tanto que la mediana no depende de los valores extremos, la media puede proporcionar en determinadas circunstancias resultados muy ambiguos. Hemos de tener presente que, al servirnos de una medida de tendencia central, tratamos de obtener una simple descripción de lo que en nuestros datos hay de "típico". Supóngase, para tomar un caso extremo, que en la serie de cinco números el dato superior fuera la de 962. La mediana seguiría siendo en nuestro caso 72, en tanto que la media subiría a $1\,241/5$, o sea 248.2. Ahora bien, ¿es este valor "típico", en alguna forma, de los datos? Ciertamente no. No se encuentra en parte alguna cerca de los datos de los cinco casos. Es verdad, por supuesto, que en un ejemplo tan extremado ninguna medida particular podría utilizarse para describir adecuadamente el caso típico, pero, como quiera que cuatro de los cinco datos se sitúan alrededor de 72, el empleo de la mediana resultaría manifiestamente menos equívoco. Podemos, pues, decir que: *siempre que una distribución es fuertemente asimétrica, esto es, siempre que hay considerablemente más casos extremos en una dirección que en otra, la mediana será por lo regular más apropiada que la media.*

La relación entre la desviación y las posiciones relativas de la

media y la mediana se indica en la figura V.1. Como quiera que puede resultar afectada por unos pocos valores extremos, la media se verá "empujada" en la dirección de la asimetría, esto es, hacia la cola. Si la distribución es perfectamente simétrica, la media y la mediana coincidirán. Sabemos que las distribuciones relativas a los ingresos suelen estar desviadas por lo regular

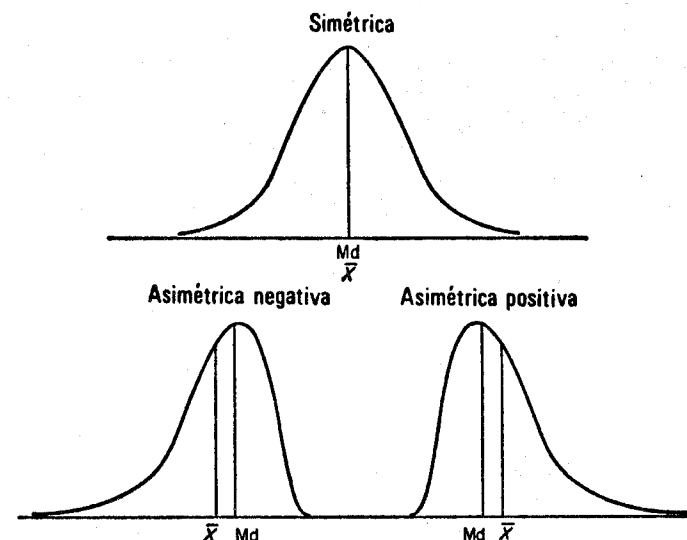


FIG. V.1. Relación entre la asimetría y las posiciones relativas de la media y la mediana

hacia los ingresos superiores, con muy pocos de ellos extremadamente altos. Resultaría, pues, muy impreciso presentar ingresos medios en el marco de una corporación o de una localidad pequeña. Por ello los datos relativos al ingreso se dan por lo regular sirviéndose de la mediana, más que de la media. Sin duda, si la distribución está muy desviada, el hecho debería mencionarse al presentar los datos. En tales casos, puede resultar útil indicar ambas cosas, la media y la mediana, pese a que esto sólo raramente se hace así en la práctica.

La media tiene una segunda propiedad que no posee la mediana: se deja manipular algebraicamente con mayor facilidad. Así, por ejemplo, precisa obtener a menudo un promedio ponderado de varios conjuntos de datos. Supóngase que tenemos los siguientes ingresos medios correspondientes a las tres localidades A, B y C:

Localidad	Habitantes	Media
A	10 000	\$ 3 518
B	5 000	4 760
C	8 000	4 122

Si el número de habitantes de las tres localidades fuera el mismo, podríamos tomar la media de esos tres datos como media general. Pero es el caso que la localidad A es dos veces mayor que la localidad B, o sea, en otros términos, que la cifra \$ 3 518 representa un doble número de casos de los que representa la cifra \$ 4 760. Si los 23 mil habitantes se hubieran puesto juntos calculándose la media general, la cifra resultante habría reflejado dicho hecho. Para obtener la media correcta, hemos de ponderar cada media separada por el número propio de casos, sumando luego y dividiendo finalmente entre el número total de éstos (23 000). Obtenemos en esta forma:

$$\bar{X} = \frac{\sum_{i=1}^k N_i \bar{X}_i}{N} \quad (\text{V.8})$$

en donde N_i y \bar{X}_i representan respectivamente el número de casos y la media de la categoría i -ésima, indicando k el número de las categorías. Tenemos, por consiguiente:

$$\begin{aligned} \bar{X} &= \frac{10\,000(3\,518) + 5\,000(4\,760) + 8\,000(4\,122)}{23\,000} \\ &= \frac{91\,956\,000}{23\,000} = \$ 3\,998.09 \end{aligned}$$

Podemos justificar fácilmente ese procedimiento de ponderación observando que la media de la categoría i -ésima fue en realidad obtenida adicionando los datos y dividiendo por N_i .² Por lo tanto, el producto $N_i \bar{X}_i$ representa la *suma* de todos los datos de dicha categoría. Así, pues, la adición de los productos y la división entre N nos da el mismo resultado que se habría obtenido si se hubieran ignorado las categorías por completo. Este tipo de manipulación algebraica de la media resulta en ocasiones muy útil. No ha de resultar difícil darse cuenta que la mediana general de los datos combinados no puede obtenerse en dicha

² Casi siempre ponderamos X_i con w_i , representando la expresión $\sum w_i X_i / \sum w_i$ nuestra media ponderada. Por lo regular hacemos la ponderación en tal forma que suma una cantidad conveniente como la unidad (esto es, $\sum w_i = 1$) o la muestra total de tamaño N , como en el ejemplo anterior.

forma. En efecto, si conociéramos los valores de los casos medios de cada una de las categorías separadas, nos faltaría todavía conocer el valor del caso medio de los datos combinados.

Obsérvese, finalmente, una diferencia importante entre la media y la mediana. El cálculo de la media requiere una escala de intervalo. En efecto, sin una escala de intervalo no tendría sentido alguno hablar de sumar marcas. Es manifiestamente necesario suponer, por ejemplo, que la suma de los números 30 y 45 equivale a la de los números 20 y 55, ya que ambos pares poseen la misma media. La mediana, en cambio, puede emplearse tanto con las escalas ordinales como con las de intervalo. La marca numérica real de la mediana carecerá de sentido, a menos que dispongamos de una escala de intervalo, pero será sin duda posible *situar* la marca media. Esto significa que, entre otros, podemos separar los casos en una o dos categorías, según que aquéllos queden por encima o por debajo de la mediana. Por lo tanto, las medidas de posición pueden emplearse con escalas ordinales, hecho que resulta muy útil para el desarrollo de pruebas que no requieren escalas de intervalo.

V.5. Otras medidas de tendencia central

Existen todavía algunas otras medidas de tendencia central, ninguna de las cuales, sin embargo, encuentra un empleo muy corriente en la investigación sociológica. Una de ellas es el *modo*, que es simplemente la marca *más frecuente*. Si, por ejemplo, tomamos las tres series de números siguientes:

- | | |
|-----|------------------------|
| (1) | 71, 75, 83, 75, 61, 68 |
| (2) | 71, 75, 83, 74, 61, 68 |
| (3) | 71, 75, 83, 75, 83, 68 |

podemos decir que la primera tiene un modo de 75, ya que hay dos términos de dicha marca, en tanto que ninguna otra aparece dos veces. No hay modo alguno en la segunda serie de números, pero los hay dos, en cambio, en la tercera (75 y 83). El modo resulta tal vez más útil cuando se da un número mayor de casos y cuando los datos han sido agrupados. En tal caso hablamos a veces de una categoría modal, tomando el punto medio de la misma como modo. En los datos agrupados que hemos utilizado, la categoría modal sería la de \$ 5 000 a \$ 5 900. En una distribución de frecuencia, el modo resultará indicado por el punto más elevado de la curva. En una distribución simétrica con un solo modo en el centro, la media, la mediana y el modo serán por supuesto, idénticos. Podemos distinguir asimismo entre distribuciones "unimodales" y "bimodales", tomando esta última la forma que aparece en la figura V.2. Al hablar de

distribuciones bimodales, no solemos por lo regular suponer que ambas cúspides tengan exactamente el mismo alto, como parecería deducirse de la definición. Hay que observar que, como quiera que el modo se refiere a la categoría con el mayor número de casos, podemos servirnos de dicho concepto tanto al describir escalas nominales, como ordinales o de intervalo. De esta manera en el caso de las escalas nominales podrá considerarse la

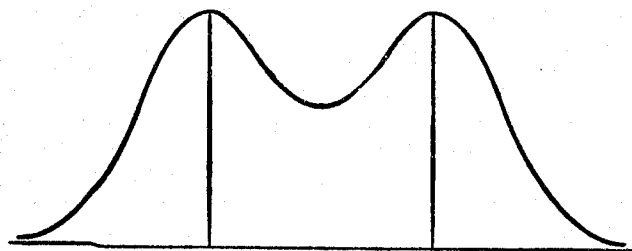


FIG. V.2. Una distribución bimodal

categoría modal como un tipo de *tendencia central*, siempre que se tenga bien presente que ello no supone un ordenamiento de categorías.

Otras dos medidas de tendencia central que prácticamente no se ven nunca en la literatura sociológica son la *media armónica* y la *media geométrica*. Se definen respectivamente por las siguientes fórmulas:

$$\text{Media armónica} = \frac{N}{\sum_{i=1}^N \frac{1}{X_i}}$$

$$\text{Media geométrica} = \sqrt[N]{(X_1)(X_2) \dots (X_N)}$$

En esta última fórmula, la N arriba del radical indica que tomamos la raíz N -ésima del producto de N datos.

V.6. Deciles, cuartiles y percentiles

Al examinar la mediana, señalamos que hay otras medidas posicionales, tales como los percentiles, que pueden utilizarse para fijar la posición de datos mayores que una proporción determinada de casos. Esas medidas, aunque no sean necesariamente medidas de tipicidad o de tendencia central, son análogas directamente a la mediana. Así, por ejemplo, en lugar de buscar un número que tenga la mitad de los datos por encima o por

debajo de sí mismo, podemos querer determinar el valor del primer cuartil, que posee la propiedad de que un cuarto de los datos sean de menor magnitud que la suya. Y en forma semejante, el tercer cuartil representa la marca que tiene por debajo de ella, en cuanto a magnitud, a los tres cuartos de los casos. Si se prefiere, se puede dividir la distribución en 10 deciles, fijando marcas que tengan una décima, dos décimas o nueve décimas de los casos con valores menores. Tal vez el lector esté más familiarizado con los percentiles, que dividen la distribución en 100 porciones de tamaño igual. Así, por ejemplo, el estudiante que falla en el nonagésimo primer percentil sabe que el 91 por ciento de los demás estudiantes tenían puntuaciones más bajas que él.

El cálculo de los deciles, los cuartiles y los percentiles es directamente análogo al de la mediana. En el caso de datos agrupados, determinaremos primero el intervalo en cuyo interior queda la medida de posición deseada. Sirviéndonos luego de los datos del cuadro V.4, obtendremos el primer cuartil localizando la posición del caso $N/4$ o 47.25-ésimo. De la columna de la frecuencia acumulativa vemos que el primer cuartil ha de situarse en algún lugar entre el intervalo de \$3 950 a \$4 950. Y como quiera que en dicho intervalo hay 38 casos, hemos de recorrer los $(47.25 - 43)/38$ de esa distancia. Así, pues, el valor del primer cuartil Q_1 será:

$$Q_1 = 3\,950 + \frac{47.25 - 43}{38} 1\,000 = 3\,950 + 112 = \$4\,062$$

Otras medidas de posición pueden calcularse en forma análoga. Obsérvese, incidentalmente, que por definición la mediana es equivalente al segundo cuartil, al quinto decil, y al quincuagésimo percentil. Si bien los deciles, cuartiles y percentiles sólo se emplean muy raramente en la investigación sociológica, conviene por lo menos conocer su sentido.

GLOSARIO

Decil
Media
Mediana
Modo
Percentil
Cuartil
Distribución asimétrica

EJERCICIOS

1. Indíquense la media, la mediana y el modo de los números siguientes: 26, 37, 43, 21, 58, 26, 33 y 45. Respuesta, 36.1; 35; 26.

21, 26, 26, 33, 33, 43, 43, 58

2. Calcúlense una media y una mediana de los datos compilados en el ejercicio 1, cap. IV. Hágase lo mismo en relación con el ejercicio 2, cap. IV.

3. Calcúlense el tercer cuartil, el cuarto decil y el septuagésimo primer percentil de los datos del ejercicio 1, cap. IV.

4. Los siguientes datos (hipotéticos) muestran la distribución del porcentaje de las familias granjeras en 60 distritos. Calcúlense la media y la mediana. Respuesta, 32.83; 32.83.

Intervalo %	Frecuencia
10-19	7
20-29	16
30-39	21
40-49	12
50-59	4
	<hr/> 60

5. Sirviéndose de los datos del ejemplo anterior, indique el lector en qué forma resultarían afectadas la media y la mediana (aumentadas, reducidas, inalteradas) si:

a) el último intervalo se ampliara de 50 a 69, permaneciendo las mismas frecuencias. Respuesta, aumentada; la misma.

b) si se añadiera un 10 por ciento a cada intervalo (haciendo los intervalos 20 a 29, 30 a 39, etcétera), con frecuencias inalteradas;

c) los intervalos permanecieran inalterados, pero pasando dos casos de la categoría 20 a 29 a la categoría 30 a 39 (haciendo que las frecuencias fueran 7, 14, 23, 12 y 4);

d) los intervalos permanecieran inalterados, pero se doblaran todas las frecuencias.

6. Un grupo de 10 muchachos y 7 muchachas participaron en un acertijo algebraico. Supóngase que la puntuación media de los muchachos fue 84 y su mediana 74, en tanto que, en relación con las muchachas, tanto la media como la mediana resultaron en 79. El maestro concluye que en esa prueba los muchachos obtuvieron un resultado mejor que las muchachas. ¿Está su conclusión justificada? ¿Por qué, o por qué no? ¿Cómo cabría explicar la gran diferencia entre la media y la mediana en los muchachos?

7. Supóngase que se ha encontrado que la edad media de los 50 gobernadores (de los Estados Unidos) es de 51.6 años, la de 100 senadores 62.3, y la de 435 diputados de 44.7. ¿Cuál es la edad media de todos esos políticos? Supóngase que las cifras anteriores indicaran medianas, ¿podría obtenerse la mediana general del mismo modo? ¿Por qué, o por qué no?

BIBLIOGRAFÍA

1. Anderson, T. R., y M. Zelditch: *A Basic Course in Statistics*, 2ª ed., Holt, Rinehart and Winston, Inc., Nueva York, 1968, cap. 5.
2. Downie, N. M., y R. W. Heath: *Basic Statistical Methods*, 2ª ed., Harper and Row, Publishers, Incorporated, Nueva York, 1965, cap. 4.

3. Hagoood, M. J., y D. O. Price: *Statistics for Sociologists*, Henry Holt and Company, Inc., Nueva York, 1952, cap. 8.

4. McCollough, C. y L. van Atta: *Introduction to Descriptive Statistics and Correlation*, McGraw-Hill Book Company, Nueva York, 1965, cap. 2.

5. Mueller, J. H., K. Schuessler y H. L. Costner: *Statistical Reasoning in Sociology*, 2ª ed. Houghton Mifflin Company, Boston, 1970, cap. 5.

6. Weinberg, G. H., y J. A. Schumaker: *Statistics: An Intuitive Approach*, Wadsworth Publishing Company, Inc., Belmont, Cal. 1962, caps. 2 y 6.

VI. ESCALAS DE INTERVALO: MEDIDAS DE DISPERSIÓN

EN LA investigación sociológica la atención se concentra en muchos casos en medidas de tendencia central. Por ejemplo, podemos querer comparar varios tipos de religión en relación con la asistencia media a la iglesia o el nivel medio de ingreso. Podemos también desear obtener, sin embargo, medidas de homogeneidad. Tal vez hayamos partido de la hipótesis que una de las religiones extraerá sus adeptos en mayor grado que las otras de una misma capa social. Sin embargo, aun si estamos interesados ante todo en comparar medidas de tendencia central, necesitamos, con todo, saber algo acerca de la dispersión en cada grupo. Nos damos cuenta intuitivamente de que, si cada religión fuera extremadamente heterogénea en cuanto al ingreso y a la asistencia a la iglesia, una diferencia determinada entre sus medias (digamos de \$ 2 000) no sería tan importante o indicativa como sería el caso si cada grupo fuera perfectamente homogéneo.

Cuando lleguemos a la estadística inductiva, estaremos en condiciones de justificar dicha intuición y de apreciar por qué las medidas de dispersión son tan importantes. En el presente capítulo vamos a concentrarnos en el mecanismo, en tanto que en el siguiente daremos una interpretación de la medida de dispersión más importante: la desviación estándar.

VI.1. El recorrido

De las distintas medidas de dispersión que vamos a examinar en este capítulo, el recorrido es con mucho el más simple. El recorrido se define como la diferencia entre la marca más alta y la más baja. Así, pues, en relación con los datos proporcionados en el capítulo anterior (72, 81, 86, 69 y 57), el recorrido sería la diferencia entre 86 y 57, o sea 29. Por lo regular solemos indicar el recorrido ya sea por medio de la diferencia real (29), o dando las dos marcas extremas, v.gr. 57 y 86. Si los datos se han agrupado, tomamos como recorrido la diferencia entre los *puntos medios* de las categorías extremas. Así, pues, si el punto medio del intervalo inferior es 2 450 y el del intervalo superior 7 450, el recorrido será de 5 000.

La simplicidad extrema del recorrido como medida de dispersión presenta a la vez ventajas e inconvenientes. En efecto, el recorrido puede resultar muy útil si se trata de obtener unos cálculos muy rápidos que puedan proporcionar una indicación bruta de la dispersión, o si los cálculos ha de hacerlos alguna

persona que no esté familiarizada con la estadística. Si los datos han de presentarse a una audiencia relativamente ingenua, el recorrido será tal vez la única medida de dispersión que aquélla esté en condiciones de interpretar fácilmente. Sin embargo, el nivel de preparación de los sociólogos está alcanzando rápidamente un punto tal, que podemos legítimamente suponer que entenderán también medidas algo más complicadas y satisfactorias. El inconveniente del recorrido es obvio: se basa exclusivamente en dos casos, que son, además, los dos casos extremos. Y como quiera que los casos extremos suelen ser raros o poco comunes en la mayoría de los problemas empíricos, nos damos cuenta que por lo regular es una cuestión de azar que obtengamos uno o dos de ellos en nuestra muestra. Supóngase, por ejemplo, que en la localidad investigada hay un millonario. Si escogemos 10 personas al azar, es probable que aquél no esté incluido entre ellas. Pero, supóngase que sí está. En tal caso el recorrido de los ingresos será extraordinariamente amplio y muy engañoso en cuanto medida de dispersión. Si nos servimos del recorrido como medida, nada sabemos acerca de la variación de las marcas entre los dos valores extremos, excepto que éstas se sitúan en algún lugar en el interior de dicho recorrido. Así, pues, como resulta del ejemplo anterior, el recorrido variará considerablemente de una muestra a otra. Por otra parte, el recorrido será por lo regular mayor en las muestras grandes que en las pequeñas, simplemente porque en los primeros tenemos más probabilidades de incluir a los casos individuales extremos. Ésta es la razón de que el recorrido no se emplee por lo regular en sociología, excepto al nivel de tipo más exploratorio.

Otra medida sumamente simple, la *razón de variación*, puede ser utilizada en el caso de los datos en grupo, lo que resulta especialmente adecuado en el caso de las escalas nominales. Consiste básicamente en una medida del grado en que se concentran los datos en la categoría modal, en lugar de que se les encuentre distribuidos uniformemente a lo largo de todas las categorías. Se define así:

$$V.R. = 1 - f_{\text{modal}}/N,$$

en donde f_{modal} se refiere al número de casos en la categoría modal, y N al número total de casos. Es evidente que esta medida resulta insensible a la distribución de casos en las categorías no modales, siendo por otra parte dependiente del proceso de categorización. Su ventaja radica en su sencillez extrema y en su atracción intuitiva, además del hecho de que en el caso de las escalas nominales no es posible hacer uso de una ordenación de categorías que permita habilitar medidas de un mayor refinamiento.

VI.2. La desviación cuartil

Otra medida empleada algunas veces en los campos de la psicología y la enseñanza, pero que raramente aparece en la literatura sociológica, es la desviación cuartil o recorrido semi-intercuartil. La desviación cuartil Q es un tipo de recorrido, pero, en lugar de representar la diferencia entre los valores extremos, se define arbitrariamente como la mitad de la distancia entre el primero y el tercer cuartiles. O en forma simbólica:

$$Q = \frac{Q_3 - Q_1}{2} \quad (\text{VI.1})$$

en donde Q_1 y Q_3 representan respectivamente al primero y tercer cuartiles. Obsérvese que la desviación cuartil mide el recorrido ocupado por la mitad central de los casos. Como quiera que Q_1 y Q_3 variarán menos de una muestra a otra que los casos más extremos, la desviación cuartil representa una medida mucho más estable que el recorrido. Por otra parte, en cambio, no saca provecho del conjunto de la información. No estamos midiendo la variabilidad entre los casos centrales ni tomamos en consideración lo que ocurre en los extremos de la distribución. De ahí, pues, que enderecemos nuestra atención a otras dos medidas que sí poseen esta propiedad deseable.

VI.3. La desviación media

Si deseamos servirnos de todos los datos, el sentido común nos sugerirá que tomemos las desviaciones de cada dato con respecto a alguna medida de tendencia central y que calculemos luego alguna especie de promedio de dichas desviaciones, con objeto de controlar el número de casos comprendidos. Sería posible tomar como medida de tendencia central la mediana o el modo, pero por lo regular tomamos la media, ya que ésta es en la mayoría de los casos la medida particular más satisfactoria. Supóngase que sumáramos simplemente las desviaciones efectivas respecto de la media. Por desgracia, como sabemos, el resultado sería siempre cero, ya que las diferencias positivas y negativas se compensan mutuamente. Esto sugiere que, para obtener una medida de dispersión alrededor de la media, hemos de deshacernos en una forma u otra de los signos negativos. Se nos ocurren inmediatamente dos métodos: 1) ignorar los signos y tomar sólo los valores absolutos de las diferencias, o 2) cuadrar las diferencias. Estos dos métodos conducen efectivamente a las dos medidas restantes de dispersión que hemos de examinar en este capítulo, a saber: la desviación media y la desviación estándar.

La desviación media se define como la media aritmética de las diferencias absolutas de cada marca con respecto a la media, o en símbolos:

$$\text{Desviación media} = \frac{\sum_{i=1}^N |X_i - \bar{X}|}{N} \quad (\text{VI.2})$$

La media de los números 72, 81, 86, 69 y 57 es 73.0. Si sustraemos 73.0 de cada uno de dichos números, ignorando los signos, y luego adicionamos los resultados y dividimos entre 5, obtenemos:

$$\frac{\sum_{i=1}^N |X_i - \bar{X}|}{N} = \frac{1 + 8 + 13 + 4 + 16}{5} = \frac{42}{5} = 8.4$$

Podemos por consiguiente decir que el promedio de los datos difiere de la media en 8.4.

Pese a que la desviación media presenta una interpretación intuitiva más directa que la desviación estándar, tiene, con todo, varios inconvenientes graves. Primero, los valores absolutos no se dejan manipular algebraicamente con facilidad. Segundo y más importante, la desviación media no es de fácil interpretación teórica ni conduce a resultados matemáticos simples. Con fines puramente descriptivos, la desviación media puede ser adecuada, pese a que, según veremos, la desviación estándar se deja interpretar más fácilmente en términos de la curva normal. Cuando lleguemos a la estadística inductiva veremos que la desviación estándar se utiliza sobre todo a causa de su superioridad teórica. Ésta es la razón de que sólo raramente encontremos en la literatura sociológica referencias a la desviación media.

VI.4. La desviación estándar

Habiendo eliminado más o menos otras varias medidas de dispersión, podemos ahora dirigir nuestra atención a la más útil y frecuente de las medidas: la *desviación estándar*. Ésta se define como la raíz cuadrada de la media aritmética de las desviaciones cuadradas con respecto a la media, o en símbolos:

$$s = \sqrt{\frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N}} \quad (\text{VI.3})$$

en donde s se emplea para designar la desviación estándar.¹ O en

¹ Algunos textos definen s con $N - 1$ en el denominador en vez de N . La razón de ello no resultará clara hasta en el capítulo xi.

palabras: tomamos la desviación de cada marca con respecto a la media, cuadramos cada diferencia, sumamos los resultados, dividimos entre el número de casos y extraemos la raíz cuadrada. Para conseguir una respuesta correcta, es indispensable que las operaciones se efectúen exactamente en el orden indicado. En nuestro ejemplo numérico la desviación estándar podría conseguirse como sigue:

X_i	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
72	- 1	1
81	8	64
86	13	169
69	- 4	16
57	- 16	256
$\bar{X} = 73.0$	0	506

$$s = \sqrt{506/5} = \sqrt{101.2} = 10.06$$

El significado intuitivo de la desviación estándar no nos aparecerá claramente hasta más adelante, cuando nos sirvamos de s para darnos las áreas bajo la curva normal. Por el momento la aceptamos simplemente como un número abstracto. Sin embargo, algunas propiedades de la desviación estándar son ya manifiestas desde ahora. Observamos, en efecto, que cuanto mayor es la dispersión alrededor de la media tanto mayor es la desviación estándar. Si todos los cinco valores hubieran sido cero, las desviaciones alrededor de la media habrían sido cero, y s también habría sido cero. Por otra parte, vemos que las desviaciones extremas con respecto a la media pesan más, con mucho, en cuanto a determinar el valor de la desviación estándar. En efecto, los valores 169 y 256 dominan las otras tres desviaciones cuadradas. Al cuadrar las desviaciones, pese a que después extraigamos la raíz cuadrada, estamos en realidad dando más peso relativo a los valores extremos todavía de lo que era el caso al calcular la media. Esto sugiere que hemos de mitigar nuestro entusiasmo inicial a propósito de la desviación estándar en cuanto "la mejor" medida particular de dispersión. Ciertamente, si hay varios casos extremos, queremos que nuestra medida lo señale. Pero si la distribución presenta unos pocos casos muy extremos, la desviación normal puede conducir a resultados engañosos, en cuanto puede ser extraordinariamente grande. En tales casos nos serviríamos probablemente como medida de tendencia central de la mediana y, tal vez, de la desviación cuartil como medida de dispersión. Sin embargo, para la mayoría de los datos la desviación normal resultará adecuada.

Es razonable preguntar: "¿por qué molestarse en extraer la raíz cuadrada al calcular una medida de dispersión?" Una respuesta fácil, aunque poco satisfactoria, sería la de decir que así es como se define la desviación estándar. Podría justificarse la extracción de la raíz cuadrada señalando que, ya que hemos cuadrado cada desviación, lo que hacemos es compensar dicho paso anterior. Sin embargo, resulta más comprensible justificar la extracción de la raíz en términos de su carácter práctico. Como quiera que, en efecto, más adelante habremos de hacer un empleo considerable de la curva normal, la desviación estándar, tal como se la ha definido, resulta ser una medida muy útil. Para otros fines nos serviremos del cuadrado de la desviación normal o *variancia*, que se define como:

$$\text{Variancia} = s^2 = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N}$$

Los matemáticos han encontrado que el concepto de variancia poseía mayor valor teórico que la desviación estándar. A partir del capítulo XVI, haremos un uso creciente de la variancia, pero de momento podemos limitar nuestra atención a la desviación estándar. Los dos conceptos son por lo demás tan fácilmente intercambiables, que podemos pasar sin dificultad del uno al otro. Que se defina la variancia como cuadrado de la desviación estándar o ésta como raíz cuadrada de la variancia, esto no reviste importancia alguna.

Cálculo de la desviación estándar de datos no agrupados. Si bien la desviación estándar puede calcularse siempre a partir de la fórmula básica que se acaba de dar, resulta a menudo más sencillo servirse de fórmulas de cálculo que no requieren la sustracción de la media de cada marca separada. En efecto, no sólo la media no será por lo regular un número entero, sino que usualmente se cometerán errores de redondeo al emplear la fórmula antes indicada. Con objeto de ver de qué modo podemos simplificar los cálculos, desarrollemos la expresión que está abajo del radical. Tenemos:

$$\begin{aligned} \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N} &= \frac{\sum_{i=1}^N (X_i^2 - 2X_i\bar{X} + \bar{X}^2)}{N} \\ &= \frac{\sum_{i=1}^N X_i^2 - 2\bar{X} \sum_{i=1}^N X_i + N\bar{X}^2}{N} \end{aligned}$$

Obsérvese que, como quiera que \bar{X} es constante, pudimos tomarla frente al signo de sumación en el segundo término del numerador. En el tercer término, a su vez, nos hemos servido del hecho de que, para toda constante k , tenemos:

$$\sum_{i=1}^N k = Nk.$$

Pero, como quiera que $\bar{X} = \sum_{i=1}^N X_i / N$, el término central del numerador se reduce a $-2\bar{X}^2$, y podemos escribir:

$$\frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N} = \frac{\sum_{i=1}^N X_i^2}{N} - 2\bar{X}^2 + \bar{X}^2 = \frac{\sum_{i=1}^N X_i^2}{N} - \bar{X}^2$$

Por lo tanto:

$$s = \sqrt{\frac{\sum_{i=1}^N X_i^2}{N} - \bar{X}^2} \quad (\text{VI.4})$$

Algunas otras fórmulas de cálculo alternativas son las siguientes:

$$s = \sqrt{\frac{\sum_{i=1}^N X_i^2}{N} - \left(\frac{\sum_{i=1}^N X_i}{N}\right)^2} \quad (\text{VI.5})$$

$$= \sqrt{\frac{\sum_{i=1}^N X_i^2 - \frac{\left(\sum_{i=1}^N X_i\right)^2}{N}}{N}} \quad (\text{VI.6})^2$$

$$= \frac{1}{N} \sqrt{N \sum_{i=1}^N X_i^2 - \left(\sum_{i=1}^N X_i\right)^2} \quad (\text{VI.7})$$

Si bien cualquiera de las formas precedentes puede utilizarse como fórmula de cálculo, la ecuación (VI.7) es la que comporta,

² La obtención de las ecuaciones (VI.6) y (VI.7) a partir de la ecuación (VI.5) se deja como ejercicio.

con todo, menos errores de redondeo, por ello se la recomienda. Sirvámonos de una de dichas fórmulas de cálculo (ec. VI.7) en el problema anterior, en donde $N = 5$.

X_i	X_i^2
72	5 184
81	6 561
86	7 396
69	4 761
57	3 249
365	27 151

En adición al número total de casos, las dos cantidades requeridas son $\sum_{i=1}^N X_i$ y $\sum_{i=1}^N X_i^2$. Ambas sumas pueden acumularse simultáneamente con las modernas calculadoras de oficina. Calculamos ahora s a partir de (VI.7):

$$s = 1/5 \sqrt{5(27\,151) - (365)^2} = 1/5 \sqrt{135\,755 - 133\,225} = 10.06$$

Nos hemos servido de este problema muy sencillo para ilustrar que la fórmula de cálculo da el mismo resultado numérico que la fórmula básica de la ecuación (VI.3). Como quiera que \bar{X} resultó ser un entero, la fórmula de cálculo ha comportado en realidad más trabajo que la fórmula original. Pero normalmente, por supuesto, esto no será así.

* *Cálculo de la desviación estándar de datos agrupados.* Si los datos han sido agrupados, podemos simplificar nuestra labor considerablemente tratando cada caso como si se hallara en el punto medio de un intervalo y sirviéndose de una medida supuesta. Sin duda introducimos con ello alguna inexactitud, pero el ahorro de tiempo es sustancial. Siguiendo una convención corriente, supongamos que $x_i = X_i - \bar{X}$. En consecuencia, las x minúsculas representan desviaciones respecto de la media, y la fórmula básica de la desviación estándar se convierte en:

$$s = \sqrt{\frac{\sum_{i=1}^N x_i^2}{N}}$$

Podemos modificar ahora la fórmula tomando en cuenta el hecho de que habrá un gran número de casos tratados todos como si tuvieran el mismo valor, esto es, uno de los puntos medios. Si multiplicamos el número de casos en cada clase por el punto

medio propio y sumamos luego los productos, nos podemos ahorrar el trabajo de sumar todos los N casos. La fórmula de la desviación estándar se convierte así en:

$$s = \sqrt{\frac{\sum_{i=1}^k f_i x_i^2}{N}} \quad (\text{VI.8})$$

en donde f_i es el número de casos del intervalo i -ésimo y k el número de intervalos.³

Supongamos ahora que anticipamos una media y tomamos las desviaciones con respecto a ésta, en lugar de respecto de la media verdadera. Mostramos en el capítulo anterior que la suma de las desviaciones cuadradas de la media será menor —que cualquier otro valor— que la suma de las desviaciones cuadradas. En particular, la suma de las desviaciones cuadradas de la media anticipada será mayor que la cifra obtenida sirviéndonos de la media verdadera, a menos, por supuesto, que aquélla coincida con ésta. Puede, pues, demostrarse que cuanto más cerca queda la media supuesta de la verdadera, tanto menor resulta la suma de las desviaciones cuadradas de la media supuesta. En otros términos: si nos servimos de una media supuesta, esperamos obtener una suma de cuadrados demasiado grande. Lo mismo que anteriormente, podemos servirnos de un factor de corrección, al que sustraemos luego del valor obtenido utilizando la media anticipada. La fórmula de la desviación estándar se convierte en tal caso en:

$$s = \sqrt{\frac{\sum_{i=1}^k f_i d_i^2}{N} - \left(\frac{\sum_{i=1}^k f_i d_i}{N} \right)^2} \quad (\text{VI.9})$$

en donde los d_i representan las diferencias entre cada marca y la media anticipada y son directamente análogos a los x_i de la ecuación (VI.8).

Antes de tomar un ejemplo numérico, examinemos la fórmula precedente con mayor atención. El segundo término debajo del radical representa el factor de corrección que ha de sustraerse de las desviaciones cuadradas de la media supuesta. Recordando la fórmula de la media expresada en términos de la media supuesta, o sea:

³ Obsérvese que *no* se elevan al cuadrado las frecuencias f_i que aparecen en el numerador de la expresión bajo el radical.

$$\bar{X} = \bar{X}' + \frac{\sum_{i=1}^k f_i d_i}{N}$$

vemos que

$$\frac{\sum_{i=1}^k f_i d_i}{N} = \bar{X} - \bar{X}'$$

y que, por lo tanto:

$$\left(\frac{\sum_{i=1}^k f_i d_i}{N} \right)^2 = (\bar{X} - \bar{X}')^2$$

De este modo, el factor de corrección resulta ser el cuadrado de la diferencia entre las medias verdadera y la supuesta. Vemos inmediatamente que, si hubiéramos anticipado la media exactamente, el factor de corrección habría sido cero. Por lo tanto, cuanto mayor sea la diferencia entre las medias verdadera y supuesta tanto mayor será el factor de corrección. Una suposición deficiente conducirá siempre al resultado correcto, pero comportará marcas numéricas mayores en ambos términos de la fórmula.

Esta puede modificarse más todavía si preferimos pensar en términos de desviaciones graduales d'_i . Lo mismo que en el capítulo v, ponemos en factor la amplitud del intervalo de cada d_i y multiplicamos el resultado final por i , una vez el proceso terminado. La fórmula se convierte así en:

$$s = \sqrt{\frac{\sum_{i=1}^k f_i d_i^2}{N} - \left(\frac{\sum_{i=1}^k f_i d_i}{N} \right)^2} = \sqrt{\frac{i^2 \sum_{i=1}^k f_i d_i'^2}{N} - \left(\frac{i \sum_{i=1}^k f_i d_i'}{N} \right)^2}$$

$$\text{Luego:} \quad s = i \sqrt{\frac{\sum_{i=1}^k f_i d_i'^2}{N} - \left(\frac{\sum_{i=1}^k f_i d_i'}{N} \right)^2} \quad (\text{VI.10})$$

$$= i \sqrt{\frac{\sum_{i=1}^k f_i d_i'^2 - \frac{\left(\sum_{i=1}^k f_i d_i' \right)^2}{N}}{N}} \quad (\text{VI.11})$$

$$= \frac{i}{N} \sqrt{N \sum_{i=1}^k f_i d_i^2 - \left(\sum_{i=1}^k f_i d_i \right)^2} \quad (\text{VI.12})$$

Obsérvese que efectivamente no hemos hecho más que sacar la amplitud i del intervalo, de debajo del radical.

Al calcular la desviación estándar de datos agrupados, podemos ahora extender el procedimiento empleado para la media,

CUADRO VI.1. Cálculo de la desviación estándar utilizando datos agrupados

Límites verdaderos	Puntos medios	f_i	d_i	$f_i d_i$	$f_i d_i^2$
\$ 1 950-2 950	\$ 2 450	17	-3	-51	153
2 950-3 950	3 450	26	-2	-52	104
3 950-4 950	4 450	38	-1	-38	38
4 950-5 950	5 450	51	0	0	0
5 950-6 950	6 450	36	1	36	36
6 950-7 950	7 450	21	2	42	84
Totales		189		-63	415

$$s = i \sqrt{\frac{\sum_{i=1}^k f_i d_i^2}{N} - \left(\frac{\sum_{i=1}^k f_i d_i}{N} \right)^2}$$

$$= 1\,000 \sqrt{\frac{415}{189} - \left(\frac{-63}{189} \right)^2} = 1\,000 \sqrt{2.196 - .111}$$

$$= 1\,444$$

añadiendo la columna $f_i d_i^2$. Aunque en realidad podríamos obtener las desviaciones cuadradas d_i^2 y multiplicar luego por f_i , resultará con todo mucho más simple multiplicar las dos últimas columnas empleadas en obtener la media (esto es: $d_i \times f_i d_i$). En efecto, habiendo multiplicado d_i por sí mismo, vemos que todos los números negativos se hacen ahora positivos.⁴ Calculemos ahora la desviación estándar de los datos agrupados utilizados en el capítulo precedente. Con fines de ilustración nos serviremos de la ecuación (VI.10), pese a que por lo regular la (VI.12) comportará menos errores de redondeo.

⁴ Obsérvese bien que la última columna del cuadro VI.1 no se obtiene elevando al cuadrado la columna $f_i d_i$, ya que al hacerlo traería consigo elevar también f_i al cuadrado.

Obtuvimos en esta forma una media de \$ 5 117 y una desviación estándar de \$ 1 444. Estos dos números pueden servir ahora para resumir los datos o para compararlos con datos de otra muestra. Según veremos más adelante que pueden emplearse también para verificar hipótesis o para apreciar medidas de población.

VI.5. El coeficiente de variabilidad

Es a veces conveniente comparar varios grupos en relación con su homogeneidad relativa, en casos en que dichos grupos tienen medias distintas. Podría, pues, resultar engañoso comparar las magnitudes absolutas de las desviaciones estándar. Cabría esperar que, con una media muy grande, podría encontrarse por lo menos una desviación estándar suficientemente grande. Así, pues, alguien podría interesarse en primer lugar por el tamaño de la desviación estándar en relación con el de la media. Esto sugiere que podemos obtener una medida de la variabilidad relativa dividiendo la desviación estándar entre la media. El resultado se ha llamado *coeficiente de variabilidad* y se designa con una V. Así, pues:

$$V = \frac{s}{\bar{X}}$$

Para ilustrar las ventajas del coeficiente de variabilidad con respecto a la desviación estándar, supóngase que un psicólogo social trata de demostrar que para todos los fines prácticos dos grupos son igualmente homogéneos en relación con la edad. En uno de los grupos la edad media es de 26, con una desviación estándar de 3. En el otro la edad media es de 38 años, con una desviación estándar de 5. Por lo tanto, los coeficientes de variabilidad son respectivamente $3/26 = .115$ y $5/38 = .132$, o sea una diferencia mucho más pequeña que la que se da entre las dos desviaciones estándar. En vista del hecho de que por lo regular la edad exacta resulta menos importante, al determinar intereses, capacidades y posición social, a medida que aumenta la edad promedio de los miembros del grupo, la comparación de los dos coeficientes de variabilidad podría resultar muy bien, en este caso, mucho menos engañosa que si se emplearan las desviaciones estándar.

Si se desea, puede utilizarse también una variancia relativa. Por desgracia, estas medidas relativas de dispersión se hallan citadas con muy poca frecuencia en la literatura sociológica. Es mucho más frecuente, en efecto, encontrar las medias y las desviaciones estándar relacionadas en columnas adyacentes.

VI.6. Otras medidas resumidas

Sólo hemos examinado dos tipos de medidas resumidas: las de tendencia central y las de dispersión. Son posibles, además, otras medidas, aunque sólo se las utiliza raramente en la investigación sociológica. Sin duda, encontramos a menudo dada la distribución de frecuencia entera, pero esto no constituye una medida particular de resumen. Resulta a veces deseable indicar en una distribución el grado de asimetría. Una de las medidas de ésta saca provecho del hecho de que cuanto mayor es la asimetría tanto mayor resulta la diferencia entre la media y la mediana. Esta medida se halla dada por la fórmula:

$$\text{Asimetría} = \frac{3(\bar{X} - Md)}{s}$$

Si la distribución está desviada hacia la derecha (grandes marcas positivas), la media será mayor que la mediana, y el resultado será un número positivo. En tanto que la distribución desviada hacia la izquierda dará un resultado negativo.

Con muy poca frecuencia, también, hallamos en sociología referencias al carácter general de las cúspides de una distribución asimétrica. Utilízase el término de *picudez* en relación con dicha medida, que examinaremos brevemente una vez que hayamos visto la curva normal. Por lo regular, los textos de estadística escritos ante todo para los estudiantes de economía se ocupan más a fondo tanto de la desviación como de la picudez. Tal vez cuando empecemos a alcanzar una mayor precisión en la descripción de las formas exactas de las distribuciones de las variables sociológicas hallaremos un mayor empleo para estas otras medidas descriptivas.

GLOSARIO

Coefficiente de variabilidad
Desviación media
Desviación cuartil
Recorrido
Desviación estándar
Variancia

EJERCICIOS

1. Calcúlense las desviaciones media y estándar de los datos indicados en el ejercicio 1, cap. v. Respuesta, 9.62; 11.59.
2. Calcúlense las desviaciones estándar y cuartil de los datos agrupados en el ejercicio 1, cap. iv. Hágase lo mismo con los del ejercicio 2, cap. iv.

3. Calcúlese la desviación estándar de los datos del ejercicio 4, cap. v. Contrólense los cálculos escogiendo una media anticipada y una fórmula de cálculo distintas. Respuesta, 10.83.

4. Indíquese en qué forma resultaría afectada la desviación normal por los cambios indicados en el ejercicio 5, cap. v.

BIBLIOGRAFÍA

1. Anderson, T. R., y M. Zelditch: *A Basic Course in Statistics*, 2ª ed., Holt, Rinehart and Winston, Inc., Nueva York, 1968, pp. 76-84.
2. Downie, N. M., y R. W. Heath: *Basic Statistical Methods*, 2ª ed., Harper and Row, Publishers, Incorporated, Nueva York, 1965, cap. 5.
3. Hagood, M. J., y D. O. Price: *Statistics for Sociologists*, Henry Holt and Company, Inc., Nueva York, 1952, cap. 9.
4. McCollough, C., y L. van Atta: *Introduction to Descriptive Statistics and Correlation*, McGraw-Hill Book Company, Nueva York, 1965, cap. 3.
5. Mueller, J. H., K. Schuessler y H. L. Costner: *Statistical Reasoning in Sociology*, 2ª ed., Houghton Mifflin Company, Boston, 1970, cap. 6.
6. Weinberg, G. H., y J. A. Schumaker: *Statistics: An Intuitive Approach*, Wadsworth Publishing Company, Inc. Belmont, Cal. 1962, cap. 3.
7. Weiss, R. S.: *Statistics in Social Research*, John Wiley & Sons, Inc., Nueva York, 1968, cap. 7.

VII. LA DISTRIBUCIÓN NORMAL

LA NOCIÓN de la distribución de frecuencia es ya familiar. El presente capítulo se ocupa de un tipo muy importante de distribución de frecuencia: la curva normal. Esta distribución es muy útil, no sólo porque un gran número de distribuciones empíricas se encuentran ser aproximadamente normales, sino debido también a su significado teórico en la estadística inductiva. En este momento, el lector no debe preocuparse por las aplicaciones en las que se emplea la curva normal. En efecto, el objeto del presente capítulo está en indicar las propiedades de la curva en cuestión y en familiarizar al lector con el empleo de cuadros basados en la misma. Esta distribución se examina en la estadística descriptiva más que en la inductiva por dos razones principales. Primero, la curva normal puede emplearse para proporcionar una interpretación de la desviación estándar. Y en segundo lugar, serán útiles al lector para familiarizarse con la distribución normal algunos capítulos antes de exponerse a pruebas estadísticas que requieren facilidad en la manipulación de la misma. Por lo tanto, cuanto mejor se comprenda la materia expuesta en este capítulo, tanto menos dificultad se experimentará más adelante.

VII.1. Distribuciones de frecuencias finitas versus infinitas

Las distribuciones de frecuencia hasta aquí examinadas comportaban un número finito de casos. De hecho, por supuesto, todas las distribuciones empíricas comportan necesariamente un número finito de casos, aunque tal vez muy grande. Sin embargo, los matemáticos consideran ventajoso a menudo pensar en términos de distribuciones basadas en un número de casos infinitamente grande. Más bien que tratar con distribuciones empíricas de aspecto anguloso, como las que ejemplifican el histograma o el polígono de frecuencia, resulta posible concebir curvas lisas basadas en un número indefinidamente grande de casos y susceptibles de ser expresadas en términos de ecuaciones matemáticas relativamente sencillas. La distribución normal es una de tales curvas. Antes de examinar esta distribución específica, convendrá estudiar la naturaleza del proceso a través del cual se desarrolla una curva lisa semejante.

Empecemos con un histograma que comprende cinco intervalos (figura VII.1a). Con fines de simplicidad supondremos que la distribución de frecuencia es simétrica. Ya vimos que si el número de intervalos aumentaba sin cambiar N , la forma del histograma tiende a hacerse irregular. Supóngase, sin embargo, que

el número de casos se ha aumentado asimismo. En tal caso, como en la figura VII.1b, será posible servirse de un mayor número de intervalos más angostos, cada uno de los cuales tenga un número suficiente de casos para mantener la regularidad. Si el número de casos sigue aumentando, pueden emplearse todavía más rectángulos, conservando, con todo, el tipo regular (fig.

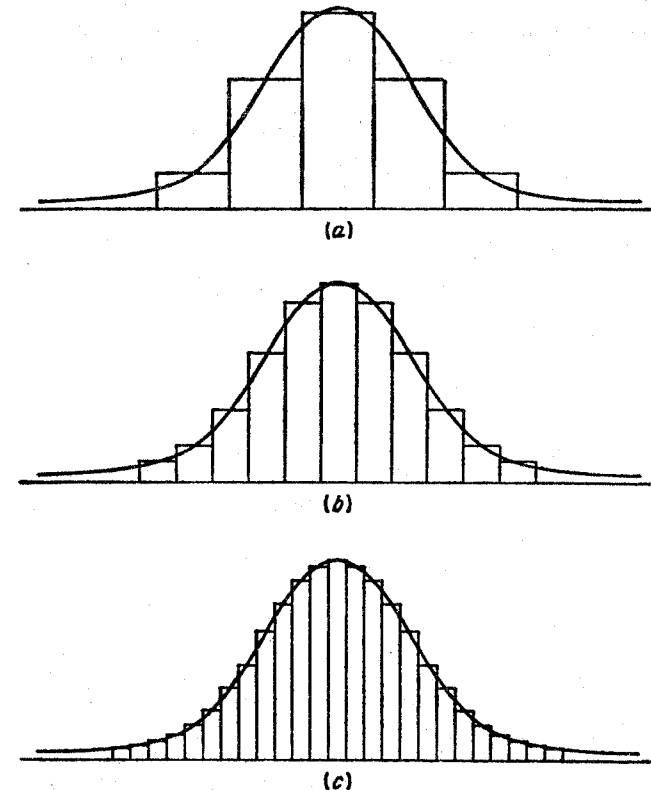


FIG. VII.1. Comparaciones de curvas lisas con histogramas de amplitudes diferentes de intervalo

VII.1c). Las curvas lisas se han trazado por los puntos medios del lado superior de cada rectángulo. Resulta claro que los rectángulos van formando aproximaciones cada vez mejores a la curva lisa a medida que el número de los mismos aumenta, esto es, a medida que disminuye el ancho de cada intervalo. Imaginemos ahora un número de casos en aumento incesante, con intervalos cada vez más angostos, hasta que los rectángulos se aproximen tan íntimamente a la curva lisa que ya no podamos

apreciar diferencia alguna entre aquéllos y ésta. Designamos la curva lisa a la que se van acercando incesantemente los rectángulos cada vez más angostos como *límite* de la distribución de frecuencia.¹ Pese a que no podamos imaginarnos un número infinito de casos, podemos, sin embargo, concebir un número tan grande de ellos, que los rectángulos se acerquen a la curva lisa con el grado de exactitud deseado.

Se recordará que el *área* de cada rectángulo puede utilizarse para representar la proporción de casos comprendidos en el in-

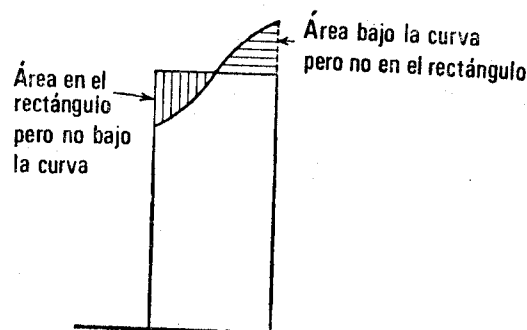


FIG. VII.2. Comparación de las áreas debajo de la curva y debajo del rectángulo

tervalo. Como ya se indicó en el capítulo IV, el área total de todos los rectángulos se suele hacer igual a la unidad. Así, pues, si la proporción de los casos del primer intervalo es .10, entonces dicho mismo número representa el área real del primer rectángulo. Observamos ahora que el área del rectángulo correspondiente puede aproximarse al área que queda debajo de la curva lisa al interior de cualquier intervalo dado. Es lo que indica la figura VII.2. A medida que el número de rectángulos aumenta, el área total de los rectángulos se convierte en una aproximación cada vez mejor al área que queda bajo la curva lisa. Esto puede verse observando que las áreas achuradas se van haciendo cada vez más pequeñas. En el límite, pues, el área debajo de la curva lisa puede obtenerse sumando las áreas de un número indefinidamente grande de rectángulos. Y como quiera que el área debajo de los rectángulos es la unidad, el área debajo de la curva lisa será asimismo igual a la unidad. El proceso que acabamos de describir es exactamente la clase de proceso que se halla en la rama de las matemáticas designada como cálculo.

¹ La noción de límite se examina también en la sección IX.1.

VII.2. Forma general de la curva normal

La curva normal es un tipo especial de curva lisa simétrica. Como quiera que la curva normal es lisa, perfectamente simétrica y se basa en un número indefinidamente grande de casos, sólo es posible aproximarse a la misma mediante distribuciones de frecuencia que comportan datos efectivos. Tiene forma de campana

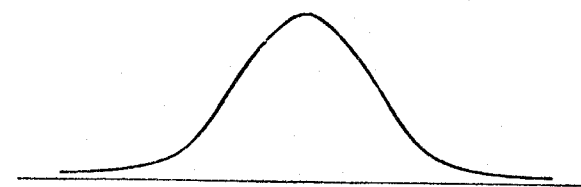


FIG. VII.3. Forma general de la curva normal

y posee cierto número de propiedades matemáticas notables, algunas de las cuales se señalarán brevemente. Como quiera que es simétrica y unimodal, su media, mediana y modo coinciden. La forma general de la distribución normal se indica en la figura VII.3.

* La ecuación matemática de la curva normal es relativamente sencilla en las normas de los matemáticos. Aunque el lector no habrá de emplear nunca dicha fórmula, ya que se han confeccionado cuadros con tal objeto, será útil, sin embargo, que la vea, para señalar y verificar algunas de las propiedades de esta distribución teórica. La fórmula es como sigue:

$$Y = \frac{1}{s\sqrt{2\pi}} e^{-\frac{(x-\bar{x})^2}{2s^2}}$$

en donde Y es la altura de la curva para un valor determinado de X . Como quiera que tanto π como e son constantes (iguales respectivamente a 3.14 y 2.72), la fórmula sólo comporta dos medidas de resumen, la media \bar{X} y la desviación estándar s .² Por lo tanto, la forma exacta de la curva normal será conocida si se nos dan los valores de dichas medidas. En otros términos: hay muchas curvas normales, una para cada combinación de la media y de la desviación estándar.

* Recordando que una cantidad afectada de un exponente negativo puede escribirse como la recíproca de dicha cantidad ele-

² Cuando lleguemos a la estadística inductiva se introducirá otra notación para la media y la desviación estándar. La fórmula de la curva normal suele escribirse en términos de una media de μ y una desviación estándar de σ .

vada a la potencia positiva, podemos escribir la fórmula como sigue:

$$Y = \frac{1}{s\sqrt{2\pi}} \left(\frac{1}{2.72(\bar{x} - x)^2/2s^2} \right)$$

en la que e ha sido sustituida por su valor numérico. Supongamos que el valor de s es fijo, y busquemos el valor de X para el

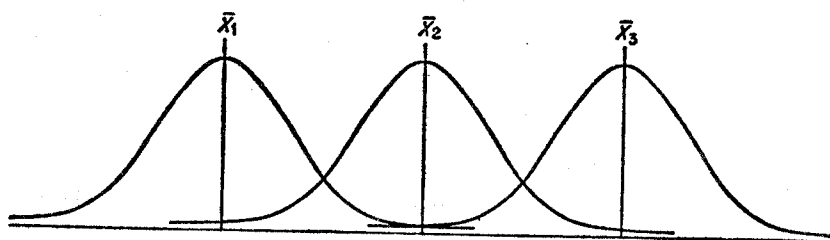


FIG. VII.4. Comparación de curvas normales de igual desviación estándar pero de medias diferentes

cual Y será un máximo. Es obvio que Y será máxima cuando el denominador incluido en los paréntesis sea mínimo. Pero dicho denominador consta de un número positivo mayor que la unidad elevado a una potencia que no puede ser negativa, ya que un número real cuadrado no puede ser nunca menor que cero. Por consiguiente, el denominador alcanzará su mínimo cuando el exponente sea cero. Y esto ocurrirá cuando X adopte el valor de \bar{X} , ya que tendremos $X - \bar{X} = 0$. Esto muestra que el modo (y, por consiguiente, la media y la mediana) es realmente \bar{X} , hecho que ya se había señalado, pero sin demostrarlo. Podemos ver, asimismo, que la ecuación da una curva que es simétrica alrededor de X . Como quiera que la cantidad $X - \bar{X}$ está al cuadrado y no puede, por consiguiente, ser negativa, las desviaciones respecto de \bar{X} en una u otra dirección producirán valores idénticos de Y .

La ecuación específica para toda curva normal particular puede obtenerse empleando los valores propios de \bar{X} y s . En la figura VII.4 pueden verse curvas normales de la misma desviación estándar, pero de medias distintas. Por otra parte, las curvas de desviaciones estándar distintas variarán en la configuración de las cúspides, tal como se indica en la figura VII.5. Cuanto menor sea la desviación normal, tanto más puntiaguda resultará la curva.

Habría que señalar que no todas las curvas simétricas en forma de campana son normales. Aunque las curvas de la figura VII.5 difieran en cuanto a las cúspides, esto se debe únicamente a diferencias en sus desviaciones normales. Todas ellas son normales en cuanto a la forma. Por regla general, las curvas simétricas unimodales pueden ser más o menos puntiagudas o apla-

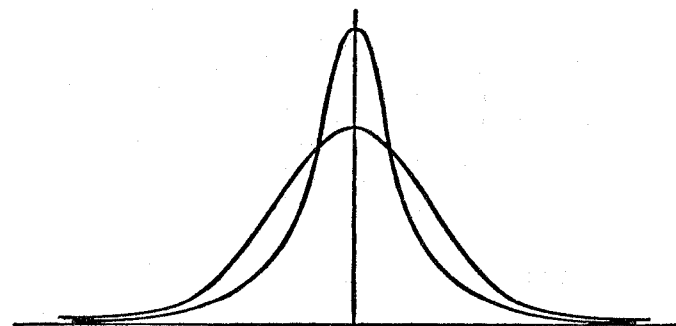


FIG. VII.5. Comparación de dos curvas normales de medias iguales pero con desviaciones estándar diferentes

nadas que la curva normal, aun siendo sus desviaciones estándar las mismas. Algunas de estas curvas pueden verse en la figura VII.6. Las que son más puntiagudas que la normal se de-

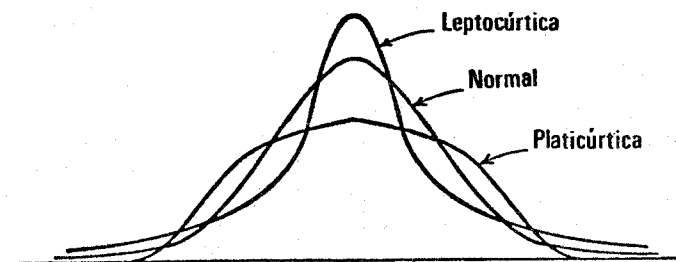


FIG. VII.6. Comparación de una curva normal con curvas de su misma desviación estándar pero distintas en cuanto a las cimas

signan como leptocúrticas y las más planas que aquélla como platicúrticas. A diferencia de la curva normal, las ecuaciones de las curvas leptocúrticas y platicúrticas tienden a comportar medidas de resumen, además de la media y la desviación estándar.

VII.3. Áreas bajo la curva normal

Con frecuencia es necesario determinar la proporción de casos

que quedan al interior de un intervalo dado. Afortunadamente, la curva normal posee una propiedad importante que hace que dicha tarea resulte relativamente sencilla. En efecto, resulta que independientemente de la media o de la desviación normal que una curva ostente, habrá un *área constante* (o proporción de casos) *entre la media y una ordenada, que es una distancia*

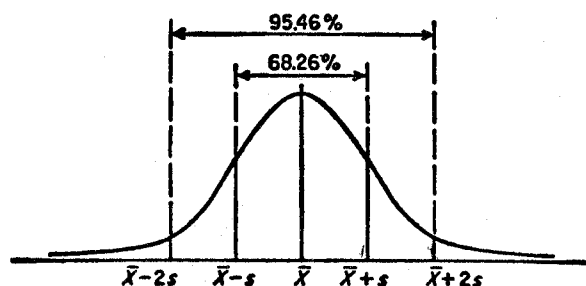


FIG. VII.7. Áreas debajo de la curva normal

cia determinada a partir de la media en términos de unidades de desviación estándar. La figura VII.7 ayuda a ilustrar el sentido de esta afirmación.

Así, pues, si vamos en una desviación estándar a la derecha de la media, encontraremos siempre .3413 del área incluida entre la media y la ordenada en dicho punto. Por consiguiente, dos veces dicha área, o .6826 estarán incluidas entre las dos ordenadas situadas a una desviación estándar a ambos lados de la media. En otros términos: un poco más de dos tercios de los casos se encontrarán siempre en el interior de una desviación estándar de la media. Y en forma análoga, el área comprendida entre la media y la ordenada a dos desviaciones estándar de aquélla será siempre .4773 y, por lo tanto, un poco más del 95 por ciento del área estará comprendido entre la pareja de ordenadas a dos desviaciones estándar a ambos lados de la media. Prácticamente, todos los casos estarán comprendidos en el interior de tres desviaciones estándar de la media, aunque la curva normal se extienda teóricamente al infinito en ambas direcciones. Por supuesto, las distancias de la media no necesitan ser siempre múltiplos exactos de la desviación estándar. Mediante un procedimiento que vamos a describir en breve, es posible determinar las áreas entre dos ordenadas cualesquiera. Por ejemplo, si nos apartamos en 1.96 desviaciones estándar a ambos lados de la media, comprenderemos casi exactamente el 95 por ciento del área, en tanto que entre las ordenadas a 2.58 desviaciones normales de la media quedará incluido el 99 por ciento del área.

Esta propiedad de la curva normal brinda una interpretación

de la desviación normal y un método para representar en forma visual el significado de esta medida de dispersión. Cierta número de distribuciones empíricas de frecuencia son lo bastante semejantes para que estas relaciones entre las áreas y la desviación normal se verifiquen razonablemente bien. Inclusive en el caso de distribuciones de ingresos, que propenden a distorsionarse en la dirección de los ingresos elevados, encontramos normalmente dos tercios de los casos en el interior de una desviación estándar de la media. Hay que tener presente, con todo, que aunque la curva normal proporciona una *interpretación* de la desviación estándar, esta propiedad no puede emplearse para *definir* lo que se entiende por desviación estándar. La definición se hace en términos de la fórmula. La propiedad en cuestión sólo se verifica en el caso de distribuciones normales o aproximadamente tales.

Resulta posible tomar cualquier curva normal y transformar sus valores numéricos de tal forma que pueda utilizarse un simple cuadro para evaluar la proporción de casos al interior de cualquier intervalo deseado. Vamos a ilustrar este proceso por medio de un ejemplo numérico. Supongamos que tenemos una curva normal con una media de 50 y una desviación estándar de 10. Busquemos la proporción de los casos en el intervalo de 50 a 65. Empezamos por determinar a cuántas desviaciones estándar se halla 65 de la media 50. Para ello tomamos la diferencia entre estos dos valores, esto es, 15, y dividimos entre la magnitud de la desviación estándar. En el presente caso el resultado es 1.5. De modo general podemos servirnos de la fórmula:

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

$$= \frac{65 - 50}{10} = 1.5$$

en donde X es el valor de la ordenada y Z representa la desviación con respecto a la media en unidades de desviación estándar.

* Antes de examinar cómo puede utilizarse el valor numérico de Z para determinar la proporción de los casos entre la media y la ordenada correspondiente a Z , permítasenos dar una interpretación alternativa de ésta. Podemos pensar en términos de una transformación efectiva de la variable X en la variable Z . En tanto que la distribución de la variable X es normal con una media de \bar{X} y una desviación estándar de s , la nueva variable, en cambio, es normal con una media de cero y una desviación estándar de uno.³ Esta desviación con una media cero y una

³ La verificación de este hecho se deja como ejercicio (véase ejercicio 3).

desviación estándar de uno se designa como *forma estándar*, y la Z se designa a menudo como la *marca*. La transformación de variables se ilustra en la figura VII.8. Sustraemos de cada X la constante \bar{X} . Al sustraer este valor constante (aquí 50) de cada X , hemos corrido cada marca original en 50 unidades a la izquierda y, por lo tanto, hemos desplazado efectivamente la curva

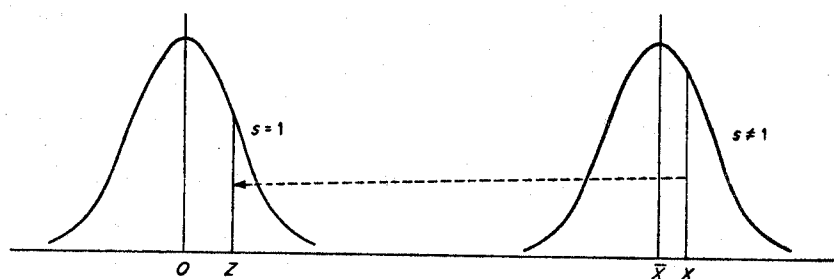


FIG. VII.8. Comparación de las formas estándar y general de la curva normal

normal original a una posición directamente sobre el origen. Esto tiene en cuenta el numerador en la expresión de Z . Dividimos ahora cada diferencia $X - \bar{X}$ entre la magnitud de la desviación estándar. Al hacerlo, o estrechamos la curva o la ensanchamos, según que su desviación estándar sea o no mayor que la unidad. Podemos, pues, pensar que hemos desplazado primero la posición de la curva normal original y que luego hemos cambiado la magnitud de la desviación estándar, de modo que quede sobre la forma estándar. Al dividir entre la desviación estándar de 10, hemos cambiado esencialmente las unidades a lo largo del eje horizontal, de modo que una distancia de 10 sobre el eje de X corresponde a la distancia de 1 sobre el eje de Z .

Independientemente de la interpretación que se dé, un valor de $Z = 1.5$ indica que la ordenada se encuentra a 1.5 desviaciones estándar de la media. En el caso de la forma estándar, esto significa, por supuesto, que la ordenada misma coincide con el valor 1.5 de la escala Z . Se han construido tablas que muestran áreas exactas para la forma estándar de la curva normal. El cuadro C del Apéndice 2 es una de ellas. Los valores de Z se dan de arriba abajo en el margen izquierdo, y horizontalmente arriba. Los dos dígitos de Z se obtienen leyendo de arriba abajo, y el tercero leyendo horizontalmente. Las cifras del cuerpo del cuadro indican la proporción del área entre la media (o sea cero) y la ordenada correspondiente a Z . En el ejemplo anterior, vemos que se hallan contenidas en dichos límites las .4332 del área. Si Z hubiera sido 1.52, el área correspondiente habría sido .4357.

VII.4. Ilustraciones suplementarias del empleo de la tabla normal

Supongamos que queremos hallar el área achurada de la curva normal indicada en la figura VII.9. En este caso el valor de Z es:

$$Z = \frac{143 - 168}{12} = \frac{-25}{12} = -2.08$$

El hecho de que Z sea negativa indica simplemente que el área achurada se sitúa a la izquierda de la media. Al utilizar la tabla

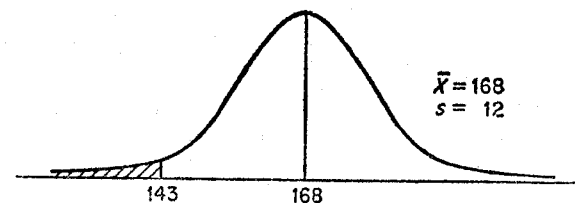


FIG. VII.9. Curva normal, con porción achurada representando el área en una sola cola

normal, el signo de Z puede ignorarse, ya que la curva es perfectamente simétrica. Del cuadro vemos que el área comprendida entre la media y una Z de 2.08 es .4812. Como quiera que el área total es la unidad, el área a la izquierda de la media ha de ser .5 (por simetría). Por consiguiente, el área achurada puede obtenerse restando el área comprendida entre la media y la ordenada del área total a la izquierda de la media. Así, pues:

$$(\text{Proporción de casos} \leq 143) = .5000 - .4812 = .0188$$

Por lo tanto, menos del 2 por ciento de los casos tienen marcas inferiores o iguales a 143.⁴ El tipo de problema ilustrado en este ejemplo es muy corriente, debido al hecho de que las comprobaciones de hipótesis casi siempre comprenden las colas de una distribución de frecuencia. Si hubiéramos querido hallar el área

⁴ En una distribución continua, la proporción de los casos que sean exactamente 143.0 será cero. Esto puede verse si imaginamos dos ordenadas extremadamente próximas una de otra. La proporción de casos entre estas dos ordenadas será también muy pequeña. Y si a continuación dejamos que las dos ordenadas se vayan aproximando indefinidamente, la proporción de los casos se hará infinitamente pequeña. Recuérdese que la línea matemática no tiene grueso. En la práctica podrá haber algunos casos con marcas de 143.0, debido a defectos de medición. Sin embargo, como quiera que estamos tratando de una distribución teórica, no importa que la ordenada ella misma se incluya o no en el intervalo. En adelante, nos referiremos simplemente al área entre dos ordenadas (pero sin comprender a éstas), o área inferior a un valor dado.

total fuera de la región definida por 168 ± 25 (como la indican las áreas achuradas de la figura VII.10), habríamos doblado simplemente el resultado anteriormente obtenido, ya que las dos áreas achuradas son exactamente del mismo tamaño.

Para tomar otro ejemplo, supongamos que necesitamos obtener el área achurada indicada en la figura VII.11. Esta área se

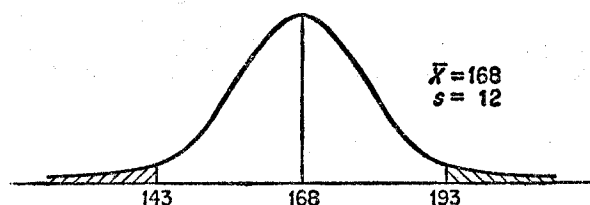


FIG. VII.10. Curva normal, con porciones achuradas presentando áreas en ambas colas

calcula hallando primero la proporción de casos entre la media y la ordenada B y sustrayendo luego la proporción de casos entre la media y la ordenada A . Las Z correspondientes a B y A son respectivamente 2.0 y 1.2. Tenemos, pues:

Proporción entre B y la media	.4773
Proporción entre A y la media	.3849
Proporción entre A y B	.0924

Por consiguiente, ligeramente más del 9 por ciento de los casos quedan entre .42 y .46. Obsérvese que si se hubiera deseado obte-

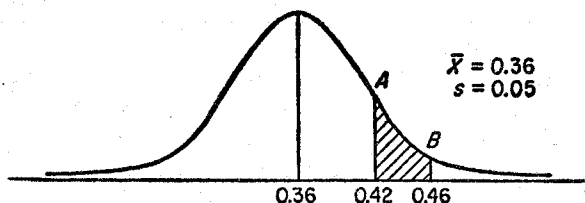


FIG. VII.11. Curva normal con porción achurada, representando el área entre dos ordenadas

ner el área entre ordenadas a ambos lados de la media, el resultado se habría obtenido más fácilmente por adición que por sustracción.

GLOSARIO

Leptocúrtico
Distribución de frecuencia límite
Curva normal
Platicúrtico
Marca estándar

EJERCICIOS

1. Ya se calcularon la media y la desviación estándar de los datos del ejercicio 1, capítulo iv. ¿Cuál fue la proporción de los casos dentro de una desviación estándar de la media? ¿Al interior de dos desviaciones estándar? ¿De tres desviaciones estándar? ¿Con cuánta aproximación corresponden dichas cifras a las que encontraríamos si la distribución fuese exactamente normal? Contéstese a las mismas preguntas en relación con el ejercicio 2, capítulo iv. Compárense y explíquense las diferencias entre los resultados de los dos grupos de datos.

2. Si la media de una distribución normal es de 80 y su desviación estándar de 12,

- ¿Qué proporción de casos se halla entre 80 y 93? Respuesta, .3606.
- ¿Qué proporción de casos se halla entre 90 y 105? ¿Entre 70 y 105? Respuesta, .1838.
- ¿Qué proporción de casos es inferior a 68?
- ¿Cuántas desviaciones estándar se necesitarían a ambos lados de la media para obtener dos colas que comprendan cada una el 2 por ciento exactamente del área total? ¿El 10 por ciento del área total? Respuesta, 2.054.
- ¿Cuál marca tiene por encima de ella el 4 por ciento de los casos? (en otros términos, sitúese la percentil 96).

* 3. Verifíquese que la forma estándar de la curva normal tiene una media de cero y una desviación estándar igual a la unidad. (Indicación: vuelva a escribirse la fórmula de la curva normal en términos de Z , aprovechando el hecho de que $Z = (X - \bar{X})/s$.)

4. Las calificaciones primarias de diversas pruebas de aptitud y actitud son tratadas a menudo por los psicólogos como escalas de intervalo. Dichas calificaciones suelen a menudo convertirse luego en calificaciones estándar con medias y desviaciones estándar convenientes. Supóngase que la calificación media primaria en un examen de admisión en la universidad es de 117 con una desviación estándar de 28.5. Supóngase, además, que esas calificaciones primarias están distribuidas normalmente.

- ¿Cuál es la proporción de calificaciones por encima de 131? ¿Debajo de 79?
- ¿Cuáles son las calificaciones primarias correspondientes a los cuartiles primero, segundo y tercero?
- *c) En los exámenes de la universidad, las calificaciones primarias se normalizan de modo que la media de la distribución normal

sea exactamente de 500 y la desviación estándar de 100. Concretamente, ¿cómo se normalizarán los grupos de datos anteriores para obtener una media de 500 y una desviación estándar de 100? (Indicación: ¿cómo se normalizaría para obtener una media igual a cero y una desviación estándar igual a la unidad?)

BIBLIOGRAFÍA

1. Downie, N. M., y R. W. Heath: *Basic Statistical Methods*, 2ª ed., Harper and Row, Publishers, Incorporated, Nueva York, 1965, cap. 6.
2. Hagood, M. J., y D. O. Price: *Statistics for Sociologists*, Henry Holt and Company, Inc., Nueva York, 1952, cap. 14.
3. Mueller, J. H., K. Schuessler y H. L. Costner: *Statistical Reasoning in Sociology*, 2ª ed., Houghton Mifflin Company, Boston, 1970, cap. 6.
4. Weinberg, G. H., y J. A. Schumaker: *Statistics: An Intuitive Approach*, Wadsworth Publishing Company, Inc., Belmont, Cal., 1962, cap. 8.
5. Weiss, R. S.: *Statistics in Social Research*, John Wiley & Sons, Inc., Nueva York, 1968, pp. 147-156.

TERCERA PARTE

ESTADÍSTICA INDUCTIVA

VIII. INTRODUCCIÓN A LA ESTADÍSTICA INDUCTIVA

EL OBJETO de este breve capítulo está en dar un bosquejo general de la estadística inductiva y, en particular, de los razonamientos que se hallan a la base de la verificación de las hipótesis estadísticas. Es muy fácil verse tan abrumado por los detalles de cada prueba particular encontrada, que resulte imposible percibir las semejanzas que todas ellas tienen en común. En tal caso, el aprendizaje de la estadística puede convertirse fácilmente en un ejercicio de "recetario" consistente en aprender de memoria las fórmulas y los procedimientos. Por ello este capítulo es muy importante y debería releerse atentamente una vez que el lector haya visto las dos o tres primeras pruebas específicas.¹

VIII.1. Estadística y parámetros

El objeto de las generalizaciones estadísticas está en decir algo acerca de diversas características de la población estudiada, sobre la base de hechos conocidos a propósito de una muestra sacada de dicha población o universo.² Designaremos las características de la población como *parámetros*, en contraste con las características de la muestra, que se designan como *estadísticos*. A estas alturas, el lector está ya familiarizado con cierto número de parámetros y estadísticas: medias, medianas, proporciones, desviaciones estándar, etcétera. Llegado aquí, el lector debería hacer una distinción precisa entre las características que se refieren a la población y las que se refieren a la muestra. Para designar las primeras suelen por lo regular emplearse las letras del alfabeto griego, en tanto que las letras latinas designan las características de la muestra.³ Así, pues, en adelante designaremos la media de la población con μ , y la de la muestra con \bar{X} ; la desviación estándar de la primera con σ , y la de la muestra con s .

Puede establecerse una distinción importante entre parámetros y estadísticas. En efecto, los parámetros son valores *fijos* referidos a la población y, por lo regular, *no se conocen*.⁴ Así, por

¹ Un momento muy adecuado para la nueva lectura será después del capítulo XI.

² Los términos *población* y universo (*universe*) suelen ser intercambiables en la literatura estadística.

³ Por desgracia, esta regla comporta cierto número de excepciones.

⁴ Los parámetros se tratarán siempre como fijos, aunque en realidad pueden variar con el tiempo. Así, por ejemplo, la edad promedio de una población variará de un momento al próximo. Por ello deberá entenderse la noción de muestras repetidas en términos de un gran número de muestras sacadas simultáneamente, y no en secuencia temporal. En muchas ocasio-

ejemplo, en cualquier momento dado, la edad o el grado promedios de todos los estudiantes de la Universidad de Harvard pueden no conocerse, pero se hallarán probablemente ser los mismos para todos los observadores. Las estadísticas varían, por otra parte, de una muestra a otra. Si se escogieran al azar 10 muestras diferentes de estudiantes universitarios, no esperaríamos que todos ellos presentaran exactamente las mismas edades promedios. Es más, desconfiaríamos si fuera así. A diferencia de los parámetros, los valores de las estadísticas de una muestra determinada se conocen o pueden calcularse. Pero *no* sabemos, sin embargo, cuán representativa sea la muestra en relación con la población, o hasta qué punto la estadística obtenida se aproxime al parámetro desconocido comparable.

Lo que nos interesa efectivamente es la población, y no en una muestra particular cualquiera. En efecto, escogemos una muestra por razón de conveniencia, pero nuestro objetivo consiste prácticamente siempre en sacar conclusiones a propósito de diversos parámetros de la población; sobre la base de estadísticas de muestras conocidas, sin duda, pero sin importancia en sí mismos. En las verificaciones de hipótesis formulamos supuestos a propósito de los parámetros desconocidos, y preguntamos a continuación cómo serían nuestras estadísticas específicas si dichos supuestos fueran correctos. Al proceder así, tratamos de decidir racionalmente si los valores supuestos de dichos parámetros son o no razonables a la vista de la evidencia de que disponemos. Por lo tanto, la verificación de hipótesis puede considerarse como una forma especial de proceso de decisión. Como quiera que los raciocinios que se hallan a la base de las hipótesis son más bien complejos, no estará por demás examinarlos aquí brevemente. En capítulos sucesivos veremos cómo se aplican a las verificaciones específicas.

VIII.2. Pasos en la verificación de una hipótesis

En ciencias sociales el término de *hipótesis* se emplea en cierto número de sentidos distintos. A veces se utiliza para designar una proposición teórica que presenta alguna remota posibilidad de verificación indirecta. Otras veces, en cambio, se emplea para designar el tipo de afirmación que puede efectivamente comprobarse estadísticamente. Con objeto de reducir la confusión será, pues, conveniente indicar cómo se emplea el término en este texto. Los criterios utilizados para definir lo que debamos en-

nes, nuestro objetivo científico consiste en realidad en deducir la naturaleza de los procesos causales que dan origen a los valores de población, los que suponemos son fijos. Al estudiar estadística parece sin embargo prudente limitarse inicialmente al concepto más simple de generalización de poblaciones fijas.

tender por verificación de una hipótesis son más bien estrictos y descartarían muchas de las llamadas "pruebas" que se encuentran en la literatura corriente en materia de ciencias sociales. Sin embargo, son adecuados a los requisitos más bien rígidos establecidos por los estadígrafos. En cuanto tales, en efecto, representan un ideal con referencia al cual puede compararse el carácter adecuado o inadecuado de cualquier comprobación real.

La hipótesis es un enunciado acerca de un acontecimiento futuro, o de un acontecimiento cuyo resultado se desconoce en el momento de la predicción, formulado de modo que pueda descartarse. O en términos más precisos, digamos que se ha comprobado una hipótesis cada vez que se han efectuado los siguientes pasos:

1. Todos los resultados posibles del experimento u observación se han *anticipado a la verificación*.⁵

2. Se ha llegado a un acuerdo, antes de proceder a la verificación, acerca de las operaciones o procedimientos a emplear en la determinación de cuáles resultados se producían efectivamente.

3. Se ha decidido previamente cuáles de los resultados implicarán, caso de producirse, el descarte de la hipótesis y cuáles su confirmación. Como resulta de lo indicado más arriba, el descarte ha de haberse tenido en cuenta como uno de los resultados posibles.

4. Se ha efectuado el experimento, o se ha observado el acontecimiento, se han registrado los resultados y se ha decidido si la hipótesis quedaba o no descartada.

Los pasos que se acaban de enumerar son muy generales. La inducción estadística tiene que ver ante todo con los pasos 3 y 4, ya que el estadígrafo ha de suponer que los dos primeros pasos se han efectuado ya. Tendremos ocasión de ver en qué forma los dos últimos pasos se hacen más específicos en una comprobación estadística. Tal vez lo más significativo de la lista anterior es el de que todas las decisiones deben ser tomadas antes de realizar la prueba. Todos los resultados posibles se dividen en dos clases, a saber: los que comportan descarte y los que no. Si eso no se hace con anterioridad a la prueba, resulta posible retener una hipótesis cambiando simplemente las reglas a medida que se avanza. Esto equivale a lo mismo que lo que haría un niño que echara a cara o cruz para decidir si va o no al cine. Éste decide, en efecto, "cara, voy; cruz, no voy". Si sale cara va al cine. Pero si sale cruz, decide hacer depender el éxito de dos cara o cruz sobre tres y sigue echándolos. En esta forma acaba siempre yendo al cine, a menos que pierda la moneda (resultado que no había anticipado).

⁵ El término *experimento* lo emplea el estadígrafo en un sentido muy amplio. Un experimento puede consistir, por ejemplo, en interrogar a un ama de casa y anotar el "sí" o el "no" a una pregunta concreta.

* Ya se indicó en el capítulo II que la prueba sólo puede hacerse acerca de una proposición formulada en conceptos que se hayan definido operativamente. El paso 2 indica que hay que ponerse de acuerdo, con anterioridad a la prueba, acerca de las definiciones operativas. A menos que sea así, resulta siempre posible retener una hipótesis, independientemente del resultado, descartando los métodos empleados. Supóngase que alguien enuncia como hipótesis suya que "cuanto más elevada sea la posición social de una persona, tanto menos probable será que sea muy etnocéntrica". Si los resultados no confirman esta proposición, podrá alegar que la medida "posición social" o "etnocentrismo" no medía lo que se la suponía medir, y que algún otro índice (que confirme su teoría) es más adecuado. Así, pues, parece preferible reservar el término de *hipótesis* para designar enunciados que se hallen al nivel operativo y puedan descartarse francamente. En efecto, si no se puede llegar de antemano a un acuerdo acerca del procedimiento, es difícil que se produzca acuerdo a propósito de los resultados. Como ya se indicó en el capítulo II, este punto de vista no niega, con todo, la importancia de la teoría, ni implica que las definiciones operativas sean las únicas necesarias para el desarrollo de la ciencia.

El tercer paso es crítico, ya que la decisión que se adopte comportará por lo regular ciertos peligros de error. En algunos casos el problema es relativamente sencillo. No todas las verificaciones de las hipótesis requieren inducción. En efecto, puede formularse una hipótesis a propósito del resultado de un acontecimiento concreto, tal, por ejemplo, un partido de fútbol. Podemos predecir, por ejemplo, que el equipo A ganará al equipo B. A condición que existan criterios para determinar si los procedimientos acordados se han seguido adecuadamente o no, las probabilidades de error en cuanto a decidir si hay que descartar o no semejante tipo de hipótesis son escasas. Sin embargo, si la información se basa en una muestra de acontecimientos sacado de una población mayor, existe mayor riesgo de error. En efecto, descartamos o dejamos de descartar la hipótesis dándonos cuenta de que, ya que nuestro juicio sólo se basa en una muestra, hemos de admitir siempre la posibilidad de error debida a la falta de carácter suficientemente representativo del mismo. Es la teoría de las probabilidades la que nos permite apreciar los riesgos de error y tomarlos en consideración al decidir acerca de los criterios que hay que emplear para descartar las hipótesis. En las próximas secciones se examinarán dos tipos de errores posibles. Podremos luego volver a la cuestión del papel que juega la estadística en las verificaciones de las hipótesis inductivas.

VIII.3. La falacia de afirmar el consecuente

A menudo no existe manera alguna de verificar nuestras proposiciones o teorías más importantes. En lugar de ello, extraemos de éstas una serie de consecuencias que deberían producirse si la proposición o teoría original fuese cierta, y es la validez de estas consecuencias la que se deja determinar por métodos empíricos.⁶ Así, pues, la prueba de la teoría original es indirecta. La teoría A implica determinadas consecuencias B, o bien, en forma simbólica, $A \Rightarrow B$. Hay que recalcar que, al pasar de A a B, se emplea más bien el razonamiento lógico o deductivo que la prueba empírica. Por consiguiente, si A es cierto, B lo ha de ser también, a condición que nuestro razonamiento al deducir A de B sea válido. Vemos luego si B se ha producido o no; si B no se ha producido (B falso), entonces sabemos también que la teoría A ha de ser falsa asimismo.

Pero, ¿qué ocurre si B resulta ser cierto? ¿Podemos decidir que A *deba* serlo asimismo? No. Si lo hacemos, cometemos la falacia de afirmar el consecuente, como los lógicos acostumbra llamarlo. Si B es cierto, podemos decir que A *puede* ser cierto, pero podría haber otro número cualquiera de teorías alternativas que implicaran también B. No podemos estar seguros de que A sea *necesariamente* cierto, a menos que podamos demostrar que no existe otra teoría alternativa válida C para la cual $C \Rightarrow B$. Por desgracia, no estamos prácticamente nunca en condiciones de hacerlo, y por ello más bien hemos de proceder por *eliminación* de teorías que por su aceptación definitiva. La buena teoría es la que no se deja eliminar, a condición, por supuesto, que se la enuncie en forma que se deje eliminar.⁷ En otros términos: ha de conducir a hipótesis que se dejen eliminar ellas mismas. Si dejamos de descartar A cuando B es cierto, corremos riesgo de equivocarnos, ya que A puede en realidad ser falso. En estadística, ese tipo de error, o sea *el error de no descartar una hipótesis efectivamente falsa*, se designa como *error de tipo II* o β .

Tal vez un sencillo ejemplo hará que el razonamiento anterior se presente como menos abstracto. Supongamos que tenemos una teoría A que consta de las tres proposiciones siguientes: 1) todas las personas se conforman a las normas de su sociedad;

⁶ En rigor este enunciado no es totalmente exacto, ya que una teoría puramente deductiva no conduce *directamente* a hipótesis comprobables. Véase [2].

⁷ El papel del experimento crítico está en poner al científico en condiciones de escoger entre varias teorías alternativas cada una de las cuales ha resistido previamente a la eliminación. Así, por ejemplo, las teorías A y A' pueden predecir ambas los acontecimientos B_1, B_2, \dots, B_k , todos los cuales se producen. Pero A puede predecir que B_{k+1} es cierto, en tanto que A' sostenga que será falso. Si B_{k+1} es efectivamente falso, entonces A puede eliminarse, y retenerse, de momento, A'.

2) una norma de la sociedad X es la de no robar; y 3) Jones es miembro de la sociedad X . Si todas las partes de la teoría son correctas, podemos deducir B , que Jones no robará. Supóngase que por alguna otra razón no estamos en condiciones de verificar directamente lo cierto o falso de A , pero que estamos en condiciones, en cambio, de averiguar la conducta de Jones. Es obvio que si Jones roba, la teoría ha de ser incorrecta, por lo menos en parte. En consecuencia, si B es falso, descartamos A . Pero, si sabemos que Jones no roba, no por ello decidiremos que la teoría sea cierta. Tal vez Jones sea simplemente más honrado que los otros. O tal vez ni siquiera sea miembro de la sociedad X . En semejante caso, si fuéramos a aceptar la teoría como correcta, correríamos un riesgo considerable de error. Llegaríamos probablemente a la conclusión de que, aunque el individuo particular en cuestión sea honrado, haríamos mejor en suspender nuestro juicio.

El absurdo del ejemplo anterior no ha de oscurecer el punto capital de que, siempre que tengamos una teoría que implica determinadas consecuencias y que éstas, pero no así aquélla, sean susceptibles de verificación, nos encontramos en la posición lógica de poder descartar la teoría, en tanto que no podemos aceptarla, en cambio, sin correr el riesgo de equivocarnos.

VIII.4. La forma de las hipótesis estadísticas

En ciencias sociales no encontramos proposiciones por el estilo de la del ejemplo anterior, por la sencilla razón de que las teorías acerca del mundo real no implican certidumbre. En lugar de considerar que si A es cierto B ha de serlo asimismo, sostenemos solamente que si A es cierto B lo será *probablemente* también. Tenemos así que aceptar la posibilidad de que B sea falso incluso cuando A es verdadero. Pero si seguimos la regla de descartar A siempre que B sea falso, corremos el riesgo de cometer otro error, esto es, el de *descartar una hipótesis cierta*. Designamos esta clase de error como *error de tipo I* o α . Sirviéndonos del ejemplo anterior, nuestras proposiciones habrán de modificarse en el sentido de decir: "*la mayoría de los individuos se conforman a las normas de su sociedad*" y "*probablemente Jones no robará*". Si Jones roba, descartamos la teoría revisada con cierto riesgo de error, ya que puede con todo ser cierta, porque es posible que Jones sea uno de los pocos miembros no honrados.

Así, pues, existen dos tipos de error que hay que tener en cuenta. El primero que examinamos (el tipo II) procede de la falacia puramente lógica consistente en afirmar el consecuente. Y cuando introducimos elementos de probabilidad en nuestra teoría, entonces admitimos un tipo adicional de error (el tipo I). Aunque hasta el presente no hayamos dicho nada todavía a pro-

pósito del razonamiento inductivo en contraste con el deductivo, se debe a la necesidad de generalizar más allá de los límites de los datos que se poseen el que debamos servirnos de semejantes enunciados de probabilidad.

¿Qué forma específica adoptan las hipótesis estadísticas? ¿A qué se parecen el A y el B ? En realidad, la teoría A consta de cierto número de supuestos acerca del carácter de la población y de los procedimientos relativos a la selección de muestras, junto con el razonamiento matemático necesario para formular enunciados de probabilidad a propósito de la de los resultados particulares de la muestra, si los supuestos adoptados son efectivamente ciertos. Por medio de estos enunciados de probabilidad decidimos con anterioridad al tiempo cuáles resultados son tan probables, que descartaríamos los supuestos A si estos resultados B no se produjeran. Razonamos, en efecto, en el sentido de que, si los supuestos son correctos, los resultados de nuestras muestras quedarán la mayor parte del tiempo dentro de un determinado recorrido de resultados. Por supuesto, sólo extraemos una muestra, pero si nuestro resultado particular cae fuera del recorrido, en lo que se denomina *región crítica*, rechazaremos los supuestos, corriendo el riesgo de cometer un error tipo I. Así, pues, el B está representado por cierto recorrido de resultados de muestras. Si los resultados quedan fuera de dicho recorrido, entonces B es falso y la hipótesis se descarta. Al decidir la extensión del recorrido a incluir bajo B , hemos de tomar en consideración (idealmente) los riesgos de errores de los tipos I y II.

Para ilustrar el proceso, supongamos que deseamos comparar muestras de empleados de oficina y de obreros de taller en relación con el porcentaje de ellos que desean para sus hijos enseñanza universitaria. Si queremos realmente demostrar que existe una diferencia entre dichos dos grupos, procedemos tratando de eliminar la hipótesis alternativa de que no existe diferencia alguna. Esto parece constituir una manera de proceder extremadamente indirecta, pero hemos de recordar que no estaremos en condiciones de demostrar directamente que sí hay diferencia. Con objeto de evitar la falacia de afirmar el consecuente, hemos de proceder a la eliminación de las falsas hipótesis. En el presente caso sólo existen lógicamente dos posibilidades: o hay diferencia o no la hay. Si la segunda posibilidad se deja eliminar, entonces podemos concluir que existe efectivamente alguna diferencia.

Establecemos, por consiguiente, la hipótesis de que el porcentaje que desea la enseñanza universitaria es el mismo en ambos grupos o poblaciones. Podemos a continuación demostrar matemáticamente que, en el 99 por ciento de todos los pares posibles de muestras, las diferencias entre las dos series de porcentajes

serían inferiores al 10 por ciento si los supuestos fueran efectivamente ciertos. En otros términos: *B* consta de diferencias de muestras que son inferiores al 10 por ciento. Y si realmente no existen diferencias entre ambas poblaciones, es sumamente probable que los porcentajes correspondientes a las dos muestras caerán dentro del 10 por ciento uno de otro. Puede, en consecuencia, decidirse que, si la diferencia entre los porcentajes de las muestras resulta ser del 10 por ciento o más, los supuestos *A* han de descartarse. Esto se hace a sabiendas de que en el 1 por ciento de las veces una diferencia de esta magnitud ocurrirá aun siendo *A* cierto. En otros términos, el riesgo de incurrir en un error de tipo I (el de descartar una hipótesis correcta) será de una probabilidad sobre ciento.

Volvamos ahora a la lista original de pasos necesarios en la verificación de las hipótesis. Ya se señaló que la inducción estadística se ocupa básicamente de los pasos 3 y 4. El investigador anticipa todos los resultados posibles de las muestras y los divide en dos clases: aquellos respecto de los cuales puede descartar sus hipótesis y aquellos respecto de los cuales no puede descartarlas. En realidad, lo que hace la estadística es proporcionar los criterios a utilizar en la división de los resultados en dos clases. Estos resultados se ponen en una u otra de las dos clases, de conformidad con los riesgos que se está dispuesto a asumir en cuanto a incurrir en los errores de tipos I y II. La mayor ventaja de los procedimientos estadísticos con respecto a los métodos intuitivos está en el conocimiento que proporcionan acerca de esos riesgos de error.

Expuesta en esta forma, la estadística no parece valer mucho la pena de preocuparse por ella. Sin embargo, el paso 3 no resulta nada fácil de efectuar con otro método cualquiera. Imagínese, por ejemplo, un experimento consistente en echar 25 cara o cruz con una moneda cuya buena manufactura se pone en entredicho. Supóngase que tratamos de decidir acerca de los resultados que, en caso de producirse, darían lugar a que llamáramos la atención a la persona que echa los cara o cruz ¿Descartaríamos la hipótesis de que el cara o cruz es correcto si salieran más de 15 caras?, ¿más de 18?, ¿o sólo si la mitad de los cara o cruz dieran caras?; ¿si se dieran 10 caras consecutivas, independientemente de los demás resultados? La teoría de las probabilidades nos permite apreciar el número de ellas que existen de obtener cualquier combinación de resultados en el supuesto de que la moneda fuese correcta. Así, pues, escogeremos aquellos resultados que, en dicho supuesto, serían prácticamente improbables.

No se espera en modo alguno que el estudiante que se enfrenta por primera vez con la inducción estadística comprenda en primera lectura todo lo que se acaba de decir acerca de los razona-

mientos relativos a la verificación de las hipótesis. Reconocemos, en efecto, que el proceso es complicado y uno de los que parecen proporcionar a los estudiantes más dificultades que cualquier otra parte de la estadística. De ahí que el estudiante deba hacer un esfuerzo especial para comprender dichos razonamientos buscando las analogías básicas con los mismos entre todas las pruebas estadísticas. Una vez que la lógica subyacente se haya penetrado a fondo, el aprendizaje de la estadística se simplifica considerablemente.

GLOSARIO

Hipótesis
Parámetro
Población
Estadística
Errores de tipo I y II

BIBLIOGRAFÍA

1. Ackoff, R. L.: *The Design of Social Research*, University of Chicago Press, Chicago, 1953, cap. 5.
2. Northrop, F. S. C.: *The Logic of the Sciences and the Humanities*, The Macmillan Company, Nueva York, 1947, caps. 7 y 8.
3. Weiss, R. S.: *Statistics in Social Research*, John Wiley & Sons, Inc., Nueva York, 1968, cap. 13.

IX. PROBABILIDAD

Todos tenemos sin ninguna clase de duda una noción intuitiva de lo que se entiende por *probabilidad*, aunque no estemos en condiciones de formular del término una definición precisa. En el lenguaje corriente hay cierto número de palabras y frases que se emplean en forma casi intercambiable con el concepto de probabilidad, tales como *posibilidades*, *perspectivas*, *ventaja*, etcétera. Estos conceptos se emplean en ocasiones en cierto número de sentidos diversos. Preguntamos, por ejemplo, "¿cuál es la probabilidad de que hoy llueva?", refiriéndonos a un acontecimiento singular (el llover hoy) que puede ocurrir o no en el futuro. El enunciado "no es probable que Jones asesinara a su suegra" se parece al anterior, pero se refiere a un acontecimiento que ha tenido ya lugar y a cuyo propósito nos falta, con todo, información suficiente para formular una afirmación categórica. Puede uno referirse a lo que sucederá a la larga: "si juegas, es probable que llegues a perder hasta la camisa". Aquí la alusión no se refiere a que uno haya de perder la camisa con un solo golpe de dados, sino a lo que ocurrirá si el experimento se repite un gran número de veces. "Un niño varón, nacido en los Estados Unidos, de padres blancos nativos, vivirá probablemente 65 años." Semejante enunciado parece referirse más al tipo generalizado de niño de los cuadros actuariales que a un Jimmy Brown concreto cualquiera.

Es obvio, sin embargo, que si hemos de hablar de la probabilidad de modo inteligente y, en particular, si hemos de hacer intervenir al matemático, el concepto ha de definirse con la precisión suficiente para que todos podamos emplearlo con el mismo sentido. Por desgracia, sin embargo, no es tan sencillo obtener una definición que satisfaga al propio tiempo al matemático y a nuestra noción intuitiva de aquello que por lo regular entendemos con el término. Según veremos, en efecto, el matemático considera necesario pensar en términos de probabilidades *a priori*, que en realidad no pueden obtenerse empíricamente y que no dependen de cualquier muestreo particular de datos. En las secciones que siguen, el concepto de probabilidad se definirá en lenguaje matemático y se examinarán algunas de sus propiedades matemáticas más importantes. Al propio tiempo trataremos de conseguir que dicha definición y dichas propiedades matemáticas parezcan razonables a la luz del empleo y la experiencia cotidianos.

IX.1. Probabilidad a priori X

En estadística nos ocupamos en establecer generalizaciones a propósito de una población compuesta por lo regular de un gran número de individuos. Semejante población puede ser una población finita realmente existente —como, por ejemplo, la población de los Estados Unidos, o los varones blancos nativos de más de 65 años— y, por lo tanto, claramente delimitable. En tal caso, tomaremos por lo regular algún tipo de muestra de la población, y el interés se dirigirá en primer término a la población por sí misma (o a algún subgrupo de ella), más que a los individuos que acontecen figurar en una cualquiera de las muestras particulares. Pero la población puede también ser una población hipotética que implique, digamos, un número ilimitado de experimentos efectuados "en condiciones similares". Por consiguiente, al estadígrafo no le interesan el acontecimiento o el individuo particulares, a no ser en la medida en que el acontecimiento o individuo en cuestión puedan ayudarle a obtener información a propósito de la población. Como quiera que este texto es un texto de estadística, nos serviremos en él del término *probabilidad* para referirnos no a acontecimientos particulares (llover hoy, Jones asesino), sino a un gran número de acontecimientos o a lo que ocurre a la larga.¹

¿Cómo podemos obtener probabilidad desde el punto de vista de acontecimientos repetidos? En primer lugar, es menester pensar en términos de un experimento ideal que se deje repetir un gran número de veces "en condiciones similares". Sin duda, las condiciones cambian en la realidad, pero ha de ser posible imaginar por lo menos que no lo hacen. En cada uno de dichos experimentos perfectos han de anticiparse todos los resultados. Así, pues, hemos de acostumbrarnos a pensar en términos de una moneda ideal que se lanza al aire un gran número de veces, en circunstancias idénticas, y con sólo dos resultados posibles (cara o cruz) en cada cara o cruz. Prescindimos del hecho de que en el proceso del lanzamiento de la moneda real podría acabar por gastarse de modo irregular, o que en ocasiones se pudiera mantener de canto. Aprendemos a concebir un juego de naipes perfectamente barajado, en el que ninguno de ellos tienda a pegarse a otro, pese a que semejante juego no se encontrará nunca en la vida real.

Llamemos todo resultado o grupo de resultados un "acontecimiento". En este caso el acontecimiento puede ser simple (que no

¹ Resulta posible tratar las probabilidades desde el punto de vista del acontecimiento singular y servirse, con todo, de las propiedades matemáticas que se examinan en la sección siguiente (véase [8]). Sin embargo, semejante tratamiento presenta por lo menos tantas diferencias conceptuales como el que empleamos en este texto.

se deja descomponer) o compuesto (una combinación de acontecimientos simples). Así, pues, el acontecimiento *A* puede ser un 6 en un solo golpe de dados; el acontecimiento *B* (compuesto) puede consistir en los resultados 2 y 4, o 6 en un solo lanzamiento, en tanto que el acontecimiento *C* (también compuesto) puede implicar la obtención de un 7 en dos jugadas. Por conven-

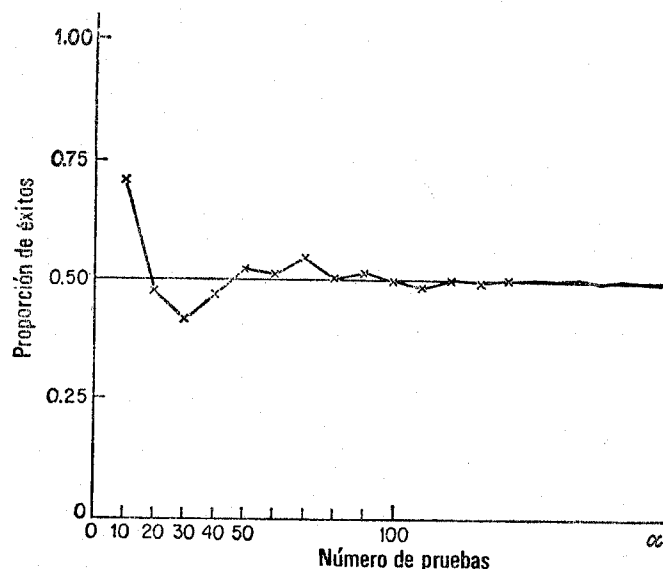


FIG. IX.1. Oscilación de la proporción de éxitos aproximándose al límite de .50

ción se utiliza el término de *éxito* cuando el acontecimiento considerado se produce, y el de *fracaso* cuando no ocurre.² Puede efectuarse el experimento un gran número de veces y obtenerse la proporción de las veces en que ocurre cada acontecimiento particular.

No estamos todavía por completo en condiciones de dar una definición formal de la probabilidad. Primero, en efecto, hay que apelar al conocimiento del lector acerca de qué ocurre empíricamente cuando un experimento como el de lanzar una moneda al aire se repite un gran número de veces. Supongamos que empezamos los lanzamientos y que a cada décimo cara o cruz anotamos la razón de los éxitos (digamos "caras") al número total de los mismos. Ahora bien, los resultados obtenidos tienden a ser semejantes a los que se indican en la figura IX.1.

² Este uso técnico de los términos *éxito* y *fracaso* no necesita ser conforme al uso general. Así, por ejemplo, el *éxito* puede indicar la contracción de la polio o la elección de un demagogo.

En los 10 primeros lanzamientos no esperamos por lo regular obtener exactamente 5 caras, ni aun con una moneda correcta. Es posible, en efecto, que el número de caras sea 7. La próxima serie de lanzamientos puede contener una larga secuencia de cruces, de modo que al final de 20 lanzamientos la proporción de caras sea de .45. La serie siguiente puede dar asimismo más cruces que caras, la próxima, ligeramente más caras que cruces, y así sucesivamente. Después de 100 lanzamientos con una moneda no sesgada esperamos que la proporción de los éxitos se sitúe alrededor de .5, en tanto que después de 1 000 lanzamientos deberíamos encontrarnos aún más cerca de dicha cifra. Así, pues, esperamos que la razón de los éxitos al número *total* de las pruebas se establezca de modo que cese de fluctuar mucho de una serie de 10 lanzamientos a otra. Después de 10 mil pruebas, inclusive si obtuviéramos 20 cruces sucesivas (acontecimiento extremadamente improbable), el efecto de ello sería negligente en la razón en cuestión.³ En cambio, si esto se hubiera producido en la tercera y cuarta secuencias, el efecto habría sido pronunciado. Por lo tanto, cuanto mayor es el número de las pruebas, tanto más se va acercando la razón a un determinado valor que los matemáticos designan como "límite". Si podemos concebir que el experimento se prolongue indefinidamente, podemos con probabilidad concebir también que la razón alcance exactamente el valor límite, o sea .5. Por tanto, nos vamos encontrando con la noción de "infinito" y que los matemáticos han hallado que éste es un concepto manifiestamente ambiguo, será preferible pensar en términos de un número de pruebas extremadamente grande.

* La noción de límite se deja definir con algo más de precisión. Decimos, en efecto, que la razón se aproxima a un límite cuando, habiendo determinado previamente qué tipo de aproximación deseamos, lanzamos la moneda un número finito de veces, hasta estar virtualmente seguros que la razón obtenida se aproxima al límite con el grado de exactitud deseado. En otros términos: escogemos primero un número muy pequeño ϵ , que represente el grado de aproximación deseado. Supóngase que ponemos $\epsilon = .0001$. Si el límite existe, hay un número finito de lanzamientos N tal, que podemos prácticamente estar seguros que la proporción de éxitos obtenida quedará en el interior de $\pm .0001$

³ Obsérvese bien que no se ha pretendido que los números absolutos de caras y cruces sean aproximadamente iguales, ni que, si se da inicialmente un exceso de caras, las cruces acabarán por compensarlo. Puede seguir habiendo un exceso de caras indefinidamente, pero la *razón* se aproximará a .5 incluso en este caso. Así, por ejemplo, si tuviéramos 35 caras y 15 cruces en los 50 primeros lanzamientos, la proporción de caras sería de .7. Un exceso de 20 caras en 100 lanzamientos (o sean 60 caras) da una proporción de .6, en tanto que el mismo exceso en 200 lanzamientos da un valor de .55.

de la verdadera probabilidad.⁴ Además, por muy pequeño que escojamos ϵ , siempre encontraremos un número finito de lanzamientos para el que ello es verdadero. Pero si no existe un límite, entonces no resultará por lo general posible.

No es en modo alguno una necesidad *lógica* el que las razones obtenidas en dicha forma se establezcan en un valor límite. En efecto, se deja por lo menos concebir que las razones en cuestión sigan fluctuando indefinidamente. Si esto fuera efectivamente así, no podríamos hablar de una sola probabilidad de caras en relación con la moneda. Sin embargo, cuando semejante límite existe, podemos definir la *probabilidad como límite de la razón de los éxitos al número total de las pruebas*. O bien, en forma más ruda, la probabilidad es la proporción de los éxitos "a la larga".

En el examen ulterior será conveniente hablar como si pensáramos en términos de probabilidades de acontecimientos singulares. Así podremos, por ejemplo, preguntar, "¿cuál es la probabilidad de obtener un 6 en un solo golpe de dados o un as rojo al sacar una sola carta de la baraja?" En realidad, al servirnos de la frase "un solo golpe de dados", tratamos simplemente de evitar el empleo de una fraseología complicada. Lo que efectivamente queremos decir es: "¿qué proporción de veces esperamos obtener un 6, a la larga, lanzando un solo dado reiteradamente?" A título de conveniencia, pues, hablaremos de un solo golpe de dados cuando en realidad queramos significar un número indefinidamente grande de lanzamientos particulares con el mismo dado.

Antes de proceder al examen de las propiedades matemáticas de las probabilidades hemos de detenernos en algunos puntos. Los experimentos de la vida real, si se repiten, parecen efectivamente seguir el tipo general examinado anteriormente y representado en forma de diagrama en la figura IX.1, o sea que nos acercamos efectivamente con ellos a un límite que se deja calcular. Esto nos conduce a hablar de la "ley de los promedios" y a esperar que la mayoría de las monedas darán caras aproximadamente la mitad de las veces o que las buenas manos alternarán en el *bridge* con las malas. Sin embargo, hay que precaverse contra esa ley de los promedios. En efecto, algunas personas han interpretado dicha ley en el sentido de que si una moneda da 10 veces caras consecutivas, entonces lo más probable es que la vez siguiente dé cruz, "a causa de la ley de los promedios". Semejante interpretación implica una predicción a propósito de un acontecimiento singular (esto es, el resultado del undécimo lanzamiento). Según veremos más adelante, sole-

⁴ El examen de los intervalos de confianza (cap. XII) ayudará a indicar que no podemos estar nunca absolutamente seguros de que la verdadera probabilidad se halle en el interior del intervalo obtenido.

mos suponer que lo que ha sucedido en los lanzamientos precedentes no tiene absolutamente importancia alguna en relación con lo que sigue.⁵ En efecto, la moneda no posee ni memoria ni conciencia. Desde el punto de vista de una estrategia inteligente, si un jugador presencia 10 caras sucesivas en diez lanzamientos, haría bien en predecir que en el undécimo volverá a salir cara, en el supuesto de que la moneda debe estar sesgada.

Debería resultar perfectamente claro que las probabilidades *a priori* tal como se definen en esta sección no pueden obtenerse *exactamente* por medios empíricos, aunque sí pueden *apreciarse*. Y esto se debe no sólo al hecho de que hemos debido imaginar experimentos llevados a cabo en condiciones ideales, sino también a la circunstancia de que ningún experimento puede repetirse indefinidamente. Sin embargo, con un número suficiente de pruebas, una probabilidad puede apreciarse con cualquier grado deseado de exactitud. Las reglas matemáticas que se dan en la sección siguiente y todos los razonamientos matemáticos que se hallan en la base de la inducción estadística se ocupan más bien de las probabilidades *a priori* que de las clases de probabilidades que pueden efectivamente obtenerse por el investigador.⁶

Así, pues, al aplicar el razonamiento estadístico a cualquier ciencia que se ocupe del mundo real, nos encontraremos en la posición lógica descrita en el capítulo VIII. Hemos de suponer alguna probabilidad *a priori* para poder aplicar el razonamiento matemático. Podemos decir que si ésta es la probabilidad *a priori* correcta, entonces determinados resultados empíricos son probables (o improbables). En esta forma, *A* es la teoría matemática, y *B* los resultados empíricos anticipados, y no existe medio alguno de verificar la teoría directamente. Si *B* resulta ser falso, podemos descartar *A*, pero, si *B* es cierto, alguna otra teoría *C* que comporte probabilidades *a priori* distintas pueda acaso explicar también los resultados. Si queremos, pues, evitar la falacia de afirmar el consecuente, habremos de adoptar probabilidades de las que en realidad sospechamos que son falsas, procediendo por eliminación. En el próximo capítulo veremos ejemplos particulares en los que así se procede.

IX.2. Propiedades matemáticas de las probabilidades*

Aunque el lector tal vez no haya de volver nunca más a calcular probabilidades, importa de todos modos que se dé cuenta de que

⁵ Esto no puede suponerse en el caso del ser humano, hecho que hay que tener presente siempre que se tomen mediciones repetidas con personas u otros animales. Véase sec. IX.4.

⁶ En rigor, el investigador sólo puede obtener *proporciones* debido al hecho de que el número de pruebas o casos será siempre finito.

en la base de todos los cuadros de los que habrá de servirse para verificar hipótesis se encuentra cierto número de propiedades azas simples de las probabilidades. En un texto como el presente no es posible profundizar mucho en la teoría de éstas. El objeto del examen que sigue es, pues, simplemente el de dar una idea de cómo operan los matemáticos con las probabilidades al poner los fundamentos de la inducción estadística. Podemos empezar identificando tres propiedades matemáticas de las probabilidades *a priori*.

La primera de ellas apenas requiere algún comentario. Como quiera que en N pruebas no podemos obtener menos de cero éxitos ni más de N , síguese que para cualquier acontecimiento A la probabilidad de que A ocurra [lo que se escribe $P(A)$] ha de ser mayor o igual a cero y menor o igual a 1. Así, pues:

$$0 \leq P(A) \leq 1$$

en donde el símbolo \leq ha de leerse como "menor que o igual a". Si $P(A) = 1$, el acontecimiento A ocurrirá con toda seguridad; si $P(A) = 0$, en cambio, entonces no es posible que A tenga lugar.

La regla de la adición. La segunda propiedad de las probabilidades es más interesante. Habida cuenta de su sencillez, tomaremos un caso especial de la regla de adición que puede enunciarse como sigue: *si los acontecimientos A y B se excluyen mutuamente, la probabilidad de obtener A o B [escrito $P(A \text{ o } B)$] es igual a la probabilidad de A más la probabilidad de B , o sea:*

$$P(A \text{ o } B) = P(A) + P(B) \text{ (si } A \text{ y } B \text{ se excluyen mutuamente) (IX.1)}$$

Por exclusión mutua entendemos que A y B no pueden tener lugar simultáneamente en el mismo experimento. Así, por ejemplo, es imposible obtener a la vez un as y un rey si se toma una sola carta de una baraja corriente. Por consiguiente, aplicando la regla de la adición a una baraja hipotéticamente perfecta tenemos:

$$P(A \text{ o } K) = P(A) + P(K) = 1/13 + 1/13 = 2/13$$

Por supuesto, pudimos haber obtenido el mismo resultado teniendo en cuenta que hay cuatro ases y cuatro reyes en la baraja y, con iguales probabilidades de selección, la probabilidad de obtener el uno o el otro de dichos naipes sería de $8/52$, o $2/13$. Y en forma análoga, la probabilidad de sacar ya sea un 5 o un 6 en un simple golpe de dados sería de $1/6 + 1/6 = 1/3$.

La regla de la adición puede hacerse extensiva a más de dos casos. Así, por ejemplo, si A , B , C ..., son todos ellos mutuamente exclusivos, entonces tenemos:

$$P(A \text{ o } B \text{ o } C \dots \text{ o } K) = P(A) + P(B) + P(C) \dots + P(K) \text{ (IX.2)}$$

Si tenemos una población compuesta de 100 personas de la clase superior, 200 de la clase superior a la media, 400 de la inferior a la media y 300 de la inferior, por ejemplo, la probabilidad de sacar una persona de la clase superior, o una de la clase superior a la media, o una de la clase inferior a la media en una sola vez será:

$$\frac{100}{1000} + \frac{200}{1000} + \frac{400}{1000} = \frac{700}{1000} = .7$$

siempre que cada persona tenga las mismas probabilidades de ser seleccionada.

Como quiera que las probabilidades son esencialmente proporciones, síguese que si tenemos todos los acontecimientos posibles, cada uno de ellos excluyendo a los demás, la suma de dichos eventos será la unidad. Así, por ejemplo, si sumamos las probabilidades de sacar un trébol, una espada, un corazón o un diamante, hemos de obtener una suma de 1. La probabilidad de que el evento A no ocurra es igual a la suma de las probabilidades de todos los eventos (mutuamente exclusivos) restantes. Por consiguiente, si sustraemos $P(A)$ de la unidad, tenemos la probabilidad de no obtener A , ya que

$$\text{si} \quad 1 = P(A) + P(B) + P(C) + \dots + P(K),$$

$$\text{entonces:} \quad 1 - P(A) = P(B) + P(C) + \dots + P(K).$$

La probabilidad de no sacar una reina, por ejemplo, es de

$$1 - \frac{1}{13}, \text{ o } \frac{12}{13}.$$

Hasta aquí sólo nos hemos ocupado de eventos que se excluyen mutuamente. Una forma más general de la regla de la adición puede enunciarse como sigue: *si A y B son dos acontecimientos cualesquiera (no necesariamente mutuamente exclusivos), entonces:*

$$P(A \text{ o } B) = P(A) + P(B) - P(A \& B) \quad (\text{IX.3})$$

en donde $P(A \& B)$ representa la posibilidad de obtener *a la vez* A y B .⁷ En el caso general, pues, la probabilidad de obtener A o

⁷ La partícula *o* tal como la emplean los matemáticos incluye la posibilidad de que A y B se verifiquen a la vez. Por consiguiente la expresión " A o B " significa " A y B , y A o B ". En orden a la anotación por teoría de grupos " A o B " significa lo mismo que $A \cup B$ en tanto que A y B significa lo mismo que $A \cap B$.

B se obtiene adicionando primero la probabilidad de A a la probabilidad de B y sustrayendo luego la probabilidad de obtener simultáneamente A y B . La razón de sustraer $P(A \& B)$ está en que la probabilidad de esta ocurrencia conjunta se ha contado dos veces: una en $P(A)$ y otra en $P(B)$. La figura IX.2 puede ayudar a comprender por qué es así.

En efecto, en la figura IX.2, las probabilidades de A y B se han representado por áreas proporcionales a sus respectivos va-

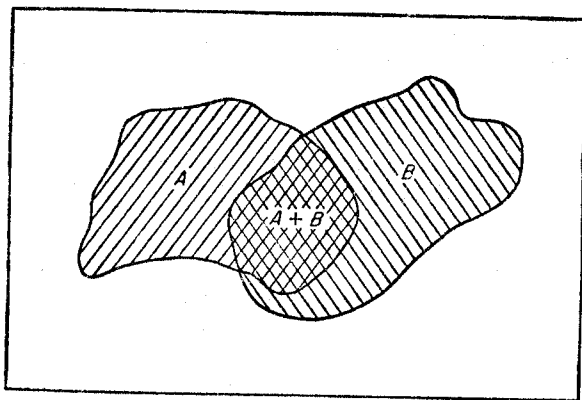


FIG. IX.2. Representación geométrica de probabilidades, con áreas proporcionales a $P(A)$, $P(B)$ y $P(A \& B)$

lores numéricos, tomándose la superficie del rectángulo como unidad. En el caso general habrá por lo regular cierto cruzamiento, es decir: A y B no serán mutuamente exclusivos. La probabilidad de obtener ya sea A o B (o ambos) está representada por el total del área achurada en cruzado. Y como quiera que el área achurada más pequeña se habrá contado dos veces, una en A y otra en B , de ahí la necesidad de sustraer $P(A \& B)$ para obtener el área total achurada en cruz.⁸

Tomemos un ejemplo numérico. Supongamos que A es el evento en que se obtenga una reina en una sola extracción, y supongamos que B es el evento de que la carta sea una espada. Entonces A y B no son mutuamente exclusivos ya que es posible obtener ambas cosas simultáneamente (o sea la reina de espadas). Por consiguiente:

⁸ El lector ha de convencerse él mismo de que, para obtener la probabilidad de A o B , pero no ambas, habremos de sustraer de $2P(A \& B)$ de $P(A) + P(B)$. Debería también tratar de extender la forma general de la regla de adición trazando una gráfica semejante para los eventos A , B y C . (Véase ejercicio 4b).

$$\begin{aligned} P(A \circ B) &= P(A) + P(B) - P(A \& B) \\ &= 4/52 + 13/52 - 1/52 = 16/52 = 4/13. \end{aligned}$$

Este resultado puede verificarse intuitivamente teniendo en cuenta que A o B podrían obtenerse extrayendo cualquier espada o una de las tres reinas restantes, o sea una de las 16 cartas consideradas. Si hubiéramos adicionado simplemente $P(A)$ y $P(B)$, la reina de espadas se habría contado dos veces. En la sección siguiente veremos una regla general para el cálculo de $P(A \& B)$, ya que no siempre resulta tan sencillo obtener dicha cantidad. Obsérvese que si los dos eventos son mutuamente exclusivos no habrá cruzamiento, y $P(A \& B)$ será igual a cero. Por lo tanto, la regla general reconduce, en esta ocasión, al caso especial de la regla de la adición examinada anteriormente.

La regla de la multiplicación. La tercera propiedad de las probabilidades nos permite obtener la probabilidad de que dos (o más) eventos ocurran conjuntamente. Podríamos enunciar esta propiedad como sigue: si A y B son dos eventos cualesquiera, la probabilidad de que se produzcan ambos es el producto de la probabilidad de que se produzca uno de ellos por la probabilidad condicional de que se produzca el otro, dado que el primer evento haya ocurrido. O en símbolos:

$$P(A \& B) = P(A)P(B|A) = P(B)P(A|B) \quad (\text{IX.4})$$

Los símbolos $P(A|B)$ y $P(B|A)$ representan lo que se designa como probabilidades condicionales. $P(A|B)$ debería leerse como "la probabilidad de A , dado que B haya ocurrido". La expresión de *probabilidad condicional* significa que admitimos que la probabilidad de A puede depender de que B ocurra o no. En otros términos: la probabilidad de A dado B puede diferir de la probabilidad de A dado que B no haya ocurrido. Así, por ejemplo, si B es el evento de que un individuo conduce el automóvil temerariamente y A el evento de que se encuentre en un accidente de tránsito, suponemos que $P(A|B)$ es mayor que $P(A)$, ya que el conducir temerariamente es causa de accidentes.

Antes de proceder a ilustrar la regla de la multiplicación, permitámonos introducir un nuevo concepto importante. Se dice de dos eventos A y B que son *estadísticamente independientes* si, y sólo si, $P(A|B) = P(A)$; y $P(B|A) = P(B)$. Así, pues, si la probabilidad de que A ocurra es la misma, independientemente de si B ha ocurrido o no, y si lo mismo es cierto respecto de B , entonces los dos eventos son independientes uno de otro. Esto significa, en lenguaje llano, que el conocimiento de que uno de los acontecimientos se ha producido no ayuda a predecir el otro. Por ejemplo: la probabilidad de sacar un as, dado que la carta

sea roja, es de $2/26$, ya que hay dos ases rojos y un total de 26 cartas rojas. Esta es numéricamente la misma que la probabilidad no condicional de sacar un as ($4/52$). Por lo tanto, el color y el valor de la carta son estadísticamente independientes. Y en forma análoga, el hecho de saber que una carta es un as no ayuda a adivinar su color. Obsérvese, de paso, que los eventos mutuamente exclusivos no son independientes. En efecto, si A y B son mutuamente exclusivos, hemos de tener siempre $P(A|B) = P(B|A) = 0$. ¿Por qué?

En el caso en que A y B sean estadísticamente independientes, tenemos $P(B|A) = P(B)$, y la regla de la multiplicación asume la forma simple de:

$$P(A \& B) = P(A)P(B) \quad (\text{si } A \text{ y } B \text{ son independientes})$$

Por lo regular encontraremos que este caso especial de multiplicación es de empleo mucho más fácil que la regla general.

Vamos a ilustrar primero la regla de multiplicación en los casos especiales en que A y B son estadísticamente independientes. Por lo regular pensamos que las repeticiones de un experimento son independientes una de otra. Así, por ejemplo, si lanzamos la moneda al aire una vez, suponemos que el resultado no afecta lo que pueda suceder en el próximo lanzamiento; la probabilidad de "cara" permanece constante de un lanzamiento al otro. En efecto, el saber que salió cara no nos ayuda a predecir el resultado del segundo lanzamiento.⁹ Por consiguiente, sirviéndonos de la regla de multiplicación, podemos calcular la probabilidad de sacar caras en dos lanzamientos sucesivos multiplicando entre sí las probabilidades de obtener cara en cualquier prueba dada. En el caso de una moneda no sesgada, la probabilidad de dos caras sucesivas será de $(1/2)(1/2) = 1/4$. Y en forma análoga, si A es el evento de sacar una carta roja, y B el evento consistente en obtener un as, entonces la probabilidad de sacar un as rojo $P(A \& B)$ será:

$$P(A \& B) = P(A)P(B) = 1/2 \times 1/13 = 1/26.$$

⁹ Suponemos que la verdadera probabilidad es conocida y que nuestra tarea consiste en predecir el resultado de cualquier prueba particular. Es cierto, por supuesto, que sin dicho conocimiento la probabilidad podría acaso *estimarse* utilizando los resultados de pruebas anteriores y sirviéndose luego de dicha estimación para predecir el futuro. Esto *no* es lo que entendemos cuando decimos que en el caso de independencia el conocimiento de un evento no nos ayuda a anticipar el otro. Así, por ejemplo, sabiendo que han salido 20 caras consecutivas, nos veríamos llevados a predecir una moneda sesgada, esto es, que la probabilidad verdadera de sacar cara es algún valor superior a .5. Y esto nos llevaría a su vez a predecir cara en ocasión del vigésimo primer lanzamiento. Sin embargo, el supuesto es de que, si existe, el sesgo es ya conocido. Por lo tanto, si se sabe que p es .8, el conocimiento de 20 caras sucesivas no nos ayudará a predecir el resultado del lanzamiento siguiente.

Tomemos dos ejemplos en los que no se da independencia. El primero de ellos comporta una situación en la que dos variables están relacionadas de tal modo que el conocimiento dé una ayuda a predecir la otra. Supóngase que tenemos los siguientes datos totalmente hipotéticos:

Carácter	Morenas	Rubias	Pelirrojas	Total
Emprendedor	300	600	300	1 200
Tímido	600	100	100	800
Total	900	700	400	2 000

Si de dicha población, arreglada por una persona ajena, se saca al azar¹⁰ una muchacha, ¿cuál es la probabilidad de que sea una pelirroja emprendedora? Como quiera que en el total de 2 mil muchachas hay 300 pelirrojas emprendedoras, la probabilidad de sacar una de dicho grupo particular es, obviamente, $300/1\,200$, o sea .15. Esta misma probabilidad se obtendrá ahora sirviéndose de la regla de multiplicación.

Supongamos que A es el evento consistente en obtener una pelirroja, y B el evento de que el carácter es emprendedor. Como quiera que hay 400 pelirrojas en conjunto, $P(A) = 400/2\,000$, o sea .2. Sin embargo, entre estas 1 200 muchachas emprendedoras hay 300 pelirrojas. Por lo tanto, si tenemos conocimiento de que el carácter es emprendedor, la probabilidad de que la muchacha sea pelirroja es de $300/1\,200$, o sea .25. En forma análoga, la probabilidad de obtener una muchacha emprendedora es de $1\,200/2\,000$, o sea .6, pero, si se sabe que el dato es el de pelirroja, la probabilidad de que la muchacha sea emprendedora es de $300/400$, o sea .75. Tenemos, pues:

$$P(A) = .2 \quad P(A|B) = .25$$

$$P(B) = .6 \quad P(B|A) = .75$$

Sirviéndonos de la regla de multiplicación llegamos a la siguiente probabilidad de obtener una pelirroja emprendedora:

$$P(A \& B) = P(A)P(B|A) = (.2)(.75) = .15$$

$$= P(B)P(A|B) = (.6)(.25) = .15$$

Para el segundo ejemplo, supongamos que hemos de calcular la probabilidad de sacar de una baraja corriente dos ases en dos

¹⁰ La muestra al azar se definirá más adelante en el presente capítulo. En una muestra al azar, todos los individuos y todas las combinaciones de individuos tienen las mismas probabilidades de ser seleccionados.

extracciones. Pongamos que A es el hecho de obtener un as en la primera extracción y B el de que saquemos un as en la segunda extracción. ¿Son A y B independientes? Esto depende de si volvemos o no a poner el as en la baraja después de la primera extracción y barajamos de nuevo antes de la segunda. Si procedemos *con sustitución*, las dos extracciones serán independientes, ya que la probabilidad de obtener un as es constante de una extracción a la próxima y que el resultado de la primera no puede afectar en modo alguno el de la segunda. En este caso,

$$P(A \& B) = P(A)P(B) = (1/13)(1/13) = 1/169.$$

Supongamos ahora que procedemos *sin reposición*, esto es, que no volvemos a colocar la primera carta en la baraja. Si aconteciera que sacáramos un as en la primera extracción, entonces la probabilidad de obtener otro sería de $3/51$, ya que sólo habría tres ases en las 51 cartas restantes. Por otra parte, si no sacáramos un as en la primera selección, la probabilidad de obtenerlo en la segunda sería de $4/51$. Por consiguiente, en este caso no tenemos independencia y habríamos de servirnos de las probabilidades condicionales para calcular $P(A \& B)$. Así:

$$P(A \& B) = P(A)P(B|A) = 4/52 \times 3/51 = 1/221.$$

Conviene advertir que la regla de multiplicación que hemos examinado podría extenderse igualmente a más de dos eventos. Así, si A , B y C son todos ellos independientes uno de otro:

$$P(A \& B \& C) = P(A)P(B)P(C)$$

Por lo que se refiere a las probabilidades condicionales, sus principios pueden aplicarse fácilmente a ciertos casos sencillos. Así, por ejemplo, si hubiéramos de sacar cuatro ases *con reposición*, podríamos calcular la probabilidad de obtenerlos como sigue:

$$P(4 \text{ ases}) = \frac{4}{52} \frac{3}{51} \frac{2}{50} \frac{1}{49} = \frac{1}{270\,725}$$

Si hay tres acontecimientos A , B y C que no son mutuamente independientes, podrá observarse la probabilidad de su ocurrencia conjunta con la siguiente fórmula:

$$P(A \& B \& C) = P(A)P(B|A)P(C|A \& B)$$

en la que $P(C|A \& B)$ se refiere a la probabilidad de C , ya que tanto A como B han ocurrido. Podemos utilizar por supuesto otras fórmulas similares, colocando en otro orden las posiciones de A , B y C . Supongamos que tenemos la población siguiente:

Actitud	Blancos		No blancos		Total
	Repúbli- canos	Demó- cratas	Repúbli- canos	Demó- cratas	
A favor del aumento de la asistencia social	50	100	25	225	400
En contra del aumento	350	200	25	25	600
Total:	400	300	50	250	1\,000

Si A es el caso en que sacamos un blanco, B aquel en que obtenemos un republicano, y C la ocasión en que la persona está a favor del aumento en la asistencia pública, y dado que sólo 50 republicanos blancos están a favor de la asistencia, tendremos $P(A \& B \& C) = 50/1\,000 = .05$.

En el propio cuadro vemos asimismo que $P(A) = 700/1\,000$; $P(B|A) = 400/700$; y que $P(C|A \& B) = 50/400$. La última de estas cifras resulta del hecho de que de entre las 400 personas que son a la vez A y B (republicanos y blancos) sólo 50 apoyan la asistencia.

Aplicando la regla de multiplicar obtendremos el resultado:

$$P(A \& B \& C) = P(A)P(B|A)P(C|A \& B) \\ = \frac{700}{1\,000} \frac{400}{700} \frac{50}{400} = \frac{50}{1\,000} = .05$$

Para verificarlo podríamos aplicar la siguiente fórmula:

$$P(A \& B \& C) = P(C)P(B|C)P(A|B \& C) \\ = \frac{400}{1\,000} \frac{75}{400} \frac{50}{75} = \frac{50}{1\,000} = .05$$

El concepto de acontecimientos estadísticamente independientes está en estrecha relación con el de la independencia entre dos (o más) *variables*, y será examinado con mayor detención en posteriores capítulos.

Ya hemos utilizado el ejemplo de la baraja, pues ésta tiene la propiedad de que los valores faciales y la secuencia son independientes, lo que supone que el conocer uno de los dos no ayuda para predecir el otro. Tanto en el ejemplo relativo al color del cabello de la muchacha con la que se va a salir, y a su conducta, y aquel en que se relacionan entre sí la raza, la preferencia política y la actitud ante la asistencia pública, consideramos necesa-

rio hacer uso de las probabilidades condicionadas para lograr resultados correctos. En estos casos afirmamos que las variables afectadas *no* son independientes, o que están correlacionadas. Para hacerlo más sencillo consideremos el ejemplo de las muchachas. Supongamos que exactamente el mismo porcentaje (60 por 100) de rubias, morenas y pelirrojas fuesen emprendedoras, en cuyo caso el conocimiento del color del cabello carecería de valor, en la predicción de la conducta. Si conservamos los mismos totales marginales, los resultados pasarían a ser:

Rasgo	Morenas	Rubias	Pelirrojas	Total
Emprendedoras	540	420	240	1 200
Tímidas	360	280	160	800
Total	900	700	400	2 000

Debe comprobarse en primer lugar que en el caso de estos datos hipotéticos no hay necesidad de emplear probabilidades condicionadas. Obsérvese además que la probabilidad (o proporción) correspondiente a cada casilla del cuadro es igual al *producto* de las dos probabilidades en los márgenes correspondientes. Si por ejemplo examinamos el cuadro superior izquierdo veremos que la probabilidad $540/2\,000 = .27$, es justamente el producto de las probabilidades que corresponden a la primera columna marginal (es decir: $900/2\,000 = .45$) y la primera fila marginal (o sea: $1\,200/2\,000 = .6$). Lo mismo es cierto para cada uno de los restantes cuadros. No importa cuántas ocasiones puedan disponerse las categorías de dos variables en una clasificación cruzada que cuente con esta propiedad, diremos que las *variables* son estadísticamente independientes entre sí. En posteriores capítulos llevaremos a cabo pruebas estadísticas relativas tanto a la independencia como a las medidas de dependencia basadas en esta sencillísima idea.

* *Nota acerca del teorema de Bayes.* Dado que $P(A \& B) = P(A)P(B|A)$ podemos resolver la probabilidad condicional, obteniendo

$$P(B|A) = \frac{P(A \& B)}{P(A)} = \frac{P(B)P(A|B)}{P(A)}.$$

Pero $P(A)$ en el denominador puede ser descompuesto en los dos términos $P(B)P(A|B) + P(\bar{B})P(A|\bar{B})$, ya que B y \bar{B} (no B) son posibilidades mutuamente exclusivas y exhaustivas. Esto nos lleva a la ecuación.

$$P(B|A) = \frac{P(B)P(A|B)}{P(B)P(A|B) + P(\bar{B})P(A|\bar{B})}$$

ecuación conocida como teorema de Bayes. Este teorema puede ser generalizado para diversas alternativas B_1, B_2, \dots, B_k , en tanto estas alternativas sean mutuamente exclusivas y exhaustivas, de modo que $\sum_{i=1}^k P(B_i) = 1$. La probabilidad de que un B_i , dado, supuesto que A ha ocurrido, puede escribirse así:

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{i=1}^k P(B_i)P(A|B_i)}$$

Es posible desde luego aplicar el teorema de Bayes siempre que se nos den todas las probabilidades condicionales e incondicionales, pero estas aplicaciones no son especialmente útiles. Puede sin embargo ser aplicado también en casos en que las "probabilidades psicológicas" hayan reemplazado los conceptos de frecuencia relativa. Hays [5] previene contra este empleo. Las aplicaciones directas del concepto bayesiano en relación con la estadística están aún relativamente poco probadas. Parece sin embargo aconsejable sugerir métodos para su empleo. Consideremos en primer lugar un problema muy sencillo. Supongamos que un individuo escoge a capricho una de dos urnas, y a continuación selecciona a capricho una bola de la urna que había elegido. La primera de las urnas contiene una mitad de bolas blancas y otra mitad negras, en tanto que la segunda contiene dos tercios de bolas blancas y un tercio de bolas negras. Sabemos que el individuo selecciona una bola blanca, y desea asignar una probabilidad al hecho de que ha seleccionado, digamos, la primera urna. Obsérvese que en este caso se trata de una especie de "probabilidad inversa", particularmente apropiada al concepto de probabilidades en el que se refleja el estado de nuestros conocimientos. Puede decirse que el individuo seleccionó o no seleccionó la primera urna, siendo las respectivas probabilidades 1 a 0. Pero si hubiéramos de hacer una apuesta, con base en el conocimiento que tenemos de que sacó una bola blanca, ¿qué ventaja estaríamos dispuestos a dar a favor de que escogiera la primera urna? Ésta es ciertamente una forma razonable de plantear el problema.

Si denominamos A al acontecimiento de la selección de una bola blanca, B al de que fue seleccionada la primera urna, y \bar{B} al acontecimiento de que fue seleccionada la segunda urna, obtendremos, al aplicar el teorema de Bayes

$$\begin{aligned}
 P(B|A) &= \frac{P(B)P(A|B)}{P(B)P(A|B) + P(\bar{B})P(A|\bar{B})} \\
 &= \frac{(1/2)(1/2)}{(1/2)(1/2) + (1/2)(2/3)} = \frac{1/4}{1/4 + 1/3} = \frac{3}{7}
 \end{aligned}$$

resultado que no habría sido posible predecir con sólo usar argumentos de sentido común. Obsérvese que puesto que las dos urnas fueron seleccionadas con las mismas probabilidades, tendremos $P(B) = P(\bar{B}) = .5$, lo que habría permitido simplificar la fórmula de Bayes.

Consideremos a continuación una clase de problema, *per se* muy alejado de la estadística, pero que es razonablemente realista desde el punto de vista de las probabilidades psicológicas implícitas en la falta de conocimientos, por parte de un observador, en relación con las frecuencias relativas u otras consideraciones que pudieran ser usadas para obtener probabilidades *a priori*. Supongamos que sabemos que un grupo de acción cuenta con cuatro medios alternativos, con costos y probabilidades de éxito diferentes. Admitamos que un observador, basándose en su apreciación de los costos relativos de los procedimientos alternativos, definidos como B_1, B_2, B_3, B_4 , les asigna las probabilidades subjetivas .4, .3, .2 y .1, respectivamente. Supongamos que calcula las posibilidades de éxito para los medios alternativos como .3, .5, .6 y .9, respectivamente. Averigua que el grupo ha tenido éxito en su acción, pero no puede determinar cuál de los medios fue el utilizado. ¿Cómo podrá valorar de nuevo su estimación original de las probabilidades de cada uno de los procedimientos, sabiendo que el éxito (A) se ha producido? Aplicando la forma más generalizada del teorema de Bayes para las primeras medias (B_1), obtenemos:

$$\begin{aligned}
 P(B_1|A) &= \frac{P(B_1)P(A|B_1)}{\sum_{i=1}^4 P(B_i)P(A|B_i)} \\
 &= \frac{(.4)(.3)}{(.4)(.3) + (.3)(.5) + (.2)(.6) + (.1)(.9)} = \frac{.12}{.48} = .25
 \end{aligned}$$

De esta forma, apoyándose en este conocimiento adicional, podrá el observador asignar al primer método la probabilidad subjetiva de .25. Utilizando cálculos similares asignaría a los restantes métodos las siguientes probabilidades subjetivas: .3125, .25 y .1875, respectivamente.

IX.3. Permutas

Es menester introducir una complicación más. Hasta aquí hemos escogido problemas muy sencillos, que casi habrían podido resolverse intuitivamente. No hace falta decir que la mayoría de los problemas de probabilidades son mucho más complejos que los que se acaban de examinar. Con objeto de operar con problemas un poco más complicados, es necesario tomar en cuenta el orden en que los acontecimientos pueden producirse. Supóngase, por ejemplo, que queremos hallar la probabilidad de obtener un as, un rey y una reina en tres extracciones con reposición. Podemos hallar la probabilidad de sacar un as en la primera extracción, un rey en la segunda y una reina en la tercera. Esta probabilidad sería de $(1/13)^3$. Pero esto representa la probabilidad de obtener un as *seguido* de un rey *seguido* de una reina. Pero hay otras posibilidades de obtener un as, un rey y una reina en tres extracciones *si no nos importa el orden de sucesión*. En realidad, estos naipes podrían obtenerse de las seis siguientes maneras: $ARR', AR'R, RAR', RR'A, R'AR, R'RA$. Puede verse que cada una de dichas posibilidades presenta las mismas probabilidades. Por lo tanto, si nos interesa la probabilidad de sacar dichas cartas *en un orden determinado cualquiera*, podemos adicionar sus probabilidades separadas (ya que son mutuamente exclusivas), con lo que obtenemos $6(1/13)^3$.

En esta forma, sirviéndonos de la regla de multiplicación, hemos referido el acontecimiento A al primer resultado, B al segundo, y así sucesivamente. En otros términos: hemos tomado en cuenta el orden, en tanto que por lo regular estamos más interesados en obtener una determinada serie de resultados. Podemos querer saber la probabilidad de cuatro ases en una mano de *bridge* o de obtener un determinado porcentaje de negros en una muestra, independientemente del orden de la extracción. Al calcular probabilidades de esta clase, *será por lo regular más sencillo determinar primero la probabilidad de cualquier orden dado de resultados*, y luego, si todos los demás órdenes son igualmente probables, podemos multiplicar simplemente el número de los órdenes posibles por la probabilidad de que ocurra uno cualquiera determinado de ellos. Obsérvese que al proceder así nos servimos tanto de la regla de multiplicación como de la de adición. Existen fórmulas concretas que permiten calcular exactamente cuántos sean los órdenes posibles en un problema determinado.

En las ocasiones en que tenemos N diferentes acontecimientos que ocurren en un orden determinado, nos referimos a ello como una *permuta* de dichos acontecimientos. En las ocasiones en que el orden carece de interés, denominaremos *combinación* a la agrupación de acontecimientos. Por ejemplo: en el caso de la combinación simple (A, R, R'), habrá seis ordenamientos dife-

rentes. Observemos cómo pueden obtenerse fórmulas para determinar el número de permutaciones en casos sencillos.

Comencemos con una situación en que todos los acontecimientos N son distintos. ¿De cuántas maneras pueden ser ordenados? Está claro que si consideramos N posiciones ordinales (por ejemplo: N sillas dispuestas en fila), la primera de aquéllas podrá ser ocupada por cualquiera de los objetos o acontecimientos. Habiendo llenado esta posición, podremos hacer lo mismo con la segunda, utilizando cualquiera de los $N-1$ acontecimientos restantes, la tercera con uno de los $N-2$, etcétera. Cuando lleguemos a la última posición sólo nos resta una posibilidad. Habrá pues:

$$N(N-1)(N-2) \dots (3)(2)(1) = N!$$

órdenes posibles; $N!$ es la expresión del largo producto de la parte izquierda de la igualdad, y se le denomina "factorial N ". Supongamos por ejemplo que tenemos 13 cartas, una de cada valor. Las volvemos de cara una por una. ¿Cuántas son las diferentes permutaciones posibles? La primera carta puede tener uno cualquiera de los trece valores. Como quiera que esta carta ya ha sido descubierta, la segunda podrá tener uno cualquiera de los doce valores restantes, siendo por tanto 13×12 las soluciones posibles para las dos primeras cartas. Continuando adelante con el montón de cartas determinaremos que habrá:

$$(13)(12)(11)(10) \dots (3)(2)(1) = 13! = 6\,227\,020\,800$$

procedimientos diferentes para ordenar las trece cartas.

Supongamos a continuación que los acontecimientos no son todos diferentes. Contamos de nuevo con trece cartas, pero dos de ellas pueden ser ases, y no distinguiremos entre los diferentes órdenes, resultando así indiferente el orden en que resulten seleccionados los dos ases. Supongamos que han sido escogidos en las posiciones quinta y undécima. Si hubieran sido distintos entre sí, y en tal caso denominamos as_1 y as_2 , para cada distinta permutación en que el as_1 apareciera antes que el as_2 , habría otra permutación idéntica en la que el as_2 precedería al as_1 . Vemos así que cuando no podemos distinguir entre estos dos ases, hay sólo la mitad de permutas en relación con el caso de que todos los acontecimientos sean distintos. Por ello el número total de permutas en este caso será $N!/2! = N!/2$.

Supongamos que los ases hubieran sido tres en vez de dos. Si se les denominase as_1 , as_2 y as_3 , observaríamos que habría habido $3! = 6$ permutas entre dichos ases, imposibles de distinguir. El número total de permutas de las trece cartas será $13!/3!$. En general, si hay N objetos, tres de los cuales no pueden ser distin-

guidos de los demás, habrá $N!/3!$. Puede generalizarse fácilmente este razonamiento, ampliándolo a más de un grupo de objetos no distintos. Supongamos que nuestras trece cartas contienen tres ases y cuatro reyes, siendo distintas las seis cartas restantes. Como quiera que los ases, caso de ser diferenciados, pueden ser ordenados de $3!$ formas, y los cuatro reyes en $4!$ formas, dividiremos $13!$ entre $3! \cdot 4!$ para llegar al número de permutas verdaderamente inconfundibles.

La regla general resulta ya obvia. Si tenemos N acontecimientos subdivididos de tal manera que el primer grupo contenga r_1 elementos no distinguibles, el segundo contenga r_2 de dichos elementos y, en general, el grupo i -ésimo contiene r_i de los mismos, tendremos un total k de tales grupos, todos distinguibles entre ellos; el número total de permutas será $N!/r_1!r_2! \dots r_k!$. Proponiendo otro ejemplo: si hay 25 niños, 6 de los cuales tienen 3 años, 8 de ellos 4 años, otros 9, 5 años, contando con uno de 6 y otro de 7 años, habrá $25!/6!8!9!1!1!$ permutas entre dichos niños, si solamente se les diferencia por sus edades.

La regla general para determinar el número de permutas de acontecimientos, no todos los cuales son distintos, presenta un caso especial muy importante, en aquellas ocasiones en que sólo hay dos clases de acontecimientos (por ejemplo: éxitos y fracasos). Si hay N acontecimientos, r de los cuales son éxitos, y $N-r$ fracasos, siendo los éxitos no distinguibles entre ellos mismos y lo mismo ocurre con los fracasos, la regla general para lograr el número de permutas se reduce a $N!/r!(N-r)!$. Si por ejemplo lanzamos 10 veces una moneda y obtenemos 6 caras, el número de disposiciones posibles de caras y cruces será $10!/6!4! = 210$. En el capítulo siguiente podremos utilizar ampliamente este caso especial cuando estudiemos la distribución binomial.

* Puede llegar a resultar tedioso el trabajar con factoriales sin recurrir a simplificaciones del cálculo. Por fortuna, al trabajar con razones entre factoriales, resulta posible llevar a cabo una cantidad considerable de cancelaciones, como en el caso del ejemplo anterior, en el que está implicada la razón $10!/6!4!$. Los siguientes son los valores numéricos de los factoriales de 1 a 20:

$1! = 1$	$11! = 3.992 \times 10^7$
$2! = 2$	$12! = 4.790 \times 10^8$
$3! = 6$	$13! = 6.227 \times 10^9$
$4! = 24$	$14! = 8.718 \times 10^{10}$
$5! = 120$	$15! = 1.308 \times 10^{12}$
$6! = 720$	$16! = 2.092 \times 10^{13}$
$7! = 5\,040$	$17! = 3.557 \times 10^{14}$
$8! = 40\,320$	$18! = 6.402 \times 10^{15}$
$9! = 362\,880$	$19! = 1.216 \times 10^{17}$
$10! = 3\,628\,800$	$20! = 2.433 \times 10^{18}$

Para valores más elevados de N resulta posible precisar los límites entre los cuales se hallará $N!$, utilizando para ello la aproximación de Stirling:

$$\sqrt{2N\pi} \left(\frac{N}{e}\right)^N < N! < \sqrt{2N\pi} \left(\frac{N}{e}\right)^N \left(1 + \frac{1}{12N-1}\right)$$

en la que $\pi \approx 3.14159$ y $e \approx 2.71828$. Los estudiantes familiarizados con el uso de logaritmos encontrarán muy conveniente trabajar con los logaritmos de los factoriales, convirtiendo así productos en sumas y razones en diferencias. Por ejemplo:

$$\log \left(\frac{8!}{3!}\right) = \log \frac{8 \ 7 \ 6 \ 5 \ 4 \ 3 \ 2 \ 1}{3 \ 2 \ 1}$$

$$= \{\log 8 + \log 7 + \log 6 + \log 5 + \log 4 + \log 3 + \log 2 + \log 1\}$$

$$- \{\log 3 + \log 2 + \log 1\} = \log 8 + \log 7 + \log 6 + \log 5 + \log 4$$

Algunos ejemplos. Estudiemos ahora algunas aplicaciones de estos principios a otros problemas de probabilidad, de naturaleza algo más complicada que la de los que hemos visto hasta ahora. Como está implícito en la introducción a esta sección, una importante estrategia general en el caso de muchos problemas en los que el orden de selección carece de importancia consiste en calcular la probabilidad de una determinada permuta, multiplicando a continuación aquélla por el número de permutas implicadas. Supongamos que deseamos, por ejemplo, obtener la probabilidad de conseguir exactamente un as y por lo menos dos reyes en cuatro tiradas, con reposición de cartas. Observaremos que esto puede realizarse al obtener bien un as y tres reyes o un as, dos reyes y alguna otra carta que no sea as ni rey. Si representamos simbólicamente estas posibilidades como *ARRR* y *ARRO* (en donde "O" representa "otra carta"), veremos que hay $4!/3! = 4$ formas de ordenar el as y los tres reyes, mientras existen $4!/2! = 12$ maneras de disponer la combinación *ARRO*. Por ser diferentes los números de las permutas en ambas situaciones es por lo que debemos mantenerlas diferenciadas. Si nuestra prueba es con reposición, la probabilidad de obtener un as en una sola tirada es de $1/13$, como lo es la de obtener un rey, en tanto que la probabilidad de sacar una O es de $11/13$. Así resulta que la probabilidad de obtener exactamente un as y dos o más reyes será:

$$4(1/13)^4 + 12(1/13)^3(11/13) = 136/28\ 561 = .0048$$

Supongamos que deseamos obtener la probabilidad de conseguir exactamente un as y por lo menos dos *corazones* en cuatro tiradas, con reposición. Aparece ahora una complicación más, ya que uno de los corazones puede ser un as. Será conveniente distinguir entre cuatro tipos de cartas: el as de corazones (*AC*), cuya probabilidad de ser seleccionado es de $1/52$; los ases no de corazones (*AC̄*), con probabilidad de $3/52$; los no ases de corazones (*ĀC*), con probabilidad de selección de $12/52$, y los no ases de corazón (*ĀC̄*) con una probabilidad de $36/52$ de ser sacados. La suma de todas estas probabilidades es naturalmente igual a la unidad ya que los tipos en cuestión son mutuamente exclusivos y exhaustivos.

Despleguemos a continuación las combinaciones que pueden producir exactamente un as y dos o más corazones, calculando el número de permutas en cada caso. Dichas combinaciones son las siguientes:

a) Exactamente dos corazones:

$$AC, \bar{AC}, \bar{AC}, \bar{AC} (4!/2!) [1/52 \cdot 12/52 \cdot 36/52 \cdot 36/52] = .02552$$

$$\bar{AC}, \bar{AC}, \bar{AC}, \bar{AC} (4!/2!) [3/52 \cdot 12/52 \cdot 12/52 \cdot 36/52] = .02552$$

b) Exactamente tres corazones:

$$AC, \bar{AC}, \bar{AC}, \bar{AC} (4!/2!) [1/52 \cdot 12/52 \cdot 12/52 \cdot 36/52] = .00851$$

$$\bar{AC}, \bar{AC}, \bar{AC}, \bar{AC} (4!/3!) [3/52 \cdot 12/52 \cdot 12/52 \cdot 12/52] = .00284$$

c) Exactamente cuatro corazones:

$$AC, \bar{AC}, \bar{AC}, \bar{AC} (4!/3!) [1/52 \cdot 12/52 \cdot 12/52 \cdot 12/52] = \frac{.00094}{.06333}$$

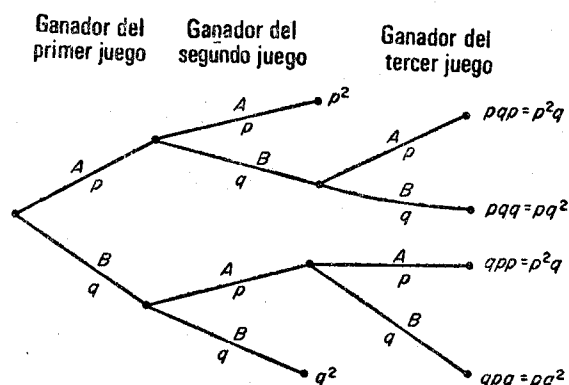
Sumando estas probabilidades de acontecimientos mutuamente exclusivos obtenemos un total de probabilidades de .063.

Consideremos por fin una situación en la que es más conveniente dibujar lo que se denomina un diagrama de árbol que represente las distintas posibilidades. Ocurre a veces que una secuencia de acontecimientos termina en puntos diferentes, según sea el desenlace de otros acontecimientos previos.

La ilustración más familiar de este hecho se observa en los eventos atléticos en los que un equipo será declarado vencedor si gana dos de tres juegos, o tal vez cuatro de siete, y en los que no hay necesidad de seguir jugando una vez que el número necesario de juegos ha sido ya ganado. Supongamos que hay dos

equipos A y B jugando una serie de "dos de tres". Supongamos también que A es el mejor equipo y que, con base en sus anteriores resultados, se le asigna una probabilidad de .6 para ganar cualquier juego dado. En un ejemplo más realista, la probabilidad de ganar cada juego puede cambiar según los resultados de los juegos precedentes, lo que podremos determinar utilizando el método que vamos a proponer. Para mayor sencillez tomemos como $p = .6$ la probabilidad de que el equipo A gane cada juego, en cuyo caso $q = .4$ representará la probabilidad del equipo B para vencer en cada juego. Se supone, por tanto, que las pruebas sucesivas son independientes. ¿Cuál es la probabilidad de que el equipo A gane la serie? ¿Cuáles son las probabilidades individuales de cada posible secuencia de ganancias y pérdidas?

Podemos expresar las posibles secuencias en el diagrama siguiente:



La rama superior del árbol representa las posibilidades, suponiendo que A ha ganado el primer juego, en tanto que la rama baja representa las correspondientes al triunfo inicial de B. Si A gana el segundo juego después de haber ganado el primero, la serie se detiene, y A gana con una probabilidad de p^2 . Sin embargo, si A gana el primer juego y B el segundo, es preciso jugar un tercer juego. Si éste es ganado por A la serie se detiene, y A gana con probabilidad de p^2q , pero si B gana el tercer juego gana la serie con probabilidad pq^2 . Esta clase de series da origen a un árbol perfectamente simétrico, aun cuando es evidentemente posible inventar competencias con *handicaps* que produzcan árboles asimétricos. Por ejemplo, el equipo A puede necesitar ganar cuatro juegos en tanto B con sólo tres vencerá.

Las probabilidades de las varias posibilidades pueden ser ya computadas como sigue:

Secuencias según las cuales
vence el equipo A ($p = .6$)

$$p^2 = .360$$

$$pqp = .144$$

$$qpp = .144$$

$$\text{Probabilidad de} \quad \text{—} \\ \text{ganar la serie} \quad .648$$

Secuencias según las cuales
vence el equipo B ($q = .4$)

$$q^2 = .160$$

$$pqq = .096$$

$$qpq = .096$$

$$\text{—} \\ .352$$

El hecho de que las probabilidades suman una unidad puede representarse algebraicamente como sigue:

$$p^2 + 2p^2q + 2pq^2 + q^2 = p^2 + 2pq(p + q) + q^2$$

$$= p^2 + 2pq + q^2 = (p + q)^2 = 1$$

IX.4. Valores esperados

Una idea, surgida probablemente en un casino de juego, tiene importantes aplicaciones estadísticas. Consiste en que si uno repite un experimento un gran número de veces, haciendo apuestas sobre los resultados, será posible calcular las ganancias (o pérdidas) esperadas, con base en diferentes suposiciones acerca de la naturaleza del juego que está siendo practicado. Para poner un ejemplo muy sencillo, supongamos que se están echando monedas a cara o cruz, apostando siempre a caras, y que cada vez que aparece una cara se gana un peso, pero se pierden 2 pesos por cada vez que sale cruz. Bajo el supuesto de que la moneda no ha sido preparada, es evidente que uno no desearía mantener tal juego durante mucho tiempo. ¿Pero cómo pueden calcularse las ganancias o pérdidas esperadas en otros casos más complejos?

En el sencillo ejemplo anterior el sentido común sugeriría multiplicar la probabilidad de cada aparición por la ganancia o la pérdida correspondiente a dicho resultado, sumando a continuación los resultados. Obtendríamos así como "ganancia" esperada la cantidad $(1)(\frac{1}{2}) + (-2)(\frac{1}{2}) = -.5$. Esto significa que, como promedio, uno habría de esperar perder 50 centavos por tirada. Las ganancias o pérdidas reales pueden por supuesto diferir de este valor esperado, pero si hubiésemos un buen número de veces, nuestra pérdida total sería aproximadamente de $.5N$, en donde N representa el número de tiradas.

Como segundo ejemplo, supongamos que lanzamos un solo dado, recibiendo un peso si sale par; perdiendo dos pesos si salen uno o tres puntos, y ganando tres pesos si salen cinco puntos. Suponiendo que todas las caras tienen las mismas posibilidades de salir, nuestras ganancias esperadas serían:

$$(-2)(1/6) + (1)(1/6) + (-2)(1/6) + (1)(1/6) + (3)(1/6) + \\ + (1)(1/6) = 1/3 = \$.333$$

por juego. En general, si hay k posibles resultados X_1, X_2, \dots, X_k , y si la probabilidad de X_i es dada por $p(X_i)$, podremos definir el valor esperado de las variables X , indicando con el símbolo

$E(X)$, como: $E(X) = \sum_{i=1}^k X_i p(X_i)$. En los ejemplos considerados hasta ahora, la X_i ha representado pagos (en pesos) para cada grupo de resultados, pero es posible concebir en términos más generales esta idea del valor esperado.

Supongamos por ejemplo que tenemos una población que contiene N individuos con puntuaciones en la X variable. Si elegimos al azar entre esta población, cada individuo tendrá una probabilidad de $1/N$ de ser seleccionado. ¿Cuál es el valor esperado de X ? En este caso tendríamos:

$$E(X) = X_1 p(X_1) + X_2 p(X_2) + \dots + X_N p(X_N) \\ = (X_1 + X_2 + \dots + X_N)(1/N) = \bar{X}$$

y obtendremos el interesante resultado de que el valor esperado de X es su media, *suponiendo que el muestreo sea aleatorio*.

A partir del capítulo siguiente nos ocuparemos ampliamente de las distribuciones de probabilidades, llamadas distribuciones por muestreo. En sentido estricto, tales distribuciones son infinitas, ya que se refieren a probabilidades que aquí definimos solamente en términos limitadores. Podemos sin embargo mencionar estas distribuciones de probabilidad como si tuviesen valores esperados interpretables como sigue: imaginemos muestreos aleatorios hechos repetidamente con una determinada población. Si tal población tiene una media a la que denominaremos con la letra griega μ , entonces $E(X) = \mu$. Desearemos encontrar también los valores esperados de otras cantidades, tales como la *muestra* media \bar{X} , la que a su vez resulta tener su valor esperado de $E(\bar{X})$ igual a μ , en el caso de muestreo aleatorio. Otra expresión de considerable interés teórico en la estadística es $E[X - E(X)]^2$, la que en el caso del muestreo aleatorio, para el cual $E(X) = \mu$, es $\sum_{i=1}^N (X_i - \mu)^2 p(X_i) = 1/N \sum_{i=1}^N (X_i - \mu)^2$, o la variancia de X . Aun cuando no haremos gran uso de la notación correspondiente a valores esperados, probablemente se encontrarán referencias a la misma en textos más avanzados, ya que en las pruebas de estadística matemática es empleada ampliamente.

IX.5. Independencia y muestreo aleatorio \times

Todas las pruebas estadísticas a examinar en este texto parten del supuesto de que hay independencia entre los acontecimientos y que, por consiguiente, las probabilidades condicionales no han de emplearse al multiplicar las probabilidades.¹¹ En otros términos: se supone que existe independencia de selección en el interior de una muestra, no teniendo la selección de un individuo influencia alguna sobre la selección de otro a incluir en la misma muestra. Sin embargo, se dan muchos casos en que se propende a violar dicho importante supuesto. De ahí que el lector deba acostumbrarse a preguntar siempre si el supuesto de independencia está o no efectivamente justificado en cualquier problema dado. Será útil, en este punto, indicar unos pocos ejemplos de situaciones en las que se corre riesgo de prescindir del supuesto en cuestión.

Los estadígrafos obtienen a menudo lo que se designa como *muestra al azar* (o *muestra irrestricta aleatoria*) con objeto tanto de satisfacer el supuesto necesario de independencia como para dar a todo individuo de la población considerada un número igual de oportunidades de figurar en la muestra. Sirviéndonos de una tabla de números al azar o algún otro arreglo por el estilo, puede obtenerse una muestra en forma esencialmente idéntica a la de extraer naipes de una baraja bien barajada o números en un juego de lotería. La muestra aleatoria posee la propiedad *no sólo de dar a cada individuo la misma oportunidad de ser seleccionado, sino también la de proporcionar a cada combinación de individuos una oportunidad igual de selección*.¹²

En rigor, como quiera que casi siempre extraemos las muestras sin reposición, el supuesto de independencia no se cumple por completo. Sin embargo, cuando la población es grande en relación con la magnitud de la muestra, podemos olvidar perfectamente la pequeña distorsión resultante de que a ningún individuo se le dé la oportunidad de ser seleccionado otra vez. Por ejemplo: si de una población de 100 mil personas se extraen 500, las probabilidades son muy pequeñas de que alguna de ellas volviera a seleccionarse en el caso de que su nombre se pusiera de nuevo en el grupo. Y en forma análoga, la diferencia es prácticamente muy pequeña si reponemos o no al extraer sólo tres cartas de una baraja; pero, si extrajéramos 35, la diferencia sería considerable. Si la muestra es relativamente grande en comparación con la

¹¹ Es lo que se verá en el caso de la binomial que se examinará en el siguiente capítulo. Sin embargo, en el caso de otras pruebas, el lector habrá de aceptar simplemente la verdad de este aserto.

¹² En el cap. XXI se distinguirá la extracción de muestras al azar de otras formas de extracción de uso corriente, tales como la sistemática, la estratificada y la de conglomerados.

población, entonces puede aplicarse a veces un factor de corrección para compensar la falta de remplazo.¹³

Pese a que los problemas resultantes de la falta de remplazo no sean graves, la falta de proporcionar a cada combinación de individuos la posibilidad de aparecer en la muestra puede traducirse en una grave violación del supuesto de independencia. Supóngase, por ejemplo, que nos dispusiéramos a clasificar los naipes corrientes en cuatro montones: uno para los tréboles, otro para las espadas, etcétera. Supóngase luego que fuéramos a seleccionar uno de dichos montones al azar. Es obvio que cada carta de la baraja tendría la misma oportunidad (1 sobre 4) de ser seleccionada, pero indudablemente todas las combinaciones no serían posibles, y no digamos ya igualmente probables. En efecto, sabiendo que el naipe de encima es una espada, sabemos que todos los demás naipes del montón son igualmente espadas.

Las muestras de área o por conglomerados empleadas comúnmente en las investigaciones sociales no cumplen el supuesto de independencia por esta misma razón. En efecto, si se seleccionan al azar 100 manzanas de casas de una población y luego se incluye en la muestra cada tercera familia de las manzanas en cuestión, es obvio que todas las combinaciones de familias no tienen la misma oportunidad de figurar en aquella. En efecto, dos familias de la misma manzana tienen mayor oportunidad de figurar en la misma muestra de lo que es el caso de dos familias en dos manzanas distintas. Como quiera que las manzanas de casas urbanas suelen ser por lo regular relativamente homogéneas en cuanto a características tales como el ingreso o la instrucción del jefe de familia, el resultado de semejante tipo de extracción de muestra será menos exacto que una selección de una muestra aleatoria del mismo tamaño. Esto puede verse intuitivamente si imaginamos una situación en que todas las manzanas sean totalmente homogéneas, como era el caso de los cuatro montones de naipes. En tal caso, en efecto, sólo necesitaríamos obtener información acerca de una vivienda en cada manzana, y el número de "casos" sería de hecho el número de las manzanas seleccionadas, esto es, un N bastante menor. Según veremos en el capítulo XXI, es posible obtener unos resultados extremadamente engañosos, si habiendo extraído una muestra semejante por conglomerados, el investigador se sirve luego de pruebas estadísticas que presuponen una extracción al azar.

Un problema análogo puede fácilmente encontrarse cuando se está interesado en los actos individuales de conducta. Supóngase, por ejemplo, que un sociólogo efectúa un experimento en el que se sirve de 30 sujetos, cada uno de los cuales formula 50 juicios distintos. Se tendrían en tal caso 1 500 juicios, y nos podríamos ver inducidos a servirnos en una prueba estadística de semejan-

¹³ Véase secc. XXI.1.

te N artificialmente ponderada, suponiendo que los 1 500 juicios en cuestión constituyeran una muestra al azar de los juicios de algún tipo de población. Pero sería manifiestamente absurdo en la mayoría de los casos suponer que los juicios de un mismo individuo son estadísticamente independientes unos de otros. En efecto, sus primeros 30 juicios afectarán probablemente a los demás, ya que a diferencia de la moneda, la persona sí tiene memoria.

Supóngase que un sociólogo se interesa ante todo en pares de personas, como unidad, más que en el individuo singular. Puede tener un grupo de 20 personas, cada una de las cuales esté en interacción con todas las demás. Tendría, en consecuencia, $(20)(19)/2$ o 190 pares de personas, pero no estaría en condiciones de considerar cada par como independiente de los otros. Es obvio que el conocimiento a propósito del par Smith-Brown suministrará probablemente alguna información sobre los pares Smith-Jones o Brown-Jones, ya que las mismas personas figuran en varios pares.

Los ecólogos, antropólogos y otros sociólogos interesados en generalizar a propósito de localidades, sociedades u otras unidades definidas espacialmente necesitan también preocuparse de la falta de independencia en una gran parte de su labor. Aquí el problema parece derivarse del hecho de que las unidades seleccionadas no son a menudo claramente distintas. En efecto, las fronteras de una sociedad o una localidad pueden ser difíciles de definir, y una unidad semejante puede pasar gradualmente a la otra, siendo las divisiones más o menos arbitrarias.¹⁴ Así, por ejemplo, si se utilizan como unidades los distritos del censo en el interior de una ciudad o los distritos territoriales en el interior de un Estado, resulta a menudo posible predecir a propósito de una unidad sobre la base de la unidad vecina. Si la cuota de delincuencia es elevada en un distrito, es probable que lo sea también en el vecino, ya que es incluso posible que las mismas bandas de delincuentes se extraigan de ambos distritos. Que "algo no está en orden" en relación con el supuesto de independencia puede percibirse intuitivamente dándose cuenta de que, cuando las unidades no son claramente distintas, sería posible ponderar el número de "unidades" a cualquier tamaño deseado, cortando simplemente el pastel en muchos pedazos pequeños. Así, por ejemplo, si no hay bastantes sociedades en el mundo para obtener significancia estadística, podría dividirse cada sociedad en 10 subregiones y obtener 10 veces más "casos".

¹⁴ Esta situación se parecería en cierto modo a la de una baraja cada una de cuyas cartas pasara insensiblemente a las otras, de modo que resultara difícil decir dónde una de ellas terminaba y empezaba la otra. O también, ¡que cada carta fuera capaz de influir los valores figurados de sus vecinas inmediatas!

En un texto como éste no es posible examinar soluciones a los problemas que comportan violaciones del supuesto de independencia. Que el autor sepa, muchos de dichos problemas no han sido resueltos satisfactoriamente. Resulta a menudo difícil apreciar la gravedad de los errores introducidos cuando *no* se cumplen supuestos requeridos, como el de independencia. Pisamos terreno firme siempre que tenemos la seguridad de que los supuestos requeridos para alguna prueba *sí* se cumplen; pero si no se cumplen, raramente resulta posible decidir exactamente en qué medida nos apartamos de dichos supuestos. Con objeto de estar seguro, el lector ha de acostumbrarse a examinar cuidadosamente todo supuesto. Si se tienen motivos para dudar de la validez de alguno, entonces habría que considerar seriamente el servirse de otro procedimiento que no lo contenga. Así, por ejemplo, podría decidirse recurrir a otra unidad de análisis, o sea a la persona, más bien que a los actos de la conducta o a los pares de personas, o bien a los delincuentes particulares más que a las tasas de delincuencia en relación con un distrito del censo.

Si bien los sociólogos y otros que se sirven de la estadística aplicada han propendido en ocasiones a prescindir de los supuestos, llegando así a conclusiones infundadas, es también posible, por otra parte, pecar de prurito excesivo de perfección. Como quiera que, en efecto, no nos las habemos nunca con situaciones tan sencillas como las de lanzar una moneda al aire o sacar naipes de una baraja perfecta, resulta siempre posible poner en tela de juicio cualquier procedimiento a título de imperfecto en relación con el ideal que se persigue. Se puede abrigar un temor tal de violar supuestos, que se prefiere prescindir por completo de toda técnica estadística. Es necesario, sobre todo en una disciplina que se caracterice por estudios exploratorios y técnicas científicas relativamente imprecisas, llegar a compromisos con la realidad. El procedimiento más indicado consistirá en hacer el menor número de compromisos posible, dentro de los límites de lo practicable.

GLOSARIO

Sucesos
Límite
Sucesos mutuamente exclusivos
Probabilidad
Muestra aleatoria
Independencia estadística

EJERCICIOS

1. En un simple lanzamiento de un dado no cargado, ¿cuál es la probabilidad de:

- ¿sacar un 6?
- ¿no sacar un 6?
- ¿sacar un 1 o un 6? Respuesta, $\frac{1}{3}$
- ¿sacar un 1 y un 6?
- ¿sacar un número impar o un 6?

2. Cuál es la posibilidad de obtener cada uno de los siguientes resultados en *tres* extracciones de un juego de naipes bien barajado:

- ¿tres sotas, con reposición? Respuesta, $\frac{1}{2^{197}}$
- ¿tres sotas, sin reposición? Respuesta, $\frac{1}{5^{525}}$
- ¿una espada, un corazón y un diamante (en cualquier orden), con reposición?
- ¿exactamente dos ases, con reposición?
- ¿por lo menos un as, con reposición? (Indicación: ¿cuál es la alternativa por al menos un as?) Respuesta, $\frac{496}{2^{197}}$
- *f) ¿por lo menos un as y por lo menos un rey, con reposición? [Indicación: en f) y en algunos de los ejercicios que siguen, será útil dividir el problema en tres pasos: 1) determinar las distintas combinaciones de cartas que dan por lo menos un as y por lo menos un rey (v.gr., un as, un rey y otra carta cualquiera: dos ases y un rey, etcétera); 2) determinar la probabilidad de obtener dichos naipes en cualquier orden particular; y 3) determinar para cada una de dichas combinaciones el número de ordenamientos posibles.]

3. Supóngase que se interroga a 1 000 novatos acerca de sus gustos musicales. Se encuentra que 400 de los estudiantes son aficionados a la música clásica, en tanto que los restantes no lo son. De estos 400 aficionados, sólo a 100 les gusta el *rock and roll*. Hay 400 personas a las que no les gusta ni un género ni otro de música, en tanto que a las restantes les gustó sólo el *rock and roll*.

- Si se escoge un estudiante al azar, de entre la población en cuestión, y si A es el acontecimiento consistente en que le gusta la música clásica y B el acontecimiento consistente en que le guste el *rock and roll*, ¿cuáles son $P(A)$, $P(B)$, $P(A|B)$ y $P(B|A)$?
- Verifíquese numéricamente que

$$P(A \& B) = P(A)P(B|A) = P(B)P(A|B)$$

- ¿Cuál es la probabilidad de seleccionar una persona a la que guste uno de los dos géneros de música, pero no ambos?
- *d) Observando que una persona puede tener uno de cuatro tipos de gusto (que le gusten los dos géneros, que no le guste ninguno, etcétera), ¿cuál es la probabilidad de que tres estudiantes seleccionados al azar como compañeros de cuarto *tengan* los mismos gustos? (Supóngase reposición). Respuesta, .10.
- *e) ¿Cuál es la probabilidad de que haya *por lo menos* dos aficionados al *rock and roll* en un corredor de ocho personas? (Supóngase selección al azar, con reposición.)

* 4. En los datos que se consignan a continuación, supóngase que A es el acontecimiento consistente en seleccionar un varón, B el acontecimiento consistente en seleccionar una persona de cultura universitaria, y C el consistente en seleccionar una persona de grado elevado de prejuicio:

Grado de prejuicio	Cultura universitaria		Cultura inferior a universitaria	
	Varones	Mujeres	Varones	Mujeres
Alto	100	50	200	250
Bajo	150	100	150	200

- a) Hállese $P(A \& B \& C)$ en una sola extracción, sin servirse de fórmula. Verifíquese que la fórmula es cierta en el caso de los datos numéricos de este ejercicio.
- b) Hágase lo mismo para $P(A \circ B \circ C)$. Será preciso desarrollar la fórmula para $P(A \circ B \circ C)$.
- c) ¿Cuál es la probabilidad de seleccionar exactamente un varón de cultura universitaria, exactamente una mujer de cultura universitaria y exactamente una persona de alto grado de prejuicio en una extracción al azar de tres personas? (Supóngase reposición.)

* 5. Los estudiantes inscritos en un curso de introducción a la sociología de la Universidad de Michigan fueron clasificados según sus aspiraciones profesionales para sí mismos o para sus cónyuges, conforme al sexo de los interrogados. Se obtuvieron los siguientes datos:

Sexo	Aspiraciones elevadas	Aspiraciones modestas	Total
Varones	43	10	53
Mujeres	71	93	164
Total	114	103	217

Supóngase que de esta población de 217 estudiantes se seleccionan aleatoriamente individuos,

- a) ¿Cuál es la probabilidad de seleccionar un estudiante de aspiraciones elevadas? ¿Cuál es la probabilidad de seleccionar un estudiante de aspiraciones elevadas, en el supuesto de que sea varón? ¿En el supuesto de que sea mujer?
- b) Supóngase que de dicha población se seleccionan individuos al azar (sin reposición), indicando por suposición en cada caso si se trata de un individuo de aspiraciones elevadas o modestas. ¿Con qué frecuencia se supondrá que tiene aspiraciones elevadas? ¿Modestas? ¿Por qué? En 217 extracciones, ¿cuántos errores se espera cometer? Respuesta, 103.

- c) Supóngase que se sabe el sexo del estudiante. Dado que es varón, ¿cuántos errores se espera cometer al asignar los 53 varones a las categorías respectivas de aspiraciones elevadas o modestas? ¿Cuántos en relación con las mujeres? Respuesta, 10; 71.
- d) ¿Cómo podría construirse un índice que mostrara la reducción proporcional de errores, si el interrogado es varón, en comparación con los errores en el caso de desconocerse el sexo? Como se verá en el capítulo xv, semejante índice puede emplearse para medir la fuerza o grado de relación entre el sexo del interrogado y sus aspiraciones profesionales.

* 6. Hágase un diagrama de árbol para calcular las probabilidades de todos los resultados posibles de una Serie Mundial (el que gane 4 juegos de un máximo de 7), suponiendo que la probabilidad de que el equipo de la Liga Nacional gane cada juego es de .6.

BIBLIOGRAFÍA

1. Alder, H. L., y E. B. Roessler: *Introduction to Probability and Statistics*, 4ª ed., W. H. Freeman and Company, San Francisco, 1968, cap. 5.
2. Feller, William: *An Introduction to Probability Theory and Its Applications*, 3ª ed., John Wiley & Sons, Inc. Nueva York, 1967.
3. Freund, J. E.: *Modern Elementary Statistics*, 3ª ed., Prentice-Hall, Inc., Englewood Cliffs, N. J., 1967, caps. 5 y 6.
4. Gelbaum, B. L., y J. G. March: *Mathematics for the Social and Behavioral Sciences*, W. B. Saunders Company, Filadelfia, 1969, caps. 2-4.
5. Hays, W. L.: *Statistics*, Holt, Rinehart and Winston, Inc. Nueva York, 1963, caps. 2 y 4.
6. Kemeny, J. G., J. L. Shell y G. L. Thompson: *Introduction to Finite Mathematics*, 2ª ed., Prentice-Hall, Inc., Englewood, Cliffs, N. J., 1966, caps. 3 y 4.
7. Mueller, J. H., K. Schuessler y H. L. Costner: *Statistical Reasoning in Sociology*, 2ª ed., Houghton Mifflin Company, Boston, 1970, cap. 8.
8. Savage, L. J.: *The Foundations of Statistics*, John Wiley & Sons, Inc., Nueva York, 1954, caps. 1-3.

X. PRUEBAS DE HIPÓTESIS: LA DISTRIBUCIÓN BINOMIAL

EN CIENCIAS sociales nos encontramos a menudo frente a simples dicotomías como la de si un individuo posee o no un determinado atributo o de si un experimento ha tenido éxito o ha fracasado. En tales casos, siempre que sea posible formular una hipótesis acerca de cierta probabilidad de éxitos, siempre que las pruebas sean independientes unas de otras y siempre que el número de éstas sea relativamente pequeño, es posible servirse de pruebas estadísticas que se comportan como una distribución binomial. Pese a que existen numerosas pruebas estadísticas más prácticas que las que se sirven de la distribución binomial, es conveniente, con todo, dedicar un tiempo considerable a dicha distribución, a causa ante todo de su sencillez. Al emplear la distribución binomial, el estudiante puede seguir en forma relativamente fácil todos los pasos que aquélla comporta, obteniendo con ello una visión de los procedimientos generales empleados en todas las pruebas estadísticas.

Es probable que al lector el presente capítulo se le antoje como excepcionalmente difícil, debido al hecho de que se exponen en el mismo cierto número de ideas nuevas en forma más bien compacta. Muchas de esas ideas vuelven a tratarse en el capítulo XI, y es tal vez preferible que el lector considere estos dos capítulos como una sola unidad, leyendo el capítulo XI antes de dominar por completo la materia del presente. En particular, se puede dejar para después la lectura de la sección X.3, que trata de diversas aplicaciones de la binomial, y de la sección X.4, sobre extensiones.

X.1. La distribución de muestreo binomial

Antes de examinar cada uno de los pasos implicados en las pruebas estadísticas, será conveniente considerar cómo se obtienen las distribuciones binomiales. Por el momento simplificaremos la cuestión limitándonos al lanzamiento de monedas. En este tipo de problemas, el número de los lanzamientos constituye la magnitud de la muestra, y nuestro interés se centra en el número de caras (éxitos) obtenidas en N pruebas.

Suponiendo que las N pruebas (lanzamientos de monedas) son estadísticamente independientes una de otra, podemos evaluar inmediatamente la probabilidad de obtener r caras y $N - r$ cruces en algún orden particular. Podemos, por ejemplo, obtener la probabilidad de conseguir r caras sucesivas seguidas de $N - r$ cruces. Supongamos que p es la probabilidad de obtener una

cara; en este caso, la probabilidad de obtener una cruz, que designamos como q , será de $1 - p$. Como quiera que las pruebas son independientes, podemos multiplicar simplemente las probabilidades incondicionales. La probabilidad de obtener exactamente r caras en el orden que se acaba de indicar será:

$$\underbrace{p \ p \ p \ \dots \ p}_{r \text{ términos}} \quad \underbrace{q \ q \ q \ \dots \ q}_{N-r \text{ términos}} = p^r q^{N-r}$$

Es obvio que en los supuestos de independencia estadística y probabilidad constante de los éxitos (v.gr., la moneda no se desgasta de modo irregular), la probabilidad de obtener cualquier otro orden particular de r caras y $N - r$ cruces será también $p^r q^{N-r}$. Por lo tanto, con objeto de obtener la probabilidad de conseguir *exactamente* r caras en *cualquier orden* sólo se necesita contar el número de maneras distintas que tenemos de obtener r caras y $N - r$ cruces. Sin embargo, por poco que N sea grande, la tarea se hace muy fastidiosa. Afortunadamente disponemos de una fórmula matemática que hace innecesaria dicha operación de contar. En efecto, el número de maneras posibles en que podemos ordenar r éxitos y $N - r$ fracasos, escrito simbólicamente como

$$\left(\frac{N}{r} \right) \text{ o como } C_r^N \text{ es:}$$

$$\left(\frac{N}{r} \right) = \frac{N!}{r!(N-r)!} \quad (\text{X.1})$$

en donde $N!$ (léase factorial N) = $N(N-1)(N-2) \dots (3)(2)(1)$ (2) (1), y lo mismo por lo que se refiere a $r!$ y a $(N-r)!$ ¹

Con fines de cálculo, la fórmula (X.1) puede simplificarse observando que algunos de los términos del numerador y el denominador se eliminan recíprocamente.¹ Como quiera que $r \leq N$, podemos escribir $N!$ como producto de dos factores, como sigue:

$$N! = [N(N-1)(N-2) \dots (N-r+1)] [(N-r) \dots (3)(2)(1)]$$

$$= [N(N-1)(N-2) \dots (N-r+1)] [(N-r)!]$$

y vemos inmediatamente que $(N-r)!$ puede eliminarse tanto del numerador como del denominador. En esta forma nos queda, pues:

¹ El símbolo $\left(\frac{N}{r} \right)$ no ha de confundirse con N/r o N dividido entre r .

$$\left(\frac{N}{r}\right) = \frac{N(N-1)(N-2)\dots(N-r+1)}{r!} \quad (\text{X.2})$$

Así, pues, si queremos encontrar el número de maneras de obtener cuatro caras en diez lanzamientos, tenemos:

$$N - r + 1 = 10 - 4 + 1 = 7$$

y por consiguiente:

$$\left(\frac{10}{4}\right) = \frac{(10)(9)(8)(7)}{(4)(3)(2)(1)} = 210$$

Obsérvese que al emplear la ecuación (X.2) se tiene el mismo número de factores en el numerador y el denominador. Esto es siempre así. Esta segunda forma es más sencilla, con fines de cálculo, que la primera. Si $r > N/2$, empezamos a tener algunos términos que aparecen tanto en el numerador como en el denominador y, por consiguiente, se eliminan recíprocamente. Por ejemplo, si $r = 6$, tenemos:

$$\left(\frac{10}{6}\right) = \frac{(10)(9)(8)(7)}{(1)(2)(3)(4)} \left[\frac{(6)(5)}{(5)(6)} \right] = 210$$

lo que nos da el mismo resultado obtenido que al calcular $\left(\frac{10}{4}\right)$.

En general puede demostrarse que

$$\left(\frac{N}{r}\right) = \left(\frac{N}{N-r}\right)$$

de modo que puede utilizarse lo mismo r que $N - r$, según cuál de ellos sea menor.

Si ahora queremos saber la probabilidad de obtener *exactamente* r éxitos en N pruebas y no nos interesamos por el orden en que ocurran, podemos multiplicar la probabilidad de obtener una secuencia particular cualquiera por $\left(\frac{N}{r}\right)$. Designando la probabilidad deseada por $P(r)$, tenemos:

$$P(r) = \left(\frac{N}{r}\right) p^r q^{N-r}$$

$$\begin{array}{lll} \text{o Probabilidad} & \text{Nº de maneras} & \text{Probabilidad} \\ \text{de } r \text{ éxitos} & \text{de obtener } r & \text{(X.3)} \\ \text{exactamente} & \text{éxitos} \times & \text{de una sucesión} \\ & & \text{dada cualquiera} \end{array}$$

Si la moneda fuera perfectamente correcta, esto es, si $p = q = 1/2$, la probabilidad de obtener exactamente cuatro caras en diez pruebas sería de:

$$P(4) = \left(\frac{10}{4}\right) \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^6 = 210 \left(\frac{1}{2}\right)^{10} = \frac{210}{1024} = .205.$$

Y en forma análoga podemos calcular las probabilidades de obtener exactamente 0, 1, 2, ..., 10 caras en 10 pruebas.

Nº de caras	Probabilidades (con $p = 1/2$)
0	1/1024 = .001
1	10/1024 = .010
2	45/1024 = .044
3	120/1024 = .117
4	210/1024 = .205
5	252/1024 = .246
6	210/1024 = .205
7	120/1024 = .117
8	45/1024 = .044
9	10/1024 = .010
10	1/1024 = .001
	<hr/> 1.000

Obsérvese que siendo r cero, la magnitud $\left(\frac{N}{r}\right)$ es indefinida y la fórmula falla. Vemos, sin embargo, que, siendo $r = 0$, sólo hay un orden posible (todo cruces). En este ejemplo, la distribución de probabilidades es perfectamente simétrica. Sirviéndose del hecho de que $\left(\frac{N}{r}\right) = \left(\frac{N}{N-r}\right)$, el lector debería convencerse por sí mismo de que $\left(\frac{N}{r}\right)$ será siempre simétrico, pero que el factor $p^r q^{N-r}$ sólo será exactamente simétrico si $p = q = 1/2$.

En el ejemplo anterior, las probabilidades se han asociado con cada uno de los 11 resultados posibles del experimento. En dicho sencillo ejemplo, sólo se daba un pequeño número de resultados concebibles, dado el supuesto de que en cada lanzamiento sólo dos de ellos eran posibles. En otros experimentos, en cambio, el número de resultados posibles puede ser muy grande y aun infinito, y puede ser necesario agrupar ciertos resultados y asociar una probabilidad con la serie entera de los mismos. Así, por ejemplo, si la moneda se hubiera lanzado al aire 1000 veces, pudimos haber calculado las probabilidades de obtener de 400 a 449, de 450 a 499 o de 500 a 549 caras.

Cuando asociamos probabilidades con cada resultado posible de un experimento, o con grupos de resultados, designamos la distribución resultante de probabilidades como distribución de muestreo. Recordando que utilizamos el concepto de probabilidad para designar el límite de la razón de los éxitos al número total de las pruebas, vemos que la *distribución de muestreo se refiere al número relativo de veces que esperamos obtener ciertos resultados en un número muy grande de experimentos.*

En el ejemplo numérico considerado, cada experimento consiste en lanzar la moneda 10 veces al aire y anotar el número de caras. Nuestros cálculos nos dicen que si efectuáramos el experimento 1 024 000 veces, podríamos esperar obtener aproximadamente (pero no exactamente) 1 000 casos en que no saliera una sola cara, 10 mil en que saliera exactamente una cara, 45 mil casos con dos caras, etcétera. Además, esperaríamos que cuanto mayor número de veces se efectuara el experimento, tanto más cerca quedarán las proporciones empíricas de estas probabilidades teóricas.

En realidad, el investigador nunca obtiene una distribución de muestreo por medios empíricos, ya que por lo regular sólo efectúa un experimento o extrae una muestra una sola vez o, a lo sumo, unas pocas veces. Importa darse cuenta de que la distribución de muestreo es una distribución hipotética, teórica, que sólo se obtendría si un experimento se efectuara un número muy grande de veces. La distribución de muestreo se obtiene aplicando razonamientos *matemáticos* o deductivos, como se hizo en el ejemplo anterior.

Como quiera que las distribuciones de muestreo no son el tipo de distribuciones que el investigador ve realmente de sus datos, las personas que no sientan afición por las matemáticas tendrán probablemente alguna dificultad en comprender el papel que juegan estas distribuciones hipotéticas en la inducción estadística. Pero, a menos que la noción de distribución de muestreo se comprenda claramente, el estudiante se encontrará prácticamente en la imposibilidad de conseguir de la estadística una comprensión que vaya más allá de la de un mero "recetario". De ahí que convenga examinar aquí en forma más sistemática los pasos que se dan al verificar una hipótesis estadística y ver exactamente cómo se emplean dichas distribuciones de muestreo.

X.2. Pasos en las pruebas estadísticas

Todas las pruebas estadísticas comportan cierto número de pasos específicos. Hay que recalcar una vez más que cada uno de dichos pasos ha de efectuarse con anterioridad a la inspección de los datos. Pueden enumerarse como sigue:

1. Formulación de supuestos.

2. Obtención de la distribución de muestreo.
3. Selección de un nivel de significación y de una región crítica.
4. Cálculo de la estadística de la prueba.
5. Tomar una decisión.

Cada uno de estos pasos se examinará con cierto detalle en el presente capítulo y luego, una vez más, en el capítulo XI, de modo que el lector pueda familiarizarse con los procesos generales que comportan las pruebas estadísticas.

1. *Formulación de supuestos.* Con objeto de aplicar la teoría de probabilidades a la obtención de una distribución de selección, el investigador ha de formular ciertos supuestos acerca de la *población* respecto de la cual va a establecer conclusiones generales y de los procedimientos de *muestreo* a emplear. Estos supuestos relativos a la población y a los procedimientos suelen corresponder por lo regular a una de las dos categorías siguientes: 1) la de aquellos de los que el investigador está relativamente seguro o está dispuesto a aceptar, y 2) la de los que le parecen más problemáticos y en los que, por lo tanto, está más interesado. Los supuestos de la primera categoría podemos ponerlos todos juntos en lo que llamaremos el *modelo*. En cuanto a los de la otra categoría, son los que el investigador desea verificar y se designan como *hipótesis*.

Por lo regular, por lo menos en las verificaciones más sencillas de que nos ocuparemos en los próximos capítulos, sólo habrá una hipótesis. Importa darse cuenta de que *desde el punto de vista de la prueba estadística misma, todos los supuestos poseen el mismo carácter lógico*. Si los resultados de la prueba aconsejan descartar los supuestos, todo lo que puede decirse, *sobre la base de la prueba misma*, es que por lo menos uno de los supuestos, y aun posiblemente todos ellos, son probablemente falsos. Como quiera que la prueba ella misma no puede suministrar información acerca de cuál de los supuestos sea erróneo, es indispensable, si los resultados han de tener algún sentido, que sólo uno de ellos sea realmente dudoso. En estas condiciones será posible descartar el supuesto en cuestión (la hipótesis) como erróneo.

Los estudiantes formulan a menudo el siguiente tipo de pregunta: "¿sobre qué base se escoge una determinada prueba estadística con preferencia a otra?" Uno de los criterios que puede darse a estas alturas es el de un modelo apropiado. En otros términos: el investigador ha de seleccionar una prueba que sólo comporte un supuesto dudoso (su hipótesis). En efecto, si una determinada prueba requiere dos o más supuestos dudosos, será difícil, por no decir imposible, decidir cuál de ellos deba descartarse. En tal caso, el estudiante deberá tratar de encontrar una prueba alternativa que no traiga consigo tantos supuestos dudosos.

Para ilustrar lo que precede con nuestro ejemplo de la moneda, la prueba binomial requiere el supuesto de que 10 lanzamientos constituyen una muestra al azar de todos los lanzamientos posibles con la misma moneda, y que todos ellos son independientes uno de otro. Suponemos, pues, que la moneda es correcta. Esto último sería por lo regular nuestra hipótesis, en tanto que lo primero constituiría nuestro modelo, ya que el interés se concentraría probablemente en saber si la moneda es o no correcta. Se concibe, sin embargo, que podamos sospechar de la persona que efectúa los lanzamientos. Si estuviéramos relativamente seguros a propósito de la moneda, por haber comprobado previamente que daba aproximadamente caras la mitad de las veces, entonces podríamos cambiar el problema y verificar una hipótesis relativa al método del lanzamiento (o sea el método de muestreo). Supóngase que no estuviéramos dispuestos a aceptar como modelo la corrección de la moneda o la corrección de la persona que efectúa el lanzamiento. En tal caso, si salen 50 caras consecutivas, decidiríamos que por lo menos uno de nuestros supuestos era indudablemente erróneo, pero no estaríamos en condiciones de decidir cuál de ellos. En general, por supuesto, ponemos toda la atención necesaria en nuestros métodos de muestreo para tener una seguridad razonable de que los supuestos relativos a los mismos son ciertos.

Para ilustrar el mismo aspecto con un ejemplo de carácter sociológico, supongamos que se nos invita a formular sólo dos supuestos en una prueba estadística determinada, esto es: 1) que en la población seleccionada las proporciones de personas de las clases media e inferior con grandes deseos de cambio de situación son las mismas, y 2) que se ha obtenido una muestra aleatoria de todas las personas. Supóngase asimismo que dichos supuestos conducen a determinadas conclusiones que no concuerdan con los hechos. Tal vez, por ejemplo, los datos de la muestra señalan un porcentaje mucho más alto de personas de la clase media con grandes deseos de cambio. Concluimos, pues, que uno u otro de los dos supuestos es probablemente erróneo. Pero ¿cuál de ellos habremos de descartar? Nos gustaría sacar la conclusión de que el falso era el primero, pero tal vez nos hayamos servido de métodos de muestreo sujetos a alguna influencia ajena. En resumen, necesitamos datos adicionales, aparte de lo que indica la prueba misma.

En este ejemplo particular, si hemos tomado todas las precauciones para asegurar la selección de una muestra al azar, podemos tomar como modelo el supuesto 2) y formular la conclusión de que el supuesto falso era probablemente el 1). Aquí nuestra propensión a aceptar el supuesto 2) se basa en nuestro conocimiento acerca de los métodos de selección empleados, o sea de nuestra metodología. En otros casos, en cambio, podemos acep-

tar ciertos supuestos sobre la base de hallazgos de investigaciones anteriores. El punto importante, sin embargo, es que *la prueba misma no sirve para identificar el supuesto o los supuestos erróneos*. En este sentido, todos los supuestos poseen el mismo carácter o grado lógico. Para poner este hecho de manifiesto y para llamar la atención del lector sobre los supuestos del modelo, tratamos la hipótesis examinada como una sola, simplemente, entre cierto número de supuestos exigidos por la prueba.

Como ya se dijo anteriormente, el investigador tiene por lo regular interés en formular una hipótesis que en realidad le gustaría poder descartar. La hipótesis efectivamente examinada se designa a menudo como *hipótesis nula* (simbolizada por H_0), por contraste con la *hipótesis de investigación* (H_1), que se formula como alternativa de H_0 . Por lo regular, aunque no siempre, la hipótesis nula enuncia que no existe diferencia entre varios grupos o que no se da relación alguna entre variables, en tanto que la hipótesis de investigación puede anticipar una relación, ya sea positiva o negativa. El investigador puede esperar en realidad que la hipótesis negativa sea errónea y se deje descartar en favor de la alternativa H_1 . No obstante, con objeto de calcular una distribución de muestreo, ha de proceder como si H_0 fuera efectivamente correcta. Supondría, por ejemplo, que la moneda no está sesgada.

Obsérvese que el supuesto de una moneda correcta proporciona una manera de calcular probabilidades exactas sirviéndose de la fórmula binomial. En efecto, si se formulara la hipótesis de que la moneda es "sesgada", nos encontraríamos con que no podíamos obtener una distribución de selección hasta después de haber especificado la hipótesis de modo más preciso. Habríamos de remitirnos a un valor específico de p , digamos de .75, por ejemplo. Rara vez estaremos en condiciones de hacerlo. Y en forma análoga, la hipótesis de investigación de que hay entre la clase media una proporción mayor de personas con grandes deseos de cambio no es tan específica como la hipótesis nula en el sentido de que no hay en absoluto diferencia alguna entre las dos clases.

2. *Obtención de la distribución de muestreo.* Habiendo formulado los supuestos necesarios, estamos ahora en condiciones de servirnos del razonamiento matemático para obtener una distribución de muestreo a la que asociamos probabilidades con resultados. Semejante distribución de probabilidades nos dirá simplemente cuán probable sea cada uno de los resultados posibles, *si los supuestos adoptados son efectivamente correctos*. Si los supuestos anteriores a propósito de la moneda y de los lanzamientos fueran realmente correctos, ya vimos que a la larga sólo podríamos esperar obtener todas las caras una sola vez sobre

1 024, sólo 10 veces sobre 1 024 obtener nueve caras, etcétera. El conocimiento de la probabilidad de un resultado particular cualquiera, al producirse por azar si nuestros supuestos fueran efectivamente ciertos, nos permite ahora tomar una decisión racional a propósito de las condiciones en las que podríamos arriesgarnos a descartar los supuestos en cuestión. Supóngase, por ejemplo, que obteníamos 10 caras en 10 lanzamientos. Existen dos posibilidades: *a)* o bien los supuestos son correctos, y éste es uno de los casos en que se produce un acontecimiento muy raro, o bien *b)* uno por lo menos de los supuestos (probablemente la hipótesis nula) es falso. Por desgracia, no podemos saber nunca cuál de las dos alternativas sea la correcta. Si lo supiéramos, en efecto, habríamos sabido de antemano acerca de los supuestos, y ya no tendría objeto alguno efectuar el experimento. Pero podemos decir que la primera alternativa es muy improbable.

Establezcamos, pues, la regla de que cada vez que obtenemos 10 caras en 10 pruebas sucesivas podemos llegar a la conclusión de que uno por lo menos de los supuestos es falso y debería descartarse. A la larga podremos eventualmente cometer errores adhiriendo a dicha regla estricta, ya que sabemos que incluso con una moneda sesgada podemos esperar obtener 10 caras una vez sobre 1 024, simplemente por casualidad. Semejante regla no nos ayudará a determinar la corrección de nuestra decisión en relación con un experimento particular cualquiera, pero las leyes de las probabilidades nos dicen exactamente qué proporción de veces podemos esperar tomar decisiones correctas *a la larga*. En cierto sentido, nuestra fe se funda más en el procedimiento que seguimos que en la decisión que formulamos en cada ocasión particular. Y dicho *procedimiento* nos proporcionará decisiones correctas la mayoría de las veces, aunque no podamos estar absolutamente seguros de decidir correctamente en una ocasión concreta cualquiera.

3. *Selección de un nivel de significación y de una región crítica.* Desde un punto de vista ideal, las decisiones del investigador deberían tomarse con anterioridad al experimento o al análisis efectivos de los datos. A partir de su conocimiento de la distribución de muestreo, selecciona un grupo de alternativas las cuales, caso de producirse, le obligarían a descartar sus supuestos. Estos resultados improbables se designan como la *región crítica*. Así, pues, divide los resultados posibles en dos categorías: *a)* aquellos en cuyo caso descartará (la región crítica), y *b)* aquellos que, de producirse, no le permitirán descartar. Con objeto de poder establecer una región crítica, ha de decidir dos aspectos además de la elección de un modelo y una hipótesis. Primero ha de decidir los riesgos que está dispuesto a asumir de incurrir en los errores de tipos I y II. Y en segundo lugar ha de decidir si

desea o no que su región crítica incluya ambas colas de la distribución de muestreo.

Según se indicó en el capítulo VIII, se han de tomar en consideración dos tipos de errores posibles. El primero consiste en descartar un grupo de supuestos en realidad correctos. El tipo de error II, en cambio, implica el no descartar supuestos en realidad falsos. Sobre la base de la distribución de muestreo pueden determinarse exactamente las probabilidades de que se produzcan determinados resultados *si los supuestos son efectivamente correctos*. Si el investigador decide que descartará cada vez que se produzca un determinado número de resultados improbables (digamos cero caras o diez caras), entonces, si los supuestos son correctos, cometerá un error de tipo I cada vez que obtenga uno de los resultados en cuestión.

La probabilidad de cometer un error de ese tipo es igual a la suma de probabilidades de cada uno de los resultados en el interior de la región. Así, por ejemplo, si la región crítica consta de cero a diez caras, la probabilidad de error de tipo I será de $2/1\,024$, o sea .002. Si se escogiera una región crítica más amplia, el riesgo de dicho tipo de error sería mayor. Supóngase que se decidiera descartar los supuestos si se obtenían cero, una, nueve, o diez caras. En este caso la probabilidad del error del tipo I sería de $(1 + 1 + 10 + 10)/1\,024$, o sea .022. La probabilidad de cometer un error de tipo I se designa como *nivel de significancia* y puede ponerse a cualquier valor deseado.

Antes de examinar los criterios posibles para decidir acerca del nivel de significancia, hay que decir algo a propósito de los errores de tipo II. En vista de nuestro examen anterior de la falacia consistente en afirmar el consecuente, es manifiestamente incorrecto sacar la conclusión de que si determinados supuestos no se dejan descartar han de ser, en consecuencia, ciertos. En efecto, otro grupo de supuestos pudo haber llevado igualmente a una distribución de muestreo con la que se podría haber llegado a conclusiones similares. Por ejemplo: si la verdadera probabilidad de caras fuera de .51 en lugar de .50, entonces la distribución de muestreo sería casi idéntica a la que calculamos. Por consiguiente, probablemente se habría elegido la misma región, y la decisión en cuanto a descartar o no habría sido la misma. Y sin embargo, en rigor, la hipótesis de $p = .5$ sería falsa y debería en realidad descartarse. Y si no estuviéramos en condiciones de hacerlo, no estaríamos sin embargo dispuestos a considerarla como la única hipótesis correcta, ya que hay un número adicional de hipótesis que tampoco podría descartarse. Decidimos simplemente que "no debíamos descartar" nuestra hipótesis.

Incluso si en un plan conservador rechazamos aceptar una hipótesis, nos gustaría, con todo, estar en condiciones de elimi-

nar cuantas falsas hipótesis hubiera. En este sentido cometemos un error siempre que dejamos de descartar una hipótesis falsa. ¿Y qué puede decirse a propósito de la probabilidad de incurrir en un error de tipo II? Desgraciadamente no es tan fácil calcular los errores de tipo II como es el caso con los de tipo I. Necesitamos, pues, aplazar nuestro examen de los mismos hasta el capítulo XIV. Conviene, sin embargo, observar un hecho importante. Y es que, para cualquier prueba dada, las probabilidades de errores de los tipos I y II son inversamente proporcionales. En otros términos: *cuanto menor es el riesgo de un error de tipo I, tanto mayor es la probabilidad de uno de tipo II*. Esto puede verse en nuestro ejemplo de los lanzamientos. El lector ha de convencerse por sí mismo que, si se elige una región crítica pequeña (digamos cero caras y diez caras), tendrá menos probabilidades de descartar *cualquier* supuesto que si se sirviera de una región más comprensiva (digamos cero, una, nueve y diez caras). En el primer caso, si bien está menos expuesto a descartar supuestos ciertos, tiene también menos probabilidades de descartar los falsos. De ahí que corra mayor riesgo de cometer un error de tipo II.

*Es, pues, imposible reducir simultáneamente los riesgos de los dos tipos de errores, a menos que se vuelva a disponer el estudio en otra forma y se seleccionen casos adicionales o una prueba estadística distinta. En la práctica ponemos la probabilidad de error de tipo I a un nivel determinado (digamos .05) y escogemos la prueba estadística que más reduce el riesgo de error de tipo II. Al elegir entre pruebas alternativas, escogemos aquella que tiene un modelo apropiado y reduce mejor el riesgo de error de tipo II.*²

La decisión en cuanto al nivel de significación a escoger depende de los costos relativos resultantes de cometer uno u otro tipo de error y debería apreciarse de acuerdo con ello. En ocasiones hay que adoptar una decisión práctica conforme al resultado del experimento. Un fabricante puede decidir instalar un equipo costoso, un investigador puede decidir extraer otra muestra y repetir su estudio, o las autoridades sanitarias pueden decidir si deben o no intentar una vacunación en masa con un nuevo suero. En otros casos no se requiere decisión práctica alguna. Un sociólogo, por ejemplo, puede publicar simplemente los resultados de su estudio en un periódico, sin tener que soportar las consecuencias de error de un tipo o del otro.

En situaciones en las que hay que tomar decisiones de carácter práctico la elección de un nivel de significación resulta particularmente difícil. En el ejemplo del lanzamiento, supóngase que la decisión comportara el dejar de seguir jugando con una moneda de cuya perfección se sospecha. Si nuestro jugador hipo-

tético tuviera la perspectiva de una esposa regañona caso de volver a la casa con los bolsillos vacíos, hará bien en dejar el juego si existe siquiera la más leve duda a propósito de la moneda. En tal caso elegiría una región crítica amplia, ya que las consecuencias de un error de tipo II (esto es, de seguir jugando si la moneda estaba realmente sesgada) serían muy lamentables. Por otra parte, si corriera el riesgo de insultar a su jefe sosteniendo que la moneda estaba sesgada, desearía estar perfectamente seguro de ello antes de adoptar su decisión. En este último caso escogería una región crítica muy pequeña, reduciendo así al mínimo el riesgo de error de tipo I. Y en forma análoga, si el costo de la vacunación en masa fuera considerable o el suero eventualmente peligroso, se desearía estar absolutamente seguro antes de aplicarlo. Desearíase hacer muy difícil descartar la hipótesis nula de que el suero no produce efecto benéfico alguno.

Si no hay más decisión práctica a tomar que la de publicar o no los resultados de un estudio, debería seguirse otra regla general. En este caso, en efecto, *el investigador debería aplicarse a demostrar a sí mismo que está en error o a obtener resultados que en realidad no desea obtener*. Por lo regular, pero no siempre, se establece una hipótesis nula que en realidad se desea descartar. Y como quiera que al investigador le gustaría estar en condiciones de descartarla, debería hacer la obtención del resultado deseado muy difícil, sirviéndose para ello de una región crítica muy pequeña.

Hay ocasiones, en cambio —y llamamos la atención del lector al respecto—, en que no se desea realmente descartar la hipótesis nula. Así, por ejemplo, la hipótesis nula puede adoptar la forma de una predicción en el sentido de que no hay diferencias religiosas o de clase en cuanto a la tasa de fecundidad. Si se desea realmente establecer tales diferencias, debería escogerse una región crítica muy pequeña, haciendo el descarte de la hipótesis nula muy difícil. Pero supóngase que el investigador desea realmente demostrar que tales diferencias no existen. Tal vez trate de demostrar que algunas teorías en boga a propósito de diferencias en materia de fecundidad son incorrectas o inadecuadas. O puede esperar que dichas diferencias no existen, de modo que no tenga que investigar desde los puntos de vista de clase o de religión al referir las tasas de fecundidad a otras variables.

En los casos que se acaban de mencionar, el investigador se halla en cierto sentido del lado falso de la hipótesis y debería estar en consecuencia interesado ante todo en reducir el riesgo de error de tipo II. En otros términos: debería preocuparse ante todo de no retener la hipótesis nula de la ausencia de diferencias si en realidad es falsa. De ahí que al escoger una región crítica pequeña no siempre se peque de conservador, haciendo así difícil descartar una hipótesis nula que en realidad se desea

² Para el examen más a fondo de esta cuestión, véase la sec. XIV.1.

retener. Los niveles de significación comúnmente empleados en la investigación estadística son los de .05, .01, y .001. En vista de lo que se acaba de exponer, el lector se habrá percatado de que dichos niveles nada tienen de sagrado o absoluto. Aunque una persona fuera normalmente conservadora en el empleo de los niveles en cuestión, estaría de todos modos en terreno más firme, si realmente no deseaba descartar la hipótesis nula, adoptando tal vez los niveles .10, .20 o incluso .30, reduciendo así el riesgo de error de tipo II.

Procede hacer una advertencia en relación con la interpretación de los resultados de las pruebas de significado, ya que es posible obtenerlos más bien falsos, aun sirviéndose del nivel .001 y cuando se desea el descarte. Las pruebas de significado nos dicen cuán probable un grupo dado de resultado sería si ciertos supuestos fueran verdaderos. Hay varios factores que determinan la probabilidad de que estemos en condiciones de descartar dichos supuestos. El primero de ellos es el grado real de falta de propiedad de los supuestos. Si, por ejemplo, la verdadera probabilidad de caras es .9, es muy probable que estemos en condiciones de descartar la hipótesis de que p sea .5, porque podemos obtener efectivamente una proporción suficientemente grande de caras para terminar en la región crítica. Por otra parte, si la verdadera probabilidad es .53, es menos probable que obtengamos los resultados extremos necesarios para el descarte.

El número de casos es otro factor importante en la determinación de cuán extremos deban ser los resultados antes de que el descarte sea posible. Con sólo 10 lanzamientos o casos ya vimos que se requieren resultados muy extremos para poder descartar. Pero si N es grande, la proporción de éxitos sólo necesita hacer la hipótesis para diferir de p en muy poco para que estemos autorizados a hacerlo. Si la moneda se lanzara al aire 10 mil veces en lugar de 10, estaríamos en condiciones de descartar la hipótesis si obtuviéramos, por ejemplo, 5 200 caras o más. En otros términos: en el supuesto de que p es exactamente una mitad, o sea .5, 5 200 caras o más en 10 mil lanzamientos serían más improbables que 10 caras en 10 lanzamientos, pese a que los resultados no sean ni con mucho tan extremos. Esto concuerda, por supuesto, con nuestra mayor fe intuitiva en grandes muestras y con la inteligencia de que, en caso de muestras muy pequeñas, podrían ocurrir con frecuencia resultados extremos debidos al mero azar. Y en forma análoga, con una selección de 10 mil personas podríamos obtener diferencias muy pequeñas en las tasas de fecundidad entre las mujeres de las clases media e inferior, y estar, con todo, en condiciones de descartar la hipótesis nula en el sentido de que no hay diferencia alguna en la población.

Con un número muy grande de casos resulta casi siempre

posible descartar cualquier falsa hipótesis que pudiéramos formular, independientemente de cuánto pueda diferir el valor de nuestra hipótesis del verdadero. Esto significa que si tenemos 10 mil casos, no deberá sorprendernos mucho que estemos en condiciones de descartar al nivel .001, y deberemos guardarnos de comunicar nuestro hallazgo como si fuera muy importante. En efecto, no debe confundirse la significación estadística con la práctica. La significación estadística sólo puede decirnos que ciertas diferencias de muestras no ocurrirían con mucha frecuencia por azar si no hubiera diferencias cualesquiera en la población. Nada nos dice, en cambio, en cuanto a la magnitud o la importancia de dichas diferencias. Por lo tanto, un factor suficientemente grande para producir diferencias estadísticamente significativas en una muestra pequeña es más digno de atención que un factor que sólo produce pequeñas diferencias de las que únicamente puede demostrarse que son significativas estadísticamente con una muestra muy grande. Si el estudio comporta un gran número de casos, nos interesamos por lo regular por otras clases de problemas que las pruebas de significación. Esta cuestión se examinará más a fondo en el capítulo xv, cuando tratemos de medidas de grado de relación. Por el momento, baste señalar que la significación estadística no implica necesariamente diferencias impresionantes o de las que son importantes para el sociólogo.

Antes de poder determinar la región crítica hay que adoptar otra clase de decisión. Hay cierto número de resultados o de grupos de resultados cuya probabilidad puede ser menor que el nivel seleccionado de significación. Así, por ejemplo, la probabilidad de obtener exactamente ocho caras es de $45/1\,024$, o sea .044. Por lo tanto, sería posible, aunque no muy sagaz, decidir rechazar la hipótesis nula en caso de darse exactamente ocho caras, y no rechazarla en otro caso. La probabilidad de error de tipo I sería así de .044. La elección de semejante región crítica, sin embargo, apenas tendría sentido teóricamente, ya que por lo regular se vacilaría aún más en cuanto a aceptar la hipótesis nula si fueran a salir nueve o diez caras, pese a que estas alternativas no correspondieran a la región crítica. Casi siempre tenemos interés en servirnos por lo menos de una cola entera de la distribución. En efecto, no nos interesa la probabilidad de obtener exactamente ocho caras, sino la probabilidad de obtener ocho o más caras, esto es, la probabilidad de obtener ocho caras o algo incluso más insólito todavía.

Pero, ¿por qué no incluir en la región crítica cero, una y dos caras, ya que dichas alternativas son tan improbables como las de ocho, nueve y diez caras? A menudo no estamos en condiciones de prever en qué dirección puedan producirse los resultados insólitos. En nuestro ejemplo sólo podemos sospechar que la moneda esté sesgada, pero podemos no tener indicio alguno en

cuanto a si está influida en favor de las caras o de las cruces. Además, esto puede acaso no importarnos. En tal caso deseáramos estar seguros y servirnos de ambos extremos de la distribución de selección. Porque, si nos sirviéramos de una región crítica de sólo ocho, nueve y diez caras, entonces, en caso de

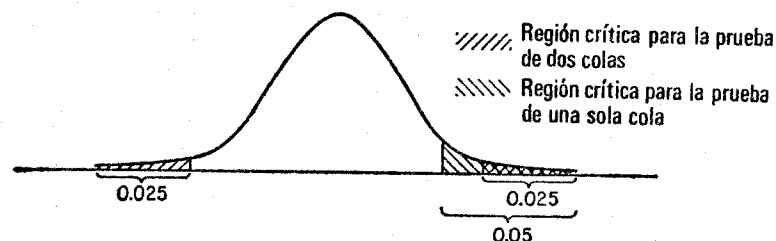


FIG. X.1. Comparación de las regiones críticas de pruebas de una sola cola y de dos colas, empleando el nivel de significación de .05

obtener exactamente una cara, nos encontraríamos en la situación desdichada de no poder descartar la hipótesis nula aun siendo falsa.

Sin embargo, hay cierto número de ocasiones en las que o estamos en condiciones de predecir la dirección de la desviación, o nos interesamos ante todo por las desviaciones en una sola dirección. Así, por ejemplo, una información previa puede habernos llevado a predecir que la moneda está alterada en favor de las caras. O podemos estar jugando cada vez a cruz, de modo que, si la moneda está afectada en favor de ésta, no necesitamos tener miedo alguno de seguir el juego. En estudios más reales, por otra parte, resulta a menudo posible prever la dirección sobre la base de la teoría o de estudios previos. Pudo haberse predicho, por ejemplo, que los católicos tendrían familias más numerosas que los protestantes. Si se está interesado en demostrar que la teoría de uno es correcta, sólo se harán pruebas de significación cuando los resultados se producen en la dirección prevista. Si se producen en sentido contrario, en cambio, no se necesita hacer prueba, ya que los datos no confirman de todos modos la teoría.

Siempre que se haya predicho la dirección, las pruebas de una sola cola serán preferibles, al mismo nivel de significación, a las de dos, ya que será posible obtener una cola mayor concentrando la región crítica entera en el extremo apropiado de la distribución de muestreo. Esta ventaja de la prueba de una sola cola se ilustra en la figura X.1 en relación con el caso de una distribución de muestreo, lisa, que ostenta la forma de una curva normal. En dicha figura, las probabilidades de cometer un error de

tipo I son las mismas en ambos casos, ya que las dos regiones críticas son del mismo tamaño (medidas en términos de áreas). Sin embargo, si los resultados ocurren efectivamente en la dirección prevista, el investigador tendrá más probabilidades de descartar la hipótesis sirviéndose de una prueba de una sola cola, ya que existe mayor probabilidad de caer en la región crítica en la dirección en cuestión. En efecto, si la verdadera probabilidad está en la dirección prevista, el riesgo de cometer un error de tipo II es menor que sirviéndose de una prueba de dos colas.

A estas alturas el lector no ha de esperar poder comprender intuitivamente las relaciones entre los errores de tipo II y las pruebas de una y dos colas. En efecto, muchas de estas nociones más bien difíciles sólo se irán aclarando una vez que se hayan examinado algunos ejemplos prácticos. El tratamiento más detallado de los errores de tipo II ha de aplazarse hasta el capítulo XIV.

Para concretar en el presente ejemplo, escojamos el nivel de .05 y sirvámonos de una prueba de dos colas. La región crítica constará en este caso de las alternativas cero, una, nueve y diez caras, ya que la inclusión de alternativas adicionales aumentaría la probabilidad de un error de tipo I más allá del nivel de .05. En nuestro ejemplo, el nivel de significación realmente empleado será el de $(1 + 1 + 10 + 10)/1024$, o sea .022. En otros casos en que la distribución de muestras es más bien continua que discreta, será posible servirse del nivel exacto deseado (v.gr., .05, .01, o .001).

4. *Cálculo de la estadística de la prueba.* Es siempre necesario calcular lo que se designa como estadística de la prueba, cuya distribución de selección ha de emplearse en la prueba. Hasta aquí sólo hemos tratado con estadísticas tales como las proporciones de las muestras, las medias y las desviaciones estándar, que son directamente comparables con las mismas cantidades en la población y pueden utilizarse como medidas para resumir los datos. La estadística de la prueba es una estadística que por lo regular no posee interés en sí misma por lo que se refiere a la descripción, pero que se emplea para verificar hipótesis. Es la estadística que contiene la distribución de muestreo la que se utiliza directamente en la prueba. En otros términos: calculamos de los datos de la muestra una cantidad que varía de modo conocido conforme a la teoría de las probabilidades. Comparamos luego su valor con la distribución de muestreo, y adoptamos una decisión evaluando la probabilidad de su ocurrencia. Por supuesto, son muchas las cantidades que pueden calcularse sobre la base de los datos de las muestras, pero sólo un pequeño número de ellas tiene distribuciones de muestreo conocidas que puedan utilizarse con fines de verificación de hipótesis.

En este ejemplo de la prueba binomial, la estadística de prue-

ba es tan sencilla, que casi no vale la pena de llamar la atención del lector al respecto. En efecto, es sencillamente el número de éxitos en N ensayos y no requiere cálculo ulterior alguno. En conexión con otros problemas, en cambio, la estadística de la prueba habrá de calcularse. En el caso de la prueba binomial, hemos dejado que r tomara todos los valores posibles, de cero a N , y hemos luego asociado probabilidades con cada valor. Supongamos que en este problema concreto, que comporta 10 lanzamientos, el número de éxitos (caras) resulta ser ocho. Poseemos ahora toda la información necesaria para adoptar una decisión.

5. *Adopción de una decisión.* Después de haber elegido su región crítica y de haber calculado su estadística de prueba, el investigador descartará o no descartará los supuestos según el resultado del experimento. Si el resultado queda dentro de la región crítica, descartará, con una probabilidad conocida de error de tipo I. Y si aquél no cae en la región crítica, no descartará los supuestos, asumiendo el riesgo de cometer un error de tipo II. En el presente ejemplo, como quiera que el resultado de ocho caras no cae en la región crítica, no descartará la hipótesis nula según la cual la moneda es insesgada.

Desde un punto de vista ideal, todas las decisiones anteriores a los pasos 4 y 5 deberían tomarse antes de proceder a la tabulación de los resultados. A menudo, en la labor exploratoria el investigador examinará primero sus datos y procederá luego a hacer pruebas de significación. Aunque esto sea a veces necesario, hay que observar, con todo, que siempre que ocurra así no se están respetando por completo las reglas del juego. En tales casos sería preferible no pretender que se está procediendo realmente a la verificación de hipótesis. De todos modos, los resultados podrían exponerse en forma sugestiva, y cualquiera que prosiga el estudio estará así en condiciones de efectuar pruebas estadísticas legítimas.

Los comentarios que preceden podrán parecer excesivamente rígidos y ambiciosos, dado el carácter de exploración de gran parte de la investigación sociológica. Sin embargo, el autor cree que es preferible fijar una "conciencia estadística" estricta, no dejando la impresión de que las cosas puedan hacerse de cualquier modo. En efecto, a menos que se adopten las decisiones con anterioridad al análisis de los datos, no puede hacerse legítimamente uso de la teoría de las probabilidades, ya que el análisis es esencialmente *ex post facto*. Y lo malo de los análisis *ex post facto* está en que el experimento puede disponerse de tal modo que el investigador no tenga manera de perder. Supóngase, por ejemplo, que haya decidido, a tientas, servirse del nivel de significado .05. Si encuentra que sus resultados son significativos al nivel de .07, puede decidir descartar sus hipótesis de todos

modos. Pero, supóngase que han sido significativos a los niveles .09 o .13 o .18, ¿dónde habrá que hacer alto? Otra manera de hacer trampa consiste en esperar hasta después del experimento para decidir si hay que servirse o no de una prueba de una cola. En tal caso, en efecto, si los resultados dan más caras que cruces, se decide simplemente que hubo de haberse utilizado una prueba de una cola, ya que subconscientemente se está anticipando una inclinación en favor de las caras. De este modo, cualquiera que sea la dirección de la desviación, se puede obtener una región crítica mayor que con una prueba de dos colas.

X.3. Aplicaciones de la binomial

La prueba del signo. Supóngase que un sociólogo se está sirviendo de un simple esquema "antes-y-después" o "sólo-después" del experimento, en el que se da un pequeño número de casos y en el que sólo está en condiciones de determinar para cada uno de ellos si su experimento ha dado o no resultados.³ Puede, por ejemplo, desear saber si la experiencia obtenida en un campo interracial es o no eficaz en cuanto a reducir los prejuicios fijos contra los negros. Somete sus investigados a una prueba de prejuicios, antes y después de la experiencia, y está en condiciones de apreciar si el tipo de prejuicio en cuestión ha disminuido o no. Indiquemos con un + ("éxito") los casos en los que el prejuicio se ha reducido, y con un - ("fracaso") aquellos en que el prejuicio ha aumentado. Si hay personas que no muestran el menor cambio, éstas quedarán excluidas del análisis. A menos que la medición haya sido muy burda, dichas personas serán relativamente pocas.⁴

La binomial requiere el supuesto de independencia de los experimentos. Por lo tanto, el sociólogo supondrá que su grupo experimental constituye una muestra aleatoria de la población a cuyo propósito se propone generalizar, y que entre los participantes no se ha producido influencia mutua alguna, o sólo poca, en cuanto a las marcas de prejuicio. Supongamos que lo que trata de establecer es que la experiencia en un campo es realmente eficaz en cuanto a reducir los prejuicios. Como quiera que esto no puede hacerse directamente, puede formular la hipótesis nula en el sentido de que la experiencia no surte efecto. Si efectivamente no produce efecto, entonces, si se sometiera a una experiencia semejante la población entera de la que se extrajo la muestra, habríamos de esperar encontrar el mismo número de

³ Para el examen de este y otros tipos de esquemas de experimento véase [6].

⁴ El problema de empate o no cambio resulta particularmente molesto en el caso de las variables ordinales, y será discutido en los caps. XIV y XVIII. Para una explicación más amplia véase Bradley [3], cap. III.

personas cuyo prejuicio se redujo y de aquellas cuyo prejuicio aumentó. En otros términos: deberíamos tener las mismas proporciones de signos más y signos menos.

Como quiera que cada miembro de la población tiene la misma posibilidad de figurar en una muestra aleatoria, la *probabilidad* de obtener un + en una extracción determinada cualquiera será de .5 bajo la hipótesis nula. Así, pues, un supuesto acerca de la *proporción* de signos + en la población permite, si se lo combina con el supuesto del azar, enunciar algo acerca de la *probabilidad* de éxitos en un ensayo determinado cualquiera. El azar asegura al propio tiempo la independencia de los experimentos. Permítasenos insistir una vez más en que *es necesario formular supuestos tanto acerca de la población como acerca del método de muestreo*. En el presente ejemplo, el interés se centra en la eficacia de la experiencia, esto es, en la proporción de éxitos entre la población. Por lo tanto, el sociólogo se asegurará de que se sirve de procedimientos correctos para obtener una muestra aleatoria.

Si en éste hay ocho personas, la distribución de muestreo de éxitos sería como sigue:

Nº de éxitos	Probabilidad
0	$1/256 = .004$
1	$8/256 = .031$
2	$28/256 = .109$
3	$56/256 = .219$
4	$70/256 = .274$
5	$56/256 = .219$
6	$28/256 = .109$
7	$8/256 = .031$
8	$1/256 = .004$
	1.000

Supongamos que el sociólogo quiere emplear un nivel de significación de .05. Como quiera que la dirección se ha anticipado, puede utilizarse una prueba de una sola cola. La región crítica puede determinarse acumulando probabilidades a partir de ocho éxitos, luego siete, etcétera, hasta que la suma resulte mayor que el nivel de significación. Por lo regular no será necesario obtener la distribución de muestreo entera, ya que en realidad sólo se emplean las colas para determinar la extensión de la región crítica. En el presente caso, la probabilidad de ocho éxitos es de .004; la probabilidad de siete u ocho éxitos es de .035, y la de seis, siete u ocho éxitos es de .144. Como quiera que la suma de las probabilidades de los resultados en el interior de la re-

gión crítica ha de ser menor que, o igual al nivel de significación seleccionado, vemos que la región crítica sólo puede constar de siete u ocho éxitos.

Supóngase que el sociólogo efectúa el experimento y encuentra que en seis casos el prejuicio ha disminuido, en tanto que ha aumentado en los otros dos. En consecuencia, no descartará la hipótesis de que el experimento no es eficaz, ya que la probabilidad de obtener dicho resultado, o inclusive uno más insólito, es mayor que .05.

La prueba en el caso de ausencia de azar. En el ejemplo anterior se supuso azar y el interés se centraba en la proporción de éxitos entre la población. En otros tipos de problemas se podrá tener información acerca de la proporción de personas que reúnen determinadas características en una población, pero puede existir una cuestión acerca de la selectividad. Así, por ejemplo, puede disponerse una prueba para ver si los profesionales están o no representados con exceso en los consejos, o si los negros no están bien representados en el jurado. Supóngase que un alcalde designa a nueve personas para una comisión, pretendiendo que son representativas, en el sentido de que todos los adultos tienen las mismas probabilidades de ser elegidos. Es sabido que el 35 por ciento de la clase laboral son oficinistas y, con todo, de los nueve miembros de la comisión seis son oficinistas; puede utilizarse una prueba binomial para determinar el grado de probabilidad de semejante distribución profesional en el supuesto de selección al azar. En este problema particular, la probabilidad de éxito por debajo de la hipótesis nula sería de .35, y la distribución de selección no sería simétrica. Consideraríamos cada una de las nueve posiciones de la comisión como *ensayo*. La probabilidad de obtener como primer comisionado a un oficinista sería de .35, y análogamente para cada una de las ocho posiciones restantes.

Otros empleos de la binomial. Además de los citados anteriormente, la binomial puede utilizarse en cierto número de otros tipos de problemas. En ocasiones pueden utilizarse medidas de posición, tales como la mediana o los cuartiles, para permitarnos verificar si una pequeña submuestra de personas es significativamente distinta de lo que esperaríamos en caso de azar. De una muestra grande es posible obtener una apreciación muy buena de la distribución de los ingresos en relación con una ciudad determinada. Si los datos sólo se han obtenido de seis armenios y si seis de estas personas se encuentran en el cuartil inferior, podemos efectuar una prueba para ver qué grado de probabilidad reviste esto, a condición, por supuesto, de adoptar las decisiones con anterioridad al experimento.⁵ Como quiera que por definición

⁵ Necesitamos tener un número muy grande de casos, con objeto de obtener una apreciación exacta de la medida de posición (v.gr., Q_1). En otro

un cuarto de la población se encontrará en el cuartil inferior, la distribución binomial proporciona la probabilidad de obtener una determinada proporción de la submuestra por debajo del cuartil de población, en el supuesto de que dicha submuestra constituya esencialmente una muestra al azar de la población mayor.

Así, por ejemplo, como quiera que la probabilidad de que cualquier persona determinada se encuentre en el cuartil inferior es de .25, la probabilidad de obtener exactamente seis armenios en el cuartil inferior sería:

$$P(6) = \left(\frac{7}{6}\right) \left(\frac{1}{4}\right)^6 \left(\frac{3}{4}\right)^1 = \frac{21}{16\,384}$$

O también

$$P(7) = \left(\frac{7}{7}\right) \left(\frac{1}{4}\right)^7 \left(\frac{3}{4}\right)^0 = \frac{1}{16\,384}$$

Como quiera que necesitamos obtener la probabilidad de conseguir seis o más éxitos, sumamos estas probabilidades y tenemos:

$$P(6) + P(7) = \frac{21 + 1}{16\,384} = .0013$$

Otro empleo de la binomial podría consistir en comprobar el carácter adecuado de una teoría que predijera correctamente la dirección de ciertas diferencias en, por ejemplo, 11 de 15 pruebas independientes. Para que dichas pruebas fueran independientes, deberían comportar muestras distintas. Así, por ejemplo, una de las muestras podría constar de jóvenes varones protestantes, otro de jóvenes muchachas protestantes, otro de varones católicos de más edad, etcétera. Cada una de las submuestras podría ser demasiado pequeña para proporcionar significación estadística separadamente, pero, si las submuestras se hubieran seleccionado independientemente, podría emplearse legítimamente una binomial para averiguar si un número suficiente de submuestras daba o no resultados en la dirección estimada. Cada submuestra constituiría en tal caso un ensayo, y la probabilidad de que en un ensayo particular cualquiera el resultado fuera en la dirección estimada sería de .5 con base en la hipótesis nula, en el sentido de que la teoría no tenía absolutamente ningún valor de estimación, esto es, en el sentido de que estima la dirección erróneamente con la misma frecuencia que lo

caso, en efecto, habrá un grado de error suficiente en dicha apreciación para requerir el empleo de una prueba de dos muestras. La razón de ello se verá claramente una vez que se hayan expuesto las pruebas de dos muestras en el capítulo XIII.

hace correctamente. Obsérvese que semejante prueba no podría emplearse si se tomaran 15 observaciones sobre la base de la misma muestra de personas.

* X.4. Extensiones del binomio

Son varios los posibles métodos para ampliar el planteamiento básico ejemplificado con el empleo de la distribución binomial. Aun cuando dichos métodos no son usados con frecuencia en las pruebas estadísticas de las ciencias sociales, debe al menos conocerse su existencia. La primera de ellas es la distribución *multinomial*, utilizable en situaciones en que se dan más de dos clases de eventos. Hemos visto ya que si hay k clases distintas de eventos, y si r_i es el número de eventos en la i -ésima clase, resultará que el número de permutas para dichos eventos vendrá dado por la expresión $N!/r_1!r_2!\dots r_k!$. Si los eventos son estadísticamente independientes y las probabilidades de obtener las distintas clases de eventos vienen dados por p_i , con $i = 1, 2, \dots, k$, y con $\sum_{i=1}^k p_i = 1$, en tal caso la probabilidad de obtener *exactamente* r_1 eventos del tipo 1, r_2 eventos del tipo 2, \dots y r_k eventos k en algún orden particular será:

$$\underbrace{(p_1 p_1 p_1 \dots)}_{r_1 \text{ términos}} \underbrace{(p_2 p_2 p_2 \dots)}_{r_2 \text{ términos}} \dots \underbrace{(p_k p_k p_k \dots)}_{r_k \text{ términos}} = p_1^{r_1} p_2^{r_2} \dots p_k^{r_k}$$

Si multiplicamos esta expresión por el número de permutaciones, obtendremos la fórmula

$$P(r_1, r_2, \dots, r_k) = \frac{N!}{r_1! r_2! \dots r_k!} p_1^{r_1} p_2^{r_2} \dots p_k^{r_k}$$

Es importante observar que esta fórmula nos da la probabilidad de obtener *exactamente* el número especificado de eventos de cada tipo. Supongamos por ejemplo que tenemos conocimiento de que una escuela contiene 50 por ciento de caucásicos, 30 por ciento de negros y 20 por ciento de orientales. ¿Cuál es la probabilidad de que el "primer equipo" de fútbol contenga exactamente 3 caucásicos, 7 negros y 1 oriental, bajo el supuesto de que la composición racial del equipo está sujeta a un proceso de selección puramente gobernado por el azar? Utilizando la distribución multinomial, tendremos:

$$P(3, 7, 1) = \frac{11!}{3!7!1!} (.5)^3 (.3)^7 (.2)^1 = .007$$

Encontramos inmediatamente una dificultad que crea complicaciones en el uso de la distribución multinomial en las pruebas estadísticas. No resulta obvio en muchos casos cómo puede especificarse sin ambigüedad un grupo de soluciones que sean más "infrecuentes" que la ya obtenida. En este ejemplo se dan varias clases de combinaciones "insólitas". Veamos las siguientes: el equipo puede no contar con negro alguno, o no incluir orientales, pero ¿qué resultado cae en la región crítica? Si es posible especificarlo, podrá idearse una prueba correcta. Si reunimos por ejemplo a los caucásicos con los orientales, podríamos interesarnos por la probabilidad de obtener siete o más negros en el equipo. Pero en este caso, como en otros muchos, estaríamos utilizando la distribución binomial y no la multinomial. Resulta posible un segundo tipo de modificación de la binomial cuando se ha estado muestreando *sin* reposición una población relativamente pequeña. Si una población de tamaño M contiene M_1 elementos del tipo 1, M_2 elementos del tipo 2, y, en general, M_i elementos del tipo i , y si los tamaños correspondientes de la muestra son N y N_i , la probabilidad de obtener exactamente N_1, N_2, \dots, N_k casos de cada tipo viene dada por lo que se denomina *distribución hipergeométrica*, a saber:

$$P(N_1, N_2, \dots, N_k) = \left(\frac{M_1}{N_1} \right) \left(\frac{M_2}{N_2} \right) \dots \left(\frac{M_k}{N_k} \right) / \left(\frac{M}{N} \right)$$

Si deseamos por ejemplo determinar la probabilidad de obtener exactamente seis espadas, seis tréboles y un diamante en una mano de bridge de tres cartas (tomadas al azar, pero sin reposición), tendríamos:

$$P(6, 6, 1) = \left(\frac{13}{6} \right) \left(\frac{13}{6} \right) \left(\frac{13}{1} \right) / \left(\frac{52}{13} \right)$$

lo que resulta un número sumamente pequeño. Tropezaríamos de nuevo con la misma dificultad al especificar las alternativas que pueden ser consideradas "más raras" que la anterior particular combinación. En el capítulo xv haremos una prueba exacta de Fisher para 2×2 tablas, basada en la distribución hipergeométrica, en la que se incluyen sólo dos tipos de eventos.

Se observará por último que la distribución binomial puede ser aproximada por otras distribuciones cuando la muestra total sea tan grande que haga que los cálculos resulten aburridos. Cuando N es grande y p tiene un valor intermedio, con el producto $Np > 5$, cabe acercarse al binomio mediante una distribución normal, en cuyo caso podremos utilizar pruebas basadas en

proporciones de éxitos. Estas pruebas serán presentadas en los capítulos XI y XIII.

Se da a veces el caso de que el tamaño de la muestra sea moderadamente grande, en tanto que p es muy pequeño (o sumamente grande). Por ejemplo: p (o q) puede referirse a un acontecimiento poco usual, tal como el contraer una enfermedad o suicidarse. Si planteamos el problema de modo que p se refiera a la probabilidad de aquel raro evento, de modo que $p < q$, y si $Np < 5$, podrá calcularse aproximadamente el binomio, haciendo uso de la *distribución de Poisson*, por medio de la siguiente fórmula:

$$P(r) = \frac{\lambda^r e^{-\lambda}}{r!}$$

en la que r se refiere al número de éxitos en N intentos: $\lambda = Np$, y e es la constante natural, aproximadamente igual a 2.718.

Hay tablas para hallar los valores de $r!$ y de $e^{-\lambda}$ (véase Spiegel [8]), mediante cuyo empleo se reduce el trabajo del cálculo.

Para ilustrar el empleo de la aproximación de Poisson, supongamos que la probabilidad de ser arrestado en una localidad determinada es de .06, pero que en un muestreo de 50 japoneses-norteamericanos adultos sólo uno de éstos ha sido arrestado. En tal caso $Np = 3.0$ y

$$P(1) = \frac{3^1 e^{-3}}{1!} = 3e^{-3}$$

De manera análoga

$$P(0) = \frac{3^0 e^{-3}}{0!} = e^{-3}$$

en la que convencionalmente definimos $0!$ como la unidad. Al objeto de obtener la probabilidad de que uno o menos de los japoneses-norteamericanos sea detenido, sumaremos $P(1)$ y $P(0)$, obteniendo

$$P(1) + P(0) = 4e^{-3} = 4(.0498) = .199$$

X.5. Sumario

Este capítulo contiene una cantidad considerable de ideas, nuevas y fundamentales, además de examinar el mecanismo de la propia distribución binomial. En el capítulo siguiente habremos de discutir de nuevo, con cierta amplitud, muchas de estas ideas complementarias, tanto bajo la forma de hipótesis acerca de las medias, como en función de otras dos distribuciones de muestras.

Podrán observarse las importantes semejanzas que se dan en los pasos orientados a probar las hipótesis, y en los conceptos generales que han sido presentados en este capítulo. Revisemos éstos de nuevo, brevemente.

Resulta en primer lugar necesario formular algunos supuestos, tanto acerca de la población que va a ser estudiada como en relación con el método de muestreo de dicha población. Utilizando tales supuestos y la teoría de las probabilidades, podremos hacer afirmaciones específicas acerca de los resultados, con referencia a la hipótesis nula. En el caso del binomio, por ejemplo, aquellos supuestos hacen posible asignar un valor numérico específico (por ejemplo, $p = .5$) a las probabilidades de éxito de una prueba determinada. Para tomar decisiones en cuanto a la región crítica (es decir: el grupo de soluciones para las cuales rechazaremos H_0), necesitamos obtener lo que se denomina una distribución del muestreo, es decir: una probabilidad de distribución que asigna una probabilidad numérica específica a cada solución o a cada grupo de soluciones.

Decidiremos a continuación lo relativo al nivel de significación, que es la probabilidad de rechazar la hipótesis nula cuando ésta es en realidad verdadera (un error tipo I). Idealmente debe hacerse esta decisión evaluando los costos de un error tipo I por comparación con los de un error tipo II, no rechazando H_0 cuando en realidad es falsa. Al decidir además hacer uso de una prueba de una o de dos colas, queda determinada nuestra región crítica. Este grupo de resultados a excluir se encuentra acumulando las probabilidades, comenzando con los resultados más extremos y moviéndose hacia el centro, hasta que la suma resultante de probabilidades sea ligeramente menor que el nivel de significación (por ejemplo, .05). Vemos entonces los datos, computamos la estadística de prueba (por ejemplo, el número de éxitos), y tomamos nuestra decisión. Si el resultado cae dentro de la región crítica nos veremos obligados a rechazar H_0 , sabiendo que cometeríamos un error tipo I con una probabilidad igual a la del nivel de significación seleccionado. Si el resultado no cae dentro de la región crítica, no rechazamos la hipótesis, corriendo en este caso el riesgo de cometer un error tipo II. Aun cuando es difícil (como veremos en el capítulo XIV) determinar con exactitud la probabilidad de cometer un error tipo II, pues ello depende de hasta qué punto sea falsa nuestra hipótesis nula, sabemos que, para una muestra de tamaño fijo, cuanto menor hagamos el riesgo de cometer un error tipo I, tanto mayor será el de cometer uno del tipo II.

GLOSARIO

Distribución binomial
Región crítica

Distribución hipergeométrica
Modelo *versus* hipótesis
Distribución multinomial
Pruebas de una y de dos colas
Distribución de Poisson
Distribución de muestras
Nivel de significación

EJERCICIOS

1. En 11 lanzamientos de una moneda insesgada, ¿cuál es la probabilidad de obtener exactamente cuatro caras? ¿Exactamente siete caras? ¿Menos de tres caras? Respuesta, $P(4) = 330/2.048$.
2. Supóngase que la moneda del ejercicio anterior está sesgada y que la probabilidad de obtener cara es en realidad .6. Sin efectuar los cálculos, indíquese de qué modo esto afectaría cada una de las probabilidades anteriores (esto es, si las aumentaría, las reduciría o las dejaría inalteradas). Respuesta, menor de $P(4)$.
3. Supóngase que se quiere verificar la hipótesis nula, en el sentido de que la moneda es insesgada, echando 11 lanzamientos. Indíquese la región crítica que se utilizaría:
 - a) para una prueba de dos colas al nivel de .05. Respuesta: 0, 1, 10 u 11 caras
 - b) para una prueba de dos colas al nivel de .10
 - c) para una prueba de dos colas al nivel de .01
 - d) para una prueba de una sola cola al nivel de .05, anticipando que $P(\text{cara}) > .5$. Respuesta: 9, 10 u 11 caras.
 - e) para una prueba de una sola cola al nivel de .10, anticipando que $P(\text{cara}) < .5$.
4. En una localidad determinada, el 10 por ciento de la población es judía. Un estudio de los consejos de directores de diversas agencias de servicios indica que de un total de siete presidentes de los consejos cuatro son judíos. ¿Qué probabilidad existe de que esto pueda deberse al azar? En éste y los demás ejercicios que comportan verificación de hipótesis, indíquense los razonamientos y enumérense los supuestos adoptados. Respuesta, $P = .0027$.
5. Un psicólogo toma 12 grupos que dispone por pares según la estatura. Tiene así seis pares de grupos, en los que cada par de un grupo constituye un grupo experimental y el otro el grupo de control. El experimento comporta un intento de aumentar la cohesión de los grupos, y el experimentador está en condiciones de apreciar si el grupo experimental es o no más coherente que el grupo de control con el que ha sido apareado. ¿Cómo puede servirse de la binomial para verificar la hipótesis nula en el sentido de que el experimento es ineficaz? En este problema han de indicarse todos los supuestos requeridos, calcularse la distribución de muestreo y proceder a la elección de una región crítica.
- *6. Supóngase que se está estudiando un pequeño grupo de 12 personas y se desea verificar la hipótesis de que cuanto mayor es el grado de conformación a las normas del grupo tanto más elevada

es la posición de la persona en el grupo. En relación con ambas variables (conformación y posición) sólo se está en condiciones de apreciar si el individuo está por encima o por debajo de la mediana. ¿Cómo se utilizaría la binomial para verificar la hipótesis nula de que no existe relación alguna entre dichas variables? No se deje de indicar el razonamiento.

* 7. Supóngase que sabemos que la probabilidad de que se cometa suicidio entre un grupo de cierta edad es .003. Se ha descubierto que en una muestra seleccionada al azar, de 1 200 indios navajos del mismo grupo de edad, no ha habido suicidios. ¿Cuán probable es que esto haya sucedido por pura casualidad?

BIBLIOGRAFÍA

1. Alder, H. L., y E. B. Roessler: *Introduction to Probability and Statistics*, 4ª ed., W. H. Freeman and Company, San Francisco, 1968, cap. 6.
2. Anderson, T. R., y M. Zelditch: *A Basic Course in Statistics*, 2ª ed., Holt, Rinehart and Winston, Inc., Nueva York, 1968, cap. 11.
3. Bradley, J. V.: *Distribution-free Statistical Tests*, Prentice-Hall, Inc., Englewood Cliffs, N. J., 1968, caps. 3 y 7.
4. Hays, W. L.: *Statistics*, Holt, Rinehart and Winston, Inc., Nueva York, 1963, cap. 5.
5. Pierce, Albert: *Fundamentals of Nonparametric Statistics*, Dickenson Publishing Company, Inc., Belmont, Cal., 1970, caps. 9 y 12.
6. Sellitz, C., M. Jahoda, M. Deutsch, y S. W. Cook: *Research Methods in Social Relations*, Henry Holt and Company, Inc., Nueva York, 1959, cap. 4.
7. Siegel, Sidney: *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill Book Company, Nueva York, 1956, pp. 36-42.
8. Spiegel, M. R.: *Theory and Problems of Statistics*, Schaum's Outline Series, McGraw-Hill Book Company, Nueva York, 1961, cap. 7.
9. *Tables of the Binomial Probability Distribution*, National Bureau of Standards, Applied Mathematics Series, Núm. 6, 1950.

XI. PRUEBAS DE MUESTRAS SIMPLES QUE IMPLICAN MEDIAS Y PROPORCIONES

EN ESTE capítulo nos ocuparemos de verificación de hipótesis acerca de las medias y las proporciones de poblaciones. La media o la proporción de una muestra obtenida de una sola de éstas se comparará con el parámetro de la hipótesis y se decidirá si ésta debe o no descartarse. El lector no tardará en descubrir que las pruebas de la forma examinada en este capítulo tienen mucho menos utilidad práctica que las que comportan varias muestras. A estas alturas, sin embargo, importa más obtener una buena comprensión de las ideas fundamentales que preocuparse excesivamente por las aplicaciones prácticas. Por desgracia, las pruebas más sencillas no siempre son las más útiles.

Se recordará que las pruebas estadísticas que incluyen la binomial se servían de la regla de la multiplicación para obtener una distribución de muestreo. Pudimos ver en esta forma exactamente de qué modo se utilizaba la teoría de las probabilidades para conseguir dicha distribución. De aquí en adelante, las consideraciones matemáticas se van haciendo cada vez más complicadas, a tal punto que, pese al hecho de que sería deseable comprender lo que hay detrás de cada argumento, el lector habrá de empezar a aceptar cada vez más enunciados con la garantía de la palabra. Sin duda, hay pruebas matemáticas disponibles, pero la mayoría de ellas necesita del cálculo o incluso de una preparación matemática considerablemente mayor.

XI.1. Distribución en muestreo de las medias

Un teorema relativamente notable, se basa en los mismos principios y reglas de probabilidades que la binomial, pero no se deja comprobar en un texto como éste. Dicho teorema puede enunciarse así: *Si de una población normal de magnitud N con una media de μ y una variancia de σ^2 se extraen reiteradas muestras al azar, la distribución de selección de las medias de las muestras será normal, con la media μ y la variancia σ^2/N .* Examinaremos cuidadosamente lo que dice el teorema del límite central.

Partimos primero de una población normal, a sabiendas, por supuesto, de que en la vida real semejante población perfectamente normal no existe. Nos imaginamos luego a nosotros mismos extrayendo de la población en cuestión un número muy grande de medias de magnitud N al azar.¹ Para cada una de dichas muestras obtenemos una media \bar{X} . Por supuesto, estas me-

¹ No se confunda el número de muestras (que es infinito) con el tamaño de cada una de ellas (N).

dias de las muestras variarán algo de una a otra de ellas, pero esperamos, con todo, que se amontonarán alrededor de la verdadera media μ de la población. Esto es lo que nos dice el teorema del límite central. Dice que si dibujamos una gráfica de la distribución de dichas muestras, el resultado será una curva normal. Por otra parte, la desviación estándar de esa distribución

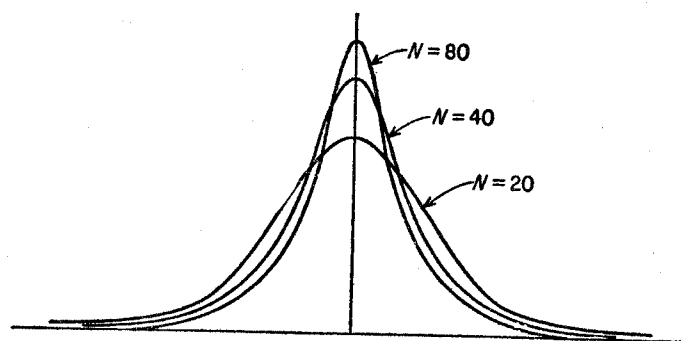


FIG. XI.1. Comparación de las distribuciones normales de muestreo para muestras de tamaño diferente

normal de las medias de las muestras será de σ/\sqrt{N} . Por consiguiente, cuanto mayor sea el tamaño escogido de muestra, tanto menor será la desviación estándar en la distribución de las muestras, esto es, tanto mayor será el agolpamiento de las medias de éstas (véase figura XI.1). Si consideramos las medias de las muestras como apreciaciones de la media de la población, podemos decir que hay cierta cantidad de error en nuestro proceso de estimación, debido a fluctuaciones del muestreo. Por consiguiente, designamos la desviación estándar de una distribución de muestreo como *error estándar*. En este caso, el error estándar de la media, indicado simbólicamente como $\sigma_{\bar{x}}$, es σ/\sqrt{N} .

El lector ha de tener claramente presente que se hallan implicadas tres distribuciones distintas, dos de las cuales acontecen ser exactamente normales. En efecto, *primero* tenemos la población, de la que se presume que es normal, con una media de μ y una variancia de σ^2 [escrita en adelante, para abreviar, como $Nor(\mu, \sigma^2)$]. En *segundo lugar*, tenemos una distribución de datos en el interior de cada muestra. Si N es grande, esta distribución será probablemente con toda razón representativa de la población y puede ser, por consiguiente, aproximadamente normal. Obsérvese que ésta es la única distribución que se obtiene en forma efectivamente empírica.² Y en *tercer lugar*, tenemos la distri-

² Como quiera que ésta es la distribución que el investigador ve efectiva-

bución de selección de una estadística (en este caso, la media). Acabamos de ver que la distribución de muestra de la media será asimismo normal, pero tendrá una desviación estándar menor que la población (a menos que el tamaño de la muestra N sea uno).

La relación entre la población y la distribución de muestreo puede verse en diagrama en la figura XI.2. Cuanto mayor sea la

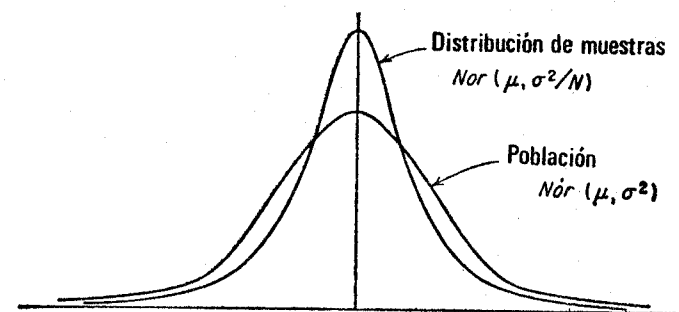


FIG. XI.2. Comparación entre las distribuciones de la población y de la muestra

magnitud N de la muestra, más puntiaguda será la distribución de selección, como puede verse en la figura XI.1. Hay que tener claramente presente que, pese a que sus desviaciones estándar se relacionan directamente, constituyen, con todo, distribuciones completamente distintas. Todos los "casos" de la distribución de muestreo son *medias* de muestras distintas. Como era cierto en el caso de la binomial y como lo será en todas las demás pruebas estadísticas, es más bien la distribución de muestreo y no la población original la que se utiliza directamente en las pruebas de significación. Los supuestos a propósito de la población pueden aparecer en el modelo. Mediante la teoría de las probabilidades los enunciados acerca de la población y de los métodos de muestreo se traducen en enunciados acerca de la distribución de muestreo.

En resumen, las medias y las desviaciones estándar de las tres clases de distribución son como sigue:

	Media	Desviación estándar
Población	μ	σ
Muestra	\bar{X}	s
Distribución de muestreo	μ	σ/\sqrt{N}

mente, puede haber propensión a confundir esta clase de distribución con la de muestreo.

El teorema del límite central concuerda con la intuición del sentido común en que, suponiendo que se hayan evitado distorsiones, puede tenerse más confianza en la apreciación de la media de una muestra grande que de una pequeña.³ Dice, en efecto, que las medias de las muestras variarán menos de una muestra a otra si N es grande. Pero constituye, con todo, un refinamiento considerable con respecto al sentido común, en cuanto proporciona una indicación de cuánta más confianza deberíamos tener si N es aumentado en una cantidad determinada. Así, por ejemplo, podemos ver que para partir el error estándar por la mitad necesitamos hacer N cuatro veces mayor. Nos dice asimismo que cuanto más homogénea es la población, para empezar, esto es, cuanto más pequeño es el valor de σ , tanto menor es el error estándar σ/\sqrt{N} y tanto mayor el agrupamiento de las medias de las muestras alrededor de la media de la población.

* Puede ofrecerse una justificación teórica de este importante teorema introduciendo la idea de las combinaciones lineales, de la que haremos uso más adelante en varias ocasiones. Una media es en realidad una función lineal simple de las puntuaciones X_i ,

ya que $\bar{X} = \frac{1}{N} (X_1 + X_2 + \dots + X_N)$. Puede mostrarse, en forma más generalizada, que si tenemos una variable Y que es una combinación lineal cualquiera de las X_i , y si esta X_i ha sido seleccionada independientemente, como ocurre cuando sacamos al azar una simple muestra, podremos obtener expresiones simples de la media (valor esperado) de Y , y para la variancia de Y . Específicamente, si

$$Y = c_1 X_1 + c_2 X_2 + c_3 X_3 + \dots + c_N X_N$$

y si las X_i son seleccionadas independientemente, entonces

$$E(Y) = c_1 E(X_1) + c_2 E(X_2) + \dots + c_N E(X_N)$$

y

$$\text{Var } Y = \sigma_Y^2 = c_1^2 \sigma_{x_1}^2 + c_2^2 \sigma_{x_2}^2 + \dots + c_N^2 \sigma_{x_N}^2$$

* En el caso de las muestras al azar, el valor esperado de cada X_i es μ . Si situamos cada $c_i = 1/N$, entonces Y pasa a convertirse en la media de la muestra, y tendremos:

³ Obsérvese que tenemos más confianza en *apreciaciones* basadas en grandes muestras; sin embargo, al descartar una hipótesis al nivel de .05, asumimos el mismo riesgo de un error de tipo I, independientemente de la magnitud de N . Como veremos dentro de poco, la extensión de la región crítica utilizada en la prueba toma en consideración la magnitud de la muestra, lo que explica la incongruencia aparente.

$$E(\bar{X}) = E(Y) = \left(\frac{1}{N} \right) [\mu + \mu + \dots + \mu] = \frac{1}{N} (N\mu) = \mu$$

y

$$\begin{aligned} \sigma_Y^2 = \sigma_X^2 &= \frac{1}{N^2} \sigma_{x_1}^2 + \frac{1}{N^2} \sigma_{x_2}^2 + \dots + \frac{1}{N^2} \sigma_{x_N}^2 \\ &= \frac{1}{N^2} [\sigma^2 + \sigma^2 + \dots + \sigma^2] \\ &= \frac{1}{N^2} (N\sigma^2) = \frac{\sigma^2}{N} \end{aligned}$$

La última fórmula es consecuencia del hecho de que la variancia de cada X_i es justo σ^2 , ya que estamos tratando en casos individuales seleccionados con igual probabilidad de entre una población con variancia σ^2 . Vista intuitivamente, la idea es que si repetimos un experimento consistente en sacar el "primer" caso un gran número de veces, la distribución de estos primeros casos será aproximadamente $N(\mu, \sigma^2)$. Lo mismo ocurriría con una sacada repetida de segundos casos, etcétera.

El teorema del límite central. Estamos ahora en la posibilidad de formular un teorema más general, conocido con el nombre de teorema del límite central, como sigue: *Si se extraen diversas muestras de magnitud N al azar de una población cualquiera (de la forma que sea) con una media de μ y una variancia de σ^2 , entonces, a medida que N crece, la distribución de muestreo de las medias de las muestras se aproxima a la normalidad, con la media μ y la variancia σ^2/N .*

Este teorema es más notable todavía que el anterior. Dice, en efecto, que por muy notable que sea la distribución de la que partimos, a condición que N sea lo bastante grande, podemos contar con una distribución de muestreo aproximadamente normal. Como quiera que es la distribución de muestreo, y no la población, aquella de que nos servimos en las pruebas de significación, esto significa que, siempre que N es grande, podemos abandonar por completo el supuesto acerca de la normalidad de la población y seguir sirviéndonos en nuestras pruebas, con todo, de la curva normal.

El lector ha de tratar de convencerse de que la ley de los grandes números tiene sentido empírico. La mejor manera de obtener una buena comprensión de lo que el teorema del límite central significa, y de convencerse al propio tiempo uno mismo de que el error estándar es realmente σ/\sqrt{N} , consiste en extraer un número de muestras de una población de media y desviación estándar conocidas, calcular las medias, hallar la desviación es-

tándar de las muestras y comparar el resultado obtenido con σ/\sqrt{N} .⁴ ¿Por qué debería la distribución de muestreo hacerse normal si la distribución original no lo es? Echemos una ojeada

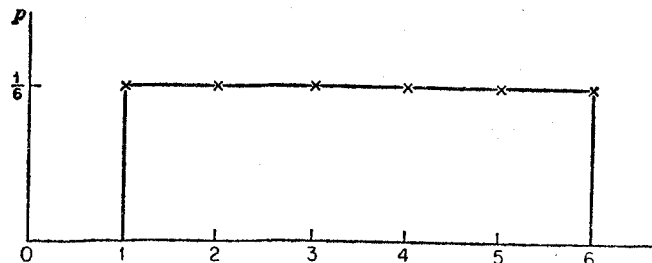


FIG. XI.3. Distribución de la población de las probabilidades de obtener caras de 1, 2, 3, 4, 5 o 6 con un dado perfecto

a una población que diste de ser normal y veamos qué ocurre a medida que vamos tomando muestras mayores.

Imagínese que estamos echando algún dado matemático ideal, con el cual las probabilidades de obtener cada una de las seis

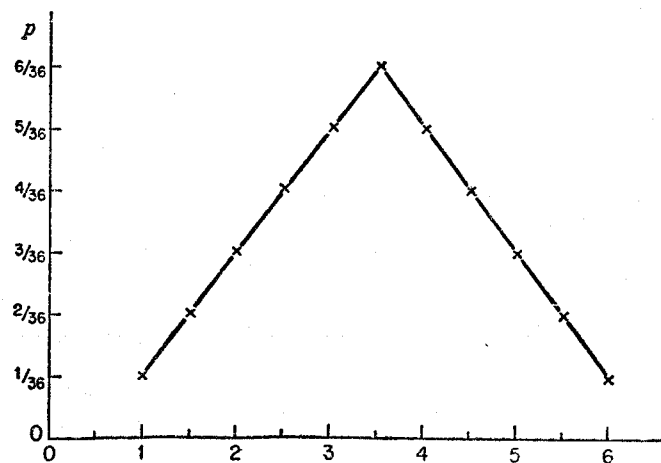


FIG. XI.4. Distribución de muestreo de las medias de las caras, con dados perfectos y muestras de tamaño 2

caras son exactamente $1/6$. La distribución de probabilidad para la jugada de un solo dado es en este caso rectangular, es decir: todos los números (de 1 a 6) tienen la misma posibilidad de ocurrir. Este tipo de distribución forma un contraste pronunciado

⁴ Véase el ejercicio 1 al final del capítulo.

con la distribución normal, en la que los valores extremos son menos probables que los que quedan más próximos a la media. Semejante distribución rectangular puede representarse como en la figura XI.3. En rigor, por supuesto, la distribución sería discreta, y no continua como parece indicarlo el diagrama.

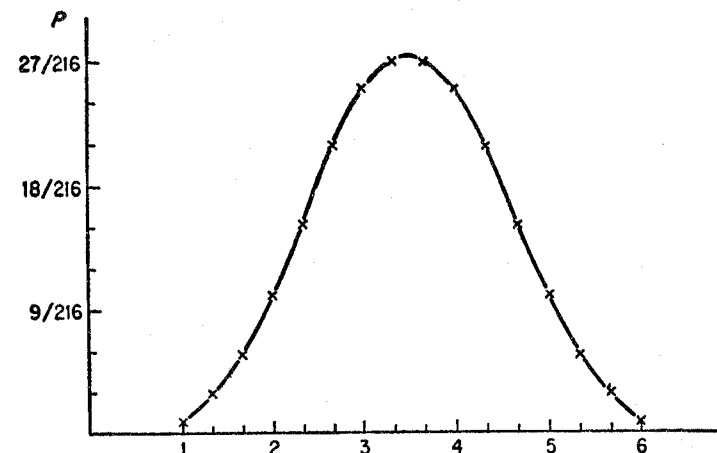


FIG. XI.5. Distribución de muestreo de las medias de las caras, con dados perfectos y muestras de tamaño 3

Si consideramos una distribución semejante como población de todas las jugadas posibles del dado, calculemos la distribución de muestreo de las medias de muestras de tamaño 2. Esto significa que hemos de echar dos dados, sumar los valores de las caras y dividir entre 2. Como bien lo saben los jugadores experimentados del "crap", esas sumas van de 2 a 12, siendo 7 el valor más probable. Al tratar de obtener las probabilidades de ocurrencia de cada una de dichas sumas, observemos primero que hay (6) (6) o sean 36 resultados posibles si los dados son distintos. Así, pues, el primer dado puede caer con cada una de las caras hacia arriba, y lo propio puede hacer el segundo. Para obtener la probabilidad de conseguir una suma de marcas de 7 y, por lo tanto, una media de 3.5, sólo necesitamos contar el número de maneras en que dicho resultado puede producirse. Obviamente, hay seis pares que pueden dar una marca de 7, a saber: (1,6), (2,5), (3,4), (4,3), (5,2) y (6,1). Una suma de 6 sólo puede obtenerse de cinco maneras distintas: (1,5), (2,4), (3,3), (4,2) y (5,1). Y en forma análoga, sólo hay una manera de obtener una suma de 12 (6,6) o una suma de 2 (1,1). Por consiguiente, la distribución de la probabilidad de las medias puede representarse como sigue:

Media	Probabilidad	Media	Probabilidad
1.0	$1/36$	4.0	$5/36$
1.5	$2/36$	4.5	$4/36$
2.0	$3/36$	5.0	$3/36$
2.5	$4/36$	5.5	$2/36$
3.0	$5/36$	6.0	$1/36$
3.5	$6/36$		$36/36$

Si se la representa en forma de gráfica, esta distribución asume la figura de un triángulo (figura XI.4).

Si se juega con tres dados, se suman las caras y se obtienen las medias, la distribución de muestreo será como sigue:

Media	Probabilidad	Media	Probabilidad
1.00	$1/216$	3.67	$27/216$
1.33	$3/216$	4.00	$25/216$
1.67	$6/216$	4.33	$21/216$
2.00	$10/216$	4.67	$15/216$
2.33	$15/216$	5.00	$10/216$
2.67	$21/216$	5.33	$6/216$
3.00	$25/216$	5.67	$3/216$
3.33	$27/216$	6.00	$1/216$
			$216/216$

Esta distribución, como puede apreciarse en la figura XI.5 empieza ya a acercarse a la forma de la curva normal, pese a que el tamaño de la muestra no sea más que 3. Después de un examen atento de las cifras anteriores, el lector estará en condiciones de comprender intuitivamente lo que ocurre y por qué se va obteniendo una curva en forma cada vez más pronunciada de campana a medida que la magnitud de N aumenta. En efecto, si bien con una sola jugada es tan probable sacar un 6 como un 3 o un 4, y de hecho dos 6 son tan probables como dos 3, sólo hay, con todo, una manera de obtener dos 6, en tanto que hay cierto número de maneras distintas de obtener un promedio de 3.0 en dos o más jugadas. En lenguaje común decimos que los grandes números tienden a ser compensados por pequeños, sobre todo si N es grande.

XI.2. Prueba para la media de la población, conociendo σ

Veamos ahora cómo el teorema del límite central y la ley de los grandes números pueden aplicarse en las pruebas estadísticas. Para empezar, vamos a tomar el modelo más simple posible con

finde de ilustración. Como quiera que algunos de los supuestos requeridos en este modelo no son prácticos, se abandonarán más adelante. Se tratará una vez más con cierto detalle cada uno de los cinco pasos examinados en el capítulo X, con objeto de que el lector se vaya familiarizando con el proceso de desarrollar pruebas estadísticas.

Problema. Supóngase que un investigador está interesado en verificar la propiedad de los procedimientos de muestreo empleados en un estudio local, realizado por entrevistadores sin experiencia. El investigador en cuestión sospecha que los ingresos correspondientes de las familias de las clases media y superior puedan haber sido subestimados, esto es, que hayan tenido mayor probabilidad de aparecer en la muestra que los de las familias de ingresos más bajos. Se dispone de datos del censo que muestran que el ingreso familiar medio de la localidad es de \$7 500 y la desviación estándar de \$1 500. Una encuesta más reducida comprende 100 familias, que se suponen seleccionadas al azar, y se encuentra que el ingreso familiar medio de esta muestra es de \$7 900. ¿Tiene razón el investigador al sospechar que la muestra estaba distorsionada?

1. **Adopción de supuestos.** Con objeto de poder servirse del teorema del límite central, hay que adoptar ciertos supuestos. Como ya se indicó anteriormente, ha de haber siempre un supuesto a propósito del método de muestreo. En el presente caso, suponemos que éste se ha hecho al azar. En realidad, éste es el supuesto que nos interesa verificar, ya que sospechamos de la habilidad de la persona que efectuó la encuesta en cuanto a dar a todas las familias una oportunidad de selección igual. Presumiblemente, estamos dispuestos a aceptar ciertos supuestos acerca de la población, a saber: que los datos del censo son más dignos. Si no podemos aceptar las cifras de éste, entonces tendremos por lo menos dos supuestos dudosos, y la interpretación de los resultados se hará excesivamente difícil. En consecuencia, nuestra hipótesis será la de muestreo al azar. En cuanto a los demás supuestos relativos a la población, éstos constituirán el modelo.

Si N no es demasiado grande, se requiere una población normal. Aquí se plantea la cuestión de saber: "¿cuán grande ha de ser N para que podamos dejar el supuesto de normalidad y servirnos del teorema del límite central?" No existe una respuesta sencilla a dicha cuestión, y la respuesta depende, entre otros: 1) de qué grado de precisión se desea acerca de la apreciación de la probabilidad de error de tipo I, y 2) de cuál grado de aproximación poseemos respecto de una población normal. Pese a que haya que ser cauto a propósito de las simples reglas prácticas, puede sugerirse que, si $N \geq 100$, el supuesto de normalidad puede casi siempre turnarse. Si $N \geq 50$ y se tiene además evidencia em-

pírica en el sentido de que la desviación con respecto a la normalidad no es importante, entonces pueden también utilizarse pruebas del tipo examinado en la presente sección. Pero si $N \leq 30$, en cambio, habría que guardarse decididamente del empleo de semejantes pruebas, a menos que se sepa que la aproximación a la normalidad es buena. Cuando se emplean muestras pequeñas, suele por lo regular carecerse de semejante información, ya que no hay casos suficientes en la muestra para indicar la forma de distribución de la población. Por lo tanto, en el caso de muestras pequeñas deberían por lo regular emplearse otros tipos de pruebas. Supongamos en el presente problema que podemos servirnos legítimamente del teorema del límite central. Como sabemos, las distribuciones relativas al ingreso suelen ser en general algo distorsionadas. Por otra parte, tenemos una muestra razonablemente grande.

Adicionalmente a los supuestos anteriores, si vamos a servirnos del teorema del límite central necesitamos aceptar asimismo las cifras del censo relativas a μ y σ y suponer una escala de intervalo. Tenemos, por tanto, los siguientes supuestos:

Nivel de medición: escala de intervalo

Modelo: población normal (puede abandonarse)

$$\mu = \$7\,500$$

$$\sigma = \$1\,500$$

Hipótesis (nula): selección al azar.

2. *Obtención de la distribución de muestreo.* Afortunadamente, la labor de obtener la distribución de muestreo nos la dan ya hecha. Como quiera que sabemos que la distribución de muestreo de las medias de las muestras es normal o aproximadamente normal, podemos ir directamente a la tabla normal. En adelante, las distribuciones de muestreo se darán siempre en forma de tablas del Apéndice 2. Importa tener presente, sin embargo, que dichas tablas se han calculado sirviéndose de la teoría de las probabilidades. Es tan fácil, en efecto, perderse en detalles de cálculo, que se llega a olvidar que cuando en nuestras pruebas estadísticas nos servimos de tablas nos estamos sirviendo en realidad de una distribución de muestreo.

3. *Elección de un nivel de significación y de una región crítica.* La elección del nivel de significación apropiado depende, por supuesto, de los costos relativos que implican los errores de tipos I y II. Si el investigador deja de descartar la hipótesis de selección al azar cuando la muestra no está efectivamente equilibrada, corre el riesgo de informar hallazgos falaces. Por otra parte, si se descarta cuando la hipótesis es realmente cierta, puede tener que repetir la encuesta, con los gastos considerables que ello

pueda acaso acarrear. Desde el punto de vista ideal, habría de adoptar una decisión racional basada en los costos de dichos dos tipos de error. En la práctica, sin embargo, esto le resultará tal vez difícil. Supongamos, por ejemplo, que se decide por un nivel de .05. A continuación debería decidir servirse de una prueba de una sola cola, ya que la dirección del sesgo se ha estimado. Si

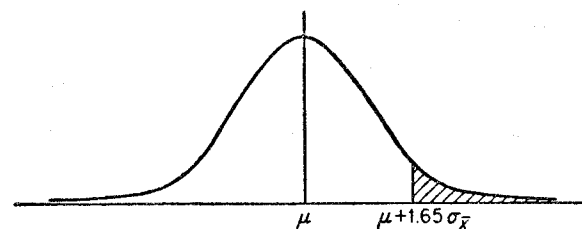


FIG. XI.6. Distribución normal de muestras, con área achurada representando una región crítica de una sola cola al nivel de significación de .05

resultara que la media de la muestra era inferior a \$7 500, difícilmente sospechará que los que realizaron la encuesta hayan sobreseleccionado los grupos de ingresos medio y superior.⁵ Dada la elección del nivel de .05 y de una prueba de una sola cola, la región crítica se determina por la tabla normal. Como quiera que solamente el 5 por ciento del área de la curva normal se sitúa a la derecha de una ordenada de 1.65 desviaciones estándar mayor que la media, sabemos que, si el resultado supera la media μ en más de 1.65 desviaciones estándar, la hipótesis ha de descartarse (véase figura XI.6).

4. *Cálculo de la estadística de la prueba.* Sabemos que si todos los supuestos son correctos, la distribución de muestreo de las \bar{X} será $Nor(\mu, \sigma^2/N)$. O sea, en los términos de nuestro ejemplo:

$$\mu = \$7\,500$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} = \frac{1\,500}{\sqrt{100}} = \$150.$$

Con objeto de poder servirnos de la tabla normal, es necesario convertir a datos estándar, o, en otros términos, obtener una estadística Z que sea $Nor(0.1)$. Anteriormente nos servimos de la fórmula:

⁵ En este problema, los datos de la muestra se han dado en realidad, y sabemos la dirección del resultado. Sin embargo, el lector debe pensar que esta decisión pueda efectuarse antes de conocer el resultado.

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

Esta fórmula es aplicable en el caso de una muestra que sea *Nor*(\bar{X} , s^2), pero no en el caso de la distribución de selección. Recordemos, pues, cada uno de los pasos de nuestro procedimiento. Hemos formulado una serie de supuestos con objeto de obtener una distribución de muestreo. Esta distribución nos indica cuán probable sería una \bar{X} determinada si nuestros supuestos fueran realmente ciertos. El sociólogo, en cambio, ha obtenido de su muestreo una sola \bar{X} , y quiere servirse de la distribución de muestreo teórica para poder apreciar la probabilidad de obtener un resultado tan insólito o más insólito que su \bar{X} particular. En efecto, al servirse de la tabla normal, opera en realidad con la distribución de muestreo. En esta distribución cada "caso" es una \bar{X} , la media es μ , y la desviación estándar es σ/\sqrt{N} . Por lo tanto, \bar{X} sustituye a X , μ sustituye a \bar{X} , y σ/\sqrt{N} sustituye a s en la fórmula anterior de Z . Tenemos, pues:

$$\begin{aligned} Z &= \frac{\bar{X} - \mu}{\sigma/\sqrt{N}} \\ &= \frac{7\,900 - 7\,500}{150} = 2.67 \end{aligned}$$

En otros términos: la media de la muestra es 2.67 errores estándar mayor que la media de la población.

5. *Adopción de una decisión.* Como quiera que \bar{X} se desvía de la presunta μ en más de 1.65 desviaciones estándar en la dirección prevista, la hipótesis debería descartarse al nivel .05. De hecho, sin embargo, habiendo calculado Z exactamente, podemos decir más que esto. En efecto, sirviéndonos de una prueba de una sola cola, la probabilidad de obtener una Z de esta magnitud o mayor es de .0038. En la práctica se recomienda calcular exactamente el nivel de significación, siempre que ello sea posible. Al hacerlo así, en efecto, indicamos que el resultado se sitúa en una región crítica más reducida todavía que la que originariamente se estableciera. Como quiera que el lector preferirá tal vez servirse de un nivel de significación distinto de aquel del autor, resulta por lo regular útil proporcionar probabilidades exactas, o lo más exactas posibles, de modo que aquél pueda sacar sus propias conclusiones en cuanto a aceptar o no los hallazgos. En el presente ejemplo, el sociólogo descartaría la hipótesis nula de que

la muestra fue al azar. Y habría de decidir a continuación si quiere o no extraer otra muestra.

XI.3 La distribución *t* de Student

En la mayoría de los casos es totalmente impráctico tratar a σ como si fuera conocida. Por lo regular incurrimos en dificultades considerables con objeto de asegurar el carácter fortuito de la muestra, ya que lo que nos interesa en primer término es la comprobación de los supuestos acerca de la población a estudiar. En pruebas de la clase que se examina en este capítulo, es probable que el lector desee verificar una hipótesis relativa a μ . Pero si así fuera el caso, ¿estará jamás en condiciones de conocer el valor de σ ? Prácticamente no. Porque si tuviera conocimiento de σ , estaría también sin duda alguna en condiciones de conocer μ , a menos, por supuesto, que alguien como, por ejemplo, su maestro de estadística le estuviera deliberadamente ocultando información. Por lo regular, pues, no conocerá los valores ni de μ ni de σ . ¿Qué puede, pues, hacer en semejante situación? Como quiera que el teorema del límite central comporta σ , no puede ignorar su valor por completo. Una solución podría parecer consistir en remplazar σ por s , desviación estándar de la muestra. De hecho, esto es lo que se hacía corrientemente antes del desarrollo de la estadística moderna. En efecto, en la fórmula de Z , σ/\sqrt{N} se remplazaba sencillamente por s/\sqrt{N} y, como quiera que s podía calcularse directamente de los datos de la muestra, no había más incógnitas en la fórmula. Y de hecho, este procedimiento da resultados razonablemente buenos cuando N es grande. Sin embargo, como habremos de ver a continuación, las probabilidades obtenidas en esta forma pueden ser totalmente falaces cuando N es relativamente pequeño. Veamos por qué es así.

Podemos construir una estadística alternativa de prueba como:

$$t = \frac{\bar{X} - \mu}{s/\sqrt{N-1}}$$

Esta estadística fue introducida por W. S. Gosset, que escribía bajo el seudónimo de "Student", y se conoce con el nombre de distribución *t* de Student. Comparando t con Z , observamos que, en tanto que los numeradores son idénticos, los denominadores difieren, en cambio, en dos aspectos, a saber: 1) tenemos un $N-1$ bajo el radical, y 2) σ ha sido remplazada por s . Con objeto de comprender estas modificaciones, examinémoslas una después de otra. Al hacerlo habremos de introducir algunas ideas nuevas.

La desviación estándar de la muestra s puede emplearse como una estimación de σ . Si bien el problema de la apreciación se tratará en el próximo capítulo, baste mencionar aquí que a menudo necesitamos que una estimación posea ciertas propiedades. Una de las propiedades de una "buena" estimación es que sea insesgada. Ahora bien, contrariamente a lo que podría suponerse, resulta que s no es una estimación completamente insesgada de σ . Puede demostrarse matemáticamente que otra cantidad, que podemos designar con $\hat{\sigma}$ y se obtiene mediante la fórmula

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N-1}}$$

es una estimación insesgada de σ .⁶ La única diferencia entre $\hat{\sigma}$ y s es el factor $N-1$ del denominador. Así, pues, pese a que el lector ha aprendido a calcular s , se encuentra ahora con que debería servirse de otra fórmula al estimar σ . En el presente problema es más bien σ/\sqrt{N} que σ la que ha de estimarse, ya que es la primera expresión que aparece en el denominador de Z . Aun siendo cierto que $\hat{\sigma}/\sqrt{N}$ sea una estimación menor de σ/\sqrt{N} , es posible, sin embargo, evitar por completo el cálculo de $\hat{\sigma}$ si s se ha obtenido ya. Obsérvese que

$$\frac{\hat{\sigma}}{\sqrt{N}} = \frac{\sqrt{\sum_{i=1}^N (X_i - \bar{X})^2 / (N-1)}}{\sqrt{N}}$$

Y recordando que \sqrt{a}/\sqrt{b} puede escribirse como $\sqrt{a/b}$, tenemos:

$$\frac{\hat{\sigma}}{\sqrt{N}} = \sqrt{\frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N(N-1)}}$$

⁶ En rigor, no es $\hat{\sigma}$, la estimación equilibrada de σ , sino que $\hat{\sigma}^2$ es una estimación equilibrada de σ^2 . No tenemos por qué preocuparnos, con todo, por esta distinción sutil. En este texto nos serviremos normalmente de un acento circunflejo (^) sobre una letra griega para indicar una estimación del parámetro. Algunos textos definen s con $N-1$ en el denominador, pero preferimos por nuestra parte mantener la distinción entre las dos fórmulas.

$$= \frac{\sqrt{\sum_{i=1}^N (X_i - \bar{X})^2 / N}}{\sqrt{N-1}} = \frac{s}{\sqrt{N-1}}$$

Así, pues, podemos tomar una estimación algo sesgada de σ , dividir entre una cantidad que es ligeramente más pequeña que \sqrt{N} , y llegar a $s/\sqrt{N-1}$ como estimación insesgada de σ/\sqrt{N} . Ésta es la razón de que aparezca $N-1$ en el denominador de t .⁷

Al sustituir Z por t , la modificación introducida sirviéndonos de $N-1$ es relativamente pequeña, pero la sustitución de σ por s , en cambio, puede revestir un significado considerable si N es pequeña. Como quiera que s varía de una muestra a otra, el denominador de t varía lo mismo que el numerador. Para un valor dado de \bar{X} , si la s de una muestra particular acontece ser demasiado pequeña, entonces t será muy grande, y si s es grande, t será relativamente pequeña. Por consiguiente, habrá una mayor variabilidad entre los valores de t que entre los valores comparables de Z . Esto significa que la distribución de muestreo de t será más aplanada que la normal. Por lo tanto, la distribución t tendrá colas mayores. Qué tan aplanada sea t dependerá de la magnitud de la muestra. Si N es muy pequeña, la distribución t será muy plana en comparación con la curva normal. En otros términos: será necesario apartarse un número mayor de desviaciones estándar de la medida para incluir el 95 por ciento de los casos. A medida que N crece, la distribución t se va acercando cada vez más a la distribución normal, pero permaneciendo siempre, sin embargo, ligeramente más achatada que ésta. Así, pues, hay una distribución t para cada magnitud de la muestra. El hecho de que la distribución t se vaya acercando a la normalidad cobra sentido intuitivamente si nos damos cuenta de que, a medida que N crece, s se convierte en una estimación muy precisa de σ , de modo que importa poco que nos sirvamos en el denominador de ésta o de aquélla.

Con objeto de servirnos de la distribución t , hay que suponer una población normal, sobre todo si N es relativamente pequeña. El cálculo de la distribución de muestreo de t requiere que el numerador $(\bar{X} - \mu)$ esté normalmente distribuido y que varíe también independientemente del denominador $s/\sqrt{N-1}$. Por lo general, no esperaríamos que haya independencia entre el numerador y el denominador, ya que s se calcula en realidad tomando

⁷ Algunos textos recomiendan el empleo de $N-1$ para muestras pequeñas y de N para las grandes. Sin embargo, semejante procedimiento parece añadir una confusión innecesaria. Por supuesto, en el caso de muestras grandes, es indiferente servirse de una cifra o de la otra.

desviaciones con respecto a \bar{X} y, por consiguiente, sería sorprendente encontrar \bar{X} y s estadísticamente independientes una respecto de otra. Conociendo la \bar{X} de la muestra, esperaríamos aumentar nuestra posibilidad de anticipar s para la misma muestra. Acontece, sin embargo, que, para las poblaciones normales y muestreo al azar, la media y la desviación estándar de la muestra son estadísticamente independientes una de otra. Como quiera que, con todo, esta propiedad no se verifica para todas las distribuciones de la población y que, por otra parte, $\bar{X} - \mu$ no estará distribuida normalmente a menos que N sea grande, de ahí que al servirnos de la prueba t hayamos de suponer una población normal.

Problema. Supóngase que se están apreciando los programas de una muestra al azar de 25 agencias de asistencia social individual seleccionadas entre la población de todas las del Estado de Nueva York. Cada una de ellas lleva un registro del porcentaje de los casos favorables, de acuerdo con un criterio uniforme. Se ha establecido una norma según la cual el porcentaje medio de éxitos sería del 60 para todas las agencias. Sin embargo, en su muestra el lector encuentra que el porcentaje medio es del 52 y la desviación estándar del 12 por ciento. ¿Tiene el lector algún fundamento para sospechar que para la población conjunta de las agencias el nivel de los éxitos está por debajo de la norma esperada?

1. Formulación de supuestos. Los supuestos necesarios pueden enumerarse como sigue:

Nivel de medición: escala de intervalo

Modelo: muestreo al azar
población normal

Hipótesis: $\mu = 60$ por ciento.

Obsérvese que no se requiere supuesto alguno a propósito de σ , ya que en realidad s se ha obtenido empíricamente y puede emplearse directamente en la prueba t . El nivel de medición, en cambio, requiere ciertos comentarios. Como quiera que cada cliente de una agencia es un éxito o un fracaso, y como quiera que las cifras obtenidas para cada agencia son porcentajes de éxitos, cabría pensar que nos hallamos simplemente en presencia de una escala nominal dicotómica, y no de una escala de intervalo. Y efectivamente, si las unidades del análisis fueran *clientes* en lugar de agencias, éste sería el caso. Recuérdese, sin embargo, que las unidades que se están estudiando son *agencias*. Se ha obtenido una marca para cada agencia (esto es, un porcentaje de éxitos), y esta marca representa legítimamente una escala de intervalo. Así, por ejemplo, una diferencia entre el 30 y el 40 por ciento es

lo mismo que una diferencia entre el 70 y el 80 por ciento. Ambas diferencias pueden convertirse en el mismo número real de clientes.

2. Obtención de la distribución de muestreo. Las distribuciones de muestreo para t se dan en el cuadro D del Apéndice 2. Como quiera que esas distribuciones difieren para cada magnitud de la muestra, el cuadro en cuestión se ha condensado de modo que sólo dé las colas de cada distribución. Al servirnos del cuadro necesitamos primero localizar la magnitud apropiada de la muestra leyendo la columna de la izquierda de arriba abajo. Estos tamaños de la muestra se dan por lo regular en términos de *grados de libertad* df (*degrees of freedom*), que en este tipo de problema es siempre $N - 1$.⁸ A continuación, hállese el nivel de significación apropiado leyendo horizontalmente arriba. Las cifras del cuerpo del cuadro indican la magnitud de t necesaria para obtener significación en el nivel designado.

3. Selección de un nivel de significación y de una región crítica. Sirvámonos del nivel de .05 y de una prueba de una sola cola. Del cuadro D vemos que para 24 grados de libertad se necesita una t de 2.064 o más para obtener significación al nivel de .05 para una prueba de dos colas. Para una prueba de una sola cola y el nivel de .05, sólo necesitamos una t de 1.711 o mayor. En el caso de pruebas de una cola simplemente partimos en dos los niveles de significación requeridos para las pruebas de dos colas. Esto se debe a que nos apartamos de la media el mismo número de desviaciones estándar, con objeto de obtener una región crítica de una sola cola de .05, que nos apartaríamos para obtener una región de dos colas de .10.

4. Cálculo de la estadística de la prueba. Si bien es cierto que la distribución de muestreo de \bar{X} es $Nor(\mu, \sigma^2/N)$ y que, por consiguiente, la distribución de Z es $Nor(0,1)$, esta información no nos sirve, con todo, de gran cosa, ya que σ no nos es conocida. En lugar de ello calculamos el valor de t , y obtenemos:

$$t = \frac{\bar{X} - \mu}{s/\sqrt{N-1}} = \frac{52 - 60}{12/\sqrt{24}} = -3.27$$

5. Decisión. Hallamos que toda t cuyo valor numérico fuera ≥ 1.711 se encontrará en la región crítica. Por consiguiente, descartamos la hipótesis de que $\mu = 60$ y, con cierto riesgo de error, vemos que el nivel real de éxitos de las agencias queda por debajo del nivel esperado. Leyendo horizontalmente en el cuadro D la hilera correspondiente a 24 grados de libertad, vemos que para una prueba de una sola cola el nivel de significación

⁸ Para el examen de los grados de libertad véase secc. XII.1.

correspondiente a un t de 3.27 cae en algún punto comprendido entre .005 y .0005.⁹

En este punto pueden registrarse varios hechos a propósito de la distribución t . Si se examina la columna correspondiente a $p = .05$ para una prueba de dos colas, se observará que a medida que la magnitud de la muestra aumenta, los valores de t disminuyen y convergen con bastante rapidez hacia 1.96, o sea el valor necesario para significación si se empleara el cuadro normal. Estos valores debieran dar una idea razonablemente buena del grado de aproximación a la curva normal para cualquier magnitud dada de la muestra. Para valores de $N - 1$ mayores que 30, se necesitará por lo regular interpolación, y para valores muy superiores a 120 habrá que servirse del cuadro normal, ya que los valores de t no se dan. Algunos textos indican arbitrariamente que sólo se necesita emplear la tabla t cuando $N \leq 30$. Pese a que esta regla práctica dé resultados razonables, la posición que aquí adoptamos es que siempre es preferible *servirse de la tabla t cuando σ es desconocida* y puede presumirse una población normal. Como quiera que el cuadro t no es de uso más difícil, parece razonable servirse de valores exactos con preferencia a aproximaciones normales. Conviene recalcar también que no hay una teoría única que se aplique a muestras pequeñas y otra, totalmente distinta, que se aplique a las grandes, como algunos textos lo dan a entender.

Como puede verse del cuadro t , las distribuciones normal y t sólo difieren considerablemente cuando la magnitud de la muestra es relativamente pequeña. Por otra parte, siempre que se emplea t hay que suponer una población normal, a menos que N sea muy grande, en cuyo caso Z puede sustituir aproximadamente a t . Por lo tanto, la prueba t reviste valor práctico en situaciones en las que se tienen muestras pequeñas y que se puede suponer una población normal. Por desgracia, cuando las muestras son pequeñas por lo regular solemos saber menos acerca de la naturaleza exacta de aquélla. Así, por ejemplo, si un investigador realiza un estudio de exploración con 17 casos, ¿tiene muchas probabilidades de estar en condiciones de aceptar el supuesto de normalidad? Probablemente no. Según veremos en el capítulo XIV, hay pruebas que pueden emplearse como alternativas de la t y que no implican el supuesto de normalidad.

XI.4. Pruebas que comportan proporciones

Hasta aquí sólo hemos considerado en este capítulo ejemplos que

⁹ Pese a que no puedan obtenerse de la tabla de probabilidades exactas, la interpolación, con todo, siempre es posible. Sin embargo, por lo regular basta indicar que p queda entre dos valores determinados, v.gr. $.0005 \leq p \leq .005$.

comportaban una escala de intervalo. Por otra parte, había que presumir también normalidad de población en el caso de muestras pequeñas. En esta sección veremos cómo puede emplearse la ley de los grandes números para abarcar pruebas que comportan proporciones, siempre que N sea bastante grande. En efecto, las proporciones se tratarán como casos especiales de las medias, de modo que nuestras consideraciones anteriores seguirán teniendo aplicación.

Supóngase que tenemos una simple escala nominal dicotómica. Podemos querer verificar una hipótesis relativa, por ejemplo, a la proporción de los varones en una población. Asignamos arbitrariamente el valor uno a los varones y cero a las hembras, y tratamos las marcas como una escala de intervalo. Aunque no se dé una unidad claramente concebida, a menos que ésta sea el atributo de "masculinidad", que se posee o no se posee, podemos, con todo, tratar dichas marcas arbitrarias como una escala de intervalo, porque sólo son dos. Si se añadiera una tercera categoría, ello ya no sería posible, sin embargo, ya que en tal caso sería necesario determinar la posición exacta de dicha categoría en relación con las de las otras dos. Lo que aquí decimos, en efecto, es que no es necesario hacer una distinción entre escalas nominales, ordinales y de intervalo en el caso de una dicotomía, ya que el problema de comparar distancias entre marcas no se plantea nunca.

Tenemos así una población compuesta por entero de unos y ceros. Es ésta una distribución bimodal, de casos concentrados todos ellos en uno de los dos puntos, que ciertamente no es normal. Pero sabemos que, si N es suficientemente grande, la distribución de muestreo de las medias de las muestras será aproximadamente $Nor(\mu, \sigma^2/N)$, independientemente de la forma de la población. Todo lo que hay que hacer, pues, es averiguar la media y la desviación estándar de esa población de unos y ceros.

Pongamos que p_u representa la proporción de varones en la población y q_u la proporción de hembras, en las que la u subíndice indica que tratamos de la población universal. Con objeto de obtener la media de los unos y los ceros de ésta, sumamos simplemente los valores y dividimos entre el número total de casos. El número de unos será así el número total de casos multiplicado por la proporción de varones. Independientemente del número de ceros, la contribución de éstos a la suma será cero. Por lo tanto, la media de la población será:

$$\mu = \frac{Mp_u}{M} = p_u$$

en donde M representa la magnitud de la población (en cuanto distinta de la magnitud N de la muestra). En consecuencia, la

media de cierto número de unos y ceros es exactamente la proporción de unos. En virtud de un razonamiento similar, $\bar{X} = p_s$, en donde p_s representa la proporción de varones en la muestra.

Sirviéndonos de la fórmula general de la desviación estándar, podemos demostrar que $\sigma = \sqrt{p_u q_u}$. Si empleamos los símbolos de los parámetros de población, la fórmula de σ se transforma en:

$$\sigma = \sqrt{\frac{\sum_{i=1}^M (X_i - \mu)^2}{M}} = \sqrt{\frac{\sum_{i=1}^M (X_i - p_u)^2}{M}}$$

Observando el numerador de la cantidad debajo del radical, vemos que habrá sólo dos tipos de cantidades que representen las desviaciones cuadradas de la media p_u . Para cada marca de uno, la desviación cuadrada respecto de la media será de $(1 - p_u)^2$, y para cada cero será de $(0 - p_u)^2$. Como quiera que en la suma de cuadrados habrá Mp_u unos y Mq_u ceros, tenemos:

$$\sigma = \sqrt{\frac{Mp_u(1 - p_u)^2 + Mq_u(0 - p_u)^2}{M}} = \sqrt{\frac{Mp_u q_u^2 + Mq_u p_u^2}{M}}$$

Si de cada término del numerador ponemos $Mp_u q_u$ en factor, obtenemos:

$$\begin{aligned} \sigma &= \sqrt{\frac{Mp_u q_u (q_u + p_u)}{M}} = \sqrt{\frac{Mp_u q_u}{M}} \\ &= \sqrt{p_u q_u} \end{aligned}$$

Obsérvese, de paso, que M se elimina tanto en la fórmula de μ como en la σ , la media y la desviación estándar de la población son independientes de la magnitud real de la población.

Por lo tanto, podemos servirnos del teorema del límite central para obtener:

$$\sigma_{\bar{X}} = \sigma_{p_s} = \frac{\sigma}{\sqrt{N}} = \sqrt{\frac{p_u q_u}{N}}$$

en donde el símbolo σ_{p_s} indica que operamos con el error estándar de las proporciones de la muestra. En nuestra nueva terminología, p_s sustituye a \bar{X} , p_u sustituye a μ , y σ_{p_s} sustituye a $\sigma_{\bar{X}}$ en la fórmula de Z . Así, pues:

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{p_s - p_u}{\sqrt{p_u q_u / N}}$$

Obsérvese que aunque parezca que tengamos una fórmula totalmente distinta de la anterior, no hay en ésta en realidad nada nuevo, excepto un cambio de símbolos. Esto es así porque hemos estado en condiciones de demostrar que las proporciones pueden tratarse como casos particulares de las medias. Conviene recalcar, con todo, que la ley de los grandes números requiere que N sea grande, con objeto de poder servirnos de la aproximación normal. Cuando N sea pequeña, la binomial constituirá una prueba más apropiada.

* Existe una relación estrecha entre esta prueba, relativa a las proporciones, y la distribución binomial. Ya se ha indicado que si N es grande, y si $Np > 5$, en donde $p < q$, podremos, mediante una distribución normal, aproximarnos a la distribución binomial. Es cierto que en el caso de la distribución binomial utilizamos *cifras* de éxitos, más que proporciones.

El valor esperado del número de éxitos resulta ser Np , y la desviación estándar del número de éxitos es \sqrt{Npq} . Para convertir cada uno de ellos en proporciones podemos dividirlos entre N , obteniendo p como valor esperado, y

$$\frac{1}{N} \sqrt{Npq} = \sqrt{\frac{Npq}{N^2}} = \sqrt{\frac{pq}{N}}$$

para la desviación estándar. Así en el caso de muestras grandes podríamos haber formulado un problema binomial en forma de proporciones, cambiando nuestros símbolos a p_u y q_u y tratando el problema de acuerdo con los procedimientos señalados en este mismo capítulo. Por ejemplo: en el caso de una prueba de signo podríamos haber utilizado la hipótesis nula de que $p_u = .5$, comparando este valor con la proporción de éxitos p_s que se halló en la muestra.

Problema. El lector está interesado en apreciar el programa de una agencia de asistencia social individual y ha extraído una muestra al azar de 125 casos de su archivo. Se ha encontrado que el porcentaje de los casos favorables es de 55, en comparación con la norma del 60 por ciento. ¿Puede sacarse de ello la conclusión que el éxito de la agencia en cuestión queda por debajo de la norma?

1. Formulación de supuestos.

Nivel de medición: escala nominal dicotómica

Modelo: muestreo al azar
Hipótesis: $p_u = .60$

Este ejemplo es deliberadamente semejante al anterior, con objeto de poner de relieve la diferencia en las unidades del análisis. Aquí, en efecto, se estudia una sola agencia, y la muestra es de clientes, que constituyen éxitos o fracasos. En el ejemplo de antes, las unidades seleccionadas eran las agencias y no los clientes, y la medida de cada agencia consistía en el *porcentaje* de casos favorables. Obsérvese que no se requiere más supuesto acerca de la población que la hipótesis, ya que se supone implícitamente que aquélla es bimodal.

2. *Obtención de la distribución de muestreo.* La distribución de muestreo será aproximadamente normal, ya que N es grande.

3. *Elección de nivel de significación y región crítica.* En gracia a la variedad, elijamos un nivel de .02 y una prueba de una sola cola.

4. *Cálculo de la estadística de la prueba.* Calculamos Z de la siguiente forma:

$$Z = \frac{p_s - p_u}{\sqrt{p_u q_u / N}} = \frac{.55 - .60}{\sqrt{[(.60) (.40)] / 125}} = \frac{-.05}{.0438} = -1.14$$

Obsérvese que en el denominador se emplean p_u y q_u con preferencia a p_s y q_s . En el caso de que el lector se viera inclinado a servirse de t en lugar de Z , observe que en la hipótesis de p_u el valor de σ está determinado por la fórmula $\sigma = \sqrt{p_u q_u}$.

5. *Decisión.* Del cuadro normal puede verse que una Z de -1.14 o menos ocurriría aproximadamente el 13 por ciento de las veces por azar, si los supuestos fueran ciertos. En consecuencia, no descartamos la hipótesis al nivel de significación de .02. Sobre la base de los datos disponibles, no se deja establecer que la agencia se halle por debajo de la norma.

GLOSARIO

Teorema del límite central
 Distribución rectangular
 Error estándar
 Distribución t

EJERCICIOS

1. Sirviéndonos del cuadro de números al azar del cuadro B del Apéndice 2 (véase secc. XXI.1 para la explicación del empleo de dicho cuadro), elijanse 10 muestras, de magnitud 4 cada una, de la población de los 65 casos dados en el ejercicio 1 del capítulo IV. Calcúlese la

media de cada una de esas 10 muestras y obténgase la desviación estándar de estas 10 medias. Se tiene ahora una apreciación bruta y ligeramente distorsionada del error estándar de la media. ¿Cómo se presenta la cifra obtenida en comparación con el error estándar conseguido sirviéndonos del teorema del límite central y empleando para ello la desviación estándar calculada en el ejercicio 2 del capítulo VI?

* 2. Verifíquese la distribución de selección de la media de tres golpes de dados del diagrama de la figura XI.5.

3. Una muestra de magnitud 50 tiene una media de 10.5 y una desviación estándar s de 2.2. Verifíquese la hipótesis de que la media de la población es de 10.0 sirviéndose: a) de una prueba de una sola cola al nivel de .05, y b) de una prueba de dos colas al nivel de .01. Hágase lo mismo con muestras de tamaños 25 y 100 y compárense los resultados. Respuesta, para $N = 50$, $t = 1.59$; sin rechazo para a) y b).

4. Supóngase sabido que el ingreso medio anual de trabajadores que en una fábrica trabajan en la línea de ensamble es de \$7 000 con una desviación estándar de \$900. El lector sospecha que los trabajadores sindicalmente activos obtendrán ingresos superiores al promedio, y toma una muestra aleatoria de 85 de dichos miembros activos, obteniendo una media de \$7 200 y una desviación estándar de \$1 000. ¿Puede decirse que los miembros activos del sindicato tengan ingresos notablemente superiores? (Empléese el nivel de .01.) Respuesta $Z = 2.05$; no rechazar.

5. Se ha establecido una lista de 200 residentes de una localidad, en edad de voto, y se ha encontrado que de dos candidatos a un cargo el candidato A obtuvo el 54 por ciento de los votos seleccionados. ¿Existe fundamento para suponer que A ganará? Empléese el nivel de .05. Enumérense todos los supuestos que hay que formular. Respuesta, $Z = 1.13$.

6. Supóngase que se ha normalizado una prueba de medición de los "deseos de uniformidad" de estudiantes universitarios en todo el país. El 50 por ciento de los estudiantes tenía puntuaciones brutas de 26 o más (las puntuaciones altas indicando deseos mayores de uniformidad). Sospechando que estos deseos serán por lo regular más grandes en el caso de adultos sin instrucción universitaria, un sociólogo extrae una muestra aleatoria de adultos de 25 años o mayores, residentes de su localidad. Encuentra: 1) que el 67 por ciento de los 257 adultos sin instrucción universitaria muestran marcas de 26 o más altas, y 2) que el 59 por ciento de 80 adultos de instrucción universitaria presentan marcas dentro de dicho margen.

a) ¿Puede deducir que las marcas de cada grupo de adultos en la localidad son significativamente más altas que las correspondientes a los estudiantes universitarios, cuya prueba ha sido estandarizada? (utilícese el nivel .001).

b) Supóngase que el sociólogo conoce la distribución exacta entera de las marcas de los estudiantes universitarios de la prueba. Sobre la base del material del presente capítulo, indíquense algunos procedimientos alternativos para verificar el significado de las desviaciones de los dos grupos de marcas de los adultos respecto de las marcas normalizadas. ¿Requieren dichos procedimientos alternativos algunos supuestos adicionales? Explíquese.

BIBLIOGRAFÍA

1. Freund, J. E.: *Modern Elementary Statistics*, 3ª ed., Prentice-Hall, Inc., Englewood Cliffs, N. J., 1967, caps. 9 y 11.
2. Hagood, M. J., y D. O. Price: *Statistics for Sociologists*, Henry Holt and Company, Inc., Nueva York, 1952, caps. 15 y 16.
3. Hays, W. L.: *Statistics*, Holt, Rinehart and Winston, Inc., Nueva York, 1963, cap. 10.
4. Wallis, W. A., y H. V. Roberts: *Statistics: A New Approach*, The Free Press of Glencoe, Ill., Chicago, 1956, caps. 11 y 13.

XII. ESTIMACIÓN DE PUNTO E INTERVALO

HASTA aquí el examen de la inducción estadística sólo ha tratado de la verificación de hipótesis. Aparte de ello puede también haber interés en la *estimación* de parámetros de población, y a este tema se dedica el presente capítulo. Después de examinar los principios que la estimación comporta, procederemos a estudiar las relaciones existentes entre las verificaciones de las estimaciones y las hipótesis. Examinaremos en dicho momento las modificaciones que requieren para ello la distribución t y las proporciones. Finalmente, estudiaremos la cuestión general de la determinación del tamaño de la muestra, ilustrando el problema mediante procedimientos de estimación.

En los dos capítulos precedentes, el lector habrá observado que en cierto número de problemas prácticos la verificación de las hipótesis es impracticable, porque no estamos en condiciones de concretar algún valor hipotético determinado para el parámetro, por ejemplo, μ . Vamos a ver ahora en qué forma los procedimientos de estimación pueden proporcionar en tales casos una alternativa muy útil de las pruebas reales. Por otra parte, el sociólogo puede eventualmente tener mayor interés en las estimaciones que en las verificaciones de hipótesis. Así, por ejemplo, en una encuesta el objetivo práctico del estudio puede consistir en estimar la proporción de personas que consumen un determinado producto o que votan en unas elecciones. O puede ser necesario estimar el ingreso mediano en una región, o el número medio de hijos por matrimonio. Sin duda, las pruebas de hipótesis concretas pueden revestir cierta utilidad en tales casos, pero la estimación constituye, con todo, el procedimiento más obvio.

Hay básicamente dos clases de estimación, a saber, la estimación del punto y la estimación del intervalo. En la primera de ellas nos interesa el mejor valor singular que pueda utilizarse para apreciar un parámetro. Así, por ejemplo, podemos apreciar que el ingreso medio en la ciudad de Nueva York es de \$ 8 500. Sin embargo, por lo regular queremos obtener también alguna idea acerca de cuán exacta sea nuestra estimación. Nos gustaría poder anticipar que el parámetro se sitúa en algún lugar de un intervalo determinado, o a un lado u otro de la estimación del punto. Así, por ejemplo, podemos querer formular un enunciado por el estilo de "el ingreso medio en la ciudad de Nueva York se sitúa entre \$ 8 000 y \$ 9 000". Estos dos tipos de estimación se examinan en las secciones que siguen a continuación.

XII.1. Estimación del punto

El problema relativo a cuál estadística deba emplearse como estimación de un parámetro parece ser absolutamente obvio y constituir materia de sentido común. En efecto, si se quiere estimar la media (o la mediana o la desviación estándar) de una población, ¿por qué no servirse de la media (o la mediana o la desviación estándar) de la muestra? Pese a que en tales casos el sentido común no nos proporcionaría resultados demasiado aberrantes, veremos, con todo, que el problema no es tan sencillo como parece. Obviamente, podríamos apreciar la media de una población en cierto número de maneras distintas. En efecto, además de la media de la muestra, podríamos servirnos de la mediana o del modo, o podríamos utilizar un número situado en medio de dos valores extremos, o podríamos utilizar como estimación el valor de la observación decimotercera. Algunos de esos procedimientos serían mejores que otros. Necesitamos, por consiguiente, criterios que nos permitan apreciar el grado de bondad de cada clase de estimación. El sociólogo, que se sirve de las estadísticas como de un instrumento aplicado, rara vez necesita preocuparse por semejantes criterios. Por lo regular, en efecto, sólo se le dice que se sirva de una estimación determinada. No obstante, vale la pena saber por lo menos de cuáles criterios se sirve el matemático al decidir cuál estimación deba emplearse. Dos de los criterios más importantes del matemático son el sesgo (*bias*) y la eficiencia. Vamos a examinarlos uno por uno. En relación con otros criterios, tales como la suficiencia, la consistencia y el principio de la máxima probabilidad, el lector ha de recurrir a textos más avanzados.

Sesgo. Se dice de una estimación que no es *sesgada* si la media de su distribución de muestreo es exactamente igual al valor del parámetro que se aprecia. En otros términos: el valor esperado a la larga de la estimación es el parámetro mismo. Obsérvese que nada se dice aquí acerca del valor de cualquier resultado de alguna muestra particular. De acuerdo con esta definición, \bar{X} es una estimación sesgada de μ , ya que la distribución de muestreo de \bar{X} tiene a μ como media o valor esperado. Esto no significa, sin embargo, que podamos esperar que algún valor particular cualquiera de \bar{X} sea igual a μ , ni sabremos nunca, en cualquier problema real, si la media de nuestra muestra corresponde o no de hecho a la media de la población. Hay que tener presente que el término *sesgo*, tal como aquí se emplea, se refiere a los resultados a la larga. En la investigación práctica, en cambio, el lector puede estar acostumbrado a servirse del término para referirse a las propiedades de la muestra particular que haya extraído.

Ya se dijo en el capítulo anterior que la desviación estándar s de la muestra es una estimación ligeramente sesgada de σ . La estadística s tiene una distribución de selección, lo mismo que la tiene \bar{X} . En otros términos, las desviaciones estándar de la muestra estarán distribuidas alrededor de las desviaciones estándar de la población real, del mismo modo en que las medias de las muestras se distribuyen alrededor de μ . Sin embargo, puede demostrarse matemáticamente que la media de la distribución de muestreo de s^2 es $[N - 1/N]\sigma^2$, y no σ^2 . Por consiguiente, s^2 es una estimación sesgada de σ^2 . Para hallar una estimación no sesgada de σ^2 , tomamos la cantidad:

$$\begin{aligned} \frac{N}{N-1} s^2 &= \frac{N}{N-1} \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N} \\ &= \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N-1} = \hat{\sigma}^2 \end{aligned}$$

Como quiera que la distribución de muestreo de s^2 es $[N - 1/N]\sigma^2$ vemos que $\hat{\sigma}^2$ tiene una distribución de muestreo de media exactamente igual a:

$$\frac{N}{N-1} \left[\left(\frac{N-1}{N} \right) \sigma^2 \right] = \sigma^2$$

Si bien la razón básica de por qué sea $\hat{\sigma}^2$ (y no s^2) la estimación no sesgada es que los matemáticos operan en esta forma, a veces; con todo, se da de ello una explicación intuitiva en término del concepto de los *grados de libertad*, término que será usado en capítulos posteriores. El número de los grados de libertad es igual al número de cantidades desconocidas menos el número de ecuaciones independientes que ligan estas incógnitas. El lector recordará que, con objeto de llegar a una solución única de un sistema de ecuaciones algebraicas simultáneas, se necesitaba el mismo número de ecuaciones que de incógnitas. Así, pues, para poder resolver en relación con X , Y y Z se necesitan tres ecuaciones que ligen entre sí estas variables. Si sólo se tuvieran dos ecuaciones, entonces se podría asignar a alguna de las variables, por ejemplo Z , cualquier valor que se nos antojara. Los valores de las otras dos variables podrían entonces ser determinados mediante las dos ecuaciones simultáneas. Si se tuvieran cinco incógnitas y sólo tres ecuaciones por resolver simultáneamente, entonces se podrían asignar valores arbitrarios a dos

cualesquiera de las incógnitas, y los valores de las demás incógnitas estarían determinados. En este caso tendríamos dos grados de libertad, ya que podemos atribuir libremente valores a dos variables cualesquiera.

Al calcular una desviación estándar de valores de la muestra, hemos de servirnos de una ecuación que ligue las N variables X

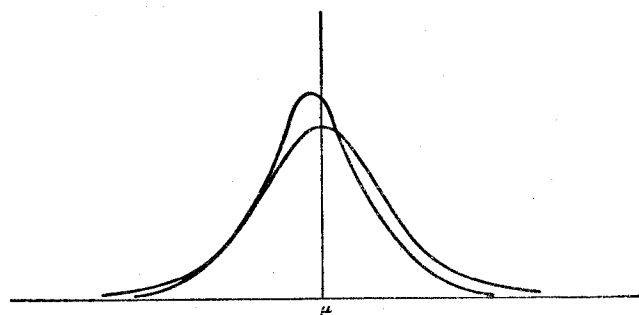


FIG. XII.1. Comparación de las distribuciones de muestreo de una estimación sesgada, con alta eficacia, y una estimación no sesgada, de eficacia menor

a la media de la muestra, o sea la ecuación $\sum_{i=1}^N X_i/N = \bar{X}$. Dado el valor de \bar{X} , podemos atribuir valores arbitrarios a $N-1$ de las X_i , y la última de éstas estará determinada por la ecuación. Como quiera que hemos perdido un grado de libertad al determinar el valor de la media de la muestra alrededor de la cual se han tomado desviaciones, para obtener nuestra estimación no sesgada de σ^2 hemos de dividir entre $N-1$ en lugar de N . Si el lector prefiere pensar en esta forma, puede considerar que hemos ajustado ligeramente el número de casos, con objeto de compensar por el hecho de que hemos tomado desviaciones respecto de la media de la muestra y no de la media de la población verdadera. Esencialmente, al calcular la media de la muestra, hemos despreciado un caso. Se verá que las estimaciones no sesgadas se obtienen frecuentemente dividiendo entre los grados de libertad, mejor que dividiendo entre el número total de casos.

Eficiencia. La eficiencia de una estimación se refiere al grado en que la distribución de muestreo está agrupada alrededor del verdadero valor del parámetro. Si la estimación no está sesgada, dicho agrupamiento puede medirse por medio del error estándar de la estimación: cuanto menor sea el error estándar, tanto mayor es la eficiencia de la estimación. La eficiencia siempre es relativa. Ninguna estimación puede ser totalmente eficiente, ya que esto implicaría que no existía error de muestreo alguno. Sin

embargo, podemos comparar dos estimaciones y decir que una de ellas es más eficiente que la otra. Supóngase, por ejemplo, que tenemos una población normal. En este caso, el error estándar de la selección al azar es σ/\sqrt{N} . Si para apreciar la media de la población se utilizara la mediana de la muestra, entonces el error estándar de la mediana sería de $1.253 \sigma/\sqrt{N}$,¹ para muestras al azar. Por lo tanto, ya que el error estándar de la media es más pequeño que el de la mediana, la media es la estimación más eficiente. Ésta es, por supuesto, la razón de que por lo regular se utilice la media con preferencia a la mediana, incluso cuando, como en el caso de una población normal, su media y su mediana son idénticas. Decimos que la media está menos sujeta a las fluctuaciones de la muestra o, en otros términos, que es más eficiente.²

De los dos criterios que acabamos de examinar, el de la eficiencia es el más importante. Si dos estimaciones tienen el mismo grado de eficiencia, escogeremos, por supuesto, la que esté menos sesgada. Ésta es la razón de que se utilice $\hat{\sigma}$ con preferencia a s . En cambio, una estimación eficiente ligeramente sesgada será preferible a otra no sesgada pero menos eficiente. Un simple diagrama ayudará a comprender el porqué de ello. En la figura XII.1, la curva puntiaguda ligeramente sesgada sería preferible, ya que, pese a que a la larga propenderíamos a subestimar el parámetro en una pequeña cantidad, tenemos con todo mayor probabilidad, en el caso de una prueba dada, de obtener una estimación de la muestra relativamente vecina del parámetro. El hecho de que sepamos que a la larga las estimaciones se irán promediando hacia la cifra correcta no nos sirve de gran consuelo si, en relación con una muestra determinada cualquiera, la aproximación tiene probabilidades de apartarse mucho del parámetro.

XII.2. Estimación del intervalo

El lector recordará que, cuando estudiaba física elemental, se le invitaba a pesar un pedazo de madera varias veces y a buscar luego el valor medio e indicar el margen de error posible. Así pudo, por ejemplo, haber indicado que el peso del pedazo de

¹ Aquí la media y la mediana de la población serían idénticas.

² No siempre es cierto que la media constituya la estimación más eficiente, aunque para la mayoría de las poblaciones, sobre todo si la desviación respecto de la normalidad no es demasiado grande, sea efectivamente así. Obsérvese que la cuestión acerca de la eficiencia relativa es totalmente distinta de la cuestión acerca de cuál medida sea la medida descriptiva más apropiada de la tendencia central. Esta última, en efecto, sólo se refiere al problema de hallar la medida singular mejor para representar los datos de la muestra.

madera era de 102 ± 2 gramos, significando que consideraba que el peso verdadero se situaba en algún lugar entre 100 y 104 gramos. Al proceder así, el lector admitía la posibilidad de error de la medición e indicaba qué grado de confianza tenía en la exactitud obtenida. Pese a que en dicho momento no se llamara expresamente su atención al respecto, el lector también habría admitido que no estaba absolutamente cierto que el verdadero valor estuviera comprendido en el intervalo obtenido. Sin embargo, si éste se hubiera ampliado, aquél habría estado más seguro de que sí estaba comprendido en el nuevo intervalo. Así, por ejemplo, habría estado prácticamente seguro de que el valor verdadero había de situarse entre 98 y 106 gramos, y se habría jugado hasta el último centavo en favor de que se encontraba entre 2 y 202 gramos. Al obtener apreciaciones de intervalo en relación con parámetros, hacemos esencialmente lo mismo que hace el físico, con la diferencia que estaremos en mejores condiciones de estimar la probabilidad exacta de error.

El procedimiento efectivo empleado para obtener una estimación de intervalo, o lo que se designa como *intervalo de confianza*, es muy sencillo y no comporta idea básica realmente nueva alguna. Estableceremos primero simplemente cómo el intervalo se obtiene, procediendo luego a examinar por qué se construye de este modo. Se decide primero acerca del riesgo de error que se está dispuesto a asumir al afirmar que el parámetro se sitúa en algún punto al interior del intervalo si en realidad no es así. Digamos que se decide estar dispuesto a admitir que se está equivocado el .05 de las veces, lo que suele designarse como *intervalo de confianza del 95 por ciento*.³ El intervalo se obtiene apartándose en ambas direcciones de la estimación del punto (v.gr. la media de la muestra) cierto múltiple de errores estándar correspondiente al nivel de confianza elegido. Así, por ejemplo, para apreciar la media μ de la población, obtenemos un intervalo como sigue (sirviéndonos del nivel de 95 por ciento):

$$\bar{X} \pm 1.96 \sigma_{\bar{X}} = \bar{X} \pm 1.96 \frac{\sigma}{\sqrt{N}}$$

en donde 1.96 corresponde a la región crítica de la curva normal, sirviéndonos del nivel de .05 y de una prueba de dos colas. Si $\bar{X} = 15$, $\sigma = 5$, y $N = 100$, el intervalo de confianza sería:

$$15 \pm 1.96 \frac{5}{\sqrt{100}} = 15 \pm 0.98$$

³ Obsérvese que en el caso de intervalos de confianza nos referimos a la unidad menos la probabilidad de error. Esto indica que tenemos "confianza" de estar en lo cierto el 95 por ciento, por ejemplo, de las veces.

En otros términos: el intervalo iría de 14.02 a 15.98.⁴

Con objeto de interpretar los intervalos obtenidos con dicho método, necesitamos volver a lo que sabemos acerca de la distribución de muestreo, en este caso la de la media. Supongamos que tenemos una distribución de muestreo normal con una media de μ y una desviación estándar de σ/\sqrt{N} . Para nuestros pro-

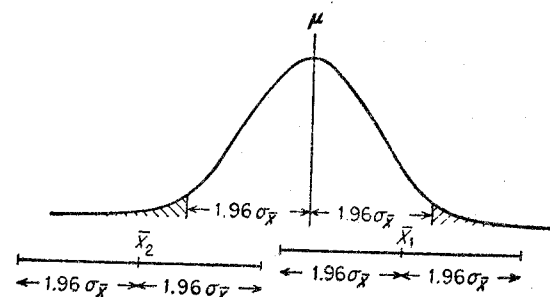


FIG. XII.2. Comparación de intervalos de confianza con la distribución de muestreo de la media, mostrando por qué los intervalos de confianza del 95 por ciento comprenden μ el 95 por ciento de las veces

pósitos hay dos clases de medias de una muestra, a saber: 1) las que no caen en la región crítica, y 2) las que sí caen en ella. Supongamos primero que hemos obtenido una \bar{X} (\bar{X}_1 de la figura XII.2) que no cae en la región crítica. Sabemos que una \bar{X} semejante ha de quedar en el interior de $1.96 \sigma_{\bar{X}}$ de μ . Si ponemos un intervalo a ambos lados de esta \bar{X} , apartándonos de ella en $1.96 \sigma_{\bar{X}}$ en ambas direcciones, debemos cruzar frente a μ , la media de la distribución de muestreo, tanto si \bar{X} está a la derecha como a la izquierda de μ . Y en forma análoga, si la \bar{X} obtenida queda al interior de la región crítica (véase \bar{X}_2 en la figura XII.2), entonces esta \bar{X} quedará a mayor distancia de 1.96 errores estándar de la μ , y el intervalo de confianza no llegará hasta ésta. Pero sabemos también que el 95 por ciento de las veces obtendremos \bar{X} que no caen en la región crítica, y sólo un 5 por ciento de las veces \bar{X} que caen en ella. En otros términos: sabemos que sólo un 5 por ciento de las veces obtendremos con este procedimiento intervalos que no comprendan el parámetro (v.gr. μ). El 95 por ciento restante de las veces el procedimiento nos dará

⁴ Estos puntos terminales del intervalo se designan a menudo como *límites de confianza*.

medias de una muestra lo suficientemente vecinas del parámetro para que los intervalos de confianza obtenidos comprendan efectivamente a éste.

Al interpretar los intervalos de confianza conviene tener presentes las siguientes advertencias. El estudiante principiante tiende a servirse de frases vagas por el estilo de "tengo un 95 por

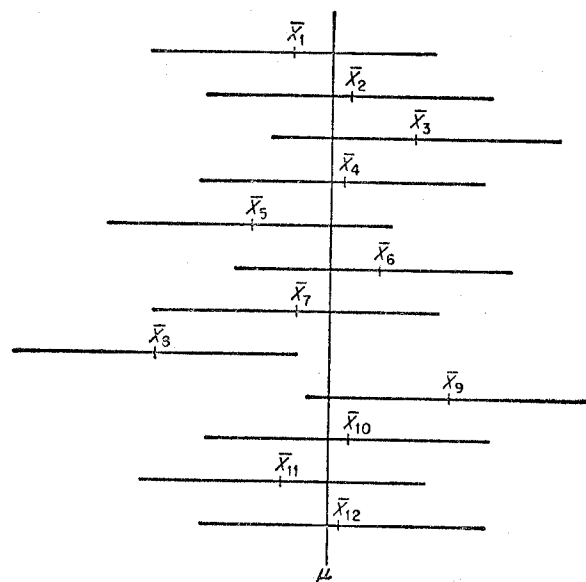


FIG. XII.3. Distribución de intervalos variables de confianza con respecto a un valor fijo del parámetro μ

ciento de confianza en que el intervalo contiene el parámetro", o bien "la probabilidad de que el parámetro quede en el interior del intervalo es de .95". Al hacerlo así puede no darse cuenta cabal de que el parámetro es un valor fijo y que son los intervalos los que varían de una muestra a otra. De acuerdo con nuestra definición de la probabilidad, la probabilidad de que el parámetro quede en el interior de un intervalo determinado cualquiera es cero o uno, ya que el parámetro está o no está en el interior del intervalo obtenido. Un simple diagrama indicando el valor fijo del parámetro, en este caso μ , y la variabilidad de los intervalos ayudará a comprender más claramente la interpretación correcta. La figura XII.3 pone de manifiesto que nuestra confianza está más bien en el procedimiento utilizado que en algún intervalo cualquiera. Podemos decir que el procedimiento es tal que, a la larga, el 95 por ciento de los intervalos obtenidos

comprenderán el verdadero parámetro (fijo). El lector ha de guardarse de concluir o suponer que el intervalo particular obtenido posee alguna propiedad especial que no poseen otros intervalos comparables obtenidos de otras muestras. Algunas veces se afirma que, si se extrajeran muestras reiteradas, el 95 por ciento de las veces las medias de estas muestras caerían en el interior del intervalo de confianza que se ha calculado (por ejemplo, 15 ± 0.98). Esto implica, por supuesto, que la \bar{X} obtenida en la muestra del investigador es igual exactamente a μ o es, por lo menos, una aproximación muy cercana a μ . En realidad, sin embargo, el intervalo particular obtenido puede ser tal que sólo unas pocas \bar{X} caigan en su interior. Nuestra confianza, como sucede siempre en la inducción estadística, no está en algún resultado particular cualquiera de la muestra, sino en el procedimiento empleado.

Es posible poner el riesgo de error a cualquier nivel deseado, sirviéndonos para ello del múltiple apropiado del error estándar. Sin embargo, el lector ha de observar que, al reducir el riesgo de error, se aumenta también necesariamente la amplitud del intervalo, a menos que se aumente simultáneamente el número de casos. Cuanto más amplio sea el intervalo, tanto menos nos dice acerca del parámetro. Decir que el ingreso mediano de las familias de Nueva York se sitúa entre \$ 1 000 y \$ 25 000 equivale a proclamar lo que todo el mundo sabe. Así, pues, el investigador se enfrenta a un dilema. Puede afirmar que el parámetro se sitúa en el interior de un intervalo muy angosto, pero la probabilidad de error será grande, o puede hacer una afirmación muy vaga, con la seguridad virtual de estar en lo cierto. Lo que se decida a hacer exactamente dependerá del carácter de la situación. Aunque convencionalmente suelen emplearse intervalos de confianza del 95 y el 99 por ciento, cabe insistir en que dichos niveles nada tienen de sagrado.

Intervalos de confianza y pruebas de hipótesis. Si bien el objeto explícito de poner intervalos de confianza de una estimación está en indicar el grado de exactitud de ésta, los intervalos de confianza constituyen también pruebas implícitas de una vasta serie de hipótesis.⁵ Son pruebas implícitas en el sentido de que las hipótesis concretas no se formulan, sino que se hallan simplemente implicadas en aquéllos. En efecto, en el intervalo de confianza tenemos una prueba implícita de todo valor posible de μ que pueda suponerse. La figura XII.4 indica de qué modo los intervalos de confianza se relacionan con las pruebas de las hipótesis.

Concentrémonos en el intervalo de confianza trazado alrededor

⁵ Conviene insistir en que, si bien la estimación de intervalo y la prueba de hipótesis comportan ideas íntimamente relacionadas, constituyen, con todo, procedimientos distintos.

de \bar{X} . Supóngase que, en lugar de haber obtenido semejante intervalo, hemos formulado hipótesis a propósito de varios valores alternativos de μ y hemos procedido a comprobarlas. Supóngase, para mayor sencillez, que el valor de σ ha sido dado y que se ha empleado el nivel de significación de .05 y se utilizó una prueba

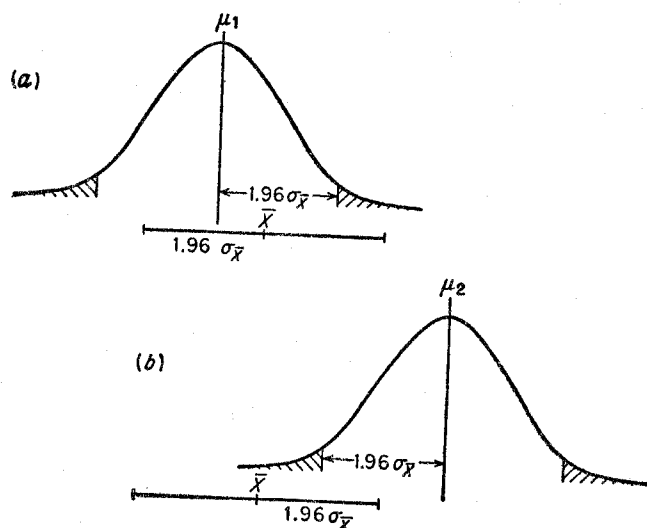


FIG. XII.4. Comparación de un intervalo de confianza del 95 por ciento con pruebas de hipótesis al nivel de .05 mostrando el no rechazo de la media hipotética μ_1 , que queda dentro del intervalo, y el descarte de la μ_2 hipotética, que queda fuera del intervalo

de dos colas. Supóngase primero que habíamos anticipado hipotéticamente un valor como el de μ_1 (figura XII.4a), que se sitúa efectivamente al interior del intervalo de confianza en cuestión. En este caso, la media \bar{X} de la muestra no caería, manifiestamente, en la región crítica, y la hipótesis no se habría descartado al nivel de .05. Por otra parte, si hubiéramos supuesto un valor al exterior del intervalo, tal como μ_2 (figura XII.4b), la distancia entre la μ_2 de la hipótesis y \bar{X} sería mayor que $1.96 \sigma_{\bar{X}}$, y esta segunda hipótesis se habría descartado. Está claro, pues, que si supusiéramos valores hipotéticos de μ que se sitúan en algún punto al interior del intervalo de confianza, no descartaríamos dichas hipótesis al nivel de significación apropiado. Y si supusiéramos valores de μ que quedan al exterior del intervalo, sabemos que estas hipótesis se descartarían.

Así, pues, habiendo obtenido un intervalo de confianza, podemos decir a simple vista cuáles habrían sido los resultados si hubiéramos verificado las hipótesis. Si el carácter de nuestro problema fuera tal que no se sugiriera hipótesis particular alguna como preferible a las otras, entonces, obviamente, la alternativa práctica, en relación con una serie de pruebas, consistiría en obtener un solo intervalo de confianza.⁶ El lector debería convencerse por sí mismo de que los ejemplos examinados en el capítulo precedente pudieron haberse tratado con igual facilidad por el método del intervalo de confianza.

Supuestos en relación con los intervalos de confianza. El empleo de intervalos de confianza no nos libera de la necesidad de formular supuestos acerca de la naturaleza de la población y del método de muestreo utilizado. Básicamente, los supuestos en el caso de un problema de intervalo de confianza son los mismos que los que se requieren para cualesquier pruebas implícitas, con la diferencia de que no es necesario, por supuesto, suponer un valor hipotético determinado para el parámetro que se estima. En este texto supondremos siempre el muestreo aleatorio. Por otra parte, si se emplea una distribución de muestreo, hemos de suponer una población normal o de tener una muestra suficientemente grande. Si se emplea, en cambio, una distribución t o cualquier otra distribución de muestreo, entonces habría que formular los supuestos usuales requeridos en pruebas comparables.

XII.3. Intervalos de confianza para otros tipos de problemas

Hasta aquí el examen de los intervalos de confianza sólo ha comprendido casos en los que el parámetro que se estimaba era la media de una población siendo σ conocida. Si el problema se cambia, las modificaciones del procedimiento son obvias, y la interpretación básica de los intervalos de confianza y su relación con las pruebas de las hipótesis siguen siendo las mismas. El intervalo de confianza de un parámetro se obtiene siempre procediendo a una estimación del parámetro en cuestión e incluyéndolo en un intervalo cuya amplitud es función del error estándar de la estimación.⁷

Si debido a no conocerse σ hay que servirse de la distribución t , recurrimos simplemente a la estimación del error estándar

⁶ Debe observarse, sin embargo, que cuando probamos una determinada hipótesis nula obtenemos un valor específico de probabilidad, tal como $P = .032$, lo que normalmente no obtendríamos en relación con un intervalo de confianza.

⁷ En algunos casos, sin embargo, como en el de los intervalos de confianza para coeficientes de correlación, la estimación del punto puede no caer exactamente en el centro del intervalo.

y sustituimos el múltiplo obtenido sirviéndonos del cuadro normal por la cifra correspondiente del cuadro *t*. En esta forma, para un intervalo de confianza de 99 por ciento para la media y 24 grados de libertad, tendríamos:

$$\bar{X} \pm 2.797 \hat{\sigma}_{\bar{X}} = \bar{X} \pm 2.797 \frac{s}{\sqrt{N-1}}$$

Si en el ejemplo de la sección XI.3 del capítulo anterior se hubiera operado con un intervalo de confianza del 99 por ciento, el resultado habría sido:

$$52 \pm 2.797 \left(\frac{12}{\sqrt{24}} \right) = 52 \pm 6.85$$

Por consiguiente, el intervalo de confianza del 99 por ciento va de 45.15 a 58.85. Vemos que este resultado concuerda con el que se obtuvo previamente (esto es, $.001 < p < .01$), por cuanto la μ supuesta de 60 cae efectivamente fuera del intervalo calculado, y sabemos, por lo tanto, que la hipótesis habría debido descartarse al nivel de .01 (en una prueba de dos colas).

Y en forma análoga, podemos obtener intervalos de confianza para las proporciones. En efecto, sustituyendo \bar{X} por p_u y σ/\sqrt{N} por $\sqrt{p_u q_u/N}$, el intervalo de confianza del 95 por ciento sería:

$$p_u \pm 1.96 \sqrt{\frac{p_u q_u}{N}}$$

Nos encontramos aquí con una dificultad que no se presentaba cuando podía anticiparse para p_u un valor determinado. En efecto, como quiera que obviamente p_u no será conocido, se hace necesario apreciar el error estándar. Pueden recomendarse a tal objeto dos procedimientos sencillos, uno de los cuales es más conservador que el otro.⁸ En primer lugar, toda vez que la magnitud de la muestra ha de ser grande para justificar el uso de las tablas normales, p_u constituirá por lo regular una apreciación razonablemente buena de p_u . Por consiguiente, si sustituimos simplemente p_u por p_u (y q_u por q_u), podemos obtener un intervalo que por lo regular será bastante parecido al correcto. Así, por ejemplo, en el caso de la sección XI.4 del capítulo precedente

⁸ Para un tercer método algo más preciso véase p. 244.

habríamos obtenido el intervalo de confianza del 98 por ciento de la manera siguiente:

$$p_u \pm 2.33 \sqrt{\frac{p_u q_u}{N}} = .55 \pm 2.33 \sqrt{\frac{(.55)(.45)}{125}} = .55 \pm 0.1037$$

Si alguien siente preocupación en utilizar una estimación del error estándar sin corregir en alguna forma el error adicional de muestreo introducido de este modo, puede servirse de un método más conservador para obtener el intervalo. En efecto, como quiera que el producto pq alcanza un valor máximo para $p = q = .5$, síguese que el intervalo de confianza más extenso posible se obtendrá sirviéndose del valor .5 como estimación de p_u .⁹ Toda vez que por lo regular se desea un intervalo angosto, nos comportamos cautamente al obtener un intervalo que es todo lo grande que pueda ser, independientemente del valor de p_u . Sirviéndonos de este método más conservador, obtenemos un intervalo algo distinto, a saber:

$$.55 \pm 2.33 \sqrt{\frac{(.5)(.5)}{125}} = .55 \pm 0.1042$$

Obsérvese que este segundo intervalo es sólo ligeramente más ancho que el primero. Siempre que $.3 \leq p \leq .7$, los dos métodos darán aproximadamente los mismos resultados.

* Si p_u resulta ser muy grande o muy pequeño, el método conservador puede dar un intervalo que sea muy amplio. Si alguien siente preocupación en usar el primer método en el que p_u es estimada por p_u , es posible combinar los dos métodos para conseguir un intervalo más razonable que siga, con todo, siendo conservador. En tal caso, nos servimos primero del método más conservador para obtener un intervalo de confianza aproximado. Supóngase que este intervalo vaya de .10 a .25, siendo p_u de .175. Estaremos entonces razonablemente seguros de que el valor real de p_u ha de situarse en algún punto al interior de dicho intervalo aproximado (y conservador). Al calcular el intervalo más exacto, tomamos ahora como estimación de p_u el valor aproximado dentro del intervalo que quede más cerca de .5. En el ejemplo numérico anterior escogeríamos el valor .25, ya que el empleo del mismo en la fórmula del error estándar nos dará un intervalo más amplio de lo que haría cualquier otro valor del intervalo .10 a .25. En otros términos: en lugar de servirnos de nuestro p_u real (esto es, de .175), escogemos el mayor valor que suponemos que p_u pueda adoptar. Por consiguiente, calculamos el intervalo de confianza del 95 por ciento como sigue:

⁹ El lector debería convencerse por sí mismo de que es así.

$$.175 \pm 1.96 \sqrt{\frac{(.25)(.75)}{N}}$$

Este intervalo será mayor, y por consiguiente más conservador, que el que se obtiene tomando bajo el radical a p , y no comporta, con todo, el empleo del valor .5, del que sospechamos que es con mucho demasiado grande.

XII.4. Determinación del tamaño de la muestra

De acuerdo con la práctica seguida de ir introduciendo pocas ideas nuevas a la vez, hemos aplazado la cuestión de cómo pueda determinarse el tamaño de la muestra con anterioridad a la recopilación de los datos. Una de las preguntas que con mayor frecuencia le ponen al estadígrafo es la de "¿cuántos casos necesito?" La respuesta depende, por supuesto, de lo que se tenga el propósito de hacer con los resultados de la muestra. Más concretamente: hay que determinar diversos hechos antes de poder dar una respuesta adecuada. Por lo regular, lo que hemos de hacer es remontarnos hacia atrás, a partir de los datos que esperamos obtener, para poder determinar el tamaño desconocido de la muestra. Hasta aquí hemos considerado el tamaño de la muestra como cantidad conocida. Las estadísticas tales como la media y la desviación estándar de la muestra pueden obtenerse de los resultados de ésta. Una vez que hemos decidido el nivel de significación de una prueba o el intervalo de confianza deseado, podemos poner todos estos valores en una fórmula y decidir la amplitud del intervalo de confianza, o bien si deba o no descartarse una hipótesis nula. Sin embargo, en la clase de problemas que estamos considerando en esta sección el tamaño de la muestra será desconocido. Esto significa que, para resolver nuestra ecuación con respecto a N , hemos de conocer todas las demás cantidades de la fórmula. Y una vez integrados todos esos valores en la ecuación, entonces la solución de N se convierte en un sencillo problema algebraico. Con objeto de ilustrar el proceso nos serviremos de un problema de intervalo de confianza.

Supóngase que queremos saber cuántos casos se requieren para estimar el número promedio de años de escuela completados por las personas de padres nacidos en el extranjero. Antes de poder dar una respuesta a esta cuestión, necesitamos obtener los siguientes elementos de información: 1) el nivel de confianza a utilizar, 2) el grado de exactitud con que deseamos apreciar el parámetro, y 3) alguna estimación razonable de los valores de cualesquier parámetros que puedan aparecer en la fórmula. Así, por ejemplo, podemos querer apreciar la media

con una aproximación de $\pm .1$ año de escolaridad y servirnos de un intervalo de confianza del 95 por ciento. Obsérvese que hay que concretar tanto la una como la otra de dichas cantidades, ya que podemos obtener siempre una aproximación de $\pm .1$ año si estamos dispuestos a admitir un gran riesgo de error. Nos serviremos ahora de estos valores en la fórmula del intervalo de confianza:

$$\bar{X} \pm 1.96 \underbrace{\frac{\sigma}{\sqrt{N}}}_{.1}$$

El conocimiento del nivel de confianza deseado nos ha permitido introducir el valor 1.96. Como quiera que deseamos una aproximación de $\pm .1$, o una amplitud total de intervalo de .2, sabemos que la cantidad de $1.96 \sigma / \sqrt{N}$ ha de ser igual a .1. Aunque el valor de \bar{X} sea desconocido, vemos inmediatamente que ello no reviste importancia alguna en este problema, ya que deseamos obtener un intervalo de cierta amplitud, independientemente del valor de \bar{X} .

Supóngase que tratamos ahora de resolver la ecuación

$$.1 = 1.96 \frac{\sigma}{\sqrt{N}}$$

respecto de N . Tenemos todavía una incógnita, o sea σ . Pero, ¿cómo podemos obtener σ antes de haber reunido los datos? La cosa es clara: hemos de estimar su valor por algún método que, en cierto sentido, vaya más allá de los datos que habremos de reunir. En esencia, hemos de formular un supuesto ilustrado en cuanto a su valor, ya sea sirviéndonos de un conocimiento experto, de los resultados de estudios previos, o eventualmente de un estudio-guía de cualquier clase que sea. Por lo regular, un estudio-guía resultará muy costoso y, por consiguiente, hay que remitirse a uno u otro de los dos métodos restantes. Incuestionablemente, el procedimiento más satisfactorio consistiría en determinar σ exactamente, pero, si esto pudiera hacerse, ya no tendría probablemente interés alguno el extraer una muestra. Obsérvese que el tipo de estimación necesario en esta clase de problema es totalmente distinto del que se empleó en apreciar σ a partir de los datos de la muestra. Por lo tanto, de nada sirve estimar σ con $\hat{\sigma}$ o sirviéndonos de la distribución t . Si de todos modos hemos de estimar, lo mismo podríamos estimar el valor de σ que el de σ o s . En el presente ejemplo, supóngase que so-

bre la base de la mejor información obtenible estimamos que σ será de aproximadamente 2.5 años. Sirviéndonos de este valor y resolviendo respecto del tamaño requerido de la muestra, tenemos:

$$.1 = 1.96 \frac{2.5}{\sqrt{N}}$$

o sea

$$\sqrt{N} = \frac{1.96(2.5)}{.1} = 49$$

y

$$N = 2401$$

Obsérvese que hemos resuelto respecto de N pasando todas las cantidades excepto \sqrt{N} a un lado de la ecuación y simplificando. Finalmente elevamos al cuadrado ambos miembros de la ecuación para eliminar el radical.

Sin duda, sólo podemos obtener un valor aproximado para el tamaño deseado de la muestra, ya que los parámetros habrán de estimarse. No tendría ciertamente sentido alguno, por ejemplo, tomar exactamente los 2401 casos. Sin embargo, semejante aproximación nos dará por lo regular resultados mucho mejores que cualquier corazonada intuitiva acerca del número de casos necesario. En las aplicaciones prácticas solemos por lo regular estudiar más de una variable a la vez, lo que complica todavía la cosa considerablemente. Estamos también limitados, generalmente, por los recursos disponibles, y a menudo hemos de acomodarnos con cualquier grado de exactitud que podamos obtener. Pero aun así, será útil a menudo calcular el tamaño necesario de la muestra a título de guía de nuestro propósito de investigación.

Si bien la cuestión de determinar el tamaño de la muestra no se examinará en los capítulos siguientes en conexión con otros procedimientos estadísticos, el lector encontrará algunos ejercicios que le imponen la estimación de N a propósito de otras clases de problemas. En todos estos casos la aplicación es obvia, aunque en ocasiones haya que recurrir abundantemente al álgebra.

GLOSARIO

Intervalo de confianza
Grados de libertad
Eficiencia de la estimación
Estimación del intervalo
Estimación del punto
Estimación no sesgada

EJERCICIOS

1. Obténganse los intervalos de confianza para los ejercicios 3, 4 y 5 del cap. XI. ¿Concuerdan los resultados obtenidos con los de ejercicios anteriores? ¿Cómo se sabe? Respuesta al ejercicio 5, 47—.61.

2. Se toma una muestra aleatoria de 200 familias de una localidad y se encuentra que en el 36 por ciento de los casos es el marido quien toma más de la mitad de las decisiones de carácter financiero. ¿Cuál es el intervalo de confianza del 99 por ciento para el porcentaje de familias en las que el marido toma más de la mitad de dichas decisiones? ¿En qué sentido concreto proporciona el intervalo pruebas implícitas de hipótesis?

3. ¿Cuántos casos se necesitarán para establecer un intervalo de confianza del 99.9 por ciento para la media si la amplitud total del intervalo de confianza no ha de rebasar \$ 500 y la desviación estándar se supone ser de \$ 1300? Respuesta, $N = 295$.

4. Si se sospecha que la proporción de propietarios de casa es de aproximadamente .75 en una determinada zona de residencia, ¿cuántos casos se necesitarán para obtener un intervalo de confianza del 95 por ciento y de amplitud no mayor a .03, expresada en términos de proporciones? Supóngase que la proporción de los propietarios de casa se aprecia en .5, ¿cuántos casos se necesitarán en este supuesto?

5. Sirviéndonos del hecho de que para poblaciones normales la distribución de muestreo de la mediana presenta un error estándar de $1.253 \sigma/\sqrt{N}$, podemos situar un intervalo de confianza alrededor de la mediana. Supóngase que en el ejercicio 3 precedente se deseaba poner un intervalo de la misma amplitud alrededor de la mediana de la muestra. Sirviéndonos de la misma apreciación de la desviación estándar, ¿cuántos casos necesitaríamos? ¿Qué revela el resultado a propósito de las eficiencias relativas de la media y la mediana? Respuesta, $N = 463$.

* 6. Se ha sostenido que el intervalo de confianza del 95 por ciento representa una serie de pruebas implícitas de *dos colas* al nivel de .05. Explíquese por qué el intervalo de confianza del 95 por ciento no representa pruebas implícitas de *una cola* al nivel de .05.

BIBLIOGRAFÍA

1. Freund, J. E.: *Modern Elementary Statistics*, 3ª ed., Prentice-Hall, Inc., Englewood Cliffs, N. J., 1967, caps. 9 y 11.
2. Hagood, M. J. y D. O. Price: *Statistics for Sociologists*, Henry Holt and Company, Inc., Nueva York, 1952, caps. 15 y 16.
3. Hays, W. L.: *Statistics*, Holt, Rinehart and Winston, Inc., Nueva York, 1963, caps. 7 y 9.
4. Wallis, W. A. y H. V. Roberts: *Statistics: A New Approach*, The Free Press of Glencoe, Ill., Chicago, 1956, cap. 14.

CUARTA PARTE
ESTADÍSTICAS BIVARIADAS Y MULTIVARIADAS

XIII. PRUEBAS DE DOS MUESTRAS: DIFERENCIA DE LAS MEDIAS Y LAS PROPORCIONES

EN EL capítulo XI se examinaron pruebas que consideraban una sola muestra. Hallamos que dichas pruebas no eran muy prácticas para el sociólogo, ya que por lo regular no es posible encontrar una hipótesis suficientemente concreta para predecir un valor para μ o p_u . Sin embargo, cuando el interés se centra en *comparaciones* entre varias categorías de muestras, resulta innecesario concretar los niveles absolutos de uno u otro de los grupos. En lugar de ello, puede probarse sencillamente la hipótesis nula de que no existe entre ellos diferencia alguna. Así, por ejemplo, sería extremadamente difícil anticipar el nivel de ingreso de los negros en Detroit o el nivel de prejuicio de los blancos en esa ciudad. Sin embargo, supóngase que nos interesaba probar la hipótesis de que el ingreso promedio de los negros es el mismo que el de los blancos nacidos en el extranjero, o que los judíos tienen para los negros el mismo grado de prejuicio que los no judíos. Este último tipo de hipótesis lo reconsideraremos aquí.

En una ciencia social como la sociología, el interés propende a centrarse en establecer *relaciones* entre variables. Esto contrasta con el tipo de la encuesta que reúne datos y en la cual, según vimos, la estimación del punto y el intervalo de un solo parámetro puede revestir importancia primordial. Cuando se establecen comparaciones entre dos muestras, tenemos la clase más simple de problema en el que dos variables pueden referirse una a otra. Hasta aquí sólo nos hemos ocupado de una sola variable a la vez. Ésta es tal vez la razón principal de que las pruebas examinadas hasta el presente no hayan sido demasiado útiles para los sociólogos. En este capítulo vamos a ocuparnos de pruebas en las que una simple variable dicotómica puede ser referida a otra variable. Así, por ejemplo, al comparar a los judíos y los no judíos por lo que se refiere al prejuicio, relacionamos de hecho a éste con la religión. Y en forma análoga, podría quererse comparar los dos sexos con respecto a "otros aspectos" o desde el punto de vista de otras características relativas a la personalidad. Las comparaciones pueden establecerse asimismo entre un grupo de control y un grupo de experimento en el que se ha introducido alguna variable. En los capítulos siguientes se examinarán pruebas que comportan más de dos muestras.

XIII.1. Prueba de la diferencia de las medias

Con objeto de extender la prueba de las medias de una muestra única a una prueba en la que pueda establecerse una compara-

ción entre las medias de dos muestras, hemos de servirnos nuevamente del teorema del límite central. Un teorema importante, derivado, puede enunciarse como sigue: *si se extraen muestras independientes al azar, de los tamaños N_1 y N_2 respectivamente, de poblaciones que son respectivamente $Nor(\mu_1, \sigma_1^2)$ y $Nor(\mu_2, \sigma_2^2)$, la distribución de muestreo de la diferencia entre las dos medias de las muestras $(\bar{X}_1 - \bar{X}_2)$ será igual a $Nor(\mu_1 - \mu_2, \sigma_1^2/N_1 + \sigma_2^2/N_2)$.* Lo mismo que en el caso de muestras individuales, este teorema puede generalizarse en el caso de muestras grandes para abarcar cualesquier poblaciones de medidas μ_1 y μ_2 y de variancias σ_1^2 y σ_2^2 respectivamente. En efecto, a medida que N_1 y N_2 aumentan, la distribución de selección de $\bar{X}_1 - \bar{X}_2$ se aproxima a la normalidad, lo mismo que antes. Examinemos ahora este teorema más de cerca.

Se hace referencia a muestras aleatorias independientes. Esto significa que las muestras han de seleccionarse *independientemente una de otra*. El hecho de que la muestra sea al azar asegura independencia en el interior de ella, en el sentido de que el conocimiento de la marca del primer individuo seleccionado no nos ayuda a predecir la marca del segundo. Esto *no* es, con todo, lo que aquí se entiende por "muestras al azar independientes". En efecto, no sólo ha de haber independencia en el interior de cada muestra (asegurada por el hecho de la selección al azar), sino que ha de haberla además *entre* las muestras. Así, por ejemplo, las muestras no pueden aparearse, como sería eventualmente el caso entre grupos de control y grupos de experimento. Si se fueran a comparar, por ejemplo, los dos sexos, no podría utilizarse la prueba de la diferencia de las medidas en muestras compuestas de parejas de marido y mujer.

El requisito de que las muestras sean independientes una de otra es sumamente importante, aunque a menudo se lo pase por alto en la investigación, particularmente cuando se maneja con una muestra en grupo. Si la muestra en conjunto es estrictamente al azar, y si se comparan dos submuestras tomadas de una misma muestra aleatoria mayor, el supuesto de independencia entre las dos submuestras en cuestión tendrá lugar, ya que todos los casos de la muestra mayor se habrán seleccionado independientemente uno de otro. Por ejemplo: si se comparan varones con hembras, deberemos hacer un muestreo general de los varones y otro muestreo, seleccionado independientemente, de todas las hembras. Es decir: la selección de Bob Jones no tiene influencia ninguna en la probabilidad de que sea seleccionada Susie Smith.

Por lo regular, en la investigación social tomamos una sola muestra mayor, aunque con fines de análisis podamos considerar los datos como procedentes de diversas muestras indepen-

dientes. En la mayoría de los casos, el problema de la falta de independencia entre las muestras no se planteará, a menos que deliberadamente las hayamos apareado. Como pueden darse circunstancias en las que el diseño del muestreo no sea tan sencillo, deberá prestarse atención a la posibilidad de que no se satisfaga el supuesto de independencia entre las muestras.

En el teorema en cuestión se nos dice que si continuáramos a seleccionar indefinidamente, seleccionando cada vez dos muestras y estableciendo una gráfica de sus medias, la distribución de selección de esta *diferencia* entre medias sería normal o aproximadamente normal. El lector ha de tratar de representarse exactamente lo que aquí ocurre. Tenga presente que, como sociólogo, él sólo obtendrá en realidad dos muestras y una sola diferencia, en tanto que aquí tratamos de la distribución hipotética de todas las diferencias posibles. Como quiera que la distribución de muestreo es para una *diferencia* entre medias de muestras, la media de la distribución de muestreo está dada por la diferencia entre dos medias de población, más bien que por cualquiera de ellas separadamente. En el caso especial en que μ_1 y μ_2 sean iguales, la media de la distribución de muestreo será cero. Si $\mu_1 > \mu_2$, esperamos que la mayoría de las \bar{X}_1 será mayor que las \bar{X}_2 , correspondientes, y que la media de la distribución de selección será por consiguiente positiva. Por ejemplo, si $\mu_1 = 60$ y $\mu_2 = 40$, la distribución de $\bar{X}_1 - \bar{X}_2$ tendrá 20 como media o valor esperado.

No es en cambio tan fácil ver por qué la variancia habría de ser $\sigma_1^2/N_1 + \sigma_2^2/N_2$, o sea la *suma* de las variancias de la distribución de muestreo de las medias separadas. Es obvio que no podría emplearse una diferencia de variancias $\sigma_1^2/N_1 - \sigma_2^2/N_2$, ya que podría obtenerse, para la distribución de muestreo, cero o una variancia negativa. En cambio, la variancia $\sigma_1^2/N_1 + \sigma_2^2/N_2$ es *mayor* que cualquiera de las dos variancias σ_1^2/N_1 o σ_2^2/N_2 . ¿Por qué es esto así? Aunque no pueda darse una justificación completa de la fórmula sin recurrir al razonamiento matemático, puede, con todo darse cierto tipo de explicación intuitiva. Fundamentalmente, esperamos que el error estándar correspondiente a la diferencia de las medias sea mayor que cualquiera de los errores estándar separados, porque tenemos ahora dos fuentes de error, o sea una en cada muestra. Así, pues, la mitad de las veces las dos \bar{X} estarán en error en sentidos opuestos. Con fines de simplificación, supongamos que $\mu_1 = \mu_2$. En este caso, si \bar{X}_1 es mayor que μ_1 y \bar{X}_2 es mayor que μ_2 , el resultado de la sustracción será una cantidad grande positiva, porque los errores son en sentidos opuestos. Por ejemplo, si \bar{X}_1 es más grande en 20 que μ_1 y \bar{X}_2 es

menor en 15 que μ_2 , la diferencia resultante, $\bar{X}_1 - \bar{X}_2$ diferirá de $\mu_1 - \mu_2$ en 35, combinando, pues, los errores implicados. Y en forma análoga, si \bar{X}_1 es pequeña y \bar{X}_2 es grande, puede resultar una diferencia negativa sustancial. En otros términos: con mucha frecuencia obtendremos diferencias relativamente grandes entre las medias de las muestras, ya que cada media variará independientemente de la otra. En consecuencia, la distribución de muestreo de una diferencia tendrá una desviación estándar mayor que cualquiera de las distintas distribuciones de muestreo separadas.

* La fórmula para el valor esperado y la variancia de $\bar{X}_1 - \bar{X}_2$ puede ser deducida utilizando una vez más las expresiones correspondientes a las combinaciones lineales. Se recordará que si $Y = c_1X_1 + c_2X_2$, tendremos $E(Y) = c_1E(X_1) + c_2E(X_2)$, y $\sigma_Y^2 = c_1^2\sigma_{X_1}^2 + c_2^2\sigma_{X_2}^2$, a condición de que X_1 y X_2 sean independientes. Si hacemos ahora que Y represente una diferencia de medias, sustituyendo X_1 por \bar{X}_1 , y X_2 por \bar{X}_2 , haciendo $c_1 = 1$ y $c_2 = -1$, tendremos, como caso especial, los resultados

$$E(Y) = E(\bar{X}_1 - \bar{X}_2) = (1)E(\bar{X}_1) + (-1)E(\bar{X}_2) = \mu_1 - \mu_2$$

y

$$\sigma_Y^2 = (1)^2\sigma_{\bar{X}_1}^2 + (-1)^2\sigma_{\bar{X}_2}^2 = \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}$$

Obsérvese que si hubiésemos formado la suma de \bar{X}_1 y \bar{X}_2 , la expresión de la variancia para dicha cantidad hubiera sido la misma que la correspondiente a su diferencia. En el capítulo xvi estudiaremos otros tipos más complejos de comparaciones en las que se incluye una generalización de esta simple comparación de dos medias de muestras.

Vamos a ver ahora un ejemplo ilustrativo del empleo de la prueba de la diferencia de las medias. El caso de las σ conocidas no lo examinaremos, ya dicho problema es obvio y más bien poco práctico. Supondremos, pues, que las σ no se conocen. Consideraremos dos casos particulares: en el primero supondremos que $\sigma_1 = \sigma_2$, en tanto que en el segundo se supondrán dos σ desiguales. Es obvio que estos dos modelos comprenden todas las alternativas posibles.

Problema. Se establece una comparación entre dos tipos de distritos, o sea entre los predominantemente urbanos y los que son fundamentalmente rurales. Los distritos en cuestión se comparan en relación con el porcentaje de personas que votan por los demócratas en una elección presidencial, con los siguientes resultados:

Distritos urbanos

$$\begin{aligned} N_1 &= 33 \\ \bar{X}_1 &= 57\% \\ s_1 &= 11\% \end{aligned}$$

Distritos rurales

$$\begin{aligned} N_2 &= 19 \\ \bar{X}_2 &= 52\% \\ s_2 &= 14\% \end{aligned}$$

¿Presentan estos datos motivos razonables para suponer que existen diferencias significativas en las preferencias electorales de dichos dos tipos de distritos? Supóngase que éstos se han seleccionado al azar de una lista de todos los distritos del Far West, y que estudios previos han mostrado que las respectivas distribuciones de población son aproximadamente normales.

$$\text{Modelo 1: } \sigma_1 = \sigma_2$$

1. Supuestos

Nivel de medición: el porcentaje de votos democráticos es una escala de intervalo

Modelo: muestras aleatorias independientes
poblaciones normales, $\sigma_1 = \sigma_2 = \sigma$.

Hipótesis: $\mu_1 = \mu_2$.

El supuesto de normalidad puede abandonarse siempre que las N sean grandes (por ejemplo, ambas sobre 50). El supuesto $\sigma_1 = \sigma_2$ puede comprobarse efectivamente por medio de la prueba F que se examinará en el capítulo xvi. Esta prueba comporta una comparación de las dos desviaciones estándar de las muestras. Si s_1 y s_2 no difieren mucho no podrá rechazarse la hipótesis de que $\sigma_1 = \sigma_2$. Si de acuerdo con los resultados de la prueba F el supuesto de desviaciones estándar iguales es razonable, será más eficaz aprovecharse de ello para apreciar el valor común de σ . Dado el supuesto de que las dos poblaciones sean normales, los supuestos adicionales de medias y desviaciones estándar iguales equivalen a sostener que las dos poblaciones son idénticas.

Como quiera que estamos interesados en saber si existe o no alguna diferencia entre los dos tipos de distritos, nuestra hipótesis nula será la de que no existe diferencia. Por lo visto, sospechamos que sí existe diferencia, y por ello formulamos una hipótesis que deseamos descartar. En este caso podemos designar legítimamente la hipótesis como hipótesis "nula", que no indica relación entre las variables "tipo de distrito" y "preferencia electoral". Se concibe que hubiéramos podido estar en condiciones de concretar que las medias de la población se espera que sea alguna constante distinta de cero. Así, por ejemplo, las hipótesis pudieron haber adoptado la forma de $\mu_1 - \mu_2 = 10$, si se hubiera anticipado que la votación en favor de los demócratas sería un

10 % superior en los distritos urbanos. Sin embargo, en ciencias sociales estamos raramente en condiciones de poder concretar tanto.

2. *Distribución de muestreo.* Nos serviremos de la distribución t , ya que las σ no se conocen y que el número total de casos es muy inferior a 120.

3. *Nivel de significación.* Escojamos el nivel de .01 y una prueba de dos colas.

4. *Cálculo del estadístico de la prueba.* Se recordará que la distribución t se calcula tomando la diferencia entre el valor obtenido de la muestra y la media de la distribución de muestreo, y dividiendo entre una estimación del error estándar de esta distribución. Nos interesa aquí la *diferencia* entre las medias de la muestra, $\bar{X}_1 - \bar{X}_2$. Como quiera que la media de la distribución de muestreo es $\mu_1 - \mu_2$, obtenemos para t la siguiente expresión:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}} \quad (\text{XIII.1})$$

en donde $\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}$ es una estimación del error estándar de la diferencia entre las medias de las muestras. Como quiera que en la hipótesis nula se ha supuesto que $\mu_1 = \mu_2$, la expresión para t se convierte, en este *caso especial*, en

$$\frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}}$$

La semejanza entre el numerador anterior y el que utilizamos en la prueba de una sola muestra es más o menos casual, o sea resultado del hecho de que, en la hipótesis nula, las μ se eliminaron. Sin embargo, *no* debe sacarse la conclusión de que la μ del primer tipo de problema se ha remplazado simplemente por la media de la *muestra* de la segunda de éstas. En realidad, la expresión $(\bar{X}_1 - \bar{X}_2)$ ha remplazado a \bar{X} , $(\mu_1 - \mu_2)$ ha remplazado a μ , y $\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}$ ha remplazado a $\hat{\sigma}_{\bar{X}}$.

Nos falta ahora evaluar $\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}$. Sabemos, por supuesto, que

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$$

Como quiera que en este caso $\sigma_1 = \sigma_2$, podemos indicar el valor común como σ , sacarlo del radical, y simplificar la expresión de $\sigma_{\bar{X}_1 - \bar{X}_2}$ como sigue:

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma^2}{N_1} + \frac{\sigma^2}{N_2}} = \sigma \sqrt{\frac{1}{N_1} + \frac{1}{N_2}} = \sigma \sqrt{\frac{N_1 + N_2}{N_1 N_2}} \quad (\text{XIII.2})$$

La variancia común σ^2 puede evaluarse ahora obteniendo una *apreciación combinada* de ambas muestras. Como quiera que las dos variancias de las muestras se basarán por lo regular en números distintos de casos, podemos obtener una apreciación de σ^2 tomando un promedio ponderado de las variancias de las muestras, poniendo cuidado en dividir entre los grados propios de libertad, con objeto de conseguir una estimación insesgada. Extrayendo la raíz cuadrada, obtenemos la estimación de $\hat{\sigma}$ como sigue:

$$\hat{\sigma} = \sqrt{\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}}$$

Puesto que: $N_1 s_1^2 = \sum_{i=1}^{N_1} (X_{i1} - \bar{X}_1)^2$, podremos sustituir $N_1 s_1^2$ por:

$$\sum_{i=1}^{N_1} x_{i1}^2, \text{ en donde } x_{i1} = X_{i1} - \bar{X}_1.$$

Si hacemos lo mismo para $N_2 s_2^2$, obtenemos

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{N_1} x_{i1}^2 + \sum_{i=1}^{N_2} x_{i2}^2}{N_1 + N_2 - 2}}$$

De este modo, si tomamos la suma de los cuadrados alrededor de la media de la primera muestra y sumamos a ella la suma de los cuadrados de las desviaciones alrededor de la media de la segunda muestra, dividiendo finalmente entre $N_1 + N_2 - 2$, obtenemos una estimación combinada de la variancia común.

Obsérvese que el símbolo $\hat{\sigma}$ se emplea ahora para representar una estimación distinta de la que vimos en los capítulos anteriores. Para indicar una estimación insesgada se emplea a menudo en la literatura estadística el símbolo " \wedge ". Como quiera que hemos perdido 2 grados de libertad, uno en cada cálculo de s_1 y s_2 a partir de \bar{X}_1 y \bar{X}_2 , los grados totales de libertad quedan en $N_1 + N_2 - 2$. Para obtener nuestra estimación, nos hemos servido

de ambas muestras, dando un mayor peso a la variancia de la mayor de ellas. Semejante estimación combinada será más eficaz que las estimaciones basadas en una u otra sola de las muestras en cuestión. A título de control del cálculo, el valor numérico de $\hat{\sigma}$ se situará por lo regular entre los de s_1 y s_2 .

Finalmente, obtenemos una estimación de $\sigma_{\bar{x}_1 - \bar{x}_2}$ tomando nuestra estimación de σ y multiplicando por $\sqrt{\frac{N_1 + N_2}{N_1 N_2}}$ como en la ecuación (XIII.2). Así:

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}} \sqrt{\frac{N_1 + N_2}{N_1 N_2}} \quad (\text{XIII.4})$$

Obsérvese que la ecuación (XIII.4) se diferencia de la ecuación (XIII.2) en que el σ de la ecuación (XIII.2) ha sido sustituido por su estimado $\hat{\sigma}$, como se define en la ecuación (XIII.3). En este punto la fórmula parece terrible. Sin embargo, el lector debería repasar los pasos algebraicos examinados anteriormente, para convencerse de que la fórmula no es tan complicada como a primera vista parece.

En nuestro ejemplo numérico obtenemos los siguientes resultados:

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{33(121) + 19(196)}{33 + 19 - 2}} \sqrt{\frac{33 + 19}{33(19)}} = (12.42)(.288) = 3.58$$

Por lo tanto,

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}} = \frac{57 - 52}{3.58} = 1.40$$

Obsérvese que nuestro estimado $\hat{\sigma} = 12.42$ cae entre $s_1 = 11$ y $s_2 = 14$.

5. *Decisión.* Como quiera que se utilizó una estimación combinada de la desviación estándar común, los grados de libertad asociados a t serán $\bar{N}_1 + \bar{N}_2 - 2$, o sea 50. Encontramos que $t = 1.40$, cuya probabilidad sería considerablemente superior a .01 si todos los supuestos fueran correctos. Decidimos, pues, no descartar la hipótesis nula al nivel de .01, y llegamos en consecuencia a la conclusión de que no se dan diferencias electorales significativas entre los distritos urbanos y rurales del Far West.

Modelo 2: $\sigma_1 \neq \sigma_2$. Vemos ahora cuáles modificaciones resultan

necesarias cuando es imposible suponer que las dos poblaciones presentan las mismas desviaciones estándar. Probablemente habremos verificado y descartado la hipótesis de $\sigma_1 = \sigma_2$. En consecuencia, ya no es posible ahora simplificar la fórmula $\sigma_{\bar{x}_1 - \bar{x}_2}$ introduciendo un valor común para σ , ni lo es tampoco formar una estimación combinada. En semejante caso, estimamos las dos desviaciones estándar (distintas) separadamente. Estimamos σ_1^2/N_1 a partir de $s_1^2/(N_1 - 1)$, y σ_2^2/N_2 sobre la base de $s_2^2/(N_2 - 1)$, con lo que obtenemos:

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{N_1 - 1} + \frac{s_2^2}{N_2 - 1}} \quad (\text{XIII.5})$$

En el ejemplo empleado anteriormente tenemos, pues:

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{121/32 + 196/18} = \sqrt{3.78 + 10.89} = \sqrt{14.67} = 3.83$$

$$\text{Y por consiguiente, } t = \frac{57 - 52}{3.83} = 1.31.$$

Así, pues, los resultados obtenidos en los dos modelos distintos no difieren grandemente.

Si bien el procedimiento empleado en el modelo 2 es más sencillo desde los puntos de vista lógico y de cálculo a la vez, la estimación de $\sigma_{\bar{x}_1 - \bar{x}_2}$ no es, con todo, tan eficaz, en él, como la que se obtuvo anteriormente. Por otra parte, aun si suponemos poblaciones normales, el modelo 2 resulta algo dudoso en los casos en que las N no son muy grandes o en que los tamaños de las muestras difieren mucho una de otra. La dificultad se hace presente al escoger el grado adecuado de libertad. Así, por ejemplo, si la primera muestra fuera excepcionalmente pequeña, sería muy falaz servirse de $N_1 + N_2 - 2$ como grados de libertad, ya que s_1 sería una estimación muy deficiente de σ_1 , ya que el valor de $s_1^2/(N_1 - 1)$ sería por lo regular mucho mayor que el de $s_2^2/(N_2 - 1)$. Esto es cierto porque no siendo muy diferentes los valores de s_1^2 y s_2^2 , los tamaños relativos de las dos fracciones vendrán fundamentalmente determinados por sus denominadores. Se ha sugerido que, a menos que las N sean grandes, es preferible servirse de la siguiente expresión para obtener una aproximación de los grados correctos de libertad:

$$df = \frac{\left(\frac{s_1^2}{N_1 - 1} + \frac{s_2^2}{N_2 - 1}\right)^2}{\left(\frac{s_1^2}{N_1 - 1}\right)^2 \left(\frac{1}{N_1 + 1}\right) + \left(\frac{s_2^2}{N_2 - 1}\right)^2 \left(\frac{1}{N_2 + 1}\right)} - 2 \quad (\text{XIII.6})$$

En esta forma obtenemos en el ejemplo anterior:

$$df = \frac{(14.67)^2}{(3.78)^2(1/34) + (10.89)^2(1/20)} - 2 = 33.89 - 2 = 31.89 \approx 32$$

Obsérvese que algunas de las magnitudes de la fórmula de los grados de libertad ya se calcularon anteriormente. De la tabla t , sirviéndonos de 32 grados de libertad, vemos que la hipótesis nula no debería descartarse al nivel de .01.

Por lo que se refiere a los supuestos, la única diferencia entre los modelos 1 y 2 es el supuesto de que $\sigma_1 = \sigma_2$. Obsérvese que nada hay en el segundo procedimiento que requiera que las desviaciones estándar sean desiguales. Si ocurre que son iguales (o casi) el segundo modelo será sencillamente el más eficaz. Parecerá tal vez que el segundo procedimiento sea preferible en general, porque no requiere el supuesto de $\sigma_1 = \sigma_2$. Sin embargo, según acabamos de ver, este modelo necesita aproximaciones para los grados de libertad. En el caso de muestras grandes, los dos métodos proporcionarán por lo regular resultados similares, si las desviaciones estándar son efectivamente iguales, ya que las dos desviaciones estándar de las muestras serán, una y otra, buenas estimaciones de la σ común.

Si se da el caso de que las σ se conocen para ambas poblaciones, entonces sus respectivos valores pueden ponerse directamente en la fórmula de $\sigma_{\bar{x}_1 - \bar{x}_2}$, ya que no se requiere estimación alguna. Puede en este caso calcularse Z y utilizarse el cuadro normal. Con las σ conocidas, no habrá necesidad, por supuesto, de distinguir entre los modelos 1 y 2. Es obvio, sin embargo, que los casos en que ambas σ sean conocidas serán extremadamente raras en la investigación práctica.

XIII.2. Diferencia de proporciones

Lo mismo que en el caso de pruebas que comportan proporciones de una sola muestra, la diferencia entre dos proporciones puede tratarse como caso particular de la diferencia entre dos medias. Si comparamos dos muestras aleatorias, independientes, en relación con las proporciones de personas afectadas de prejuicios, podemos formular la hipótesis nula de que las proporciones p_{u_1} y p_{u_2} , respectivamente, de personas con prejuicios son iguales en las dos poblaciones. Como quiera que ya se demostró en el caso de proporciones que $\sigma_1 = \sqrt{p_{u_1}q_{u_1}}$ y $\sigma_2 = \sqrt{p_{u_2}q_{u_2}}$, síguese que las desviaciones estándar de las dos poblaciones han de ser iguales. Por lo tanto, el siguiente ejemplo se sirve esencialmente de los mismos procedimientos empleados en el primer modelo, en el caso de la prueba de diferencia de las medias.

Problema. Supóngase que se establece una comparación a propósito de los hábitos de recreación entre trabajadores de línea de ensamble y personas cuyo trabajo no consiste en una mera repetición ni se halla sujeto al ritmo de la máquina. Supongamos que el investigador sospecha que los trabajadores de línea de ensamble serán más propensos a escoger formas de recreación del tipo de espectador "pasivo". En una muestra aleatoria de 150 trabajadores de ensamble en una determinada fábrica se encuentra que el 57 por ciento dan preferencia a las formas de recreación pasivas. En una segunda muestra, seleccionada asimismo al azar, el 46 por ciento de los trabajadores, sobre 120, indican también preferencia por las formas de recreo pasivas. ¿Existe al nivel de .05 diferencia significativa alguna entre ambos grupos?

1. Supuestos.

Nivel de medición: el tipo de recreación como dicotomía

Modelo: muestreo al azar independiente

Hipótesis: $p_{u_1} = p_{u_2}$ (implica $\sigma_1 = \sigma_2$)

2. *Distribución de muestreo.* Como quiera que ambas N son relativamente grandes, la distribución de muestreo de las diferencias entre las proporciones será aproximadamente normal, con la media $p_{u_1} - p_{u_2} = 0$, y una desviación estándar de:

$$\sigma_{p_{s_1} - p_{s_2}} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} = \sqrt{\frac{p_{u_1}q_{u_1}}{N_1} + \frac{p_{u_2}q_{u_2}}{N_2}}$$

en donde q_{u_1} y q_{u_2} son iguales, respectivamente, a $1 - p_{u_1}$ y $1 - p_{u_2}$.¹

3. *Nivel de significación y región crítica.* El problema especifica que hemos de servirnos del nivel .05. Resulta indicada una prueba de una sola cola, ya que la dirección de la diferencia se anticipa. Por consiguiente, cualquier valor positivo superior a 1.65 indicará que los resultados son tan improbables, con dichos supuestos, que la hipótesis nula ha de descartarse.

4. *Cálculo de la estadística de la prueba.* Como quiera que por hipótesis tenemos $p_{u_1} = p_{u_2}$, síguese que $\sigma_1 = \sigma_2 = \sigma$, pudiendo emplearse la fórmula especial:

$$\sigma_{p_{s_1} - p_{s_2}} = \sigma \sqrt{\frac{N_1 + N_2}{N_1 N_2}}$$

Anteriormente, en la prueba de proporciones de una sola mues-

¹ Si las muestras son pequeñas, nos servimos de la prueba de Fisher, descrita en el capítulo xv.

tra, pudo prescindirse de la estimación de σ , ya que el valor de p_u se suponía. Ahora, en cambio, la hipótesis enuncia simplemente que $p_{u_1} = p_{u_2}$, pero sin especificar, con todo, cuál sea el valor real de estas proporciones. Esta es la razón de que necesitemos una estimación combinada del error estándar. En lugar de buscar un promedio ponderado de las dos variancias de las muestras, que es lo que hicimos antes, podemos obtener una estimación ligeramente menor, calculando una estimación combinada (\hat{p}_u) de p_u . Encontramos luego \hat{q}_u por sustracción. Ya que:

$$\sigma = \sqrt{p_u q_u}$$

podemos poner

$$\hat{\sigma} = \sqrt{\hat{p}_u \hat{q}_u}$$

Así, pues

$$\hat{\sigma}_{p_{s_1} - p_{s_2}} = \hat{\sigma} \sqrt{\frac{N_1 + N_2}{N_1 N_2}} = \sqrt{\hat{p}_u \hat{q}_u} \sqrt{\frac{N_1 + N_2}{N_1 N_2}} \quad (\text{XIII.8})$$

Con objeto de obtener \hat{p}_u , se toma un promedio ponderado de las proporciones de las muestras de la manera siguiente:

$$\hat{p}_u = \frac{N_1 p_{s_1} + N_2 p_{s_2}}{N_1 + N_2} \quad (\text{XIII.9})$$

Obsérvese que el numerador de esta expresión no es más que el número total de los individuos de ambas muestras que prefieren formas de recreación de tipo pasivo. Así, en el caso de nuestro ejemplo numérico, obtenemos:

$$\hat{p}_u = \frac{150(.57) + 120(.46)}{150 + 120} = .521$$

Por lo tanto, $\hat{q}_u = 1 - \hat{p}_u = .479$

$$\begin{aligned} \text{y} \quad \hat{\sigma}_{p_{s_1} - p_{s_2}} &= \sqrt{(.521)(.479)} \sqrt{\frac{150 + 120}{(150)(120)}} \\ &= (.4996)(.1225) = .0612 \end{aligned}$$

Y de ahí que

$$Z = \frac{(p_{s_1} - p_{s_2}) - 0}{\hat{\sigma}_{p_{s_1} - p_{s_2}}} = \frac{.57 - .46}{.0612} = 1.80$$

5. *Decisión.* Como quiera que con una prueba de una sola cola la probabilidad de obtener un valor de Z igual o mayor que 1.80 es de .036, siempre que la hipótesis nula sea efectivamente correcta, podemos descartar esta hipótesis al nivel de .05. Concluimos, pues, que existe una diferencia significativa en relación con la preferencia de tipos de recreación pasiva entre las dos clases de trabajadores de la fábrica considerada.

Hay que mencionar aquí que existen diversas clases alternativas de pruebas, la más importante de las cuales es la de la χ al cuadrado, que se examinará en el capítulo xv, que pueden utilizarse en lugar de la prueba de la diferencia de las proporciones. Como quiera que, en efecto, el empleo de la prueba de la diferencia de las proporciones está limitado a dos muestras y una variable dicotómica, ésta no resulta tan práctica como la prueba χ al cuadrado, que puede aplicarse lo mismo a tres o más muestras. Sin embargo, una de las ventajas de la prueba de la diferencia de las proporciones es que, mediante modificaciones adecuadas, se la puede utilizar en el caso de muestras de áreas o por conglomerados. Desafortunadamente, las modificaciones en cuestión no tienen cabida en el marco del presente texto.

* *Diferencia de diferencias de proporciones.* Podemos ampliar fácilmente el principio de una prueba para una diferencia de proporciones (o medias) hasta abarcar una diferencia de diferencias, o incluso una diferencia de diferencias de diferencias. Supongamos, por ejemplo, que tenemos datos relativos tanto a trabajadores como a trabajadoras, y que deseáramos comparar los sexos en orden a la *relación* entre los trabajos realizados y las preferencias recreacionales. Tal vez encontraríamos en el caso de los hombres una diferencia tal como la que acabamos de ilustrar, pero ninguna en el caso de las mujeres. O tal vez la dirección de la diferencia pueda resultar contraria entre ambos sexos. Ampliando esta ilustración podríamos desear agregar el dato relativo a las edades. En tal caso puede concebirse que tendríamos una diferencia de diferencias (entre hombres y mujeres) en el caso de los trabajadores jóvenes, y un resultado distinto para los trabajadores adultos. Puede observarse que estamos anticipando problemas que tal vez surjan cuando manejemos *más de dos variables*, y cuando las diferentes variables puedan causar peculiares efectos combinados. En tales casos se afirma que hay "interacción" entre las variables, o que sus efectos unidos son no aditivos. En los capítulos xvi y xx tendremos oportunidad de estudiar con más detalle estos tipos de posibilidades.

En el muy sencillo ejemplo en el que deseamos comparar las diferencias de proporciones entre hombres y mujeres, supongamos que p_{u_1} y p_{u_2} representan las proporciones de población para hombres, como en el anterior ejemplo. Tendremos entonces dos

proporciones semejantes, p_{u_3} y p_{u_4} que representarán las mujeres, y podríamos hacer una prueba similar de la hipótesis nula, tal como, para las mujeres $p_{u_3} = p_{u_4}$. Pero podemos probar asimismo la hipótesis más compleja de que las diferencias (de población) para los sexos son también idénticas. Nuestra hipótesis nula pasa así a ser

$$p_{u_1} - p_{u_2} = p_{u_3} - p_{u_4} \quad \text{o} \quad (p_{u_1} - p_{u_2}) - (p_{u_3} - p_{u_4}) = 0$$

Expresado de otra manera, estamos sentando la hipótesis de que la *relación* entre clase de trabajo y preferencias recreativas (medida por una diferencia de proporciones), es igual para ambos sexos. Una hipótesis alternativa podría consistir en que la diferencia es mayor entre los hombres que entre las mujeres.

Podemos utilizar de nuevo el principio de las combinaciones lineales, planteando

$$Y = c_1 p_{s_1} + c_2 p_{s_2} + c_3 p_{s_3} + c_4 p_{s_4}$$

En cuanto a la hipótesis nula que estamos considerando, haremos $c_1 = c_4 = 1$, y $c_2 = c_3 = 1$, resultando (siempre que se trate de muestras seleccionadas independientemente)

$$E(Y) = E(p_{s_1}) - E(p_{s_2}) - E(p_{s_3}) + E(p_{s_4}) = (p_{u_1} - p_{u_2}) - (p_{u_3} - p_{u_4})$$

y

$$\sigma_Y^2 = \frac{p_{u_1} q_{u_1}}{N_1} + \frac{p_{u_2} q_{u_2}}{N_2} + \frac{p_{u_3} q_{u_3}}{N_3} + \frac{p_{u_4} q_{u_4}}{N_4}$$

podemos ya formar Z , como sigue:

$$Z = \frac{(p_{s_1} - p_{s_2}) - (p_{s_3} - p_{s_4})}{\sqrt{\frac{p_{u_1} q_{u_1}}{N_1} + \frac{p_{u_2} q_{u_2}}{N_2} + \frac{p_{u_3} q_{u_3}}{N_3} + \frac{p_{u_4} q_{u_4}}{N_4}}}$$

y usar el cuadro normal en forma directa. Como el denominador contiene las incógnitas p_{u_i} y q_{u_i} , podemos estimárselas mediante las correspondientes p_{s_i} y q_{s_i} , fijando conservadoramente cada grupo como igual a .5.

Es importante advertir que la expresión para la variancia de Y comprende cuatro N_i diferentes, las que aparecen como denominadores en fracciones separadas. Como quiera que los productos $p_{u_i} q_{u_i}$ se encuentran normalmente cerca del valor .25, veremos

que el valor de cada fracción será primordialmente función del tamaño de la submuestra. En un terreno práctico, si hay una submuestra muy pequeña, ésta puede dominar la expresión correspondiente a la variancia de Y , y por tanto también al denominador de Z . De esta manera, y para lograr un máximo de eficacia, desearemos usar submuestras del mismo tamaño aproximado. Si una submuestra es muy pequeña, podrá no resultar significativa la prueba anterior, por razón de ser grande el denominador de Z , resultando además injustificada la aproximación normal.

Puede seguirse exactamente el mismo procedimiento en relación con las diferencias entre las medias, por ejemplo $(\bar{X}_1 - \bar{X}_2) - (\bar{X}_3 - \bar{X}_4)$. Habremos sin embargo de aplazar este asunto hasta que en el capítulo XVI abordemos las comparaciones generales entre k medias.

XIII.3. Intervalos de confianza

En el caso de problemas de una sola muestra, ya vimos que la construcción de un intervalo de confianza constituye a menudo un procedimiento más práctico que la verificación de las hipótesis. En la investigación social, sin embargo, los intervalos de confianza raramente se emplean como alternativas de pruebas de dos muestras. La razón de ello reside en que nos interesa por lo regular establecer la existencia de una relación entre dos variables, esto es, de una diferencia significativa. En tanto que interesa menos, en cambio, la magnitud efectiva de la diferencia en cuestión. El sociólogo, en efecto, raramente trata de sacar la conclusión de que la diferencia entre dos medias se sitúa entre 17 y 28, por ejemplo. Por lo regular, se da por satisfecho si encuentra alguna diferencia significativa cualquiera. Este hecho revela indudablemente la falta de madurez de las ciencias sociales y la preponderancia de los estudios exploratorios. Es posible que, a medida que las hipótesis se vayan haciendo más precisas, aumente también la necesidad de los intervalos de confianza en los problemas de dos muestras.

El procedimiento empleado para el establecimiento de intervalos de confianza es una extensión directa del que se examinó anteriormente. Se toman simplemente los resultados de las muestras, en este caso una *diferencia* entre sus medias, y se sitúa un intervalo alrededor de $\bar{X}_1 - \bar{X}_2$, que sea un múltiplo adecuado del error estándar. Así, por ejemplo, si se deseaba un intervalo de confianza del 95 por ciento, lo obtendríamos como sigue:

$$(\bar{X}_1 - \bar{X}_2) \pm 1.96 \sigma_{\bar{X}_1 - \bar{X}_2}$$

Si se requiriera una estimación del error estándar y de la distribución t , la fórmula se modificaría de la manera usual.

XIII.4. Muestras dependientes: pares asociados

En ocasiones resulta ventajoso concebir un estudio en el que las muestras no sean independientes una de otra. Uno de los tipos más comunes de los problemas de esta clase es aquel en que los casos de las dos muestras se han asociado por pares. Puede haber grupos de control y grupos experimentales, en los que los miembros se hayan apareado desde el punto de vista de algunas características importantes. O puede emplearse un simple esquema "antes y después", en el que las mismas personas se comparan antes y después de haberse introducido alguna variable experimental. En este último caso, las "dos" muestras constan de los mismos individuos. Es obvio que semejantes muestras no son independientes una de otra. En efecto, el conocimiento de las marcas de los primeros miembros de cada par (primera muestra) ayudaría a predecir las de los segundos. De hecho, el objeto del apareamiento, o de servirse dos veces de los mismos individuos, consiste en controlar las más variables posibles, aparte de la experimental. Se persigue hacer las dos muestras lo más iguales posibles, o sea mucho más que si se hubieran seleccionado independientemente.

En relación con semejantes problemas, el investigador podría verse tentado a usar una prueba de diferencia de las medias. Sin embargo, habría de ser obvio que este procedimiento no estaría justificado, ya que no tenemos $2N$ casos (N en cada muestra) que se hayan seleccionado independientemente. Como quiera que, en efecto, las muestras se han apareado deliberadamente, cualesquier peculiaridades de las muestras se darán probablemente lo mismo en la otra. En realidad, sólo se tienen N casos independientes, siendo cada "caso" un par de individuos, uno de cada uno de las muestras. Por consiguiente, si tratamos cada pareja de individuos, como un solo caso, podemos legítimamente proceder a efectuar pruebas estadísticas, a condición que se cumplan los demás supuestos requeridos. En lugar de efectuar una prueba de diferencia de las medias, podemos proceder por comparación directa por pares, obteniendo una marca de diferencia para cada par. Si nos servimos de la hipótesis nula de que no existe diferencia alguna entre las dos poblaciones, suponiendo así que la variable experimental no produce efecto alguno, podemos establecer simplemente la hipótesis de que la media de las diferencias por pares (μ_D) es cero. El problema se reduce así a una verificación de una sola muestra de la hipótesis $\mu_D = 0$.

Problema. Supóngase que un grupo de acción se propone influir a los electores urbanos para que voten en favor de unas

propuestas de viviendas populares en las próximas elecciones. Se aparean cuidadosamente las ciudades del Estado en relación con variables que se suponen ser significativas, y se emplean dos métodos distintos de ejercer influencias sobre los electores. El método del grupo A comporta un procedimiento indirecto consistente en influir sobre los elementos directivos de las ciudades, pero sin apelar directamente a la masa. En las ciudades del grupo B, en cambio, la organización actúa como grupo de presión, apelando, como organización ajena, directamente al elector. Las cifras siguientes indican los porcentajes de votos en favor de la fluorización. ¿Es uno de los métodos superior al otro?

Núm. del par	Grupo A, %	Grupo B, %	Diferencia, %
1	63	68	5
2	41	49	8
3	54	53	- 1
4	71	75	4
5	39	49	10
6	44	41	- 3
7	67	75	8
8	56	58	2
9	46	52	6
10	37	49	12
11	61	55	- 6
12	68	69	1
13	51	57	6
			52

1. Supuestos

Nivel de medición: El porcentaje de los votos es una escala de intervalo

Modelo: muestreo aleatorio
diferencias de población distribuidas normalmente

Hipótesis: $\mu_D = 0$.

Hay que suponer que los pares que figuran en las muestras han sido seleccionados al azar de alguna población de pares. Como se verá más abajo, este supuesto plantea algunas veces un problema difícil de interpretación. Como quiera que son las diferencias de cada par las que nos interesan directamente, hay que suponer que la población de todas las diferencias posibles está distribuida normalmente. Si N fuera grande, podría prescindirse de este supuesto.

2. Distribución de muestreo. Como quiera que no se da la desviación estándar de las diferencias de la población, hay que re-

currir a la distribución t , con $N - 1$, o sean 12 grados de libertad. Obsérvese que éstos representan la mitad de los grados de libertad que se habrían utilizado si la prueba de la diferencia de las medias (con $\sigma_1 = \sigma_2$) hubiera sido posible.

3. *Nivel de significado y región crítica.* Sirvámonos del nivel de .05 y de una prueba de dos colas. Por consiguiente, con 12 grados de libertad, si $t \geq 2.179$, descartaremos la hipótesis nula.

4. *Cálculo de la estadística de la prueba.* Primero hallamos la media de las diferencias de la muestra sumando las de la columna de diferencias y dividiendo entre $N (= 13)$. Se obtiene además la desviación estándar de la muestra de las diferencias.

$$\bar{X}_D = 52/13 = 4.0$$

$$s_D = \sqrt{\frac{\sum (X_D - \bar{X}_D)^2}{N}} = \sqrt{\frac{328}{13}} = 5.023$$

Por consiguiente:

$$t = \frac{\bar{X}_D - \mu_D}{s_D / \sqrt{N - 1}} = \frac{4.0 - 0}{5.023 / \sqrt{12}} = 2.76$$

Obsérvese que una vez que se ha obtenido la columna de diferencia, dejamos de prestar atención a las restantes columnas. Este mismo principio es de aplicación en situaciones más complejas, en las que por ejemplo podemos tener una diferencia de diferencias por cada par. (Ver ejercicio 5.)

5. *Decisión.* Con 12 grados de libertad, una probabilidad de .02 corresponde a una t de 2.681. Decidimos, en consecuencia, descartar la hipótesis nula y, observando la dirección de la diferencia, concluimos que el método B es superior al A .

XIII.5. Comentarios a propósito de los esquemas experimentales y pruebas de significación

Pese a que no sea posible profundizar mucho en un texto como éste en cuestiones de la planificación de experimentos, unos breves comentarios tienen con todo aquí su lugar adecuado.² El lector podrá acaso haberse preguntado a sí mismo cómo sea que preferíamos siempre servirnos de muestras asociadas, en lugar de muestras independientes. Indudablemente, se pierden con aqué-

² Para más detalles acerca de los esquemas experimentales, véase cualquier texto corriente sobre métodos de investigación. Véase en particular [8], capítulo IV.

llas algunos grados de libertad y, como quiera que el empleo de las muestras asociadas implica partir los casos por la mitad (en relación con la prueba), ¿es que no se pierde más, con ello, de lo que se gana? Todo esto depende de cuán acertados estemos en el apareamiento de los casos. Por supuesto, el objeto de la asociación está en reducir las diferencias debidas a variables extremas. Esto significa que un apareamiento cuidadoso debería reducir considerablemente cada una de las diferencias por pares. En otros términos: cuanto mejor sea el apareamiento, tanto menor será la desviación estándar de las diferencias. Así, pues, si bien el número de casos se reduce, la s_D debería reducirse asimismo. Si se obtiene una fuerte reducción de la desviación estándar de las diferencias en relación con la pérdida de casos, entonces salimos ganando al aparear. Como quiera que, por lo regular se perderán casos en los procedimientos de apareamiento (véase *infra*), la conclusión lógica es la siguiente: no se aparee, a menos de estar completamente seguro de haber localizado las variables significativas importantes. Si el lector está estudiando la delincuencia y aparea conforme al color del pelo, se verá probablemente más apurado que si no apareara en absoluto.

Los textos sobre métodos suelen por lo regular mencionar el hecho de que es probable que con el procedimiento de apareamiento se perderá un número considerable de casos. O sea que habrá que eliminar muchos casos, porque no hay casos similares con los que se dejen aparear. Semejante reducción puede resultar desastrosa en el caso del supuesto de la muestra aleatoria. En efecto, un sociólogo puede eventualmente partir de una muestra aleatoria de 1 000 casos y terminar con 200 que se dejen aparear. Al proceder así, es probable que se sesgue fuertemente su muestra final, eliminando la mayoría de los casos más extremos o poco comunes, difíciles, efectivamente, de aparear. En esta forma resulta a menudo difícil determinar el carácter de la población a cuyo propósito se está generalizando. Por ello hay que proceder con la mayor prudencia al generalizar los resultados. Por lo tanto, este tipo de esquema es probablemente más útil en estudios en que el interés por generalizar respecto de una población finita concreta, tal como la de los blancos nativos en Chicago, es secundario.

En conexión con semejante reducción de casos y las dificultades en cuanto a generalizar a una población concreta, se sostiene a menudo que no hay verdadero interés en la población misma, ya que el objeto fundamental del investigador consiste en establecer "relaciones entre variables". Así, por ejemplo, un psicólogo puede acaso empezar sirviéndose de aquellos novatos varones blancos que siguen un curso de introducción a la psicología y se prestan voluntariamente como sujetos de estudio. Puede producirse mayor muestreo todavía, a medida que algunos sujetos se

van eliminando en el proceso de apareamiento. Supóngase que se encuentra entonces una relación entre la variable experimental y alguna variable dependiente. Se propenderá, en este caso, a sacar la conclusión que la misma relación subsistiría independientemente de la población estudiada, esto es, concluir que se trata de una relación universal. Si ello resulta efectivamente ser así, el sociólogo puede muy bien afirmar que no tiene interés alguno por extender la generalización a cierta población finita cualquiera. Pero, ¿sobre cuál base puede suponer que la relación hallada en una población tan restringida es cierta asimismo en relación con otras poblaciones? Obviamente, antes de poder hacer legítimamente semejante afirmación, el experimento ha de efectuarse sobre un gran número de poblaciones muy distintas. Pese a que en un experimento cuidadosamente dispuesto se puede obtener el control de cierto número de variables, prodúcese casi siempre una pérdida correspondiente del grado en que los resultados se pueden generalizar a poblaciones más extensas.

En el agrupamiento por pares resulta indicado seleccionar al azar en el interior de cada par echando una moneda al aire para decidir cuál miembro del par deba asignarse al grupo experimental y cuál al grupo de control. Semejante procedimiento confiere mayor contenido lógico a la interpretación de los resultados, en el sentido de que cabe excluir la autoselección. Así, por ejemplo, en el intento de influir sobre los electores en materia de vivienda popular, supóngase que se permitía a las autoridades locales elegir aquel de los dos tipos de influencia que preferían o que creían iba a resultar más eficaz en su localidad particular. Es posible, en estas condiciones, que todas o la mayoría de las localidades con cierto tipo de autoridades fueran objeto del método indirecto, en tanto que las de otro tipo de dirigentes se verían tratadas por el método directo. Tendríamos así una variable incontrolada (el tipo de autoridades), cuyos efectos se confundirían irremediablemente con los de la variable experimental. Concretamente, supóngase que el grupo *B* resultaba tener el porcentaje más elevado de votos favorables, pero que al propio tiempo dicho grupo tenía las autoridades más democráticas, debido al hecho que éstas tendían a favorecer la aplicación a sus respectivas localidades del método indirecto. ¿Cómo podríamos saber si la diferencia en la votación se debía efectivamente a la superioridad del método *B* y no, acaso, a las diferencias entre las autoridades de los dos grupos de localidades?

Podría alegarse que el tipo de autoridades hubo de haberse controlado en el proceso de apareamiento, de modo que dos localidades de uno cualquiera de los pares tuviera el mismo tipo. Sin embargo, es obviamente imposible controlar en el proceso de apareamiento todas las variables operantes, no sólo debido a dificultades prácticas, sino a causa de nuestros conocimientos limi-

tados acerca de cuáles variables son efectivamente las más importantes. En algún punto habremos de admitir que puede haber variables importantes, muchas de las cuales el investigador no conoce y que no se han controlado en el proceso de apareamiento. Y es precisamente en dicho momento cuando confiamos en la selección al azar, o sea en las leyes de la probabilidad, esperando que los efectos de las variables incontroladas se habrán neutralizado mutuamente. Así, por ejemplo, con una *N* mayúscula, esperamos que, en números redondos, la mitad de las localidades de autoridades más democráticas habrán quedado en el grupo *A*, y la otra mitad en el grupo *B*. Y lo mismo acontecerá con otras variables incontroladas.

En los esquemas experimentales *ex post facto*, en las que el investigador sólo entra en función *después* de haberse efectuado el experimento y en las que, por lo tanto, no ha tenido oportunidad de efectuar tales asignaciones al azar, la posibilidad de autoselección nunca puede descartarse. Ni nos ayudan las leyes de las probabilidades a apreciar los efectos de la variable experimental en comparación con los efectos posibles de variables respecto de las cuales los grupos no se han apareado. Una de las mayores ventajas de los experimentos de laboratorio sobre los llamados "naturales", o *ex post facto*, está precisamente en ese control al azar de la autoselección posible.

Sugiérense a menudo otros métodos de asociación de muestras, a título de alternativas del método por pares. Por lo regular, tales métodos alternativos presentan la ventaja de atenuar la reducción de los casos, pero conducen a dificultades cuando se llega al análisis estadístico. Uno de dichos métodos comporta la asociación por distribuciones de frecuencia. Así, por ejemplo, puede ponerse atención en que los dos grupos sean similares en relación con el ingreso medio, la edad media, la distribución general del ingreso, etcétera. En esta forma, los grupos resultan comparables en relación con dichas medidas de resumen, aunque algún individuo no tenga en el otro grupo contrapartida exacta alguna con la que se lo pueda aparear. En ese tipo de esquema violamos claramente una vez más el supuesto de independencia; pero, que el autor sepa, no existe modo simple alguno de servirse de una prueba estadística que sea a la vez eficaz y no comporte algún supuesto en entredicho. Se podrían aparear casos lo mejor posible y proceder como acaba de indicarse, pero el apareamiento conducirá indudablemente a un esquema inoperante. Sin duda, no sería legítimo servirse de una prueba de diferencia de medias de $N_1 + N_2 - 2$ grados de libertad.

Pruebas de significación y generalizaciones a poblaciones. Se ha suscitado un amplio debate en la bibliografía sociológica en relación con la adecuación de las pruebas de significación en aquellas

ocasiones en que uno trata con la población íntegra. (Ver especialmente [3], [7], [9] y [10].) Puede, por ejemplo, contarse con datos correspondientes a todos los condados o estados de los Estados Unidos o de una región en particular. Si así ocurre, no habrá una población más extensa en relación con la cual se desee generalizar, pudiendo ser difícil concebir el proceso de generalización involucrando una extrapolación a un universo más amplio de probabilidades, o a estos mismos casos bajo circunstancias análogas. En este caso resultarían inadecuadas las pruebas de significación, ya que no habría implícito ningún error en el muestreo.

La actitud que uno adopte en esta cuestión depende en primer lugar de si está satisfecho con generalizaciones a poblaciones fijas, o si desea sacar conclusiones acerca del proceso causal que pueden haber generado los datos de población. En este texto hemos conceptualizado el problema como si nuestro único objetivo fuese el de deducir partiendo de poblaciones fijas, pero es evidente que cuando deseamos relacionar nuestros hallazgos con análisis *teóricos* nuestros objetivos no son nunca tan sencillos. El problema de sacar deducciones causales partiendo de datos no experimentales, basados bien sea en muestreos o en la totalidad de las poblaciones, es demasiado complicado para su examen en un texto elemental como éste. Sin embargo, hay un procedimiento para obtener las pruebas de significación mucho más compatible con las explicaciones teóricas en lo que se refiere a *por qué* se ha hallado una relación particular.

Supongamos, por ejemplo, que, habiendo usado la totalidad de los 50 estados, hemos hallado una diferencia entre los del norte y los del sur, o bien entre los que tienen gobernadores republicanos o demócratas. Normalmente no nos conformaríamos con hacer una simple descripción de tales diferencias, sino que querríamos ofrecer una explicación, relacionada tal vez con las diferencias regionales o políticas. Admitamos que hemos advertido que los estados del sur gastan una proporción relativamente mayor de sus presupuestos en supercarreteras, pero menor en educación superior. Antes de que podamos hacer declaraciones acerca de que nuestra explicación deberá orientarse a buscar factores causales determinantes de esta diferencia regional, habremos de pensar en un escéptico hipotético que establezca el planteamiento de una sencilla explicación alternativa de nuestro hallazgo, a saber: la "causalidad".

Podría, en efecto, decirnos: "Afirma usted que ha encontrado una diferencia achacable a características regionales. Yo podría haber utilizado una tabla de números al azar para dividir los 50 estados. O bien, podría haberlos ordenado alfabéticamente con base en la tercera letra de sus nombres. Si yo pudiese probar que tal proceso, basado o casi basado en el azar, hubiera producido una

diferencia tan grande o mayor que la suya, resultaría que su explicación no era más plausible que la mía."

Obsérvese que aquí no se habla de una generalización a una población mayor que la total de los 50 estados. El argumento gira alrededor de los procesos que pueden haber generado diferencias entre subpoblaciones ordenadas de distintas maneras. Es evidente que si hubiese sido posible obtener diferencias tan grandes como las diferencias regionales al hacer uso de una tabla de números elegidos al azar, y siendo la *teoría* del escéptico mucho más simple que la nuestra, no tendría objeto adentrarse más en los datos. Si adoptamos este punto de vista en relación con el proceso de la generalización, tiene sentido el hacer pruebas de significación, incluso cuando se cuente con datos correspondientes a la totalidad de la población. Parecería como si la mayoría de los sociólogos tuviera presente este más amplio objetivo, orientándose a decir algo acerca de los procesos causales, y por ello plantearan pruebas dirigidas a eliminar la alternativa del simple "proceso casual". Sin embargo, debe insistirse que la prueba de significación *no* excluye muchas otras clases de explicaciones alternativas, tal como la que, por ejemplo, introduce variables adicionales como causas comunes de las dos variables bajo estudio. En el capítulo XIX volveremos a este, más dificultoso, problema.

EJERCICIOS

1. Se seleccionan al azar 50 distritos electorales en una ciudad. Se encuentra que 20 de ellos están atendidos por centros de la localidad, en tanto que los restantes no lo están. Se comparan los porcentajes de delincuencia en esos dos tipos de distritos y se obtienen los siguientes datos (que se indican en el número de delinquentes por 1 000 adolescentes):

Medida	Con centro	Sin centro
Magnitud de la muestra	20	30
Media	27	31
Desviación estándar(es)	6	8

Efectúese una prueba de significación de la diferencia entre los dos tipos de distritos (nivel de .01), sirviéndose *a)* del modelo 1, y *b)* del modelo 2. ¿Cómo se presentan unos respecto de otros los resultados? Respuesta, *a)* $t = 1.87$; no rechazo.

2. Una muestra al azar de mujeres casadas que siguen viviendo con sus maridos ha sido objeto de selección, clasificándose a las mujeres en "satisfechas" o "insatisfechas" con sus respectivas vidas maritales. Se comparan luego los dos grupos de mujeres en relación con el tiempo de sus matrimonios, con los siguientes resultados:

Tiempo del matrimonio (redondeado al año)	Satisfechas f_1	Insatisfechas f_2
0-2	34	10
3-4	41	16
5-9	50	23
10-14	39	25
15-19	18	14
20-39	15	16
Total	197	104

¿Existe alguna diferencia significativa entre estos dos grupos al nivel de .01?

3. Supóngase que se espera encontrar que la diferencia entre los ingresos medios anuales de muestras de médicos y dentistas sea de unos \$ 500 (esto es, $\bar{X}_1 - \bar{X}_2 = 500$). Se aprecia que las desviaciones estándar son respectivamente de \$ 1 900 y \$ 1 600. Se planea seleccionar en la muestra total el mismo número de médicos que de dentistas. ¿Cuántos casos se necesitarán para establecer significación entre los ingresos medios de doctores y dentistas al nivel de .05? Supóngase que se quiere tomar un número doble de médicos que de dentistas. ¿Cuántos casos se necesitarán en este último supuesto? Respuesta, .95 de cada uno.

4. Se ha clasificado una muestra aleatoria de estudiantes universitarios como "dirigidos por otros" y "dirigidos por sí mismos". Se encuentra que el 58 por ciento de los alumnos avanzados son dirigidos por otros, en tanto que pertenece a esta categoría el 73 por ciento de los alumnos novatos. En la muestra total figuran 117 alumnos avanzados y 171 alumnos novatos. ¿Es esta diferencia significativa al nivel de .001?

* 5. Supóngase que se ha dispuesto un experimento de antes —y— después con grupo de control. En otros términos: se han relacionado dos grupos por pares y se han tomado medidas de ambos grupos

Par	Grupo de control		Grupo experimental	
	Antes	Después	Antes	Después
A	72	75	66	77
B	61	60	61	65
C	48	37	43	49
D	55	64	55	53
E	81	76	76	91
F	50	59	52	68
G	42	49	40	51
H	64	55	65	74
I	77	75	67	79
J	69	78	64	63

antes y después del experimento. Empléese la prueba t en relación con la efectividad de la variable experimental: a) sirviéndose solamente de las marcas de "después" e ignorando las de "antes"; b) empleando las marcas "antes" y "después" en el grupo experimental únicamente, y c) utilizando los cuatro juegos de marcas. (Indicación: ¿Cómo pueden emplearse las cuatro marcas para descartar los efectos sobre la variable experimental de factores ajenos susceptibles de haber afectado ambos grupos? Compárense las ventajas y los inconvenientes de los métodos a) y b). ¿Cuáles son las ventajas de c) respecto de a) y b)? Respuesta, a) $t = 1.25$, sin rechazo.

* 6. En el cuadro XV.4 del capítulo xv se encontrarán algunos datos relacionando las puntuaciones que los niños reciben por su habilidad, esfuerzo y clase social.

a) Teniendo en cuenta tan sólo la clase media, hágase una prueba para ver si la relación entre esfuerzo y grado varía según el nivel de habilidad del estudiante.

b) Amplíese esta prueba para ver si la "interacción" probada mediante a) difiere según sea la clase social del estudiante.

Nota: En realidad, en b) se estará buscando una interacción de una interacción, o lo que se denomina una interacción de segundo orden.

BIBLIOGRAFÍA

1. Alder, H. L., y E. B. Roessler: *Introduction to Probability and Statistics*, 4ª ed., W. H. Freeman and Company, San Francisco, 1968, caps. 8 y 10.
2. Downie, N. M., y R. W. Heath: *Basic Statistical Methods*, 2ª ed., Harper and Row, Publishers, Incorporated, Nueva York, 1965, caps. 11 y 12.
3. Gold, David: "Statistical Tests and Substantive Significance", *American Sociologist*, Vol. 4 pp. 42-46, 1969.
4. Goodman, L. A.: "Modifications of the Dorn-Stouffer-Tibbetts Methods for 'Testing the Significance of Comparisons in Sociological Data'", *American Journal of Sociology*, Vol. 66, pp. 355-359, 1961.
5. Hagood, M. J., y D. O. Price: *Statistics for Sociologists*, Henry Holt and Company, Inc., Nueva York, 1952, cap. 19.
6. Hays, W. L.: *Statistics*, Holt, Rinehart and Winston, Inc., Nueva York, 1963, cap. 10.
7. Kish, Leslie: "Some Statistical Problems in Research Design", *American Sociological Review*, Vol. 24, pp. 328-338, 1959.
8. Selltitz, C., M. Jahoda, M. Deutsch y S. W. Cook: *Research Methods in Social Relations*, Henry Holt and Company, Inc., Nueva York, 1959, cap. 4.
9. Selvin, H. C., "A Critique of Tests of Significance in Survey Research", *American Sociological Review*, Vol. 22, pp. 519-527, 1957.
10. Winch, R. F., y D. T. Campbell: "Proof? No. Evidence? Yes. The Significance of Tests of Significance", *American Sociologist*, Vol. 4, pp. 140-143, 1969.

XIV. ESCALAS ORDINALES: PRUEBAS NO PARAMÉTRICAS DE DOS MUESTRAS

HASTA aquí no hemos tenido ocasión de examinar pruebas de significación que comportaran escalas ordinales, pese a haber señalado en el capítulo II que éstas son muy frecuentes en ciencias sociales. En el presente capítulo vamos a ocuparnos de pruebas de dos muestras que pueden emplearse con escalas ordinales, pruebas comparables directamente con las que comportan diferencias de medias y proporciones, tales como las vimos en el capítulo anterior. Por lo tanto, las pruebas examinadas en este capítulo pueden emplearse para relacionar variables de escala ordinal con las que comportan una escala nominal dicotómica. En los capítulos sucesivos veremos luego pruebas que permiten relacionar una escala ordinal con una escala nominal de cualquier número de categorías o con otra escala ordinal.

Las pruebas que se examinan en el presente capítulo se designan a menudo como *no paramétricas*, o como *pruebas libres de distribución*, por cuanto no requieren el supuesto de una población normal. En realidad, tanto un término como el otro son algo equívocos. No queremos significar, en efecto, que comportan distribuciones las pruebas que no tienen parámetros. Ni puede estar una población "libre distribución". De hecho, ambos términos se emplean para designar una vasta categoría de pruebas que no requieren el supuesto de normalidad ni algún otro supuesto que especifique la forma exacta de la población. Sin duda, en todas las pruebas no paramétricas se requieren algunos supuestos acerca de la naturaleza de la población, pero por lo general, con todo, dichos supuestos son más débiles y menos restrictivos que los que necesitan las pruebas paramétricas. Por lo demás, ya nos hemos encontrado con algunas pruebas no paramétricas. Así, por ejemplo, la prueba binomial, la del signo y la de diferencia de proporciones no requieren el supuesto de normalidad, ya que todas ellas se refieren a escalas nominales dicotómicas. A diferencia de estas pruebas no paramétricas particulares, aquellas de las que nos ocupamos en este capítulo comportan todas ellas escalas ordinales, lo que permite servirse de un nivel de medición algo más alto. En el siguiente capítulo se examinarán dos pruebas no paramétricas adicionales, que sólo comportan, una y otra, escalas nominales.

¿Cuál es la ventaja de las pruebas no paramétricas en comparación con una prueba tal, por ejemplo, como la de la diferencia de las medias? Al servirnos de la prueba *t* en relación con una prueba de diferencia de medias, vimos que era indispensable servirnos no sólo de una escala de intervalo, sino también de una

población normal. Sin duda, podría prescindirse del supuesto de normalidad en el caso de muestras grandes, pero alegábase, con todo, que precisamente cuando las muestras son pequeñas el supuesto de normalidad resulta más dudoso. Por consiguiente, esperamos encontrar que las alternativas no paramétricas de la prueba de la diferencia de las medias sean más útiles siempre que se dé una de las dos condiciones siguientes: 1) que no podamos servirnos legítimamente de una escala de intervalo, pero estando justificado, con todo, el ordenamiento de las marcas, o 2) que la muestra sea pequeña y la normalidad no pueda presumirse. Como quiera que estas pruebas no paramétricas comportan supuestos más débiles que la prueba de la diferencia de las medias, pueden acaso no sacar provecho de toda la información disponible. Así, pues, si puede emplearse legítimamente una escala de intervalo y si el supuesto de normalidad puede hacerse en el caso de muestras pequeñas o abandonarse en el caso de las grandes, la prueba de la diferencia de las medias será por lo general preferible a las pruebas no paramétricas.

¿En qué sentido podemos decir que una prueba sea preferible a otra? ¿Cuáles criterios se emplean para adoptar semejante decisión? En primer lugar, como ya se indicó anteriormente, si una prueba nos impone ciertos supuestos dudosos que no se dejen verificar en sí mismos, no será tan satisfactoria como la que no nos los imponga. Si todas las demás condiciones fueran iguales, lo que prácticamente nunca ocurre, escogeríamos siempre la prueba que requiriera los supuestos más débiles. Y si los resultados de la prueba aconsejaran descartar, podríamos tanto más fácilmente considerar la hipótesis nula como único supuesto falso. Por desgracia, sin embargo, el problema no es tan sencillo. Si lo fuera, en efecto, siempre nos serviríamos de procedimientos no paramétricos. Resulta por lo regular que la prueba que requiere supuestos más firmes es también más fuerte, en el sentido que su empleo comporta un riesgo inferior de error de tipo II. Tenemos, pues, dos criterios que actúan en sentidos opuestos y han de evaluarse en consecuencia. O sea que las pruebas no paramétricas requieren supuestos más débiles, pero son menos fuertes. Obtendremos una idea más clara de lo que se entiende por supuestos "fuertes" y "débiles" cuando lleguemos a las pruebas concretas no paramétricas que pueden utilizarse como alternativas de la prueba de la diferencia de las medias. Antes, sin embargo, hemos de examinar la cuestión de cómo se aprecia la fuerza relativa de una prueba.

*XIV.1. Fuerza y eficiencia de la fuerza

La *fuerza* de una prueba se define como $1 - (\text{probabilidad de error de tipo II})$, o sea como $1 - \beta$. Así, pues, la fuerza de una

prueba es inversamente proporcional al riesgo de dejar de descartar una hipótesis falsa. Cuanto más capaz es una prueba de eliminar falsas hipótesis, tanto mayor es su fuerza relativa. Como ya se indicó, es mucho más difícil apreciar el riesgo de error de tipo II que de tipo I. Para ello, en efecto, no sólo hemos de co-

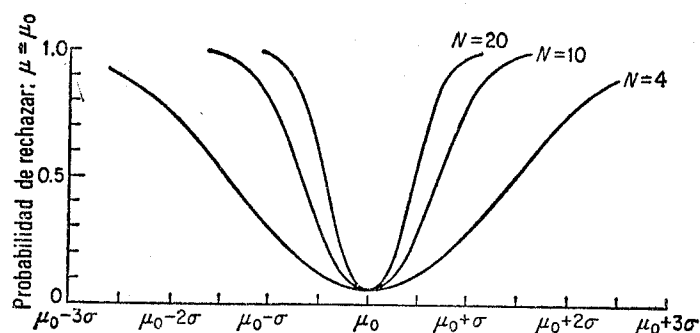


FIG. XIV.1. Funciones de potencia para pruebas de dos colas, con $\alpha = .05$, para muestras de tamaño variable. (Con la autorización de W. J. Dixon y F. J. Massey, "Introduction to Statistical Analysis", McGraw-Hill Book Company, Nueva York, 1957, fig. 14.6, p. 252.)

nocer la forma exacta de la población, sino que hemos de saber también el grado en que el parámetro tomado como hipótesis difiere del verdadero valor. En otros términos: la probabilidad de un error de tipo II, y con ella la fuerza de la prueba, depende de cuál hipótesis alternativa sea efectivamente correcta. Por estas razones, de hecho, raramente calculamos en la investigación aplicada las probabilidades de errores del tipo II. Sin embargo, como ya se dejó entrever anteriormente, la fuerza de una prueba ha de emplearse al apreciar su eficiencia relativa. Pueden hacerse varias pruebas alternativas que comporten el mismo riesgo de error de tipo I. Nos servimos, pues, de los riesgos relativos de cometer errores de tipo II para seleccionar una prueba que sea la más apropiada en un determinado conjunto de condiciones. Si bien el problema de determinar la fuerza de una prueba es bastante complejo y va más allá del propósito del presente texto, podemos indicar con todo, de modo general, lo que semejantes comparaciones comportan. Para ello necesitamos introducir la noción de la función de la fuerza.

La forma general de una función de la fuerza para una prueba de dos colas puede verse en la figura XIV.1. Semejante función nos da la fuerza de una prueba en relación con las distintas alternativas correctas posibles de la H_0 . O en forma más correcta:

supóngase que hemos tomado como hipótesis un determinado valor μ_0 para la media de la población. Supóngase, sin embargo, que la verdadera media de la población se sitúa en realidad a dos errores estándar de la media de la hipótesis. Es obvio, en este caso, que H_0 es falsa y debería descartarse. Como quiera que la fuerza de una prueba es $1 - \beta$, dicha fuerza nos da en realidad la probabilidad de descartar H_0 cuando ésta es falsa. Y esta última probabilidad, y no la probabilidad de error, nos es dada por el alto de la curva. Si la verdadera media se sitúa a dos errores estándar de μ_0 , la probabilidad de descartar H_0 puede determinarse hallando la altura de la curva, en dicho punto, en el eje de las X. Así, pues, los valores a lo largo del eje de las X indican los valores correctos posibles de μ , en tanto que los del eje de las Y indican las probabilidades de descartar H_0 .

Obsérvese que si el valor correcto de la media es efectivamente μ_0 (y que por lo tanto cometeríamos error descartando H_0), la altura de la función de la fuerza viene dada por el nivel de significación de la prueba. ¿Por qué? Obsérvese asimismo que si el valor correcto de μ no queda demasiado distante de μ_0 , la fuerza de la prueba, según la indica la altura de la curva, es menor que en el caso en que el verdadero valor es totalmente distinto de μ_0 . Esto nos dice que nuestro riesgo de error de tipo II es relativamente grande cuando el valor tomado como hipótesis no queda demasiado lejos del valor correcto, pero que, si nos hemos apartado del blanco en un grado considerable, tendremos una probabilidad mucho mayor de descartar nuestras hipótesis falsas. Esto concuerda con el argumento intuitivo que formulamos anteriormente en conexión con la binomial. Y corresponde asimismo a nuestros intereses prácticos. En efecto, si nuestra hipótesis nula es casi correcta, no nos preocupa mucho que dejemos de descartarla, pese a que, desde el punto de vista técnico, estemos en error al proceder en esta forma. Cuando H_0 es sustancialmente incorrecta nos interesa verdaderamente descartarla.

* Para generar la altura de la función de fuerza en cualquier punto dado situado en el eje horizontal, necesitaremos estar listos para suponer la forma que tenga la distribución del muestreo. En este caso particular suponemos que la distribución del muestreo de \bar{X} es $Nor(\mu, \sigma^2/N)$. Si la media verdadera de μ se encuentra a la derecha de la media supuesta μ_0 , como aparece en la figura XIV.2, la distribución real del muestreo (alrededor de μ) se hallará a la derecha de la distribución supuesta del mismo (alrededor de μ_0). Usamos desde luego la distribución supuesta del muestreo para determinar la región crítica, ya que desconocemos la μ verdadera. Supongamos que la región crítica resulta ser el grupo de X menores que a pero mayores que b . Para determinar la fuerza de la prueba debemos evaluar la probabilidad

real de caer dentro de la región crítica, puesto que la media verdadera es μ y no μ_0 . Esto se logra calculando la superficie sombreada colocada *bajo la distribución real de la muestra*, situada en el diagrama a la izquierda de a y a la derecha de b . Vemos que cuando μ y μ_0 están alejadas, dicha superficie es casi la unidad, pero cuando μ y μ_0 están muy próximas se aproxima a α (por ejemplo .05), en su límite más bajo.

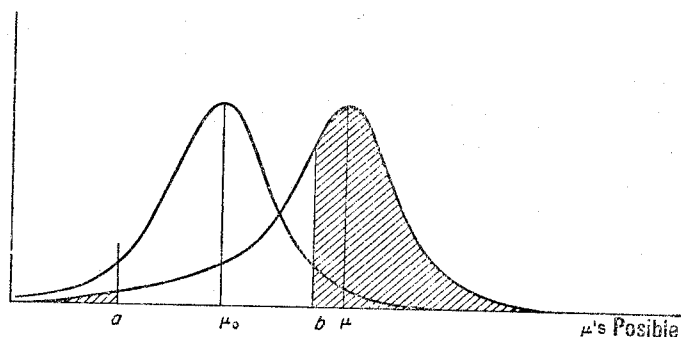


FIG. XIV.2. Derivación de la fuerza como función de $(\mu - \mu_0)$.

Con objeto de dar una indicación mejor de cómo se emplean en realidad las funciones de la fuerza, podemos comparar la función de la fuerza de una prueba de dos colas (figura XIV.1) con las de algunas de una sola cola. Supongamos, una vez más, que H_0 predice que la verdadera media es μ_0 . Obsérvese la prueba de una sola cola en la que nos hemos servido como región crítica de la cola superior o positiva. Si el verdadero valor de μ es efectivamente mayor que μ_0 , la mayoría de las medias de la muestra sacadas de la población serán también mayores que μ_0 , y tendremos mayores probabilidades de terminar en dicha región crítica de una sola cola que si nos hubiéramos servido de una prueba de dos colas al mismo nivel de significación. En otros términos: si μ queda efectivamente a la derecha de μ_0 , tenemos mayores probabilidades de descartar H_0 con una prueba de una sola cola en dicha dirección. Esto significa, por supuesto, que la fuerza de esta prueba particular de una sola cola será mayor para valores de μ en dirección positiva. Pero supóngase que el verdadero valor de μ queda en realidad a la izquierda de μ_0 . En tal caso, la mayoría de las \bar{X} quedará a la izquierda de μ_0 , y muy pocas de ellas caerán en la región crítica del extremo opuesto (o positivo) del continuo. En este caso, por consiguiente, no estaremos prácticamente nunca en condiciones de descartar H_0 , y la fuerza de la prueba de una sola cola será efectivamente muy débil. Y es obvio que el tipo opuesto de modelo se producirá en el caso

de pruebas de una sola cola con regiones críticas en las colas inferiores o negativas.

Las funciones de la fuerza de pruebas de una y de dos colas pueden compararse como en la figura XIV.3. En resumen, vemos que la prueba de una sola cola será más fuerte que la correspondiente de dos colas (sirviéndonos del mismo nivel de significación) para alternativas que se hallan en dirección de la región crítica, pero será mucho menos fuerte si el parámetro queda en realidad en dirección opuesta a la que se anticipó. Por consiguiente, el riesgo de error de tipo II es considerable si se efectúa una prueba de una cola y se yerra al anticipar la dirección. En tal caso, los datos tampoco pueden emplearse de cualquier modo para apoyar la teoría. Por lo tanto, probablemente no se tendrá interés alguno en seguir adelante con la prueba, a menos que, con fines de exploración, se quiera averiguar si una teoría totalmente opuesta tendría o no mérito alguno.

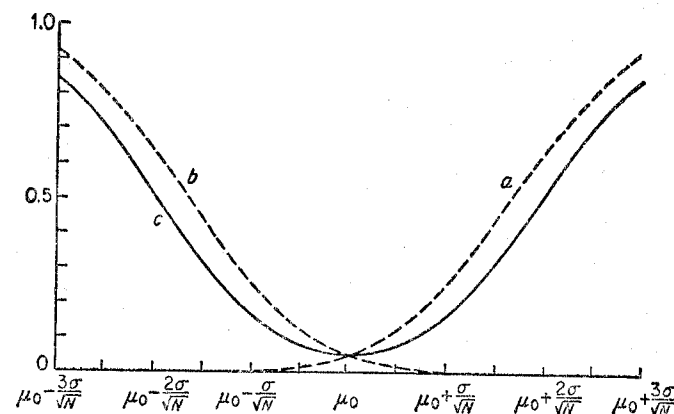


FIG. XIV.3. Comparación de funciones de potencia para pruebas de una y dos colas, con $\alpha = .05$. a) Rechace si $Z > 1.645$. b) Rechace si $Z < -1.645$. c) Rechace si $Z > 1.96$ o si $Z < -1.96$. (Con la autorización de W. J. Dixon y F. J. Massey, "Introduction to Statistical Analysis", McGraw-Hill Book Company. Nueva York, 1957, fig. 14.5, p. 249.)

Al comparar las pruebas de una y de dos colas, hemos visto que una prueba determinada puede ser más fuerte en relación con ciertas alternativas y menos fuerte en relación con otras. En términos generales, esto puede ocurrir también al comparar dos clases muy distintas de pruebas. Así, por ejemplo, no tardaremos en ver que una prueba no paramétrica particular puede ser más fuerte que otra en un determinado conjunto de circuns-

tancias, pero menos en otras. Es este hecho lo que hace que sea difícil desarrollar generalizaciones relativamente simples a propósito de la superioridad de una prueba respecto de otra. Y la situación se complica además por el hecho de que una prueba podrá ser fuerte en relación con muestras grandes, pero relativamente menos en el caso de muestras pequeñas. Por supuesto, la fuerza de cualquier prueba determinada aumentará con el tamaño de la muestra, ya que para cualquier nivel de significación determinado el aumento del tamaño de la muestra hace posible descartar la hipótesis nula con menores desviaciones respecto de los valores anticipados. Hemos visto, por ejemplo, que el error estándar de la media decrece a medida que crece N y que, por consiguiente, a medida que N aumenta, la media de la muestra ha de quedar más cerca del valor tomado como hipótesis para que podamos retener H_0 . Lo que decimos, pues, es que podemos descartar más fácilmente una hipótesis falsa cuando N es grande. Sin embargo, aunque la fuerza de una prueba pueda aumentar al aumentar N , la *tasa* del aumento de fuerza puede no ser la misma para todas las pruebas. Por lo tanto, una prueba de fuerza relativamente débil con una N pequeña puede acaso "alcanzar" a otra, de modo que la primera sea en realidad más fuerte en el caso de muestras grandes.

Con objeto de comparar la fuerza relativa de dos pruebas, podemos preguntarnos cuántos casos se necesitarían con la primera para obtener la misma fuerza que con un número determinado de casos de la segunda. Por lo regular comparamos la fuerza de una prueba determinada con la de la alternativa más fuerte. En el caso de las tres primeras pruebas no paramétricas examinadas en este capítulo, la alternativa más fuerte será la prueba t para la diferencia de las medias. Se emplea comúnmente el término de *eficiencia de la fuerza* para designar la fuerza de una prueba determinada en relación con su alternativa más fuerte. Si designamos la eficiencia de la fuerza de una de dichas pruebas no paramétricas como del 95 por ciento, queremos decir que la fuerza de la prueba no paramétrica sirviéndose de 100 casos es aproximadamente la misma que la de la prueba t sirviéndose de 95 casos, si el modelo empleado en la prueba t es correcto.

Como quiera que es necesario suponer una determinada forma de la población para poder evaluar la fuerza de una prueba, nos imaginamos, en la ilustración anterior, que tenemos en realidad un nivel de medición de escala de intervalo y que las dos poblaciones son normales en cuanto a la forma. Al determinar la eficiencia de la fuerza de la prueba no paramétrica, nos estamos fundamentalmente preguntando a nosotros mismos cuánto nos costará el dejar de aceptar el supuesto de normalidad si semejante supuesto fuera de hecho legítimo. Aquí vemos que el hecho de dejar de aceptar dicho supuesto y nuestro empleo consecuen-

te de la prueba no paramétrica nos costaría cinco casos adicionales por encima de los 95 utilizados en la prueba de la diferencia de las medias. Con una pérdida de eficiencia tan pequeña, es probable que seguiríamos adelante con la prueba no paramétrica si tuviéramos la menor duda respecto de los supuestos requeridos por aquélla. Por otra parte, si la eficiencia de la fuerza sólo fuera del 60 por ciento y si los alejamientos respecto de la normalidad no fueran demasiado grandes (o si N fuera grande) nos serviríamos probablemente de la prueba de la diferencia de las medias.

Como ya se indicó en el capítulo precedente, cuando las muestras son pequeñas necesitamos preocuparnos más por el supuesto de normalidad. En el caso de N pequeña no será por lo regular posible traducir enunciados de eficiencia de la fuerza en comparaciones de tamaños exactos de muestras, ya que estas últimas cantidades han de ser siempre enteras. Así, por ejemplo, con 95 por ciento de eficacia, una muestra de tamaño 10 que se sirviera de la prueba no paramétrica sería equivalente de forma aproximada a una de 9.5 que se sirviera de la prueba t . Pese a que semejante enunciado no tenga sentido desde el punto de vista operacional, ayuda, por lo menos, a establecer comparaciones.

Antes de terminar esta sección, conviene recordar una vez más que la eficiencia de la fuerza de una prueba determinada puede depender del tamaño de la muestra seleccionada. Puede ser muy eficaz en relación con muestras pequeñas, pero mucho menos eficaz en el caso de muestras grandes.

XIV.2. La prueba de las secuencias (runs) de Wald-Wolfowitz

En la prueba de las secuencias, así como en las otras dos pruebas que se examinarán en este capítulo a continuación, suponemos que tenemos dos muestras aleatorias independientes y que el nivel de medición es por lo menos una escala ordinal. En las tres pruebas en cuestión, nuestra hipótesis nula será que las dos muestras se han extraído de la misma población continua (o de poblaciones idénticas). La dimensión subyacente se supondrá ser continua, y no discreta, aunque admitamos que puedan resultar datos ligados entre sí, debido a la imperfección del instrumento de medición. La hipótesis de que las dos muestras se hayan tomado de la misma población es en realidad muy similar a nuestro supuesto en la prueba de la diferencia de las medias. En efecto, como ya se indicó anteriormente, cuando juntamos los supuestos de normalidad, de variancias iguales y de medias también iguales, suponemos en realidad que las dos poblaciones son idénticas. En el caso de la prueba de las secuencias, ponemos la hipótesis de que las dos poblaciones presentan exactamente la misma forma y pueden por consiguiente tomarse como iguales. Sin embargo, no necesitamos especificar la naturaleza de dicha

forma. Ésta podrá ser normal o no serlo. Por lo tanto, hacemos un conjunto de supuestos más débil que el que se requiere en la prueba de la diferencia de las medias, o sea, más débil en el sentido de que la prueba de la diferencia de las medias (con σ iguales) requiere todos los supuestos de la prueba de las secuencias, con el supuesto, *además*, de normalidad y el empleo de una escala de intervalo.

En la prueba de la diferencia de las medias nuestro interés se centra en diferencias de la tendencia central más que en las diferencias de dispersión o de forma. La prueba de las secuencias, en cambio, verifica esencialmente todas esas posibles diferencias simultáneamente. Como veremos en seguida, su empleo principal está en la verificación de diferencias de dispersión o de forma, ya que, para la verificación de diferencias de la tendencia central, hay pruebas no paramétricas más eficaces. Obsérvese, de paso, que la hipótesis nula no se ha establecido en términos de medias o de desviaciones estándar, sino más bien en términos de diferencias cualesquiera. Eso se aplica también a las pruebas no paramétricas a examinar en el presente capítulo. Con las escalas ordinales no tiene sentido, por supuesto, pensar en términos de medias y de desviaciones estándar.

El principio básico implicado en la prueba de las secuencias es muy sencillo, lo mismo que los cálculos. Tomamos primero los datos de ambas muestras y ordenamos los datos de los más altos a los más bajos, prescindiendo de que provienen de muestras distintas. Si la hipótesis nula es correcta, confiamos en que las dos muestras estarán bien mezcladas. En otros términos: no contamos con una gran serie de datos de la primera muestra seguida por otra larga serie de datos de la segunda. Así, por ejemplo, si designamos las muestras como A y B, esperamos que la ordenación resultará más o menos como sigue:

ABBABAAABABBABBAABAAB

y no como

AAAAAAAAABBBBBBBBBBBB

Con objeto de comprobar hasta qué punto las dos muestras están mezcladas una vez ordenadas, contamos simplemente el número de series continuas que se producen. La *secuencia* se define como serie continua de datos de la misma muestra. En el primero de los dos ejemplos anteriores tenemos una secuencia de una sola A, seguida de una serie de dos B, luego una sola A, una sola B, una serie de tres A, etcétera. El número total de secuencias es, por lo tanto, de 14. En el segundo ejemplo, en cambio, las A están agrupadas en la mitad inferior del continuo, y sólo tenemos cuatro secuencias o *runs*. Por lo regular, el cómputo de las secuencias se facilitará evitándose además errores, trazando una

línea debajo de los datos de la primera muestra y una raya arriba de las de la segunda. En esta forma sólo necesitamos contar el número de rayitas separadas. Si el número de las secuencias es grande, como en el primer ejemplo, entonces las dos muestras estarán tan bien mezcladas que no estaremos en condiciones de descartar la hipótesis nula. Por otra parte, un número reducido de secuencias significa probablemente que la hipótesis es incorrecta y debería descartarse. La distribución de muestreo de las secuencias puede utilizarse para establecer la región crítica de la que nos servimos para descartar la hipótesis nula.

Problema. Supóngase que unos jueces han ordenado 19 organizaciones sociales de acuerdo con el prestigio de las mismas, atribuyendo una puntuación de 1 a la de mayor prestigio y de 19 a la inferior. Diez de dichos grupos restringen la admisión a los no judíos, en tanto que los otros 9 admiten también a éstos. Suponiendo que dichas organizaciones sociales se han seleccionado al azar sobre la base de una lista de todas las demás organizaciones sociales de la localidad, ¿podemos llegar a la conclusión de que en la población se da una diferencia significativa de prestigio entre las organizaciones sociales restrictivas y las no restrictivas?

Admisión restrictiva: Rangos 1, 2, 4, 5, 6, 7, 9, 11, 14, 17 ($N_1 = 10$)

Admisión no restrictiva: Rangos, 3, 8, 10, 12, 13, 15, 16, 18, 19 ($N_2 = 9$)

1. Supuestos.

Nivel de medición: el prestigio como escala ordinal

Modelo: muestras aleatorias independientes

Hipótesis: las muestras se han extraído de poblaciones con las mismas distribuciones continuas.

2. **Distribución de muestreo.** Si tanto N_1 como N_2 son menores o iguales a 20, la distribución de muestreo exacta del número de secuencias r está dada en el cuadro E del Apéndice 2. Para N mayores, la distribución de muestreo de r es aproximadamente normal, con la

$$\text{media} = \mu_r = \frac{2N_1N_2}{N_1 + N_2} + 1 \quad (\text{XIV.1})$$

y la

$$\text{desviación estándar} = \sigma_r = \sqrt{\frac{2N_1N_2(2N_1N_2 - N_1 - N_2)}{(N_1 + N_2)^2(N_1 + N_2 - 1)}} \quad (\text{XIV.2})$$

Obsérvese que, aunque no se suponga la normalidad de la población, la distribución de muestreo de r será aproximadamente nor-

mal, incluso con N pequeñas. Como habremos de ver en seguida, cierto número de estadísticas de prueba no paramétricas poseen también esta propiedad. Obsérvese asimismo que las fórmulas de la media y del error estándar sólo comportan los tamaños de las muestras y no requieren, por lo tanto, que procedamos a apreciar los parámetros de la población, como era el caso con la prueba de la diferencia de las medias. La simplicidad comparativa de las fórmulas de las distribuciones de muestreo de las estadísticas no paramétricas se debe en parte al hecho que, como quiera que las marcas se han ordenado y han de tomar siempre, por consiguiente, los valores numéricos 1, 2, 3, ..., N , las magnitudes tales como la suma y la desviación estándar de los órdenes dependen únicamente del número de casos empleado.

3. *Nivel de significación y región crítica.* Como quiera que el cuadro E, Apéndice 2, sólo da el número de secuencias necesarias para el descarte al nivel de .05, nos vemos reducidos, en relación con muestras pequeñas, a dicho nivel de significación, pese a que pueden encontrarse cuadros más completos en [9]. Obsérvese que la prueba de las secuencias no toma en consideración la dirección de la relación entre el prestigio y la restricción de admisión. Por otra parte, cuando nos servimos de la distribución de muestreo de r , sólo estamos interesados en una cola, ya que sólo podemos descartar la hipótesis nula si hay un pequeño número de secuencias (independientemente de la dirección de la diferencia).¹ En sentido estricto, pues, empleamos la prueba de las secuencias como prueba de una sola cola, pese a que no se haya anticipado la dirección de la relación. La misma situación se nos presentará con la prueba de Mann-Whitney, que se examina en la sección siguiente, así como en otras pruebas importantes de las que habremos de ocuparnos en capítulos subsiguientes. Con objeto de evitar ambigüedades, distinguiremos, por lo tanto, entre pruebas de una sola cola y las situaciones en las que la dirección se haya anticipado. Hasta aquí semejante distinción no era necesaria, ya que todas las pruebas de una sola cola comportaban predicciones en relación con la dirección.

En el caso de distribuciones de muestras normales ya hemos visto que, cuando la dirección fue prevista, cortábamos a la mitad un nivel significativo al utilizar una sola cola de la distribución de muestreo. En el caso de las pruebas en secuencia y en el de otras diversas aplicaciones, habremos de confiar en otro tipo de justificaciones al cortar por mitad los niveles de significación cuando la dirección ha sido predicha. En el curso del presente ejemplo

¹ Sin embargo, hay otras aplicaciones de la prueba de las secuencias en las que pueden emplearse las dos colas. Así, por ejemplo, puede haber acaso demasiadas secuencias si las muestras se han mezclado más bien artificialmente que al azar, y este hecho puede utilizarse en una prueba del grado de aleatoriedad.

supongamos que no hay diferencia alguna en la población de las organizaciones sociales en relación con el prestigio de las organizaciones restrictivas y no restrictivas. Llamemos A al acontecimiento consistente en haber logrado resultados significativos a un nivel de, por ejemplo, .05 sin haber predicho la dirección. Claramente $P(A) = .05$. Llamemos ahora B al acontecimiento que consiste en que la dirección de la diferencia de las muestras es la predicha, suponiendo que no se den diferencias algunas en la población. En tal caso, $P(B) = .5$ si prescindimos de la probabilidad de que la diferencia sea exactamente igual a cero.

Como A y B serán normalmente dos acontecimientos separados, tendremos que la probabilidad de lograr significación al nivel .05, sin predecir la dirección y la probabilidad de predecir correctamente la dirección, vendrá dada por $P(A \& B) = P(A)P(B) = (.05)(.5) = .025$. Podrá utilizarse este mismo principio en cuantas ocasiones la distribución de muestreo de una estadística de prueba sea, o bien simétrica, o bien insensible a la dirección de una diferencia. Si hubiéramos estado, por ejemplo, interesados en comparar tres muestras (como vamos a hacer en los dos capítulos próximos), y si hubiéramos podido predecir el orden exacto de estas diferencias (por ejemplo $\bar{X}_1 > \bar{X}_2 > \bar{X}_3$), la probabilidad de obtener diferencias en este orden exactamente sería de $1/6$, bajo el supuesto de que $\mu_1 = \mu_2 = \mu_3$, pudiendo en tal caso dar justificadamente como $1/6$ el nivel de significación sin haber predicho la dirección. Por supuesto que este procedimiento se presta al razonamiento *ex post facto*, y sólo puede aplicarse a condición de que las predicciones hayan precedido al examen de los datos.

Los números en el cuadro nos dan el número de secuencias que brindarán significación al nivel de .05, suponiendo que no se haya predicho la dirección. Cualquier valor de r , por lo tanto, que sea igual o menor que la cifra del cuadro nos indicará que tenemos tan pocas secuencias que bien podemos rechazar la hipótesis nula a este nivel. Como el número de casos en las dos muestras es de diez y nueve, respectivamente, veremos que podremos rechazarla si obtenemos seis o menos secuencias.

4. *Cálculo de la estadística de la prueba.* Si disponemos las organizaciones por orden de prestigio y trazamos líneas debajo de los datos de la primera muestra y arriba de las del segundo, vemos que se dan 12 secuencias.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19

Pese a que el número de casos es un poco pequeño para que se verifique la aproximación normal, podemos con todo seguir adelante con el cálculo, sirviéndonos de esta aproximación para

ilustrar su empleo y comparar los resultados con los que se obtienen sirviéndose del cuadro E del Apéndice 2. Como de costumbre, calculamos el valor de Z , que nos dirá a cuántas desviaciones estándar se sitúa el número de secuencias obtenido respecto de la media o número de secuencias esperado con la hipótesis nula. Así, pues,

$$\mu_r = \frac{2(10)(9)}{10 + 9} + 1 = 10.47$$

y

$$\sigma_r = \sqrt{\frac{2(10)(9)[2(10)(9) - 10 - 9]}{(19)^2(18)}} = 2.11$$

Tenemos por consiguiente:

$$Z = \frac{r - \mu_r}{\sigma_r} = \frac{12 - 10.47}{2.11} = .725$$

Como quiera que el número de secuencias obtenido es efectivamente mayor que la media o el número esperado, no necesitamos proseguir, ya que para el rechazo de la hipótesis se necesitan números pequeños de secuencias. Si el número de secuencias hubiera sido menor que el esperado, habríamos buscado el valor de Z en el cuadro normal, usando el cuadro como si estuviéramos haciendo una prueba de dos colas (es decir: rechazando al nivel .05 si $Z \leq -1.96$).

5. *Decisión.* Ya que el número de secuencias ha resultado ser mayor que seis, cifra indicada en el cuadro E, decidimos no descartar la hipótesis nula al nivel de .05. Como acabamos de ver, el empleo de la aproximación normal nos conduce también a la misma conclusión. Sobre la base de nuestros datos, concluimos, en consecuencia, que no existe diferencia entre los dos tipos de organización en lo relativo al prestigio.

Empates. En los datos anteriores no ha habido dos organizaciones que tuvieran marcas empatadas. El supuesto de la continuidad subyacente excluye teóricamente la posibilidad de empates, ya que dos marcas no serán nunca exactamente iguales. Sin embargo, debido a imperfecciones de medición, y semejantes imperfecciones se darán casi seguramente en la mayoría de la investigación social, en la práctica semejantes empates se presentan. Obsérvese que si dos organizaciones de la misma muestra hubieran estado empatadas en relación con las marcas de prestigio, la prueba de las secuencias no habría sido afectada. Pero, supóngase que los empates se producían entre las muestras. En tal caso, el número de secuencias puede resultar considerablemente

afectado, según la forma en que los empates se rompan. Supóngase, por ejemplo, que dos organizaciones (de muestras distintas) hubieran estado ligadas por lo que se refiere a la octava y novena posiciones. Si estas posiciones se hubieran desplazado del orden anteriormente empleado, habríamos obtenido 10 secuencias en lugar de 12. En otros términos, obtendríamos 10 secuencias o 12, según el orden empleado. Y como quiera que dicho orden sería totalmente arbitrario, podríamos encontrarnos con que a veces descartamos y otras veces dejamos de descartar la hipótesis nula. El procedimiento más seguro que podemos utilizar en el caso de empates consiste en contar el número de secuencias sirviéndonos de todos los medios posibles de romperlas. Y si todos los ordenamientos conducen a la misma decisión (la de descartar o de no descartar), entonces nos podemos adherir a la misma con seguridad. Pero si conducen a decisiones distintas, será posible resolver el problema echando una moneda al aire, pero tal vez el procedimiento más seguro consista, en tal caso, en suspender el juicio. Bradley recomienda [1] un inteligente procedimiento consistente en facilitar al lector la gama de probabilidades obtenidas al romper empates por todos los métodos posibles. Es evidente que si se da un gran número de órdenes con empates, la estadística de la prueba no deberá ser usada.

XIV.3. La prueba de Mann-Whitney o de Wilcoxon

Otra prueba no paramétrica que puede emplearse en las situaciones en que la prueba de las secuencias resulta apropiada es una prueba que parece haber sido inventada independientemente por cierto número de personas y se conoce comúnmente con el nombre de prueba de Mann-Whitney o de Wilcoxon. Esta prueba requiere exactamente los mismos supuestos que la de las secuencias y, lo mismo que ésta, comporta un procedimiento muy simple. Combinamos nuevamente los datos de las dos muestras y las ordenamos de 1 a 19. Centramos a continuación nuestra atención en la segunda muestra (o en la que sea menor). Tomando cada dato de la segunda muestra, contamos el número de datos de la primera muestra que tengan un orden mayor. Una vez hecho esto con cada uno de los datos de la segunda muestra, sumamos los resultados, que nos dan la estadística U . La distribución de muestreo de U puede obtenerse exactamente si las N son pequeñas, o se puede obtener con aproximación por medio de una curva normal en el caso de muestras mayores. Si U es excepcionalmente pequeña o excepcionalmente grande, podemos descartar el supuesto de que las dos muestras se hayan extraído de la misma población.

Una forma alternativa de exactamente la misma prueba puede emplearse con la aproximación normal. En lugar de obtener U

directamente, podemos contar la suma de los órdenes de cada una de las muestras. Procedemos luego en forma análoga a la de la prueba de la diferencia de las medias. Tomamos una diferencia de las sumas de los órdenes para cada muestra y sustraemos de dicha diferencia una cantidad que representa la diferencia esperada con la hipótesis nula. Esta diferencia de diferencias, análoga a $(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)$, se divide luego por el error estándar para obtener Z . La analogía no es perfecta, ya que tratamos con sumas de órdenes y no con sus medias, pero el paralelo con la prueba de la diferencia de las medias es perfectamente claro. Aquí también, un valor numérico grande de Z conducirá al rechazo. Vamos a ilustrar ahora el procedimiento de Mann-Whitney sirviéndonos del mismo ejemplo anterior. Compararemos a continuación la eficacia de la fuerza de esta prueba con la de la prueba de las secuencias.

Problema. El mismo de la prueba de las secuencias.

Admisión restrictiva: órdenes 1, 2, 4, 5, 6, 7, 9, 11, 14, 17 ($N_1=10$)

Admisión no restrictiva: órdenes 3, 8, 10, 12, 13, 15, 16, 18, 19 ($N_2=9$)

1. **Supuestos.** Los mismos que en la prueba de las secuencias.

2. **Distribución de muestreo.** La distribución de muestreo de U se encontrará en el cuadro F del Apéndice 2, si ni N_1 ni N_2 son mayores que ocho, y en el cuadro G, si una de las N queda entre 9 y 20 y la otra entre 1 y 20. Nótese que los dos cuadros tienen diferente formato: El F tiene en la parte alta diferentes combinaciones de N_1 y N_2 , con los valores de U en el margen inferior izquierdo y con los valores de las probabilidades en el cuerpo del cuadro. Así, si $N_2=6$ y $N_1=4$, siendo siempre N_2 el mayor de los dos tamaños de las muestras, y si $U=5$, veremos que la probabilidad de obtener $U \leq 5$ es de .086, con dirección predicha. Los otros cuadros del G, por otra parte, corresponden a diferentes niveles de significación, apareciendo los valores críticos de U en el cuerpo del cuadro. Así, para $\alpha=.001$, con dirección predicha, para $N_1=13$ y $N_2=10$ (no siendo N_2 necesariamente mayor que N_1), observaremos que un valor de U igual o menor que 17 supondrá significación. Para N mayor, la distribución de muestreo de U será aproximadamente normal, con la

$$\text{media} = \mu_U = \frac{N_1 N_2}{2} \quad (\text{XIV.3})$$

y la

$$\text{desviación estándar} = \sigma_U = \sqrt{\frac{N_1 N_2 (N_1 + N_2 + 1)}{12}} \quad (\text{XIV.4})$$

3. **Nivel de significación y región crítica.** Con fines de comparación, seguimos sirviéndonos del nivel de .05, sin predecir la dirección de la relación.

4. **Cálculo de la estadística de la prueba.** La estadística U puede calcularse por uno de los dos métodos siguientes. Con N pequeña será relativamente sencillo calcular U siguiendo el procedimiento implicado en la fórmula de definición. Centrándonos en cada uno de los nueve grupos de la segunda muestra, contemos el número de casos de la primera muestra que tienen menor prestigio y, por lo tanto, mayores marcas de orden. Como quiera que la primera organización de la segunda muestra se ha clasificado como tercera en prestigio, hay en la primera muestra ocho grupos con marcas de prestigio inferiores. Y en forma análoga, el segundo grupo de la segunda muestra se clasifica como octavo, de modo que hay cuatro grupos en la otra muestra con marcas de prestigio inferiores. Prosiguiendo el proceso para cada una de las organizaciones restantes de la muestra 2 y sumando, obtenemos:

$$U = 8 + 4 + 3 + 2 + 2 + 1 + 1 + 0 + 0 = 21$$

Obsérvese que si hubiéramos seguido el mismo procedimiento, pero centrandó nuestra atención en los grupos de la primera muestra, habríamos obtenido:

$$U' = 9 + 9 + 8 + 8 + 8 + 8 + 7 + 6 + 4 + 2 = 69$$

Cualquiera de estas dos cantidades podría emplearse para verificar la significación de la relación, pero, como quiera que las tablas se han establecido en términos del valor menor de U , siempre nos servimos de la menor de las dos cantidades en cuestión. No será necesario calcular U y U' , ya que una vez obtenido uno de los valores el otro puede calcularse sirviéndose de la fórmula:

$$U = N_1 N_2 - U' \quad \text{o} \quad U' = N_1 N_2 - U \quad (\text{XIV.5})$$

En este caso nos serviríamos como estadística de prueba del valor 21.

Si el número de los casos es relativamente grande o si existen empates, será probablemente más conveniente obtener U sumando los órdenes de las muestras separados, designando estas sumas de órdenes como R_1 (rango) y R_2 y sirviéndose de las fórmulas:

$$U = N_1 N_2 + \frac{N_2 (N_2 + 1)}{2} - R_2 \quad (\text{XIV.6})$$

o bien

$$U' = N_1 N_2 + \frac{N_1(N_1 + 1)}{2} - R_1 \quad (\text{XIV.7})$$

según cuál de ellas resulte más conveniente. Sumando los órdenes obtenemos así:

1	3
2	8
4	10
5	12
6	13
7	15
9	16
11	18
14	19
17	
<hr/> R ₁ = 76	<hr/> R ₂ = 114

A título de control habríamos de tener

$$R_1 + R_2 = \frac{N(N + 1)}{2}$$

o bien

$$76 + 114 = \frac{19(20)}{2} = 190$$

en donde N representa el número total de casos en *ambas* muestras. Por lo tanto:

$$U = 10(9) + \frac{9(10)}{2} - 114 = 90 + 45 - 114 = 21.$$

* Las sumas de órdenes R_1 y R_2 pudieron haberse empleado directamente al hacer la prueba, no siendo necesario en tal caso calcular U . Ya que las tablas exactas para las N pequeñas suelen darse en términos de U , por lo regular resultará ventajoso pensar en términos de estadística U . Pero el empleo de las sumas de órdenes puede utilizarse eurísticamente para señalar la semejanza de la prueba Mann-Whitney con la de la diferencia de las medias. Una pequeña operación algebraica convencerá al lector de que podemos tomar las ecuaciones de (XIV.3) a (XIV.7) y obtener el resultado de que, para la aproximación normal, la estadística

$$Z = \frac{R_1 - R_2 - (N_1 - N_2)(N + 1)/2}{\sqrt{N_1 N_2 (N + 1)/3}} \quad (\text{XIV.8})$$

será aproximadamente $Nor(0,1)$. Expresando Z en esta forma, observamos que el numerador consta de la diferencia $R_1 - R_2$, junto con un término que resulta ser el valor esperado o a largo plazo de dicha diferencia en la hipótesis nula. Dicho factor de corrección es necesario, por supuesto, ya que tratamos con una diferencia de *sumas*, y no de medias, lo que nos obliga a tomar en cuenta el hecho de que, por lo regular, las dos N no serán iguales. Si N_1 y N_2 son iguales, observamos que el segundo factor en cuestión se convierte en cero, quedándonos simplemente como numerador $R_1 - R_2$. Vemos en esta forma la semejanza con la prueba de la diferencia de las medias, en la que el numerador se reducía a $\bar{X}_1 - \bar{X}_2$ en el caso de la hipótesis nula de que no había diferencias. Por lo tanto, cabría concebir la prueba de Mann-Whitney como prueba de la diferencia de los órdenes *sumados*.

5. *Decisión.* Sirviéndonos del cuadro G del Apéndice 2, vemos que al nivel de .05, si la dirección no se ha anticipado, necesitamos una U de 20, o *más pequeña*, para poder descartar la hipótesis nula. De ahí que apenas dejemos de descartar la de que no hay diferencia entre los dos tipos de organizaciones. Obsérvese, sin embargo, que si la dirección se hubiera predicho de antemano, habríamos necesitado una U de 24, o menos, al nivel de .05. Vemos, de paso, que a pesar de llegarse a la misma conclusión con las pruebas de las secuencias y de Mann-Whitney, estuvimos, con todo, mucho más cerca del descarte con la segunda que con la primera. Por lo tanto, si H_0 fuera realmente falsa, tendríamos en este caso un riesgo menor de error de tipo II que con la prueba de las secuencias.

Si nuestra N hubiera sido mayor, podríamos habernos servido de la aproximación normal. Con objeto de ilustrar el procedimiento, podemos calcular Z en relación con los datos anteriores. Obtenemos así:

$$Z = \frac{U - N_1 N_2 / 2}{\sqrt{N_1 N_2 (N_1 + N_2 + 1) / 12}} = \frac{21 - 45}{\sqrt{10(9)(20) / 12}} = -1.96$$

Si hubiéramos remplazado U por U' (= 69), hubiéramos obtenido

$$Z = +1.96$$

* Si nos hubiéramos servido de la ecuación (XIV.8), habríamos obtenido asimismo:

$$Z = \frac{76 - 114 - (10 - 9)(20)/2}{\sqrt{10(9)(20)/3}} = -1.96$$

Así, pues, el empleo de la aproximación normal conduce a la conclusión de que, sin la dirección predicha, a duras penas podríamos descartar al nivel de .05. Por supuesto que las tablas exactas son preferibles a la aproximación normal siempre que estén a nuestro alcance.

Empates. Si ocurren empates hemos de suponer una vez más que se deben a imperfecciones de medición y que las distribuciones subyacentes son en realidad continuas. Si los empates tienen lugar en el interior de las clases, éstas no tendrán, por supuesto, efecto alguno sobre la U , y podemos proceder como anteriormente. Y si los empates tienen lugar entre clases, damos a cada uno de los casos el promedio de las marcas que habría tenido de no existir aquéllos. Así, pues, si dos organizaciones están empatadas en los órdenes octavo y noveno, cada una de ellas recibe una marca de $(8 + 9)/2$, o sea 8.5. Si la décima organización hubiera estado asimismo empatada con los dos grupos anteriores, cada uno de ellos habría recibido el orden $(8 + 9 + 10)/3$, o sea 9.0. Al calcular U , se producirá probablemente ahora menos confusión si nos servimos del método de la suma de los órdenes. En efecto, el factor de corrección comporta el error estándar de U y, por consiguiente, aparece en el denominador de Z . La fórmula revisada se convierte así en:

$$Z = \frac{U - N_1 N_2 / 2}{\sqrt{[N_1 N_2 / N(N-1)] [(N^3 - N) / 12 - \sum T_i]}} \quad (\text{XIV.9})$$

en donde $N = N_1 + N_2$ y $T_i = (t_i^3 - t_i) / 12$, siendo t el número de observaciones empatadas en relación con un orden determinado.

Al calcular $\sum T_i$, observamos primero todos los casos en los que se dan empates. Tal vez dos grupos estén empatados en relación con las marcas octava y novena, y tres en relación con las marcas inferiores. En este caso tenemos una t de dos y una de tres. O sea:

$$\begin{aligned} \sum T_i &= T_1 + T_2 = \frac{t_1^3 - t_1}{12} + \frac{t_2^3 - t_2}{12} \\ &= \frac{2^3 - 2}{12} + \frac{3^3 - 3}{12} = \frac{6}{12} + \frac{24}{12} = 2.5 \end{aligned}$$

y

$$Z = \frac{21 - 45}{\sqrt{\frac{10(9)}{19(18)} \left(\frac{19^3 - 19}{12} - 2.5 \right)}} = -1.964$$

Esta corrección de los empates sólo puede emplearse con la aproximación normal, ya que las tablas exactas se han calculado sin tener en cuenta los empates. Por lo regular, el efecto del factor de corrección será despreciable salvo si el número de empates es muy grande.² Si el número de empates es extremadamente grande deberá usarse probablemente la prueba de Smirnov (ver más abajo) como alternativa a la de Mann-Whitney.

Comparación entre las pruebas de Mann-Whitney y de las secuencias. Para ambas pruebas la hipótesis nula es que las dos muestras se han extraído de poblaciones iguales. Por lo regular, nuestro interés se centra en las diferencias de tendencia central, como en el caso de la prueba de la diferencia de las medias. En ocasiones, sin embargo, podremos estar más interesados en las diferencias de dispersión o de forma. A título de enunciado general, podemos decir que la prueba de Mann-Whitney será más fuerte que la de las secuencias, siempre que las mayores diferencias entre las dos poblaciones sean con respecto a la tendencia central, en tanto que la segunda será más fuerte en aquellas situaciones en que las poblaciones sólo difieren ligeramente en cuanto a la tendencia central, pero sustancialmente, en cambio, en dispersión o en forma.

Un simple ejemplo servirá para ilustrar este punto. Supóngase que tenemos dos poblaciones de medianas iguales, pero, en un caso, con una distribución muy homogénea y, en el otro, muy heterogénea. Podríamos, en tal caso, esperar resultados como los siguientes:

Muestra 1	Muestra 2
5	1
6	2
7	3
8	4
9	13
10	14
11	15
12	16
<hr/> R ₁ = 68	<hr/> R ₂ = 68

² Parecería como si la corrección de los empates redujera siempre el denominador sin cambiar el numerador, pero debemos tener presente que dichos empates harán normalmente que U y U' se aproximen, lo que se traducirá a su vez en una disminución del numerador.

En este ejemplo extremo, la prueba de Mann-Whitney no conduciría a descartar la hipótesis nula (la cual es manifiestamente falsa), porque R_1 es exactamente igual a R_2 . Sirviéndonos de la prueba de las secuencias, en cambio, estaríamos manifiestamente en condiciones de descartar, porque sólo tendríamos tres secuencias. Como quiera que dejar de descartar significa cometer un error de tipo II, vemos que en este caso la fuerza de la prueba de las secuencias es mayor que la de la Mann-Whitney. En la mayoría de los casos, tenemos más probabilidades de hallar diferencias en la tendencia central, con diferencias relativamente menores en dispersión. El lector hará bien en convencerse por sí mismo que en el caso de tales poblaciones tenemos probabilidades de obtener un número relativamente grande de secuencias hacia el centro de la distribución. Y en relación con semejantes datos, la prueba de las secuencias será mucho menos fuerte que la de Mann-Whitney. En relación con la mayoría de las aplicaciones sociológicas, la prueba de Mann-Whitney parece ser la más útil de las dos.

* Si se ha conseguido un nivel de escala de intervalo y se suponen legítimamente poblaciones normales, pudo haberse efectuado la prueba t para la diferencia entre las medias. En tales condiciones, ¿cuánto perderíamos sirviéndonos de la prueba de Mann-Whitney, cayendo para ello hacia atrás en cuanto al nivel de medición y sirviéndonos de un modelo más débil? La evidencia está en que, en el caso de muestras medianas y grandes, la eficacia de la fuerza de la prueba de Mann-Whitney es aproximadamente del 95 por ciento en comparación con la de la t . La eficacia de la fuerza es asimismo muy grande en el caso de muestras pequeñas, pese a que los valores numéricos exactos no sean fáciles de obtener. Bradley [1] observa que en general la eficacia de muchas pruebas no paramétricas, entre ellas la de Mann-Whitney, es relativamente mayor para las muestras pequeñas que para las grandes. Así, pues, la prueba de Mann-Whitney constituye una alternativa muy fuerte de la prueba t . En vista del hecho de que requiere supuestos mucho más débiles, debería emplearse en aquellas situaciones en que existe alguna duda razonable de la legitimidad ya sea de la escala de intervalo o de la normalidad. Se sabe menos, en cambio, a propósito de la eficiencia de la fuerza de la prueba de las secuencias. Smith [8] ha encontrado eficiencias de aproximadamente el 75 por ciento en varios ejemplos empíricos, en los que los tamaños de las muestras eran de alrededor de 20 y las poblaciones normales presentaban desviaciones estándar iguales. Bradley [1] observa que la eficacia de la prueba de secuencias con muestra grande, es, por comparación con la prueba t , de aproximadamente un tercio, en igualdad de condiciones.

XIV.4. La prueba de Kolmogorov-Smirnov

La prueba de Kolmogorov-Smirnov, que designaremos simplemente como prueba de Smirnov, es otra prueba no paramétrica de dos muestras, que requiere los mismos supuestos que las pruebas de las secuencias y de Mann-Whitney. La fuerza de la prueba Smirnov es en general difícil de evaluar, pero en aquellas situaciones en que la población difiere solamente en relación con la tendencia central, dicha fuerza parece estar comprendida entre las de las pruebas de las secuencias y la de Mann-Whitney. (Bradley [1], pp. 291-292.) En un sentido estricto, la prueba Smirnov tampoco supone empates, pero, como veremos, el procedimiento es muy conveniente en las situaciones en que se da un buen número de empates, como resultado de haber agrupado los datos en categorías ordenadas.

En la investigación sociológica nos servimos con mucha frecuencia de variables que son en realidad escalas ordinales, pero en relación con las cuales los datos se han agrupado, con todo, en tres o más categorías grandes. Si se dan cuatro o más categorías ordenadas de esta clase, la prueba de Smirnov resultará particularmente útil, en tanto que el número de empates prohibiría probablemente el empleo de la de Mann-Whitney. Un sociólogo puede acaso haber dividido los residentes de una localidad en seis clases sociales, tratando a todas las personas de una clase como ligadas a los demás miembros de la misma con respecto a la característica general. O pueden haberse ordenado las ocupaciones según la condición de las personas asignándose a todas las de la misma clase de ocupación marcas empatadas. Tal vez se haya encontrado una variable de comportamiento que dé una escala de Guttman con siete tipos de respuestas. En todos estos ejemplos podemos querer concebir la variable como continua en realidad, pero el instrumento de medición ha sido excesivamente imperfecto y ha proporcionado datos que se hallan agrupados en un número relativamente pequeño de categorías ordenadas. Lo mismo que en el caso de las escalas de intervalo, cuanto más sutiles sean las distinciones y cuanto mayor sea el número de las categorías empleadas, tanto menos información se pierde.

El principio que se halla en la base de la prueba de Smirnov es muy sencillo. Si la hipótesis nula de que se han extraído muestras aleatorias independientes, de poblaciones idénticas, es correcta, entonces esperamos que las distribuciones de frecuencia acumulada de las dos muestras sean fundamentalmente similares. La estadística de la prueba empleada en la prueba de Smirnov es la diferencia máxima entre las dos distribuciones acumuladas. Si dicha diferencia es mayor de lo que se esperaría por azar con la hipótesis nula, esto significa que la diferencia entre las distribuciones se ha hecho tan grande que decidimos

descartar la hipótesis. Podemos tomar la diferencia máxima ya sea en una sola dirección (si ésta se ha anticipado) o en ambas direcciones.

Problema. Supóngase que hemos dividido una muestra al azar de varones adultos de una localidad en seis clases sociales y los hemos clasificado al propio tiempo según sus aspiraciones bajas o altas de cambio. Estas dos últimas categorías pueden considerarse como muestras aleatorias independientes de las poblaciones más amplias de varones adultos con aspiraciones bajas o respectivamente altas, ya que una muestra total completamente al azar asegura la independencia entre las submuestras que podamos escoger. Supóngase que hemos anticipado que los de aspiraciones de cambio elevadas tenderán a ocupar una posición de clase superior a los de aspiraciones bajas. ¿Podemos concluir que los resultados son significativos al nivel de .01?

Clase	Aspiraciones bajas	Aspiraciones altas
Baja inferior	58	31
Baja superior	51	46
Media inferior	47	53
Media superior	44	73
Alta inferior	22	51
Alta superior	14	20
Total	236	274

1. *Supuestos.* Los mismos que se requieren en la prueba de Mann-Whitney y la de las secuencias.

2. *Distribución de muestreo.* La distribución de muestreo de D , o sea la diferencia máxima entre las distribuciones acumulativas, puede darse exactamente en el caso de N pequeñas (≤ 40), si $N_1 = N_2$ ([7], p. 129). Este caso no lo trataremos, ya que con N relativamente pequeñas puede emplearse, en lugar de la prueba de Smirnov, la de Mann-Whitney, y porque en la mayoría de los ejemplos sociológicos no solemos por lo regular obtener muestras exactamente del mismo tamaño. Si las dos muestras son mayores que 40 y si no se ha anticipado la dirección, necesitaremos un valor de D que sea por lo menos tan grande como

$$1.36 \sqrt{\frac{N_1 + N_2}{N_1 N_2}}$$

para poder descartar al nivel de .05. En relación con los niveles de .01 y .001, el coeficiente de 1.36 puede remplazarse por 1.63

y 1.95 respectivamente. En el caso del nivel de .10, el coeficiente correspondiente es de 1.22.

Si la dirección se ha anticipado, podemos servirnos de la aproximación de la χ -cuadrada. La estadística de la prueba χ -cuadrada (χ^2) se considerará en el capítulo siguiente, y la tabla de la misma resultará más familiar en dicho momento.³ Entretanto, la fórmula de la aproximación es como sigue:

$$\chi^2 = 4D^2 \frac{N_1 N_2}{N_1 + N_2} \quad (\text{XIV.10})$$

en donde los grados de libertad asociados a la χ -cuadrada son siempre dos en esta particular aplicación. Si bien al emplear la aproximación de la χ -cuadrada se suponen distribuciones continuas de la población, si los datos son en realidad discretos y dan lugar, en consecuencia, a grandes números de empates, las probabilidades obtenidas quedarán, en caso de desearse el descarte, en sentido conservador. En otros términos: las verdaderas probabilidades serán menores que las calculadas.

3. *Nivel de significado y región crítica.* El problema requiere el nivel de significado de .01. Ya que se ha anticipado la dirección, nos serviremos de la aproximación de la χ -cuadrada.

4. *Cálculo de la estadística de la prueba.* Obtenemos primero las distribuciones de frecuencia acumulada de cada una de las muestras (véase cuadro XIV.1), expresando los valores de F como *proporciones* de las magnitudes totales de las muestras. Así, pues, el primer valor inscrito en la columna de las F de la muestra 1 será $58/236$, o sea .246; el segundo será $109/236$, o .462, y así sucesivamente. Las últimas anotaciones de cada columna serán, por supuesto, la unidad. Formamos ahora una columna de las diferencias, $F_1 - F_2$, y localizamos la diferencia mayor con el signo positivo, ya que anticipamos mayores porcentajes de las clases inferiores con aspiraciones bajas, o sean mayores valores de F_1 . Este valor de D resulta ser de .187, como lo indica la flecha. A continuación calculamos el valor de la χ -cuadrada sirviéndonos de la ecuación (XIV.10).

5. *Decisión.* Obsérvese que cuanto mayor sea el valor de D , tanto mayor será la χ -cuadrada. Por lo tanto, sólo necesitamos saber cuán grande deba ser ésta para descartar la hipótesis nula. Recurrimos, pues, al cuadro de la χ -cuadrada (cuadro I del Apéndice 2), buscamos los grados de libertad de arriba abajo en el margen izquierdo, y el nivel de significación, arriba, de izquierda a derecha, y vemos que, con 2 grados de libertad, corresponde al nivel .01 el valor 9.210. Esto significa que si la hipótesis nula fuera cierta, obtendríamos una χ -cuadrada de esta magnitud, o

³ Debido a ello tal vez se podrá aplazar el estudio de la prueba de Smirnov hasta después de haber leído el cap. xv.

CUADRO XIV.1. Cálculos para la prueba de dos muestras de Smirnov

Clase	Aspiraciones de cambio		Diferencias $F_1 - F_2$		
	Altas F_1	Bajas F_2			
Debajo de la baja superior	58	.246	31	.113	.133
Debajo de la media inferior	109	.462	77	.281	.181
Debajo de la media superior	156	.661	130	.474	.187 ←
Debajo de la alta inferior	200	.847	203	.741	.106
Debajo de la alta superior	222	.941	254	.927	.014
Total	236	1.000	274	1.000	

$$\chi^2 = 4D^2 \frac{N_1 N_2}{N_1 + N_2} = 4(.187)^2 \frac{236(274)}{236 + 274} = 17.74$$

mayor acaso, menos del uno por ciento de las veces. Y ya que obtuvimos una χ -cuadrada de 17.74, vemos que podemos descartar la hipótesis nula. Esta misma prueba de la χ -cuadrada puede emplearse en relación con muestras pequeñas cuando se ha anticipado la dirección; y si se tiene interés en descartar la hipótesis nula, la aproximación de la χ -cuadrada será en realidad conservadora. En otros términos: las probabilidades obtenidas con este método serán mayores que las reales.

Si no se hubiera anticipado la dirección, entonces necesitaríamos, para obtener significación al nivel de .01, un valor de D que sea por lo menos igual o superior a

$$1.63 \sqrt{\frac{N_1 + N_2}{N_1 N_2}} = 1.63 \sqrt{\frac{236 + 274}{236(274)}} = 1.63(.0888) = .145$$

En este caso obtenemos D tomando la diferencia mayor, independientemente del signo. Y como quiera que este valor es el mismo que el de la D utilizada anteriormente (.187), vemos que podemos también descartar la hipótesis nula.

XIV.5. La prueba de Wilcoxon de pares asociados y órdenes provistos de signo

Las tres pruebas no paramétricas examinadas hasta aquí, en el presente capítulo, requerían que las dos muestras se selecciona-

ran independientemente una de otra. Se recordará que, al asociar pares, no podíamos servirnos de la prueba de la diferencia de las medias. En lugar de ello, tratábamos cada par como caso singular y obteníamos una marca de diferencia para cada uno de ellos. Procedíamos luego como si tuviéramos una sola muestra, y verificábamos la hipótesis nula de que $\mu_D = 0$. El lector recordará además que, al servirnos de la prueba de los signos, pudimos también haber empleado pares asociados, teniendo sólo en cuenta el signo de la diferencia y verificando la hipótesis nula con ayuda de la distribución binomial. En la prueba de los signos, habíamos de dejar de lado toda la información que poseyéramos acerca de la magnitud de las diferencias implicadas. Por otra parte, la prueba más fuerte, o sea la prueba t , requería no sólo una escala de intervalos, sino, además, el supuesto de una población normal de las marcas de las diferencias. La prueba de Wilcoxon de pares asociados y órdenes provistos de signo, en cambio, combina algunas de las características de esas dos pruebas y se sitúa, en cuanto a eficacia de la fuerza, entre ambas.

Como veremos en seguida, la prueba de Wilcoxon requiere un nivel de medición ligeramente superior al de la escala ordinal. Necesitaremos, en efecto, una escala métricamente ordenada, en la que sea posible ordenar no sólo las marcas mismas, sino además las *diferencias* entre ellas. Como quiera que las escalas métricamente ordenadas se encuentran raramente en la investigación sociológica, este requisito equivale prácticamente a que necesitemos una escala de intervalo. Sin embargo, ya que la prueba de Wilcoxon no presupone una población normal, la examinaremos junto con las demás pruebas de dos muestras no paramétricas en el presente capítulo. La eficiencia de la fuerza de esta prueba es sustancialmente mayor que la de la prueba de los signos, circunstancia que no debe sorprender, ya que ésta obtiene ventaja de tan poca información disponible. Si los supuestos de la prueba t son efectivamente ciertos, entonces la eficacia de la fuerza de la prueba de Wilcoxon es aproximadamente del 95 %, tanto para muestras pequeñas como grandes. De ahí que resulte particularmente útil en situaciones en las que tenemos un nivel de medición de escala de intervalo, pero en las que la magnitud de la muestra es con todo demasiado pequeña para justificar el supuesto de normalidad.

En esencia, la prueba de Wilcoxon comporta la misma hipótesis nula empleada en la prueba de los signos y también en la prueba t para pares asociados. La hipótesis nula sostiene que no existen diferencias entre las marcas de las dos poblaciones. Al servirnos de esta prueba, obtenemos primero las marcas de la diferencia para cada par. Estas diferencias se ordenan, prescindiendo de sus signos respectivos. Así, pues, una diferencia de -6 se ordenará por encima de la de $+3$. Una vez ordenados en esta

forma los valores absolutos de las diferencias, *asignando siempre el rango 1 a la menor diferencia numérica*, volvemos atrás y anotamos los signos. Finalmente, obtenemos las sumas de los órdenes de las diferencias, de las positivas y de las negativas. Si la hipótesis nula es correcta, esperamos que la suma de los órdenes de las diferencias positivas será aproximadamente igual a la de los órdenes de las diferencias negativas. Si éstas difieren mucho entre sí en magnitud, entonces la hipótesis nula puede descartarse. Formamos la estadística T , que es la *menor* de las sumas en cuestión. Nos servimos a continuación de tablas exactas de la distribución de selección de T si la N es pequeña, y de una aproximación normal si es grande.

CUADRO XIV.2. Cálculos de la prueba de Wilcoxon de pares asociados

Nº del par	Grupo A	Grupo B	Diferencia	Orden de la diferencia	Órdenes negativos
1	63	68	5	(+) 6	
2	41	49	8	(+) 10.5	
3	54	53	-1	(-) 1.5	1.5
4	71	75	4	(+) 5	
5	39	49	10	(+) 12	
6	44	41	-3	(-) 4	4
7	67	75	8	(+) 10.5	
8	56	58	2	(+) 3	
9	46	52	6	(+) 8	
10	37	49	12	(+) 13	
11	61	55	-6	(-) 8	8
12	68	69	1	(+) 1.5	
13	51	57	6	(+) 8	
Total					13.5

Con fines de comparación, sirvámolos de los mismos datos utilizados en el caso de la prueba t correspondiente. El cuadro XIV.2 repite dichos datos y proporciona al propio tiempo los cálculos necesarios para la prueba de Wilcoxon. Obsérvese que, al ignorar los signos, algunas de las marcas de las diferencias resultan empatadas en cuanto a la magnitud. En tal caso, damos una vez más a las marcas empatadas el valor promedio que habrían tenido de no estarlo.⁴ Así, por ejemplo, tenemos dos diferencias de tama-

⁴ Otro procedimiento algo más conservador consistiría en romper los empates de tal manera que se obtuviera el mayor valor posible de T . Los pares cuyas puntuaciones tuviesen una diferencia de 0 exactamente (es decir: ningún cambio) deberían ser eliminados del análisis.

ño 1. Como quiera que atribuimos a las diferencias más pequeñas los rangos inferiores, cada una de aquéllas obtiene una marca de orden o rango de 1.5. En la quinta columna hemos indicado el signo asociado a cada orden, entre paréntesis, a la izquierda del mismo. Vemos a simple vista que la suma de los órdenes negativos será inferior a la de los positivos. Por lo tanto, obtenemos T sumando estos órdenes negativos. No es menester retener los signos negativos al buscar el valor de T en el cuadro, ya que los valores se hallan siempre dados como positivos. Así, pues,

$$T = 1.5 + 4 + 8 = 13.5$$

Formalicemos ahora lo que hemos hecho, siguiendo los pasos en la forma habitual.

1. Supuestos.

Nivel de medición: escala métricamente ordenada (las marcas de las diferencias pueden ordenarse).

Modelo: muestras aleatorias.

Hipótesis: la suma de los órdenes positivos = a la de los órdenes negativos en la población.

2. **Distribución de muestreo.** La distribución de muestreo de T para $N \leq 25$ se da en el cuadro H del Apéndice 2. En relación con muestras mayores, la distribución de T es aproximadamente normal, con:

$$\text{media} = \mu_T = \frac{N(N+1)}{4} \quad (\text{XIV.11})$$

$$\text{y desviación estándar} = \sigma_T = \sqrt{\frac{N(N+1)(2N+1)}{24}} \quad (\text{XIV.12})$$

3. **Nivel de significación y región crítica.** Lo mismo que en el caso de la prueba t , nos serviremos del nivel de .05, sin anticipar la dirección del resultado.

4. **Cálculo de la estadística de la prueba.** Éste se halla ya efectuado en el cuadro XIV.2, que nos da una T de 13.5.

5. **Decisión.** El cuadro H del Apéndice 2 da valores críticos de T para $N \leq 25$. Ya que T representa la menor de las dos sumas de órdenes, necesitamos valores numéricos pequeños de T para descartar la hipótesis nula. Así, pues, podremos descartar H_0 siempre que T sea igual o inferior a los valores dados en el cuerpo del cuadro. Vemos que con una N de 13 necesitamos una T de 17, o menor, para poder descartar al nivel de .05. Vemos también que se necesitaría una T de 13 o menos para el descarte al

nivel de .02. Al servirnos de la prueba t en el capítulo anterior, se habrá observado que, al nivel de .02, sólo logramos descartar allí con muy poco margen; aquí, en cambio, estamos ligeramente por encima de dicho nivel, pero los resultados de ambas pruebas son, con todo, muy similares.

Si bien nuestra N es muy pequeña, podemos de todos modos servirnos de la aproximación normal con fines de ilustración. Obtenemos:

$$Z = \frac{T - N(N+1)/4}{\sqrt{N(N+1)(2N+1)/24}}$$

$$= \frac{13.5 - 13(14)/4}{\sqrt{13(14)(27)/24}} = \frac{13.5 - 45.5}{\sqrt{204.75}} = -2.24$$

Como quiera que una Z de -2.24 corresponde a $p = .025$, resulta que volvemos a llegar a la misma conclusión. El valor de T es mucho menor que el que esperaríamos debido al azar, y podemos en consecuencia descartar la hipótesis nula. Debe observarse que la anterior aproximación normal no contiene una corrección explícita de los empates, no debiendo por tanto ser usada en los casos en que el número relativo de empates es extremadamente grande.

XIV.6. Resumen

En el presente capítulo hemos examinado cuatro pruebas no paramétricas distintas. En los capítulos sucesivos veremos otras. El lector habrá observado, sin duda, que todas esas pruebas no paramétricas comportan hasta aquí ideas muy simples y considerablemente menos cálculos que la prueba de la diferencia de las medias, por ejemplo. Ésta es una razón más en favor de nuestra tesis en el sentido de que, en el futuro, los sociólogos se servirán mucho más de estas pruebas no paramétricas. Por desgracia, en un texto general no se puede hacer mucho más que examinar unas pocas de esas pruebas brevemente. Algunas de las pruebas examinadas en el presente capítulo tienen además algunas otras aplicaciones que no se han examinado. Así, por ejemplo, la prueba de las secuencias puede emplearse como prueba del carácter fortuito. La prueba de Smirnov, por su parte, puede utilizarse como prueba de una sola muestra para comparar las frecuencias observadas con las que se han anticipado teóricamente. En algunos casos, pueden obtenerse intervalos de confianza empleando procedimientos no paramétricos. Por lo tanto, una vez que se haya familiarizado con las pruebas tratadas en este texto, el lector propenderá tal vez a consultar obras más especia-

lizadas. Afortunadamente, muchos de estos procedimientos no paramétricos pueden comprenderse fácilmente, aun por parte del lector sin gran preparación matemática. Es una suerte, también, que cierto número de esos procedimientos hayan sido resumidos en textos recientes de Siegel [7], Bradley [1] y Pierce [5]. El lector podrá consultar asimismo con provecho la extensa bibliografía sobre métodos no paramétricos compilada por Savage [6].

Tanto en este capítulo como en el anterior hemos observado que es necesario distinguir entre muestras que fueron seleccionadas independientemente, y aquellas que han sido pareadas o en las que se incluyen comparaciones de puntuaciones correspondientes a los mismos individuos. De esta manera, la independencia, o la falta de ella, entre muestras, es una de las consideraciones que deben hacerse cuando se escoge entre distintos procedimientos estadísticos. En el caso de muestras pareadas formamos una simple puntuación para cada par, y a continuación utilizamos el dato como si se tratara de una sola muestra. Cuando las muestras han sido tratadas independientemente, no siendo tal vez iguales por otra parte los tamaños de las muestras, formulamos la hipótesis nula, suponiendo que ha habido muestreos independientes de las mismas poblaciones, y que la distribución de nuestra estadística de la prueba (Z , t , r , U o D) se basó en dicho supuesto. Estos principios se amplían fácilmente a más de dos pruebas. En los capítulos xv y xvi observaremos comparaciones entre tres o más muestras seleccionadas independientemente, cuando la segunda variable puede ser una escala nominal, ordinal o de intervalo. Aun cuando no nos centraremos en el examen de ejemplos más complejos, en los que haya implícitas más de dos muestras pareadas, podrá verse en el ejercicio 5 del capítulo anterior y en el ejercicio 5 de este mismo capítulo, que la ampliación es directa. La idea básica es la de que uno obtiene una sola puntuación para cada par (la que puede resultar de una diferencia de diferencias o alguna otra función más compleja), procediendo a continuación como si se hubiese tratado de una simple muestra de tamaño N , en la que N representa el número de pares (o de tríos, etcétera).

En el presente capítulo abordamos por primera vez un problema de tipo general, a saber: el del criterio que ha de aplicarse al escoger entre procedimientos estadísticos alternativos. Nos hemos centrado especialmente en el concepto de la eficacia de las fuerzas relativas de las pruebas por el hecho de que algunas de éstas requieren supuestos más fuertes que otras. No debe sin embargo el lector quedarse bajo la impresión de que el problema es tan sencillo como aquella distinción da a entender. Ya se ha hecho notar que en la mayoría de los casos prácticos no se conoce lo suficiente acerca del valor real de los parámetros como para basar en tal conocimiento unas conclusiones definitivas. Hay ade-

más otra cuestión, más técnica, que no hemos discutido. En ella están implicadas las sensibilidades relativas de las pruebas en orden a la violación de los supuestos requeridos. Por ejemplo: ¿qué perjuicio se causa si se utiliza una prueba de diferencia de medias cuando la población tiene una forma especificada como no normal? ¿Qué ocurre si se viola el supuesto de las escalas de intervalo? Los estadísticos emplean la expresión *robustez de una prueba* cuando aluden a su sensibilidad ante diversas clases de distorsiones. La robustez resulta particularmente difícil de evaluar cuando son *varias* las distorsiones, o los supuestos no cumplidos, cuya aplicación es simultánea. Aun cuando los procedimientos paramétricos, tales como la prueba de la diferencia de medias, pueden parecer razonablemente robustos bajo muchas situaciones, hay diferencias de opinión en cuanto a lo aconsejable que resulte utilizar tales pruebas cuando se dispone de alternativas no paramétricas.

Nuestra posición es la de que cuando no se pueden aplicar criterios claros lo prudente es utilizar varias pruebas diferentes, tanto paramétricas como no paramétricas, publicando los dos grupos de resultados para que el lector pueda sacar sus propias conclusiones. Habitualmente se hace esto dando, en notas al pie de la página, los resultados de una segunda prueba, comentando las razones por las cuales las conclusiones no resultaron idénticas. Cuando hay una prueba (o estimación) no paramétrica disponible, cuya fuerza sea casi tan alta como la de un procedimiento paramétrico comparable, tal como la prueba de Mann-Whitney como alternativa a la prueba *t*, parecería preferible confiar más bien en el procedimiento no paramétrico. Encontraremos empero muchos procedimientos paramétricos multivariados para los que no existe una alternativa no paramétrica satisfactoria. Antes que usar una alternativa débil o teóricamente no satisfactoria, resulta preferible, si tal es el caso, recurrir decididamente a los procedimientos paramétricos, conscientes de que no se podrán obtener con ellos resultados definitivos. No es posible, en pocas palabras, dar una simple respuesta dogmática a la pregunta: ¿Qué clase de prueba o medida es la más apropiada?

GLOSARIO

- Prueba no paramétrica
- * Eficiencia de fuerza
- * Función de fuerza
- * Fuerza de una prueba

EJERCICIOS

1. Se ha clasificado cierto número de iglesias protestantes de una localidad como: 1) predominantemente de clase alta o clase media

alta, o 2) predominantemente de clase media baja o clase baja. Se ordenan según el grado de formalismo de sus servicios, con los siguientes resultados:

Clase alta o clase media alta: órdenes 1, 2, 3, 6, 7, 8, 11, 13, 14, 15, 17, 21, 25

Clase media baja o clase baja: órdenes 4, 5, 9, 10, 12, 16, 18, 19, 20, 22, 23, 24, 26, 27.

Sirviéndose del nivel de .05, ¿puede establecerse una diferencia significativa: a) con la prueba de las secuencias, y b) con la de Mann-Whitney? ¿Qué prueba preferiría el lector? ¿Por qué? Respuesta, a) $r = 14$, no rechazar; b) $U = 52$, no rechazar.

2. En el cuadro 18.3 se dan datos de los grados de popularidad de los miembros de un grupo de un campo de trabajo de verano. Considérese a las personas con los grados de participación del 1 al 8 como "activas" en las discusiones del grupo, poniendo a las demás en la categoría de "inactivas". ¿Existe al nivel de .05 diferencia significativa alguna entre las personas "activas" y las "inactivas" por lo que se refiere a la popularidad? Empléense sucesivamente las pruebas de las secuencias y de Mann-Whitney.

3. Supóngase que se ha logrado ordenar las ocupaciones urbanas por grados descendentes, sirviéndose de las categorías generales de profesional y directivo, empleado, obrero calificado, semicalificado y no calificado. El investigador ha preguntado a todos los padres de familia si son o no partidarios del aumento de los beneficios de la seguridad social a expensas del contribuyente. Los resultados son como sigue:

Nivel de ocupación	Partidarios	Contrarios
Profesional y directivo	46	97
Empleado	81	143
Obrero calificado	93	88
Obrero semicalificado	241	136
Obrero no calificado	131	38
Total	592	502

¿Existe alguna relación significativa entre la ocupación y la actitud al nivel de .001? Respuesta, $D = .282$, $P < .001$.

4. Resolver el ejercicio 2 del capítulo XIII utilizando la prueba de Smirnov. Comparar estos resultados con los de la prueba *t*.

5. Efectúense todas las indagaciones del ejercicio 5 del capítulo XIII, sirviéndose de la prueba de Wilcoxon de los pares asociados y los órdenes provistos de signo. ¿Cómo se comparan entre sí los resultados de las dos pruebas? Respuesta, a) $T = 14.5$, no rechazar; c) $T = 11$, no rechazar.

* 6. Verifíquese que la ecuación (XIV.8) es equivalente desde el punto de vista algebraico a la otra fórmula de Z dada en la página 273.

BIBLIOGRAFÍA

1. Bradley, J. V.: *Distribution-free Statistical Tests*, Prentice-Hall, Inc., Englewood Cliffs, N. J., 1968, caps. 1-3, 5, 11 y 13.
2. Dixon, W. J., y F. J. Massey: *Introduction to Statistical Analysis* 3ª ed., McGraw-Hill Book Company, Nueva York, 1969, cap. 17.
3. Freund, J. E.: *Modern Elementary Statistics*, 3ª ed., Prentice-Hall, Inc., Englewood Cliffs, N. J., 1967, cap. 13.
4. Hays, W. L.: *Statistics*, Holt, Rinehart and Winston, Inc., Nueva York, 1963, cap. 18.
5. Pierce, Albert: *Fundamentals of Nonparametric Statistics*, Dickenson Publishing Company, Inc. Belmont, Cal., 1970, cap. 14.
6. Savage, I. R.: "Bibliography of Nonparametric Statistics and Related Topics", *Journal of the American Statistical Association*, vol. 48, pp. 844-906, 1953.
7. Siegel, S.: *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill Book Company, Inc., Nueva York, 1956, caps. 5 y 6.
8. Smith, K.: "Distribution-free Statistical Methods and the Concept of Power Efficiency", en L. Festinger y D. Katz (eds.) *Research Methods in the Behavioral Sciences*, The Dryden Press, Inc., Nueva York, 1953, pp. 536-577.
9. Swed, F. S., y C. Eisenhart: "Tables for Testing Randomness of Grouping in a Sequence of Alternatives", *Annals of Mathematical Statistics*, vol. 14, pp. 66-87, 1943.
10. Walker, H. M. y J. Lev: *Statistical Inference*, Henry Holt and Company, Inc. Nueva York, 1953, cap. 18.

XV. ESCALAS NOMINALES: PROBLEMAS DE CONTINGENCIA

EN EL presente capítulo vamos a estudiar las relaciones entre dos o más escalas nominales. Ya vimos que el caso de dos escalas nominales dicotómicas podía tratarse como un problema que comportara una diferencia de proporciones. Resulta a menudo deseable servirse de un procedimiento de prueba más general, que nos ponga en condiciones de averiguar las diferencias que haya entre tres o más muestras, o de comparar dos (o más) muestras con respecto a una variable de más de dos categorías. La prueba de la χ -cuadrada que vamos a examinar en la próxima sección nos permite establecer relaciones entre escalas nominales con cualquier número de categorías. Se introducirán al propio tiempo algunos conceptos nuevos. Hasta aquí sólo nos hemos ocupado de pruebas acerca de la *existencia* de una relación entre dos variables. En este capítulo se presentarán algunas medidas indicativas de la *fuerza* o grado de relación. Se examinarán al propio tiempo procedimientos empleados para el control de una o más variables.

XV.1. La prueba de la χ -cuadrada

La prueba de la χ -cuadrada es una prueba muy general que puede emplearse cuando deseamos apreciar si unas frecuencias obtenidas empíricamente difieren significativamente o no de las que se esperarían bajo cierto conjunto de supuestos teóricos. La prueba general presenta muchas posibilidades de aplicación, la más común de las cuales, en ciencias sociales, es la relativa a los problemas de "contingencia" en los que dos variables de escala nominal se han clasificado por comparación de una con otra.¹ Supóngase, por ejemplo, que se han relacionado una con otra la confesión religiosa y la filiación política y que los datos se han resumido en el siguiente cuadro de contingencia de 3×3 :

Partido	Protestantes	Católicos	Judíos	Total
Republicanos	126	61	38	225
Demócratas	71	93	69	233
Independientes	19	14	27	60
Total	216	168	134	518

¹ En relación con otro empleo de la χ -cuadrada, véase el ejercicio 3 al final del capítulo.

Obsérvese que si las frecuencias se convirtieran en porcentajes, podríamos decir que, en tanto que el 58.3 por ciento de los protestantes son republicanos, sólo prefieren este partido el 36.3 por ciento de los católicos y el 28.4 por ciento de los judíos. Se nos podría entonces ocurrir preguntar si esas diferencias eran o no significativas desde el punto de vista estadístico. Como quiera que se tienen tres confesiones religiosas y tres categorías de preferencia política, no podemos servirnos de una simple prueba de las diferencias de las proporciones. Sin embargo, sirviéndonos de la prueba de la χ -cuadrada, podemos establecer esencialmente la misma clase de hipótesis nula que anteriormente. Podemos suponer, en efecto, que no existe diferencia alguna entre las tres confesiones religiosas. Esto equivale a decir que las proporciones de republicanos, de demócratas y de independientes deberían ser las mismas en cada uno de dichos grupos. Partiendo, pues, del supuesto de que la hipótesis nula es correcta y de que las muestras son aleatorias e independientes, podemos calcular un conjunto de frecuencias que podrían esperarse, dados los totales marginales en cuestión. En otros términos, podemos calcular el número de protestantes de los que esperaríamos fueran republicanos y comparar esta cifra con la que se ha obtenido en realidad. Si la diferencia y las diferencias correspondientes a las otras casillas son considerables, probablemente sospechemos de la hipótesis nula.

Hay que obtener, pues, alguna medida de la diferencia entre las frecuencias observadas y las esperadas. Existe, por supuesto, una gran cantidad de medidas, pero necesitamos una con respecto a la cual la distribución de muestras sea conocida y esté tabulada. Por ello nos servimos de una media designada como de la χ -cuadrada (χ^2), que se define como sigue:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} \quad (\text{XV.1})$$

en lo que f_o y f_e se refieren respectivamente a las frecuencias observadas y esperadas para cada casilla.² O en otras palabras: la χ -cuadrada se obtiene tomando primero el cuadrado de la diferencia entre las frecuencias observadas y esperadas para cada casilla. Dividimos dicha cifra entre el número de casos *esperados* en cada casilla, con objeto de normalizarla, de modo que las mayores contribuciones no provengan siempre de las casillas mayores. Y la suma de todas esas cantidades no negativas para todas las casillas es el valor de la χ -cuadrada.

² Con objeto de reducir la confusión hemos abandonado el índice i , suponiéndose, con todo, que estamos sumando los resultados de todas las casillas.

Obsérvese que cuanto mayores son las diferencias entre las frecuencias observadas y las esperadas, tanto mayor es el valor de la χ -cuadrada. Esta sólo será cero si todas las frecuencias observadas y esperadas son idénticas. Podemos proceder a una verificación de la hipótesis nula buscando la distribución de muestreo de la χ -cuadrada. Dificilmente anticiparemos que las frecuencias observadas y las esperadas sean exactamente las mismas. Sin embargo, si el valor de la χ -cuadrada resulta mayor de lo que al azar se anticiparía, estaremos en condiciones de descartar la hipótesis nula siguiendo el procedimiento habitual.

Problema. Podemos servirnos del ejemplo puesto anteriormente, pero simplificándolo, de manera que obtengamos una tabla de 2×2 . La extensión del mismo al caso general resultará después muy sencilla. Supongamos, pues, que se han combinado los católicos y los judíos y que se ha prescindido de los independientes. Tenemos así el siguiente cuadro:

Partido	Protestantes	Católicos y judíos	Total
Republicanos	126	99	225
Demócratas	71	162	233
Total	197	261	458

Importa observar que las cifras de cada casilla son en realidad frecuencias y no porcentajes. Si las cifras dadas son porcentajes, hay que convertirlas en frecuencias, ya que, desde el punto de vista estadístico, la prueba de la χ -cuadrada comporta una comparación de frecuencias y no de porcentajes.

1. Supuestos.

Nivel de medición: dos escalas nominales

Modelo: muestras aleatorias independientes

Hipótesis: no existen diferencias entre las poblaciones profesionales en relación con la preferencia política.

Por supuesto, el nivel de medición *puede* ser más elevado. En efecto, las pruebas de la χ -cuadrada se utilizan con frecuencia con escalas ordinales e inclusive, en ocasiones, con escalas de intervalo. Sin embargo, según vimos en los capítulos precedentes, se dispone en tales casos de pruebas más fuertes que se emplearán por lo regular con preferencia a la χ -cuadrada. Una vez más, hay que suponer independencia entre las muestras para servirse de la prueba de la χ -cuadrada. La magnitud de la mues-

tra ha de ser relativamente grande, porque la χ -cuadrada, según la define la fórmula, tiene una distribución de muestreo que sólo se aproxima a la del cuadro si N es grande.³

La hipótesis nula puede formularse en cierto número de modos equivalentes. Decir que no hay diferencia entre grupos confesionales en materia de preferencia política equivale esencialmente a decir que no hay diferencia alguna entre la filiación religiosa y la preferencia electoral. Hay que tener presente, sin embargo, que semejante afirmación sólo se aplicaría a las variables tales como se las haya definido operativamente; en este caso, por ejemplo, la preferencia política y la religión se definirían como variables dicotómicas. Podría también enunciarse la hipótesis nula enumerando las diversas proporciones que se suponen iguales. Si bien este último método sea tal vez el más preciso, puede resultar con todo muy embarazoso en el caso general.

2. *Nivel de significación.* Supongamos que queremos demostrar una diferencia y que deseamos ser extremadamente cautos. Nos serviremos, en consecuencia, del nivel de .001. Supóngase asimismo que no se ha anticipado la dirección de la diferencia.

3. *Distribución de muestreo.* Las distribuciones de muestreo de la χ -cuadrada están dadas en el cuadro I del Apéndice 2. Obsérvese que las distribuciones difieren de acuerdo con los grados de libertad implicados. La determinación de los grados de libertad se examinará más abajo. Como quiera que, independientemente de la dirección de la relación entre la confesión y la preferencia política, nuestro interés está en saber si la χ -cuadrada obtenida es o no mayor de lo que se esperaría al azar, sólo nos ocupamos de la cola mayor de la distribución. La cola menor, que consta de valores muy pequeños de la χ -cuadrada, no se suele emplear por lo regular en los problemas de contingencia.

4. *Cálculo de la estadística de la prueba.* Lo primero que hacemos en el cálculo de la χ -cuadrada es obtener las frecuencias esperadas. La hipótesis nula dice que no hay preferencias de la gente en cuanto a la votación. Por lo tanto, independientemente de cuál sea el verdadero número de republicanos en cada una de las poblaciones confesionales, esperaríamos que, a la larga, habría la misma proporción de aquéllos en ambas muestras. Como quiera que la proporción de republicanos en la muestra combinada es de 225/458, o sea .4913, esperaríamos la misma cifra en cada una de las dos muestras confesionales. Así, pues, anticiparíamos en cada uno de ellos los mismos porcentajes de republicanos y de demócratas. Podemos obtener luego el número esperado de republicanos entre los protestantes multiplicando .4913 por el número total de protestantes de la muestra. En esta forma, el número anticipado de protestantes republicanos sería (.4913)

³ Para un examen más detallado de este problema véanse las pp. 299-301.

(197) = 96.8. Las demás frecuencias anticipadas pueden calcularse en forma análoga. Por lo regular se recomienda retener por lo menos una cifra decimal al calcular las frecuencias esperadas. De modo que en el caso anterior no redondearíamos a 97.

Antes de pasar adelante, conviene observar que las frecuencias esperadas también pueden obtenerse razonando en forma inversa, esto es, en términos de la proporción de republicanos que esperaríamos que fueran protestantes. Toda vez que la proporción de protestantes en la muestra combinada es de 197/458, o sea .4301, podemos obtener la frecuencia anticipada de republicanos protestantes como sigue: (.4301)(225) = 96.8. El lector ha de acostumbrarse a obtener las frecuencias esperadas en ambas formas, a título de control de los cálculos.

Una vez que nos hayamos acostumbrado al procedimiento, encontraremos probablemente más sencillo servirnos de una simple fórmula como la que se describe a continuación. Si designamos las casillas y los totales marginales como

a	b	$a + b$
c	d	$c + d$
$a + c$ $b + d$		N

entonces la frecuencia esperada puede obtenerse multiplicando los dos marginales correspondientes a la casilla en cuestión y dividiendo entre N . Así, por ejemplo, la cifra esperada para la casilla a sería

$$(a + b)(a + c)/N = (225)(197)/458 = 96.8$$

El empleo de este último procedimiento reduce todo error de redondeo que podría introducirse dividiendo primero (para obtener la proporción) y multiplicando luego.

Se observará que este procedimiento de multiplicar marginales para dividirlos entre el número total de casos, viene a ser básicamente el mismo que se examinó en el capítulo IX en relación con la independencia de dos variables. Esto pone de relieve el hecho de que las frecuencias esperadas son computadas sobre la base del supuesto de que las variables no están relacionadas, en tanto que las frecuencias observadas nos muestran el grado en que se viola este supuesto. Recuérdese que si los eventos (o variables) A y B son estadísticamente independientes, el conocer el valor de uno no nos ayudará a predecir el otro. Si las frecuencias observadas y las esperadas son exactamente iguales, ello significaría, en nuestro ejemplo, que el conocer las diferencias religiosas de una persona no nos permitiría predecir sus inclinaciones políticas.

Por convención, ponemos por lo regular las frecuencias esperadas entre paréntesis, debajo de las frecuencias realmente obtenidas para cada casilla, tal como se indica a continuación:

Partido	Protestantes	Católicos y judíos	Total
Republicanos	126 (96.8)	99 (128.2)	225
Demócratas	71 (100.2)	162 (132.8)	233
Total	197	261	458

Los cálculos para la χ -cuadrada pueden resumirse en un cuadro como el XV.1. Obsérvese que la cantidad $f_o - f_e$ tiene el mis-

CUADRO XV.1. Cálculos de la χ -cuadrada

Casilla	f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$(f_o - f_e)^2 / f_e$
a	126	96.8	29.2	852.64	8.808
b	99	128.2	-29.2	852.64	6.651
c	71	100.2	-29.2	852.64	8.509
d	162	132.8	29.2	852.64	6.420
Total	458	458.0			30.388

mo valor para cada casilla. El lector debería convencerse por sí mismo de que esto será siempre así en el caso de tablas de 2×2 , pero que no se deja con todo generalizar a otros casos. El hecho de elevar este valor al cuadrado tiene por efecto la eliminación de las cantidades negativas. Importa que se empleen en el denominador las frecuencias *esperadas*, y no las observadas. En efecto, estas últimas variarán de una muestra a otra, y pueden incluso ser iguales a cero.

Resulta a menudo más conveniente servirse de una fórmula de cálculo que no requiera la sustracción efectiva de cada frecuencia esperada de su correspondiente observada. Desarrollando el numerador en la expresión de la χ -cuadrada y uniendo los términos obtenemos:

$$\begin{aligned}\chi^2 &= \sum \frac{(f_o - f_e)^2}{f_e} = \sum \frac{f_o^2 - 2f_o f_e + f_e^2}{f_e} \\ &= \sum \frac{f_o^2}{f_e} - 2 \sum f_o + \sum f_e\end{aligned}$$

Pero, toda vez que tanto $\sum f_o$ como $\sum f_e$ son iguales a N , tenemos:

$$\chi^2 = \sum \frac{f_o^2}{f_e} - N \quad (\text{XV.2})$$

Sirviéndonos de esta fórmula, que comporta una sola sustracción, obtenemos el mismo resultado que anteriormente (véase cuadro XV.2).

CUADRO XV.2. Cálculo de la χ -cuadrada sirviéndose de la fórmula

Casilla	f_o^2	f_o^2 / f_e
a	15 876	164.008
b	9 801	76.451
c	5 041	50.309
d	26 244	197.620
Total		488.388
$\chi^2 = 488.388 - 458$		$= 30.388$

En el caso de una tabla de solamente 2×2 , resulta posible expresar la χ -cuadrada como simple función de las frecuencias de las casillas y de los totales marginales. Si se designan las casillas como anteriormente, tenemos:

$$\chi^2 = \frac{N(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)} \quad (\text{XV.3})$$

Si bien este cálculo requiere la multiplicación de números grandes, el empleo de los logaritmos lo simplificará con todo considerablemente. Vemos el paso, de la ecuación (XV.3), que la χ -cuadrada será cero cuando el producto diagonal ad sea exactamente igual al producto bc . Este hecho puede emplearse como método rápido para saber si es o no necesario seguir adelante con la prueba de significación. Si los productos diagonales son casi iguales, la χ -cuadrada será demasiado pequeña para proporcionar significación. Estos productos diagonales sirven asimismo para determinar la dirección de la relación sin que tengamos que molestarnos en calcular los porcentajes. El mayor de los dos productos indica, en efecto, cuál de las diagonales contiene la mayoría de los casos.

* Tanto las anteriores fórmulas para χ (chi) al cuadrado, como el procedimiento para calcular frecuencias esperadas, son sufi-

cientes en la mayoría de los casos, pero puede resultar útil conocer una versión algo distinta, aplicable al caso $r \times c$ en general, conveniente para quienes deseen proseguir el tema de la χ al cuadrado en otros textos más avanzados. Esta formulación alternativa será utilizada más adelante para obtener el límite superior de χ al cuadrado en el caso general $r \times c$. Por otra parte, esta forma alternativa para la fórmula no requiere el cálculo explícito de las frecuencias esperadas.

Sea N_{ij} = número observado en (i, j) -ésima casilla del cuadro, y e_{ij} = número esperado (bajo H_0) en la casilla (i, j) ,

para $i = 1, 2, \dots, r$; y $j = 1, 2, \dots, c$.

Sea $N_{i.} = \sum_{j=1}^c N_{ij}$, para $i = 1, 2, \dots, r$ (total de filas), y

$N_{.j} = \sum_{i=1}^r N_{ij}$, para $j = 1, 2, \dots, c$ (total de columnas).

Así podremos expresar χ al cuadrado como sigue

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(N_{ij} - e_{ij})^2}{e_{ij}}$$

pero puesto que

$$e_{ij} = N \frac{N_{i.}}{N} \frac{N_{.j}}{N} = \frac{N_{i.} N_{.j}}{N}$$

la fórmula computadora (XV.2) pasa a ser

$$\chi^2 = N \left[\sum_{i=1}^r \sum_{j=1}^c \frac{N_{ij}^2}{N_{i.} N_{.j}} - 1 \right]$$

y así vemos que no hay necesidad de computar explícitamente las frecuencias esperadas.

5. *Decisión.* Antes de servirnos del cuadro de la χ -cuadrada, hemos de determinar los grados de libertad asociados a esta estadística de prueba. En los problemas anteriores, los grados de libertad dependían siempre del número de los casos seleccionados. En los problemas de contingencia, en cambio, dichos grados sólo dependen del número de *casillas* del cuadro. Al calcular las frecuencias esperadas, pudo observarse que no es necesario calcular valores para cada casilla, ya que la mayoría de ellas podían obtenerse por sustracción. Y de hecho, en la tabla de 2×2 sólo

necesitamos calcular una de las frecuencias esperadas, y las otras quedan automáticamente determinadas. Esto es así porque, para calcular las frecuencias esperadas, nos servimos de los totales marginales de nuestra muestra. En otros términos: si ponemos el valor de una casilla cualquiera, los demás valores están perfectamente determinados, ya que las frecuencias esperadas han de tener los mismos totales marginales que las observadas. Por lo tanto, sólo tenemos un grado de libertad.

Habiendo, pues, averiguado que en la tabla de 2×2 sólo hay un grado de libertad, busquemos en el cuadro de la χ -cuadrada a lo largo de la hilera correspondiente a un grado de libertad hasta encontrar el nivel de significación deseado. Vemos en esta forma que al nivel de .001 le corresponde una χ -cuadrada de 10.827. Esto significa que, si todos los supuestos son efectivamente correctos, obtendremos un valor de la χ -cuadrada igual o mayor que ése una vez entre mil. En otros términos: sólo muy raramente diferirán las frecuencias observadas y las esperadas en una cantidad que dé una χ -cuadrada ≥ 10.827 , si no hubiera relación alguna entre la confesión religiosa y la preferencia en cuanto al voto (tal como se ha definido operativamente en este problema). Y como quiera que hemos obtenido para la χ -cuadrada un valor igual a 30.388, concluimos que la hipótesis nula puede descartarse al nivel de .001. Vemos, de paso, que, si N es grande, no es nada difícil llegar a obtener significación al nivel de .001.

Pese a que sólo nos ocupáramos de valores grandes de la χ -cuadrada, la dirección de la relación no se anticipó en el ejemplo anterior. Independientemente de si los protestantes presentaban más probabilidades de ser republicanos o demócratas, el resultado habría sido una χ -cuadrada grande si los porcentajes eran también grandes. En otros términos, la estadística de la prueba es aquí indiferente a la dirección de la relación, ya que comporta los cuadrados de las desviaciones y, por consiguiente, no puede ser negativa. Podemos sacar partido de las predicciones relativas a la dirección partiendo simplemente por la mitad el nivel de significación obtenido. En efecto, si la χ -cuadrada es lo bastante grande para dar significación al nivel de .10 sin anticipar dirección, el resultado será también significativo al nivel de .05, a condición, por supuesto, que la dirección de la relación se haya fijado de antemano.

Si el nivel de significación deseado no puede obtenerse exactamente de la tabla de la χ -cuadrada, se conseguirá una aproximación satisfactoria extrayendo la raíz cuadrada de la χ -cuadrada y recurriendo a la tabla normal. Así, por ejemplo, sabemos que una χ -cuadrada de 3.841 con un grado de libertad corresponde al nivel de .05 si no se ha adivinado la dirección. La raíz cuadrada de esta cifra es 1.96, que es el valor de Z necesario para obtener

significación con la tabla normal. Ésta, sin embargo, sólo puede emplearse en el caso de problemas de contingencia de 2×2 .

Caso general. En el caso general de la tabla de contingencia con r hileras y c columnas, los supuestos y cálculos para la χ -cuadrada sólo requieren una ligera modificación. La hipótesis nula de "ausencia de diferencias" o "ausencia de relación" implica ahora que cada población tendrá las mismas proporciones para cada una de las categorías de la segunda variable. Las frecuencias esperadas pueden obtenerse exactamente en la misma forma que anteriormente, pero tendremos ahora rc casillas, y los grados de libertad serán distintos.

Supóngase que nos servimos del mismo problema anterior, pero en su forma original, o sea la de una tabla de 3×3 . Observemos de paso que esta tabla nos proporciona mayor información que la de 2×2 , en la que los católicos y los judíos se combinaron en una sola categoría. Podemos, por lo tanto, esperar resultados que difieran algo de aquellos obtenidos anteriormente. Calculando las frecuencias esperadas por uno cualquiera de los métodos anteriormente sugeridos, obtenemos:

Partido	Protestantes	Católicos	Judíos	Total
Republicanos	126 (93.8)	61 (73.0)	38 (58.2)	225
Demócratas	71 (97.2)	93 (75.6)	69 (60.2)	233
Independientes	19 (25.0)	14 (19.4)	27 (15.6)	60
Total	216	168	134	518

Puede construirse una tabla de cálculo lo mismo que anteriormente (véase cuadro XV.3).

Para determinar los grados apropiados de libertad, observamos que, una vez las dos primeras frecuencias esperadas inscritas en la primera columna, la tercera se halla determinada por sustracción. Y lo mismo es cierto de la segunda. Todas las frecuencias esperadas de la tercera columna estarán determinadas a partir de los totales de la hilera. En términos generales: para cada una de las primeras $c - 1$ columnas será posible llenar todas las casillas menos una, o $r - 1$. La columna final estará, pues, siempre perfectamente determinada. Por lo tanto, el número de los grados de libertad de la tabla de contingencia de $r \times c$ puede expresarse por medio de la fórmula

$$df = (r - 1)(c - 1)$$

CUADRO XV.3. Cálculo de la χ -cuadrada para una tabla de contingencia de 3×3

Casilla	f_o	f_e	f_o^2	f_o^2/f_e
<i>a</i>	126	93.8	15 876	169.254
<i>b</i>	61	73.0	3 721	50.973
<i>c</i>	38	58.2	1 444	24.811
<i>d</i>	71	97.2	5 041	51.862
<i>e</i>	93	75.6	8 649	114.405
<i>f</i>	69	60.2	4 761	79.086
<i>g</i>	19	25.0	361	14.440
<i>h</i>	14	19.4	196	10.103
<i>i</i>	27	15.6	729	46.731
Total	518	518.0		561.665

$\chi^2 = 561.665 - 518 = 43.665$

Obsérvese que esta fórmula da un grado de libertad en el caso especial en que $r = c = 2$.

Toda vez que son 4 los grados de libertad asociados a nuestra tabla de 3×3 , vemos que para el rechazo al nivel de .001 se requiere una χ -cuadrada de 18.465. Rechazamos, por consiguiente, la hipótesis nula. Obsérvese que si para rechazar se requiere un valor mayor de la χ -cuadrada, es porque hay muchas más casillas que contribuyen a dicho valor. Como quiera que la χ -cuadrada representa una suma y no un promedio, esperaríamos que, en igualdad de condiciones, cuanto mayor sea el número de casillas, tanto mayor será la χ -cuadrada. El hecho de que el valor de la χ -cuadrada requerido para obtener significación aumente con los grados de libertad no debería sorprendernos.⁴

Corrección de continuidad. Ya se indicó que la prueba de la χ -cuadrada requiere una N relativamente grande debido al hecho de que la distribución de muestreo de la estadística de la prueba sólo se aproxima a la distribución de muestreo dada en la tabla de la χ -cuadrada si N es grande. Plántese, pues, naturalmente la cuestión de cuán grande debe ser N para que podamos servirnos de dicha prueba. La respuesta depende del número de casillas y de los totales marginales. Generalmente, cuanto menor sea el número de casillas y cuanto más aproximadamente iguales sean todos los totales marginales, tanto menor podrá ser N . Los criterios normalmente utilizados para decidir si el número de casos es o no suficiente, implican las frecuencias esperadas de cada casilla. Siempre que cualquiera de estas frecuencias sea

⁴ Obsérvese que esto era al revés en el caso de la distribución t . ¿Por qué?

aproximadamente de cinco o menor, se recomienda proceder a alguna clase de modificación, como se indica a continuación.

Se supone que la distribución de la χ -cuadrada es continua. En realidad, sin embargo, si el número de casos es relativamente pequeño, resulta imposible que el valor calculado de la χ -cuadrada tome muchos valores distintos. Esto es así porque las frecuencias observadas han de ser siempre números enteros. Al corregir con fines de continuidad, nos imaginamos que las frecuencias observadas pueden tomar efectivamente todos los valores posibles y nos servimos de los que quedan a una distancia de media unidad a uno y otro lado del entero obtenido, lo que dará los resultados más conservadores. En el caso de la tabla de 2×2 , la corrección de continuidad puede hacerse muy fácilmente. Esta corrección consiste ya sea en añadir o sustraer .5 de las frecuencias observadas, con objeto de reducir el tamaño de la χ -cuadrada. La versión corregida de la ecuación (XV.3) es la siguiente:

$$\chi^2 = \frac{N \left(|ad - bc| - \frac{N}{2} \right)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Para apreciar el efecto de la corrección de continuidad, podemos ver los siguientes cuadros:

(A)	7	13	20
	(10)	(10)	
	8	2	10
	(5)	(5)	
	15	15	30
	$\chi^2 = 5.40$		

(B)	7.5	12.5	20
	(10)	(10)	
	7.5	2.5	10
	(5)	(5)	
	15	15	30
	$\chi^2 = 3.75$		

En el cuadro B hemos corregido por razones de continuidad reduciendo las diferencias entre las frecuencias observadas y esperadas en media unidad. Hemos supuesto que había entre 6.5 y 7.5 casos en la casilla superior de la izquierda, y hemos tomado el número de 7.5, porque es el valor más cercano, al interior de dicho intervalo, de la frecuencia esperada de 10.0. En este ejemplo, la corrección de continuidad reduce el nivel de significación de aproximadamente .02 a algo más de .05. Es obvio, por lo demás, que las correcciones de continuidad producirán menos efecto cuando las frecuencias esperadas sean mayores. Toda vez que semejante corrección comporta en realidad un esfuerzo adicional muy pequeño y que, por otra parte, al proce-

der así actuamos en sentido conservador, se recomienda efectuarla siempre que en cualquier casilla la frecuencia esperada descienda por debajo de 10. Con muestras muy pequeñas, incluso esta corrección produce resultados engañosos. Para las tablas de 2×2 se dispone de una prueba alternativa que se examina en la sección siguiente.

En el caso de la tabla general de contingencia, las correcciones de continuidad no son fáciles de hacer. Si el número de casillas es relativamente grande y si solamente una o dos de las casillas tienen frecuencias esperadas de 5 o menos, entonces recomiendase, por lo general, seguir adelante con las pruebas de la χ -cuadrada, sin preocuparse mucho por tales correcciones. En cambio, si el número de casillas es pequeño, la única alternativa práctica consistirá tal vez en combinar las categorías de modo que dichas casillas resulten eliminadas. Por supuesto, las categorías sólo pueden combinarse si ello posee teóricamente algún sentido. Así, por ejemplo, si hubiera una categoría "de otras confesiones" que constara de un número tan grande de grupos confesionales que la categoría no tuviera teóricamente sentido alguno, tal vez sería preferible excluir a dichas personas por completo del análisis aunque, como regla general, no es buen sistema el de excluir datos de un análisis.

*XV.2. La prueba exacta de Fisher

En el caso de tablas de 2×2 en las que N es muy pequeña, es posible servirse de una prueba desarrollada por R. A. Fisher, que nos da probabilidades exactas, y no aproximadas. Si designamos las casillas y los marginales de la tabla de 2×2 de la siguiente manera:

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	N

podemos conseguir la probabilidad de obtener *exactamente* esas frecuencias en la hipótesis nula de que no hay diferencias en las proporciones de las poblaciones. Esta probabilidad nos está dada por la fórmula:

$$P = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{N!a!b!c!d!}$$

Esta fórmula de probabilidad puede obtenerse utilizando la distribución hipergeométrica para el cálculo de probabilidades sobre la base de muestreo *sin reposición*. En esta prueba, como en

algunas otras pruebas no paramétricas, podremos entender el problema como si éste contuviera repetidas muestras de una "población" de tamaño N . Tratamos así nuestra muestra obtenida como si se tratara de una población real, e imaginamos en este ejemplo que las categorías de nuestros casos les dan cabida en una de las cuatro casillas. Como hay $a + c$ individuos en la primera columna, $a + b$ en la primera fila, y así sucesivamente, ¿cuál será la probabilidad de que de los $a + b$ individuos de la primera fila correspondan exactamente a a la primera columna y b a la segunda? Nos imaginamos haber muestreado $a + b$ individuos al azar pero sin reposición, colocándolos en la primera fila, con los restantes cayendo por necesidad en la segunda fila. En efecto, resulta que imaginamos que llenamos las casillas por un proceso esencialmente al azar, y preguntamos cuál hubiera sido la exactitud de los resultados si hubiese sido seguido tal proceso.

Aplicando la fórmula para la distribución hipergeométrica dada en la sección X.4, veremos que la probabilidad de obtener exactamente a y b casos en las dos casillas de la fila superior vendría dada por

$$P(a, b) = \frac{\left(\frac{a+c}{a}\right) \left(\frac{b+d}{b}\right)}{\left(\frac{N}{a+b}\right)}$$

Escribiendo cada uno de los términos en función de factoriales, y simplificando, obtenemos:

$$P(a, b) = \frac{\frac{(a+c)!}{a!(a+c-a)!} \frac{(b+d)!}{b!(b+d-b)!}}{\frac{N!}{(a+b)!(N-a-b)!}} = \frac{\frac{(a+c)! (b+d)!}{a!c! b!d!}}{\frac{N!}{(a+b)!(c+d)!}} = \frac{(a+c)!(b+d)!(a+b)!(c+d)!}{N!a!b!c!d!}$$

Puede comprobarse fácilmente que se habría conseguido el mismo resultado si hubiéramos concebido el problema como orientado a seleccionar una muestra de $a + c$ individuos, asignándolos a continuación a la primera columna.

Obsérvese que hay nueve factoriales en esta fórmula de P . Por lo tanto, la tarea de calcularla sería formidable. Por otra parte, como quiera que normalmente se está interesado en obtener la

cola entera de la distribución de muestreo y no la probabilidad de averiguar exactamente los resultados obtenidos, habría que añadir, a esta probabilidad primera, las probabilidades de obtener incluso más resultados poco corrientes en la misma dirección.

Un sencillo ejemplo numérico ilustrará lo que esto significa. Supóngase que hemos obtenido la siguiente tabla de 2×2 :

3	9	12
12	5	17
15	14	29

Si suponemos que los marginales permanecen fijos, vemos inmediatamente que hay tres resultados (en la misma dirección) que son incluso más difíciles de obtenerse. Son los siguientes:

2	10	12
13	4	17
15	14	29

1	11	12
14	3	17
15	14	29

0	12	12
15	2	17
15	14	29

Obsérvese que podemos llegar a las tablas sucesivas reduciendo cada vez en uno las casillas a y d y aumentando en uno las casillas b y c , hasta llegar a la tabla final, en la que la casilla a está vacía.

Supongamos que la casilla a es siempre la que contiene el menor número de casos, ya que siempre tendremos la posibilidad de disponer las tablas en tal forma.⁵ Sirvámonos del símbolo P_0 para designar la probabilidad de obtener exactamente cero casos en la casilla a (dados los marginales en cuestión), en la hipótesis nula; pongamos que P_1 representa la probabilidad de obtener exactamente un caso en la casilla a , P_2 la de obtener exactamente dos casos, etcétera. Así, pues, en este problema particular hemos de obtener la suma de las probabilidades

$$P_0 + P_1 + P_2 + P_3$$

para calcular la probabilidad de obtener tres o menos casos en la casilla a . Y ya que nos estamos sirviendo de una prueba de

⁵ En raros casos cambiará la dirección de la relación si se sigue la regla de que la casilla a sea siempre la más pequeña. Por ejemplo, si las dos distribuciones marginales son muy desiguales, la regla tal vez no se aplique. Así, si a , b , c y d son 1, 2, 3 y 7, respectivamente, el producto ad ($= 7$) es mayor que el producto bc ($= 6$). Si uno reduce entonces a hasta 0, las casillas resultantes serán 0, 3, 4 y 6, y se producirá una inversión de dirección, puesto que $bc > ad$. Deben ser vigiladas tales inversiones y, en caso de que se produzcan, deberá denominarse como a la casilla más pequeña en la menor de las dos diagonales.

una sola cola, habremos de doblar el nivel de significación obtenido, si no estuviéramos en condiciones de poder predecir la dirección.⁶

Será mucho más conveniente que calcular cada una de las P_i de la fórmula anterior, que comporta productos de factoriales, obtener P_0 directamente y obtener luego las probabilidades restantes como funciones de P_0 . Con objeto de distinguir entre las varias combinaciones posibles de los valores numéricos de a , b , c y d en el caso de marginales fijos, sirvámonos de un subíndice k para designar la magnitud de la casilla más pequeña a . Así, por ejemplo, si hay k individuos en la casilla a , designaremos las cantidades de las diversas casillas como $a_k (=k)$, b_k , c_k y d_k . Toda vez que se supone que los marginales permanecen fijos, si disminuimos a_k y d_k en uno, hemos de aumentar b_k y c_k también en uno. Podemos ahora simplificar la fórmula de P_0 , ya que $a_0 = 0$ y, por consiguiente, $a_0! = 1$ (por definición), $(a_0 + b_0)! = b_0!$, y $(a_0 + c_0)! = c_0!$. O sea que cierto número de factoriales se eliminan, dejándonos con:

$$P_0 = \frac{(c_0 + d_0)!(b_0 + d_0)!}{N!d_0!}$$

El numerador consta ahora solamente de los factoriales de dos de los marginales, en lugar de los cuatro, y el denominador sólo comporta $N!$ y $d_0!$. El valor de d_0 puede obtenerse de la última de las tablas anteriores. Por lo tanto, en este ejemplo, $(c_0 + d_0) = 17$, $(b_0 + d_0) = 14$, $N = 29$, y $d_0 = 2$. P_0 puede calcularse ahora sirviéndonos de una tabla de logaritmos de factoriales, o bien escribiendo los factoriales y simplificando.

Con objeto de calcular los valores de P_1 , P_2 y P_3 necesitamos ahora una fórmula general de P_{k+1} en función de P_k . Ya que los marginales se suponen fijos, tenemos:

$$P_{k+1} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{N!(a_k+1)!(b_k-1)!(c_k-1)!(d_k+1)!}$$

debido al hecho de que, al añadir uno a la casilla a , lo añadimos también a la casilla d y lo sustraemos tanto de b como de c . Si dividimos ahora P_{k+1} entre P_k , prácticamente todos los términos desaparecen. En efecto, los numeradores de ambas probabilidades son idénticos, ya que todos ellos comportan los mismos marginales. El factorial de N se elimina. Y nos queda:

⁶ En un sentido estricto, la prueba de Fisher deberá ser usada probablemente sólo en el caso en que previamente se hubiera predicho la dirección, ya que las dos colas casi nunca serán perfectamente simétricas.

$$\frac{P_{k+1}}{P_k} = \frac{a_k!b_k!c_k!d_k!}{(a_k+1)!(b_k-1)!(c_k-1)!(d_k+1)!}$$

Pero $a_k!/(a_k+1)!$ es igual a $1/(a_k+1)$, y lo mismo por lo que se refiere a $d_k!/(d_k+1)!$. O sea, pues, $b_k!/(b_k-1)! = b_k$, y $c_k!/(c_k-1)! = c_k$. Por consiguiente:

$$\frac{P_{k+1}}{P_k} = \frac{b_k c_k}{(a_k+1)(d_k+1)}$$

o sea

$$P_{k+1} = \frac{b_k c_k}{(a_k+1)(d_k+1)} P_k$$

y los factoriales fastidiosos han desaparecido. Por lo tanto, podemos servirnos de esta fórmula para obtener P_1 a partir de P_0 . Una vez obtenida P_1 podemos calcular P_2 , y así sucesivamente.

Volviendo a nuestro ejemplo numérico, obtenemos P_0 como sigue:

$$P_0 = \frac{14!17!}{29!2!} = .17535 \times 10^{-5}$$

Y por consiguiente:

$$P_1 = \frac{b_0 c_0}{(a_0+1)(d_0+1)} P_0 = \frac{12(15)}{1(3)} (.17535 \times 10^{-5}) = 10.521 \times 10^{-5}$$

Al calcular P_2 hemos de cuidar de servirnos de a_1 , b_1 , c_1 y d_1 , y no de las cifras empleadas para obtener P_1 . Tenemos, así:

$$P_2 = \frac{b_1 c_1}{(a_1+1)(d_1+1)} P_1 = \frac{11(14)}{2(4)} (10.521 \times 10^{-5}) = 202.529 \times 10^{-5}$$

Y análogamente:

$$P_3 = \frac{b_2 c_2}{(a_2+1)(d_2+1)} P_2 = \frac{10(13)}{3(5)} (202.529 \times 10^{-5}) = 1\,755.252 \times 10^{-5}$$

Obsérvese que cada uno de los factores del numerador va disminuyendo en 1, al calcular P_{k+1} a partir de P_k , en tanto que los del denominador van aumentando cada vez en una unidad. Sumando las probabilidades tenemos, pues:

$$P_0 + P_1 + P_2 + P_3 = (.175 + 10.521 + 202.529 + 1\,755.252) \times 10^{-5} \\ = 1\,968.48 \times 10^{-5} = .0197$$

Por lo tanto, la probabilidad de obtener tres o menos individuos en la casilla a es, con la hipótesis nula, de .02, y tomaremos nuestra decisión de rechazar o no la hipótesis nula en consecuencia.

Debido a que la prueba de Fisher es exacta, merece preferencia respecto de la prueba de la χ -cuadrada corregida con fines de continuidad. Y como quiera que por lo regular la prueba de la χ -cuadrada dará probabilidades algo más bajas que la prueba de Fisher, si lo que se desea en realidad es rechazar la hipótesis nula, obraremos, al servirnos de ésta, en sentido conservador. En otros términos, si nos servimos de la prueba de la χ -cuadrada, puede ser que lleguemos a probabilidades que en realidad sean demasiado pequeñas, lo que nos llevaría acaso a la conclusión de que la hipótesis nula deba descartarse cuando en realidad no sea así. Si la frecuencia mínima esperada es sensiblemente superior a 5 y si se emplea la corrección de continuidad, las dos pruebas darán aproximadamente los mismos resultados. Aun logrando evitar el empleo de factoriales en el caso de la prueba de Fisher, se echa de ver que, si la frecuencia menor de la casilla es mayor que 5, los cálculos necesarios podrán resultar muy fastidiosos. De ahí que se encuentre que dicha prueba resulta más práctica en el caso de N muy pequeñas, o siempre que el tamaño de la muestra sea moderado y uno o más de los marginales sean muy pequeños. En los casos en que ambos, $(a + b)$ y $(c + d)$ son ≤ 30 , existen tablas en (3) que simplifican considerablemente el empleo de esa prueba exacta.

XV.3. Medidas de la fuerza de la relación

Hasta aquí sólo nos hemos ocupado de la cuestión de saber si existía o no una relación entre variables. Hemos establecido hipótesis nulas en el sentido de que no se daba relación alguna, y hemos tratado de descartarlas. Pero, cuando estamos en condiciones de descartar, ¿qué es lo que hemos logrado? Designamos una relación como estadísticamente significativa cuando hemos establecido, bajo el riesgo de error de tipo I, que *sí existe* una relación entre las dos variables. Sin embargo, ¿quiere esto decir que la relación es significativa en el sentido de ser una relación fuerte o importante? No necesariamente. En efecto, la cuestión de la fuerza de la relación es totalmente distinta de la de su existencia. En esta sección vamos a ocuparnos de diversas medidas de grados de asociación que ayudan a contestar la segunda de las preguntas.

A primera vista podría parecer razonable tratar de establecer la fuerza de la relación observando simplemente el nivel de significación conseguido con una prueba. Así, por ejemplo, podría discurrirse en el sentido de que si una prueba es significativa al nivel de .001 y otra al nivel de .05, la primera sería la más fuerte

de las dos. Pero, ¿es esto necesariamente así? El examen de los dos niveles de significación nos dirá en cuál caso podemos estar más seguros de que la relación existe. Así, en el primero de los dos casos citados estaríamos casi seguros de que *existe* efectivamente una relación, pero no lo estaríamos tanto en el segundo. Hemos de recordar, no obstante, que el nivel de significación alcanzado depende del tamaño de las muestras usadas. En efecto, como se indicó anteriormente, si las muestras son muy grandes, resulta por lo regular muy fácil establecer significación, aun en el caso de una relación muy superficial. Esto significa, de hecho, que, cuando las muestras son grandes, decimos en realidad muy poca cosa al afirmar que hemos establecido una relación "significativa". En el caso de muestras grandes, es mucho más importante preguntar, "dado que existe una relación, ¿cuál es su fuerza?"

Con objeto de ilustrar lo que se acaba de decir, veamos un poco más de cerca cierta propiedad de la χ -cuadrada. Al hacerlo, el lector deberá tener presente que los mismos principios se aplican exactamente a otras clases de pruebas de significación. Preguntémonos qué sucede con la χ -cuadrada cuando el número de casos aumenta. Con fines de ilustración podemos tomar la siguiente tabla de 2×2 .

30	20	50
20	30	50
50	50	100

La χ -cuadrada de esta tabla resulta ser exactamente 4.0. Supongamos ahora que se duplican los tamaños de las muestras, manteniendo las mismas proporciones en cada casilla. Obtendríamos así:

60	40	100
40	60	100
100	100	200

y la χ -cuadrada sería 8.0, o sea una cifra exactamente doble de la anterior. Examinando la fórmula de la χ -cuadrada, resulta muy fácil demostrar que, si las proporciones de las casillas permanecen inalteradas, la χ -cuadrada varía directamente con el número de casos. Si duplicamos el número de éstos, duplicamos aquélla, y si triplicamos los primeros, triplicamos la segunda. Supóngase que el número de casos inicial se multiplica por el factor k . Entonces, como quiera que las proporciones de las casillas permanecen inalteradas, toda nueva frecuencia observada será exactamente k veces la anterior, y lo mismo por lo que se refiere a las

frecuencias esperadas. La nueva χ -cuadrada puede, pues, expresarse como:

$$(\chi^2)' = \sum \frac{(kf_o - kf_e)^2}{kf_e} = \sum \frac{k^2(f_o - f_e)^2}{kf_e} = k \sum \frac{(f_o - f_e)^2}{f_e}$$

Así, pues, el valor de la nueva χ -cuadrada es exactamente k veces el de la primitiva.

Las implicaciones de este hecho pueden destacarse por medio de otra ilustración. Supóngase que obtenemos los siguientes resultados al relacionar las diferencias de sexo con la tolerancia respecto de conductas anómalas:

Tolerancia	Varones	Mujeres
Alta	26	24
Baja	24	26

En este caso la χ -cuadrada es 0.16, y estaremos en lo cierto informando que la relación no es significativa. Supóngase, sin embargo, que el estudio fue muy ambicioso y que se reunieron datos correspondientes a 10 000 casos, con los siguientes resultados:

Tolerancia	Varones	Mujeres
Alta	2 600	2 400
Baja	2 400	2 600

La χ -cuadrada es ahora 16.0, o sea un valor altamente significativo desde el punto de vista estadístico. Sin embargo, si hubiéramos expresado los resultados en términos de porcentajes, la cosa se habría presentado como mucho menos interesante. Si dijéramos que el 52 por ciento de los varones era altamente tolerante, en tanto que sólo correspondía a dicha categoría el 48 por ciento de las mujeres, nos criticarían con razón por destacar las diferencias aparentemente insignificantes tanto desde el punto de vista teórico como del significado práctico. Este ejemplo ilustra un punto muy importante. En efecto, una diferencia puede ser interesante estadísticamente sin serlo en ningún otro sentido. En el caso en que se seleccionaron 10 000 casos, podemos estar bien seguros que hay cierta relación superficial, que produciría una relación significativa desde el punto de vista estadístico.

Vemos, pues, que si una muestra es pequeña, se requiere una relación mucho más manifiesta para obtener significación. Por lo tanto, con las muestras pequeñas las pruebas de significación son mucho más importantes. En tales casos es posible que digamos mucho cuando podemos establecer significación. El nivel de significación depende de dos factores, a saber: de la fuerza

o grado de la relación y del tamaño de las muestras. Puede obtenerse significación con una relación muy fuerte y muestras muy pequeñas o, inversamente, con una relación muy débil y muestras muy grandes. En la mayor parte de la investigación social, nuestro interés primordial está no tanto en hallar variables relacionadas unas con otras, sino en localizar relaciones importantes. Aunque conviene recalcar que no todas las relaciones fuertes son importantes (*v.gr.* la relación entre las edades respectivas del marido y la mujer), para que una relación sea de alguna importancia práctica ha de ser por lo menos moderadamente fuerte. Una vez que ha sido establecida la existencia de una relación, el investigador debería preguntarse siempre, "¿cuán fuerte es?"

¿Cómo se mide, pues, la fuerza de una relación? Estamos buscando una medida descriptiva que nos ayude a resumir la relación de tal modo que podamos comparar varias relaciones y llegar a una conclusión respecto de cuál sea la más fuerte. Desde el punto de vista ideal, nos gustaría tener alguna clase de interpretación operativa de la medida que nos atrae intuitivamente. Por convención, los estadígrafos han adoptado la costumbre de concebir medidas que tengan la unidad por límite superior, y cero o bien menos uno (-1.0) como límite inferior. Muchas relaciones sólo pueden alcanzar su límite de 1.0 (o -1.0) cuando la relación es perfecta, y adoptan el valor de cero cuando entre las variables no existe relación alguna, o sea cuando son independientes. Vamos a examinar a continuación algunas medidas que pueden utilizarse con las tablas de contingencia, procediendo a apreciar sus propiedades.

Antes de entrar en el examen de varias medidas de asociación que pueden emplearse con las tablas de contingencia, habría que mencionar, por lo menos, el procedimiento relativamente sencillo y obvio de indicar diferencias en términos de porcentajes. Es posible, sin la menor duda, obtener una indicación muy buena del grado de relación entre dos variables dicotómicas comparando porcentajes. Así, por ejemplo, si el 60 por ciento de los varones seleccionados se clasifican como altamente tolerantes, en tanto que sólo se pone en tal categoría el 30 por ciento de las mujeres, tenemos una diferencia del 30 por ciento entre los dos grupos. ¿Por qué, pues, no servirnos de una medida semejante como medida de la fuerza de la relación? Si comparamos individuos de las clases media e inferior, por ejemplo, desde el punto de vista de la tolerancia, y sólo obtenemos una diferencia del 20 por ciento, podemos afirmar una relación más fuerte entre el sexo y la tolerancia que entre ésta y la clase.

En el caso especial de la tabla de 2×2 , los porcentajes pueden efectivamente compararse en tal forma, y la extensa familiarización con los porcentajes, en contraste con otros tipos de

medidas, hablaría ciertamente en favor de estas comparaciones.⁷ Pero, ¿qué pasará con la tabla general de $r \times c$? Aquí el uso de los porcentajes puede dificultarle al lector apreciar a primera vista cuán fuerte sea la relación. Supóngase, por ejemplo, que se utilizaban tres clases con los siguientes resultados: clase superior, 70 por ciento altamente tolerante; clase media, 50 por ciento altamente tolerante, y clase inferior, 30 por ciento altamente tolerante. Tenemos ahora una distancia del 40 por ciento entre las clases superior e inferior, o sea una diferencia numéricamente mayor que la que existe entre los varones y las mujeres. Por otra parte, por lo regular esperamos una diferencia mayor cuando sólo se consideran los extremos. Supóngase que se hubieran tenido cinco clases, ¿qué clase de diferencias de porcentajes esperaríamos ahora, y cómo compararíamos los resultados con los de la tabla de 2×2 ? Y para introducir una idea más, supóngase que nos sirviéramos de cuatro categorías de tolerancia. Es obvio que se hace difícil establecer comparaciones de una tabla a otra. Necesitamos, pues, una medida única de resumen, que tenga los mismos límites superior e inferior, independientemente del número de casillas.

Medidas tradicionales basadas en la χ -cuadrada. Ya se observó que la χ -cuadrada es directamente proporcional a N . Podemos servirnos de este hecho para construir varias medidas de asociación. En el caso de las dos tablas de contingencia

30	20	50		60	40	100
20	30	50	y	40	60	100
50	50	100		100	100	200

deseamos una medida que tenga el mismo valor para cada una de las tablas, ya que, cuando expresamos los resultados en términos de porcentajes, éstos son los mismos en ambos casos. En otros términos: diríamos probablemente que los grados o fuerzas de la relación son idénticos en los dos grupos de datos, y que la única diferencia está en la magnitud de las muestras. Aunque el valor de la χ -cuadrada sea el doble en la segunda tabla de lo que es en la primera, observamos, con todo, que, si se la divide en cada caso entre el número total de los casos, los resultados son idénticos. Esto sugiere que la expresión χ^2/N o algún múltiplo de la misma nos daría una de las propiedades que buscamos en nuestra medida, o sea la de dar el mismo resultado cuando las proporciones de casillas comparables son idénticas.

⁷ Veremos otra ventaja de los porcentajes cuando estudiemos declives en el capítulo xvii. Como ya se indicó en el caso de las pruebas para diferencias de diferencias en proporciones, una diferencia de proporciones puede ser considerada como un caso especial de declive.

Obsérvese que el valor de χ^2/N , o ϕ^2 según se la escribe comúnmente, es 0 cuando entre las variables no existe relación en absoluto. Resulta que, en el caso de tablas de 2×2 (o $2 \times k$), ϕ^2 tiene también la unidad por límite superior cuando la relación entre las dos variables es perfecta. Supóngase, en efecto, que hubiéramos obtenido la siguiente tabla:

50	0	50
0	50	50
50	50	100

Puede verificarse fácilmente que, en este caso, la χ -cuadrada es 100 y, por consiguiente, ϕ^2 es 100/100, o sea 1.0. Ocurrirá siempre que, cuando dos casillas opuestas diagonalmente sean ambas cero, el valor de la χ -cuadrada en una tabla de 2×2 sería N , y por lo tanto ϕ^2 será la unidad. Es obvio que, en el ejemplo considerado, la relación es perfecta. Si el sexo se relacionara en él con la tolerancia, podríamos decir que todos los varones son altamente tolerantes y todas las mujeres altamente intolerantes. En una terminología con la que no habremos de tardar en familiarizarnos, podemos decir que el todo de la variación en materia de tolerancia se explica por el sexo o está asociado con él.⁸

En la tabla general de $r \times c$, ϕ^2 puede alcanzar un valor considerablemente mayor que la unidad. Por lo tanto, se han desarrollado diversas otras medidas que son asimismo simples funciones de χ^2/N , pero que tienen también como límite superior la unidad. La primera de éstas, designada como la T de Tschuprov, se define como:

$$T^2 = \frac{\chi^2}{N \sqrt{(r-1)(c-1)}} = \frac{\phi^2}{\sqrt{(r-1)(c-1)}}$$

Aunque el límite superior de T sea la unidad, este límite sólo puede alcanzarse cuando los números de hileras y columnas son iguales. En otros términos: T ha de ser siempre menor que la unidad en una tabla de 2×3 o de 3×5 . Si hay considerablemente más hileras que columnas (o viceversa), el límite superior de T puede quedar muy por debajo de la unidad. Para corregir este hecho, podemos siempre dividir el valor obtenido de T entre la máxima T posible para números dados de hileras y columnas. Sin embargo, como quiera que disponemos de medidas más satisfactorias, no necesitamos examinar este procedimiento de corrección.

⁸ Esto supone, por descontado, que la tolerancia se toma como variable dicotómica.

* Podemos mostrar que el límite superior de ϕ^2 es $\text{Min}(r-1, c-1)$, utilizando la fórmula:

$$\chi^2 = N \left[\sum_{i=1}^r \sum_{j=1}^c \frac{N_{ij}^2}{N_{i.} N_{.j}} - 1 \right]$$

Obsérvese que:

$$\frac{N_{ij}^2}{N_{i.} N_{.j}} \leq \frac{N_{ij}}{N_{i.}} \quad \text{para } i = 1, 2, \dots, r$$

y

$$\frac{N_{ij}^2}{N_{i.} N_{.j}} \leq \frac{N_{ij}}{N_{.j}} \quad \text{para } j = 1, 2, \dots, c$$

Por tanto

$$\sum_{i=1}^r \sum_{j=1}^c \frac{N_{ij}^2}{N_{i.} N_{.j}} \leq \sum_{i=1}^r \sum_{j=1}^c \frac{N_{ij}}{N_{i.}} = \sum_{i=1}^r 1 = r$$

y

$$\sum_{i=1}^r \sum_{j=1}^c \frac{N_{ij}^2}{N_{i.} N_{.j}} \leq \sum_{j=1}^c \sum_{i=1}^r \frac{N_{ij}}{N_{.j}} = \sum_{j=1}^c 1 = c$$

Así:

$$\sum_{i=1}^r \sum_{j=1}^c \frac{N_{ij}^2}{N_{i.} N_{.j}} \leq \text{Min}(r, c)$$

y de allí:

$$\chi^2 \leq N[\text{Min}(r, c) - 1] = N[\text{Min}(r-1, c-1)]$$

Por tanto:

$$\phi^2 \leq \text{Min}(r-1, c-1)$$

Hay otra medida, introducida por Cramér y que designaremos como V , que se define como sigue:

$$V^2 = \frac{\chi^2}{N \text{Min}(r-1, c-1)} = \frac{\phi^2}{\text{Min}(r-1, c-1)}$$

en donde $\text{Min}(r-1, c-1)$ designa $r-1$ o $c-1$, según cuál de ellas sea menor (valor mínimo de $r-1$ y $c-1$). Si bien V no se utiliza corrientemente en la bibliografía social, con todo parece ser preferible a T , en cuanto puede alcanzar la unidad aun cuando los números de hileras y columnas no sean iguales. Como puede

verificarse fácilmente, V y T son equivalentes siempre que $r=c$. De otra forma, siempre será V algo mayor que T . Por supuesto, ambas medidas son equivalentes de ϕ en el caso de 2×2 . Y veremos también que V y ϕ serán idénticas en el caso de $2 \times k$.

Otra medida de asociación basada en la χ -cuadrada es el coeficiente de contingencia de Pearson, C , que está dado por:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

Al igual que las otras medidas, C se hace cero cuando las variables son independientes. Sin embargo, el límite superior de C depende del número de hileras y columnas. En el caso de 2×2 , el límite superior de C^2 se convierte en $N/(N+N)$, ya que χ^2 puede alcanzar un valor máximo de N . Por lo tanto, el límite superior de C es .707. Si bien el límite superior aumenta a medida que aumenta el número de hileras y columnas, dicho límite siempre es menor que la unidad. De ahí que C sea algo más difícil de interpretar que las otras medidas, a menos que se introduzca una corrección dividiendo entre el valor máximo de C para números particulares de hileras y columnas. En el caso de la tabla 2×2 , por ejemplo, la C obtenida habría de dividirse entre .707.

Las medidas anteriores de la fuerza de la relación se basan todas ellas en la χ -cuadrada. Como quiera que por lo regular el valor de la χ -cuadrada se habrá calculado previamente con objeto de verificar el significado, todas las medidas en cuestión requieren en realidad muy poco cálculo adicional. Pero por otra parte, no existe razón particular alguna en cuya virtud una medida de asociación haya de basarse en la estadística de la prueba correspondiente. En efecto, puede demostrarse que todas las medidas basadas en la χ -cuadrada son algo arbitrarias en su esencia y sus interpretaciones dejan mucho que desear. Así, por ejemplo, todas ellas confieren mayor peso a las columnas o hileras de marginales más pequeños que a las de marginales mayores [2]. Sin embargo, como quiera que tanto la prueba T como la C se encuentran con frecuencia en la bibliografía, el lector debería familiarizarse con sus propiedades.

La Q de Yule. Otra medida de uso corriente es la Q de Yule, que es también un caso especial de la medida γ (gamma) que se discutirá en el capítulo XVIII en relación con las escalas ordinales. Esta medida sólo puede emplearse con la tabla de 2×2 y se define como sigue:

$$Q = \frac{ad - bc}{ad + bc}$$

en donde a , b , c y d se refieren a las frecuencias de las casillas. Obsérvese que, una vez elevado al cuadrado y multiplicado por N , el numerador es el mismo que en la expresión de la χ^2 -cuadrada. Lo mismo que en el caso de las demás medidas, Q desaparece cuando las variables son independientes, o sea, cuando los productos diagonales ad y bc son iguales. A diferencia de ϕ^2 , sin embargo, Q alcanza sus límites de ± 1.0 cuando una cualquiera de las casillas es igual a cero. Con objeto de comprender el carácter de las circunstancias en cuya virtud Q pueda ser igual a la unidad en tanto que ϕ^2 queda por debajo de dicho valor, tomemos los siguientes ejemplos:

30	0	30	40	0	40
20	50	70	10	50	60
50	50	100	50	50	100

Mientras Q adopta el valor de la unidad en estas dos tablas, los valores correspondientes de ϕ^2 , en cambio, son de .429 y .667 respectivamente. En ambos casos sería imposible que desaparecieran dos casillas diagonalmente opuestas, debido al carácter de los marginales. De ahí que ϕ^2 sólo pueda adoptar el valor de uno cuando se verifican determinadas condiciones en relación con los marginales. En la tabla de 2×2 , los marginales de la primera variable han de ser idénticos a los de la segunda.⁹ Cuanto mayor sea, pues, la discrepancia entre los marginales de las hileras y las columnas, tanto menor es el límite superior de ϕ^2 .

Plantéase ahora la cuestión de saber si queremos o no considerar una relación como "perfecta" cuando sólo desaparece una de las casillas. Al parecer, la respuesta a esta cuestión debería depender, entre otras cosas, de la manera como estén formadas las categorías de las dos variables. Por lo regular es posible concebir un problema en términos de una variable independiente y una variable dependiente. Parecería, pues, razonable sostener que, para que una relación sea perfecta, los marginales de la variable dependiente habrían de "convenir" naturalmente a los de la variable independiente. Supóngase, por ejemplo, que hubiera 60 protestantes y sólo 40 católicos y judíos. En este caso, para que la relación fuera perfecta, esperaríamos que todos los 60 protestantes votaran republicano y todos los 40 restantes votaran en favor de los demócratas. Los marginales serían así los mismos para ambas variables, y tanto ϕ^2 como Q serían iguales a la unidad. Por otra parte, si la mitad de la muestra votara republicano y la otra mi-

⁹ Esto no significa que los marginales hayan de comportar una partición de 50-50. Significa, en efecto, que si uno de los marginales se parte en 70 y 30, el otro ha de estar también partido de 70 y 30. Las correcciones de marginales desiguales son asimismo posibles, pero, como se desprende del examen que sigue, habrá que ser cauto en el empleo de tales correcciones.

tad demócrata, entonces, aunque todos los votos republicanos provinieran de los protestantes, no podríamos decir que la relación era perfecta, ya que 10 de los protestantes habrían votado demócrata. En tal caso, los marginales de la variable dependiente no coincidirían con los de la independiente, y ϕ^2 sería inferior a la unidad. Por lo tanto, en tal caso ϕ^2 parecería ser la medida más apropiada, ya que Q tomaría el valor de la unidad a pesar de la relación imperfecta entre las dos variables.

Ocurre en ocasiones que los marginales de la variable dependiente son fijos, en virtud del método empleado al establecer las categorías. Así, por ejemplo, si la variable dependiente fuera en realidad continua pero se hubiera hecho dicotómica en la mediana, entonces los dos grupos de marginales no podrían ser idénticos, a no ser que los marginales de las variables independientes estuvieran también partidos en 50 y 50. Por ejemplo: si la preferencia confesional se hubiera referido a las marcas del conservadurismo político dividiendo en dos a la mediana, entonces ϕ^2 no podría alcanzar la unidad (en el supuesto de la misma partición confesional anterior). En tal caso, Q podría resultar una medida más apropiada, ya que tiene en cuenta el hecho de que los marginales de la variable dependiente se han fijado por completo en virtud del método de investigación.

La tau de Goodman y Kruskal. Cierta número de otras medidas de asociación susceptibles de emplearse con las tablas de contingencia han sido presentadas por Goodman y Kruskal [5], [6] y [7]. La mayoría de ellas comportan lo que se ha designado como interpretaciones probabilistas. Como quiera que tienen un sentido intuitivo que permite interpretar valores intermedios entre cero y uno, estas medidas podrán parecer superiores a las que se basan en la χ^2 -cuadrada.

Con objeto de ilustrar una de estas medidas, la τ , (tau), tomamos un ejemplo numérico. Designaremos las escalas nominales relacionadas una con otra como A y B , y tomaremos a B como variable dependiente.

	B_1	B_2	B_3	Total
A_1	300	600	300	1 200
A_2	600	100	100	800
Total	900	700	400	2 000

Supongamos ahora que se nos da una muestra (o población) de 2 000 personas y se nos pide clasificarlas en una de las tres categorías B_1 , B_2 o B_3 , de tal modo que terminemos exacta-

mente con 900 casos en B_1 , 700 en B_2 y 400 en B_3 . Supóngase primero que no sabemos nada acerca de los individuos que nos van a ayudar en esta tarea. Si los individuos nos son dados en un orden totalmente al azar, podemos calcular muy fácilmente el número de errores que podemos esperar cometer al asignar los individuos a una de las tres categorías en cuestión.

Como quiera que hemos de asignar 900 individuos a B_1 , en tanto que 1 100 de cada 2 000 no corresponden en realidad a dicha clase, podemos esperar cometer a la larga 900(1 100/2 000), o sean 495 errores. En forma análoga, hemos de asignar 700 individuos a B_2 , en tanto que de cada 2 000 los 1 300 no corresponden a ella. De ahí, pues, que al colocar a los individuos en B_2 podamos esperar cometer 700(1 300/2 000), o sea 455 errores. En otros términos, de los 700 que ponemos en dicha categoría sólo podemos esperar que se clasifiquen correctamente 700 - 455, o sean 245 individuos. Por supuesto, no esperamos cometer exactamente 455 errores, pero ésta es, con todo, la cifra que obtendríamos si promediáramos nuestros errores a la larga. Finalmente, esperaríamos cometer 400(1 600/2 000) o 320 errores al asignar los individuos a B_3 . Obsérvese que, pese a que hagamos a esta categoría una asignación menor, nuestro riesgo de error es superior al de las dos categorías precedentes, ya que sólo el 20 por ciento de los individuos corresponde a ella. Por lo tanto, en conjunto, al colocar los 2 000 individuos, esperaríamos cometer:

$$495 + 455 + 320 = 1\,270$$

errores. Nuestro promedio no sería muy bueno.

Pero supóngase ahora que se nos proporcionaba alguna información adicional acerca de cada individuo, diciéndonos si está en A_1 o en A_2 . Y nos preguntamos si el hecho de conocer las clases A nos ayudará a reducir el número de errores cometidos al asignar los individuos a las categorías B . Si las variables A y B son estadísticamente independientes, sabemos que el conocimiento de A no nos ayudará a predecir B . En este caso, pues, esperamos cometer exactamente los mismos errores en que incurrimos cuando no poseíamos información alguna acerca de A . Por otra parte, si la relación entre A y B fuera perfecta, estaríamos en condiciones de anticipar B con perfecta precisión conociendo A . La medida que vamos a desarrollar nos indica la reducción proporcional de errores siendo A conocida.

Veamos cómo calculamos el número de errores anticipados conociendo A . Si se nos da el hecho de que el individuo corresponde a la A_1 , podemos servirnos de las cifras de la primera columna. Hemos de poner ahora exactamente 300 de los 1 200 individuos en B_1 , los 600 restantes proviniendo de A_2 . Ya que de los 1 200 individuos de A_1 900 no corresponden en realidad a B_1 , podemos

esperar cometer 300(900/1 200) o 225 errores. Y en forma análoga, con los 600 individuos de A_1 que ponemos en B_2 podemos esperar cometer 300 errores, siendo el número de errores correspondiente a $B_3 = 225$. Tomamos ahora los 800 individuos de A_2 y asignamos 600 de ellos a B_1 y 100 de los 200 restantes a cada una de las categorías B_2 y B_3 . Al proceder así, podemos esperar cometer 150, 87.5 y 87.5 errores respectivamente. Adicionando las dos cantidades de A_1 y A_2 , vemos que podemos esperar cometer un total de 1 075 errores, si A es conocida.

Definimos la medida τ_b como reducción proporcional de errores. Así, pues:

$$\tau_b = \frac{\text{número de errores con } A \text{ desconocida} - \text{número de errores con } A \text{ conocida}}{\text{número de errores con } A \text{ desconocida}}$$

$$\tau_b = \frac{1\,270 - 1\,075}{1\,270} = \frac{195}{1\,270} = .154$$

En otros términos: nos hemos evitado 195 errores del número total esperado de 1 270, y los hemos reducido en un 15.4 por ciento. Si τ_b hubiera resultado ser .50, podríamos dar así la interpretación muy simple de que el conocimiento de A reduciría el número de errores a la mitad, en tanto que un valor de .75 equivaldría a reducir el número de los errores a un cuarto, y así sucesivamente. En el caso de ϕ^2 en cambio, semejante interpretación sencilla no es posible (véase [2]). Si hubiéramos querido interpretar las clases B a partir de las A , habríamos designado la medida correspondiente como τ_a . Por lo general, τ_a y τ_b no tendrán los mismos valores numéricos. ¿Por qué?

En el caso del cuadro 2×2 puede demostrarse que $\tau_a = \tau_b = \phi^2$. Esto nos indica que se dan dos tipos de dificultades en la anotación. Obsérvese que algunas de nuestras medidas (C , Q , T y V) vienen indicadas mediante letras latinas, en tanto que otras (ϕ y τ) lo son mediante letras griegas. Si fuéramos consecuentes deberíamos reservar las letras griegas para los parámetros de población calculados mediante muestras estadísticas. Por desgracia, una vez que los signos vienen siendo usados en forma generalizada, resulta difícil normalizar su empleo, y lo mejor que el lector puede hacer es tomar nota de la inconsistencia. Por otra parte, ciertas medidas aparecen elevadas al cuadrado, en tanto otras no lo están. Vemos especialmente en el caso 2×2 que el símbolo τ , no elevado al cuadrado, es equivalente a ϕ^2 , el que en este caso es igual a T^2 y V^2 . Así, en el caso del cuadro más general puede parecer razonable comparar τ con los otros coeficientes al cuadrado, aunque observando que no serán idénticos. En general

puede esperarse que los valores numéricos de τ sean *menores* que los coeficientes no elevados al cuadrado ϕ , T y V . Si hubiera que pensar en función de ciertas magnitudes absolutas, considerándolas pequeñas, medianas o grandes (por ejemplo: un valor inferior a .3 es "pequeño"), fácilmente podría incurrirse en error a menos que se reconocieran claramente las diferencias entre las medidas.

Lambda. Existe otra medida, lambda (λ) que es muy semejante a τ y que igualmente es asimétrica con respecto a A y B . Tomando a B como la variable dependiente con la que se hacen predicciones, obsérvese que el número esperado de errores se reducirá si se nos permite colocar a *todos* los individuos en la mayor de las categorías B_i (véase ejercicio 5, capítulo ix). En el ejemplo anterior esto habría supuesto colocar los 2 000 casos en B_1 en lugar de limitarnos a 900. Si lo hiciéramos así cometeríamos 1 100 errores, ya que hay un total de 1 100 casos en B_2 y B_3 . Obsérvese que éstos son menos errores que los que hicimos en el caso del denominador de τ_b . Supongamos que sabemos la categoría de A a la que pertenece el individuo. Si se nos permite asignar la totalidad de los 1 200 individuos de A_1 a B_2 , la fila que contiene el mayor número de individuos A_1 , cometeremos solamente $300 + 300 = 600$ errores. De manera análoga, si colocamos a todos los 800 individuos A_2 en la categoría B_1 , cometeremos sólo 200 errores. Conociendo, pues, la categoría A , y si se nos permite hacer estas distribuciones menos restrictivas, podremos esperar cometer 800 errores. Formaremos una medida λ_b , de "reducción proporcional en el error", como sigue:

$$\lambda_b = \frac{1\,100 - 800}{1\,100} = .273$$

Vemos que lambda es más fácil de calcular que tau; que supone una reducción no restrictiva de errores, y que en este ejemplo tiene un valor numérico considerablemente mayor que el de tau. Tiene sin embargo la indeseable propiedad de poder dar un valor numérico igual a cero en casos en que todas las demás medidas consideradas no serán cero, y cuando no desearíamos referirnos a las variables como no correlacionadas o estadísticamente independientes. Tal cosa puede ocurrir simplemente porque una de las B marginales sea mucho mayor que el resto, de tal manera que cualquiera que sea la categoría A , la decisión será siempre de colocar todos los individuos (para todo A_i), en la misma categoría B . Si por ejemplo las categorías B_1 y B_2 hubiesen sido combinadas en el anterior ejemplo hipotético, la decisión hubiera sido siempre la de colocar a todos los individuos en la categoría B_1 y B_2 y no en la de B_3 , de tal manera que la

resultante λ_b hubiera sido cero. Por la misma razón, aun cuando un simple total marginal (por ejemplo, B_1) no domina al resto, es probable que algunas de las categorías menos numerosas no entren en absoluto en el círculo de lambda. En el ejemplo anterior la decisión nunca resulta en la asignación de individuos a B_3 . Si se hubiera contado con una fila más, B_4 , también con un número relativamente pequeño de casos, la lambda medida podría haber sido indiferente a la distribución de casos entre B_3 y B_4 . Por estas razones se prefiere a tau sobre lambda en aquellos casos en que los totales marginales no son de aproximadamente la misma magnitud.

XV.4. Control de otras variables

Hasta aquí el examen de las pruebas de significación y de medidas de asociación sólo han comportado dos variables a la vez. En la mayoría de los problemas prácticos, en cambio, es necesario controlar una o más variables adicionales, que pueden ya sea enturbiar una relación o crear una relación espuria. Si bien es a menudo cierto que las generalizaciones en materia de ciencias sociales suelen establecerse en términos de sólo dos variables, se supone con todo casi siempre, implícitamente, que las variables relevantes se consideran como controladas. Con objeto de subrayar este hecho se emplea a menudo la frase "en igualdad de condiciones". Desde el punto de vista ideal, una hipótesis habría de enunciarse en forma que se entienda claramente cuáles variables han de controlarse. A medida que una disciplina va progresando hacia su madurez, las generalizaciones se hacen más calificadas, indicando las condiciones exactas en las que puede esperarse que se realicen. En las etapas iniciales de su desarrollo, sin embargo, resulta a menudo imposible saber cuáles son las variables relevantes que se necesita controlar. Ésta es la razón de que en ciencias sociales las proposiciones no se enuncien a menudo en forma que sugieran cuáles variables deban controlarse. No obstante, el lector debería acostumbrarse a buscar siempre las variables eventualmente posibles de controlar, aunque no se le haya invitado expresamente a hacerlo.

Según veremos más adelante, hay varios métodos posibles de control estadístico. El que se examina en el presente capítulo es tal vez el más directo y el que más se parece al experimento de laboratorio, en el que las variables de control se mantienen efectivamente constantes por medios físicos. En los experimentos de laboratorio se mantiene una variable de control a un valor constante, en tanto que las otras variables se relacionan entre sí. Así, por ejemplo, mientras se examina la relación entre la presión y el volumen, la temperatura se mantiene acaso a 70° F. Y si se encuentra una relación entre estas variables, puede resultar

posible enunciar su carácter con mucha mayor precisión que si la temperatura no se hubiera controlado. Sin embargo, el científico no estará autorizado a enunciar una generalización como de realización constante, a menos que la misma relación se verifique exactamente para todas las temperaturas. Realizará, sin duda, toda una serie de experimentos, cada uno de ellos a una temperatura diferente. Es muy probable que encuentre que la relación en cuestión sólo tiene lugar dentro de cierto margen de temperaturas. En estas condiciones habrá de especificar su generalización de modo que diga: "La relación entre la presión y el volumen es tal y cual, a condición que la temperatura se mantenga entre -100 y 600° F." Con suerte podrá hallar un factor de corrección que le permita enunciar de nuevo su proposición en forma que se aplique a un margen mayor de temperaturas. Y exactamente el mismo tipo de razonamiento se aplicará al control de variables adicionales. Podrían efectuarse controles simultáneos de diversas variables, manteniendo cada una de ellas a un valor fijo, y efectuando luego experimentos ulteriores con distintas combinaciones de valores de las variables de control. Si varios controles actuaran simultáneamente, se requerirá un número mucho mayor de experimentos análogos.

Existe cierta semejanza entre el procedimiento para lograr el control estadístico, que vamos a examinar a continuación, y un experimento de laboratorio en el curso del cual las variables son manipuladas físicamente y mantenidas constantes en diferentes niveles. Existe sin embargo una diferencia fundamental, que resulta vital, relacionada con la forma en que el observador interpreta los resultados. Cuando controlamos estadísticamente, llevamos a cabo manipulaciones con lápiz y papel, en el curso de las cuales ajustamos puntuaciones, o hacemos pasar a los individuos de uno a otro cuadro, pero en realidad no estamos manejando sus puntuaciones reales. Cuando, por ejemplo, estamos "controlando" estadísticamente un IQ, esto no significa que manejemos las constantes de inteligencia del individuo afectado. Podemos ajustar las puntuaciones de los IQ, restando de unas y sumando a otras, de manera que podamos pretender que son iguales entre sí, pero no podremos manipular la inteligencia real de una persona en forma que pueda compararse con los controles que gobiernan la temperatura o la presión en un experimento de laboratorio.

Este tipo de control y ajuste hipotético es muy conveniente, y no deberemos desconcertarnos si el mundo real coincide con lo que estamos haciendo. Si un cambio real en la inteligencia pudiera afectar nuestra relación en un sentido determinado, pero al mantenerla constante en un experimento nos fuera posible deducir la relación verdadera entre otras dos variables "con la inteligencia mantenida en nivel constante", resultarían justifica-

das nuestras manipulaciones con papel y lápiz. Debe reconocerse claramente que tales "controles" a base de lápiz y papel pueden ser realizados sobre *cualquier* variable de la que tengamos medidas (y categorías), incluso aquellas que son causalmente *dependientes* de las variables que estamos estudiando y aquellas que de manera espuria estén relacionadas, por razones extrañas, con alguna variable.

Los controles estadísticos son básicamente mucho más fáciles de realizar que los verdaderos controles, por lo que el margen de flexibilidad para su aplicación razonable es mucho mayor. Se requiere fundamentalmente una *teoría* que justifique la aplicación de tales controles, teoría en la que están implícitos supuestos acerca de la estructura causal del sistema de variables. Aunque el tema escapa al interés de un texto general sobre estadística, resulta necesario formular aquí unas palabras de cautela, ya que muchos malos entendidos, en relación con las operaciones de control estadístico, se han traducido en una aplicación ciega de variables de control sin apoyo en una teoría que lo justifique.

Volviendo al ejemplo de la relación entre las preferencias religiosas y los partidos políticos, se pueden controlar estadísticamente variables tales como el sexo y la clase social. Para mantener constante el sexo pueden, por ejemplo, ser considerados solamente los votantes varones. Si se observa que la relación se da en el caso de los varones y por separado en el de las hembras, podrá decirse que es aplicable al sexo, ya que habremos examinado ambas categorías de la variable "sexo". Es posible sin embargo que se observe la relación en el caso de los varones pero no en el de las hembras; en tales circunstancias habrá que calificar la generalización, volviendo nuestra atención a las causas por las cuales la relación existe para un sexo y no para el otro. Puede verse que el controlar las variables relevantes no sólo nos permite una prueba más rigurosa de una hipótesis, sino que nos suministra una mayor penetración en el caso en que se encuentre que la relación difiere de una categoría de la variable de control a la otra.

Algunas veces será conveniente controlar diversas variables a la vez. Debido a la escasez de los casos, se hace necesario con frecuencia controlar las variables relevantes una por una, perdiéndose, sin embargo, en esta forma cierta cantidad de información. Supóngase, por ejemplo, que se hubiera prescindido del sexo y se hubiera introducido un control en relación con la clase social de los electores. Examinaríamos, pues, cada clase social, para ver si la relación subsistía siempre. En contraste con este procedimiento, pudimos haber controlado simultáneamente desde los puntos de vista de la clase y del sexo, tomando todas las combinaciones posibles de las variables de control (v.gr. va-

rón de la clase inferior, mujer de la clase inferior, varón de la clase media, etcétera) y estudiando la relación en cada combinación de las categorías de control. Se concibe que la relación pueda verificarse acaso para todas las combinaciones, con excepción de la correspondiente a las mujeres de la clase inferior. Si esto fuera así, nos veríamos conducidos a investigar las peculiaridades de este subgrupo particular.

Con objeto de ilustrar el proceso, tomemos otro ejemplo concreto. Supóngase que tenemos los siguientes datos correspondientes a escolares: ambiente de la clase, cuota de inteligencia, grado escolar y la aplicación de cada niño. Convendrá resumir los datos en términos de una tabla maestra como la del cuadro XV.4.

CUADRO XV.4. Cuadro maestro para correlacionar cuatro variables

Inteligencia	Grados	Clase media		Clase baja		Totales
		Aplicación elevada	Aplicación baja	Aplicación elevada	Aplicación baja	
Alta	Alto	60	40	40	18	158
	Bajo	20	24	16	38	98
Baja	Alto	40	24	6	2	72
	Bajo	24	12	32	54	122
Totales		144	100	94	112	450

Obsérvese que un cuadro como éste contiene las casillas suficientes para que los cuatro tipos de información (clase, IQ, grados y aplicación) puedan ser, si así conviene, reconstruidos para cada individuo, es decir, que sabemos cuántas son las personas en las que se da la misma combinación de rasgos (por ejemplo: clase baja, IQ elevado, aplicación baja y grados altos). Si deseamos una información menos detallada podremos combinar los datos formando agrupaciones más amplias. Podemos por ejemplo reunir a los estudiantes de la clase media con los de la clase baja, manteniendo tan sólo la distinción relativa al IQ, la aplicación y los grados. Pero si se nos facilitase tan sólo una información menos detallada no nos sería posible recobrar el total de la información más que volviendo a hacer el análisis. Por tal razón un cuadro maestro tal como el XV.4 debe ser utilizado como cuadro de trabajo, sacando de él los datos para preparar una serie de otros cuadros separados.

Será en general más conveniente hacer el cuadro maestro de tal manera que la variable dependiente aparezca en la columna extrema de la izquierda, en tanto que la variable independiente más interesante aparezca en la fila baja del encabezamiento, lo que se traducirá en subcuadros con las frecuencias que están siendo comparadas directamente. En el cuadro XV.4, por ejemplo, tenemos cuatro subcuadros en cada uno de los cuales se relacionan las aplicaciones y los grados. Todos los individuos del subcuadro de la parte superior izquierda son de la clase media y tienen elevado IQ, y así sucesivamente. La exacta distribución de filas y columnas no tiene una importancia decisiva, ya que es bien claro que se las puede ordenar de acuerdo con la relación de intereses (tal como se hace en el cuadro XV.5).

CUADRO XV.5. Serie de tablas de contingencia que relacionan dos variables con dos controles simultáneos

Grados	Aplicación elevada		Aplicación baja	
	IQ alto	IQ bajo	IQ alto	IQ bajo
Clase media				
Alto	60	40	40	24
Bajo	20	24	24	12
Clase baja				
Alto	40	6	18	2
Bajo	16	32	38	54

Supóngase que sospechamos una propensión de los maestros en favor de la clase media, que se traduciría en la tendencia a dar buenas notas a los niños de la clase media, independientemente de su capacidad y aplicación, y buenas notas a los niños de la clase inferior solamente cuando muestran capacidad y aplicación a la vez. Anticiparíamos, en tal caso, que las notas habrían de ser por lo regular mejores para los niños de la clase media, controlando la inteligencia y el esfuerzo a la vez, excepto, posiblemente, en el caso de niños de gran capacidad y aplicación. Anticiparíamos asimismo que las relaciones entre las notas por una parte y la capacidad y la aplicación por la otra habrían de ser más fuertes en la clase inferior que en la media. En otros términos, si los niños de la clase media reciben siempre buenas notas, no debería haber relación (o sólo una relación superficial), en esta clase, entre las notas por una parte y la capacidad o la aplicación por la otra. Fijémonos en la relación entre las notas y la capacidad y averiguemos si es o no más fuerte en la clase

inferior. En este caso necesitaremos controlar el esfuerzo. En ambas clases habrá estudiantes aplicados y no tan aplicados. Por lo tanto, podemos construir cuatro tablas de contingencia como las del cuadro XV.5.

Comparamos ahora las dos clases con respecto a la existencia y la fuerza de la relación, considerando separadamente a los alumnos de aplicación elevada y baja respectivamente. La dirección de la relación puede también observarse en cada caso, ya sea calculando los porcentajes o comparando los productos diagonales. Calculando la χ -cuadrada y la ϕ para cada tabla, obtenemos los resultados del cuadro XV.6. Vemos en esta forma que las relaciones no son significativas por lo que se refiere a los niños de la clase media, pero que en cuanto a los de la clase inferior, en cambio, existe una relación positiva moderadamente fuerte en ambas categorías de aplicación entre la capacidad y las notas. Observamos asimismo que la relación es algo más fuerte en el caso de los estudiantes más aplicados.

CUADRO XV.6

Clase	Aplicación	χ -cuadrada	Nivel de significación	ϕ
Media	Alta	2.565	no significativa	.133
	Baja	.188	no significativa	.043
Baja	Alta	28.064	$p < .001$.546
	Baja	15.582	$p < .001$.373

El lector habrá sin duda observado el efecto pronunciado del control sobre el número de casos que figuran en cada casilla. En lugar de tener sólo cuatro casillas, en efecto, tenemos cuatro veces dicho número al servirnos de dos variables de control dicotómicas. Si se hubiera añadido un tercer control simultáneo, por ejemplo, el sexo, habríamos tenido 32 casillas en lugar de 16. Y si cualquiera de las variables hubiera comportado más de dos categorías, el número de las casillas habría aumentado. Así, pues, si bien los controles simultáneos pueden en teoría añadirse indefinidamente, el número de casos ha de ser muy grande para controlarse con este método. Una alternativa consistiría en reducir simplemente el carácter de la población y generalizar sólo respecto de los varones de la clase media de educación universitaria, o de algún otro subgrupo correspondiente. Podría seleccionarse luego una muestra mucho mayor de este subgrupo. Por lo general, si se ha de emplear el control simultáneo, resulta necesario seleccionar aquellos dos o tres controles que se presentan como más prometedores. Es posible, por supuesto, servirse de la prue-

ba exacta de Fisher cuando el número de casos de cada casilla se hace muy pequeño; pero hay que recordar que será en tal caso necesario tener un alto grado de relación para obtener significación. Debido a esta atenuación de los casos, el mero hecho de que una relación se haga no significativa al introducir controles no constituye una prueba suficiente de que la variable de control produce efecto. Habría que calcular y comparar siempre medidas del grado de relación.

En los casos en que difieran las *relaciones* entre una categoría de una variable de control y la siguiente, tendremos un ejemplo de lo que se denomina *no aditividad o interacción estadística*. Ya se examinó esta posibilidad al tratar de la prueba para una diferencia de diferencias en las proporciones, y volveremos al tema de manera más detallada en los capítulos XVI y XX. Siempre que se sospeche la posibilidad de una interacción, *deberá* hacerse una prueba estadística que la localice, antes de seguir adelante. Como inevitablemente habrá algunas diferencias leves en las relaciones entre una muestra y la siguiente, la pregunta básica por formular en tales pruebas será la de si las muestras de interacción son lo suficientemente grandes como para que aquélla haya ocurrido por casualidad, incluso en ausencia de interacción entre la población. En este ejemplo, y dado el caso de que todas las variables han sido dicotomizadas, podrá hacerse una prueba sencilla de una diferencia de diferencias en proporciones, tal como sugiere el capítulo XIII. Como están siendo consideradas simultáneamente dos variables de control, puede incluso darse el caso de que se produzca lo que se denomina una interacción de segundo orden, o una diferencia de diferencias de diferencias. Por ejemplo: la diferencia entre las relaciones de aplicación elevada y aplicación baja puede ser mayor entre los niños de la clase baja que entre los de clase media.

Si se observa que la interacción tiene significación estadística, y es además lo bastante grande como para tener significación sustantiva, resultará necesario cualificar las generalizaciones haciendo una referencia específica a la categoría de control. Habría que decir, por ejemplo: "Se encontró una relación entre grados y habilidad en el caso de los niños de clase baja, pero no en los de clase media." A partir de dicho punto deberán estudiarse separadamente las restantes relaciones entre los dos niveles de clase. Si la interacción es por el contrario estadísticamente insignificante, o tan pequeña que pueda ser ignorada, aun siendo estadísticamente significativa, podrá deducirse razonablemente que las relaciones son básicamente similares entre las categorías de control. Estaremos en tal caso en la posibilidad de simplificar considerablemente el análisis, reuniendo los resultados separados. Veamos a continuación qué tipos específicos de simplificación resultan posibles en el caso de datos categorizados.

Podemos en primer lugar reunir las pruebas de chi al cuadrado en una sola prueba global, a condición de que aquéllas estén basadas en muestras al azar seleccionadas independientemente. El procedimiento es extremadamente sencillo; bastando sumar los distintos valores de chi al cuadrado y también los grados de libertad, evaluando el resultado de la manera habitual. Supongamos por ejemplo que en el caso de cuatro cuadros 2×2 , las chi cuadradas resultantes fueron 2.1, 3.3, 2.7 y 2.9. La suma de estos valores es 11.0, y la de los grados de libertad, 4. En el cuadro vemos que una chi cuadrada de 11.0, con 4 grados de libertad resulta significativa al nivel de .05. Así, aun cuando ninguno de los valores separados de chi al cuadrado fuera significativo, podemos hacer uso del hecho de que el reunir los resultados tiene significación teórica. Estamos en efecto diciendo que si una relación se repite aproximadamente cada vez, pero la probabilidad de los resultados separados es en cada caso mayor de .05, podremos preguntarnos cuál sería el resultado de tal combinación de resultados si no hubiese relación en cualquiera de los cuatro cuadros.

Obsérvese que los resultados de semejante operación de reunión podrían muy bien diferir de la relación total entre dos variables *sin control alguno*. Al juntar los resultados, obtenemos esencialmente una relación *promedia dentro* de las categorías de la variable o las variables de control. Si hubiéramos prescindido simplemente de la variable o las variables de control, los efectos de semejantes controles habrían permanecido oscuros por completo. En tanto que, al unificar, efectuamos una sola prueba de χ -cuadrada de la relación conjunta entre dos variables, controlando en relación con las variables adicionales.

Y en forma análoga, podríamos desear obtener una sola medida de asociación calculando un promedio ponderado de las medidas basado en las cuatro tablas separadas. Un método que se ha sugerido para tal objeto consiste en el empleo de ponderaciones que sean proporcionales al número de los casos de cada tabla. Así, por ejemplo, podríamos multiplicar cada τ , por el número de casos de la tabla, sumar los resultados y dividir, finalmente, entre el número total de casos de las cuatro tablas. Terminaríamos así con una sola prueba de significación y una sola medida de asociación que representarían un promedio de los resultados de las cuatro tablas.

Otro simple procedimiento para obtener una media ponderada es el que describiremos brevemente. (Para mayores detalles véase Rosenberg [12].) El procedimiento consiste, básicamente, en estandarizar todas las categorías de control, mediante la obtención de un promedio ponderado de proporciones (o porcentajes). Supongamos haber obtenido separadamente los resultados siguientes, para hombres y mujeres:

	Varones				Hembras			
	Protes- tantes	Cató- licos	Judíos	Total	Protes- tantes	Cató- licas	Judías	Total
Republicanos	180	80	20	280	100	50	10	160
Demócratas	90	80	50	220	60	30	70	160
Independientes	30	40	30	100	40	20	20	80
Total	300	200	100	600	200	100	100	400

Comenzaremos por transformar las cifras anteriores en proporciones, totalizando a 1.00, ya que la variable independiente aparece en la parte alta de cada cuadro. Los resultados serán los siguientes:

	Varones			Hembras		
	Protes- tantes	Cató- licos	Judíos	Protes- tantes	Cató- licas	Judías
Republicanos	.60	.40	.20	.50	.50	.10
Demócratas	.30	.40	.50	.30	.30	.70
Independientes	.10	.20	.30	.20	.20	.20
Total	1.00	1.00	1.00	1.00	1.00	1.00

Aceptando que deseamos oscurecer las diferencias entre estos dos cuadros, utilizando para ello un promediado, podremos formar un promedio ponderado, multiplicando cada proporción de las contenidas en el cuadro de varones por .6, ya que son 600 los varones en un total de 1 000 individuos en la muestra. De manera análoga podemos ponderar cada cifra en el cuadro de las hembras, multiplicándola por .4. Los resultados serán los siguientes:

	Protestantes	Católicos	Judíos
Republicanos	.56 (.36 + .20)	.44 (.24 + .20)	.16 (.12 + .04)
Demócratas	.30 (.18 + .12)	.36 (.24 + .12)	.58 (.30 + .28)
Independientes	.14 (.06 + .08)	.20 (.12 + .08)	.26 (.18 + .08)
Total	1.00	1.00	1.00

en el que cada proporción de las que aparecen en el cuadro derivado es igual a la suma de las dos proporciones ponderadas (como se indica en los paréntesis), que a su vez figuraban en los

cuadros anteriores. Como la suma de las ponderaciones es de 1.0, también lo será la de las proporciones en cada columna del cuadro derivado. Los resultados pueden ser presentados también bajo la forma de porcentajes.

Este procedimiento para controlar mediante la obtención de promedios ponderados es, como se verá, muy generalizado. Hemos estandarizado el número de protestantes, católicos y judíos, de tal manera que sus tamaños relativos en las muestras de varones y de hembras pierdan significación. Si hubiese habido controles simultáneos para variables adicionales, habríamos podido ampliar este procedimiento de manera directa. Así, si hubiéramos deseado controlar según clases sociales, usando tres niveles, habríamos obtenido seis cuadros, uno para cada categoría sexo-clase. Después de haber vigilado si se produce interacción, y habiendo resuelto que ninguna diferencia importante podrá resultar oscurecida por la aplicación del procedimiento, podríamos asignar de nuevo gravámenes W_i a cada uno de los cuadros de control, haciendo $\sum W_i = 1.0$, obteniendo así un solo cuadro combinado, como en el ejemplo anterior.

Al sustituir así varias medidas y pruebas separadas por una sola medida y una sola prueba, nos enfrentamos a los problemas que se encuentran siempre que se emplean estadísticas de resumen. Concentramos nuestros datos, de modo que resulten menos estadísticos, pero, por otra parte, corremos el riesgo de distorsionar nuestros resultados. Por ejemplo: si una de las cuatro tablas en cuestión diera una χ -cuadrada grande y un grado de relación muy alto, en comparación con las demás, entonces el combinar los resultados, con lo que dicho hecho resulta oscurecido, puede revelarse como sumamente engañoso. O sea que, como siempre, las manipulaciones estadísticas no pueden constituir nunca un sustituto del sentido común.

Algunas de las ideas examinadas en esta sección, en particular las relativas a la reunión de los resultados de tablas separadas, son indudablemente nuevas y podrán parecer algo confusas de momento. Será útil, por lo tanto, volver a repasar esta sección, una vez que el lector se haya enfrentado al material de los capítulos xvi al xx. En dicho momento, en efecto, se habrán examinado ya y comparado diversos tipos de procedimientos de control.

EJERCICIOS

1. Calcúlese la χ -cuadrada para los datos del ejercicio 5 del capítulo ix. Tomando las aspiraciones profesionales como variable dependiente B , ¿cuál es el valor de τ_b ? ¿Cómo se compara el valor de τ_b con el de la medida que se calculó en la parte d) del ejercicio 5?

2. En el ejercicio 3 del capítulo xiv nos servimos de la prueba de Smirnov. Tomando los mismos datos, ¿a qué conclusión llegamos al servirnos de la prueba de la χ -cuadrada? En relación con esos datos

particulares, ¿cuál prueba se preferirá? ¿Por qué? Calcúlense ϕ , T , V , C , τ_b y λ_b .

*3. La prueba de la χ -cuadrada puede emplearse en general para comparar frecuencias observadas y teóricas. En particular, puede utilizarse para verificar la hipótesis nula de que los datos de la muestra se han seleccionado al azar de una población normal. Las frecuencias observadas se comparan con las que se habrían anticipado en caso de ser la distribución efectivamente normal, con la misma media y desviación estándar que se han calculado de los datos de la muestra.

Una vez obtenidos los valores de \bar{X} y de s , podemos servirnos de los verdaderos límites y de la tabla normal para dar las frecuencias esperadas dentro de cada intervalo. Los grados de libertad serán $k - 3$, en donde k representa el número de intervalos. Se perderá un grado de libertad, ya que el total de las frecuencias esperadas ha de ser N ; los otros dos grados de libertad que se han perdido se deben a la necesidad de utilizar \bar{X} y s a título de apreciaciones de los parámetros reales μ y σ . Teniendo estos hechos presentes, verifíquese si los siguientes datos se apartan o no significativamente de la normalidad: Respuesta $\chi^2 = 2.53$, sin rechazar.

Intervalo	Frecuencia
0.0-9.9	7
10.0-19.9	24
20.0-29.9	43
30.0-39.9	56
40.0-49.9	38
50.0-59.9	27
60.0-69.9	13
	208

4. En un estudio reciente, H. L. Wilensky [14] encontró, al controlar la condición socioeconómica, una relación general entre la actividad sindical por una parte y la orientación política y la preferencia electoral por la otra. Los datos de 15 miembros negros tendían a apoyar este hallazgo general en relación con la preferencia electoral. Siete de los ocho negros que eran miembros inactivos del sindicato no siguieron la "línea" de éste al votar en 1948, en tanto que, de los siete miembros sindicalmente activos, cinco votaron de acuerdo con la sugerencia del sindicato. Averigüese si se da o no una relación significativa, sirviéndose: a) de la prueba exacta de Fisher, con dirección anticipada, y b) de la χ -cuadrada corregida con fines de continuidad con dirección anticipada. Respuesta: a) $p = .035$; b) $\chi^2 = 3.22$, $p < .05$.

5. Utilice los datos que siguen (disponiendo los cuadros en otra forma, si es necesario) para obtener información acerca de la precisión de los enunciados a), b) y c). Allí donde sea adecuado, calcúlense medidas del grado de relación y control de las variables relevantes.

a) Las mujeres tienen menos prejuicios que los hombres, independientemente de la religión que profesen o de la clase social a que pertenezcan.

- b) Los grados de relación entre la confesión y el prejuicio contra los negros dependerán de la clase social de la persona "afectada de prejuicio".
- c) La razón de que los judíos aparezcan como menos afectados de prejuicio, en la tabla, que los no judíos se debe al alto porcentaje de mujeres y de personas de la clase superior en la muestra relativa a los judíos.

Religión	Sexo	Grado del prejuicio contra los negros				Totales
		Elevado		Bajo		
		Clase superior	Clase inferior	Clase superior	Clase inferior	
No judíos	Varones	14	30	15	16	75
	Mujeres	8	13	9	7	37
Judíos	Varones	13	7	22	15	57
	Mujeres	18	9	33	21	81
Total						250

6. Utilizando los datos del anterior ejercicio 5, constrúyanse cuadros que relacionen la religión con los prejuicios, con controles simultáneos para sexo y clase social. Suponiendo despreciable la posible interacción, normalícenlos estos resultados de forma que la relación entre religión y prejuicio, con controles, pueda ser presentada en un solo cuadro 2×2 .

*7. Supongamos que se espera llevar a cabo una prueba chi al cuadrado con un cuadro 2×2 , en que se relaciona la preferencia religiosa (protestante-católico), con la preferencia política (republicano-demócrata). Se planea tomar muestras al azar, del mismo tamaño, de protestantes y católicos, y se predice la dirección, esperando que la proporción de protestantes que son republicanos resultara de .60 aproximadamente, en tanto que la proporción de católicos que son republicanos será a su vez de .40, más o menos.

¿Cuántos casos resultarán necesarios si se requiere establecer significación al nivel de .05?

BIBLIOGRAFÍA

1. Anderson, T. R., y M. Zelditch: *A Basic Course in Statistics*, 2ª ed., Holt, Rinehart and Winston, Inc., Nueva York, 1968, cap. 9.
2. Blalock, H. M.: "Probabilistic Interpretations for the Mean Square Contingency", *Journal of the American Statistical Association*, vol. 53, pp. 102-105, 1958.
3. Bradley, J. V.: *Distribution-free Statistical Tests*, Prentice-Hall, Inc., Englewood Cliffs, N. J., 1968, cap. 8.
4. Downie, N. M., y R. W. Heath: *Basic Statistical Methods*, 2ª ed., Harper and Row, Publishers, Incorporated, Nueva York, 1965, cap. 14.

5. Goodman, L. A., y W. H. Kruskal: "Measures of Association for Cross Classifications", *Journal of the American Statistical Association*, vol. 49, pp. 732-764, 1954.
6. Goodman, L. A., y W. H. Kruskal: "Measures of Association for Cross Classifications, II: Further Discussion and References", *Journal of the American Statistical Association*, vol. 54, pp. 123-163, 1959.
7. Goodman, L. A., y W. H. Kruskal: "Measures of Association for Cross Classifications, III: Approximate Sampling Theory", *Journal of American Statistical Association*, vol. 58, pp. 310-364, 1963.
8. Hagood, M. J., y D. O. Price: *Statistics for Sociologists*, Henry Holt and Company, Inc., Nueva York, 1952, cap. 21.
9. Hays, W. L.: *Statistics*, Holt, Rinehart and Winston, Inc., Nueva York, 1963, cap. 17.
10. McCarthy, P. J.: *Introduction to Statistical Reasoning*, McGraw-Hill Book Company, Nueva York, 1957, cap. 11.
11. Mueller, J. H., K. Schuessler, y H. L. Costner: *Statistical Reasoning in Sociology*, 2ª ed. Houghton Mifflin Company, Boston, 1970, cap. 9.
12. Rosenberg, Morris: "Test Factor Standardization as a Method of Interpretation", *Social Forces*, vol. 41, pp. 53-61, 1962.
13. Siegel, Sidney: *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill Book Company, Nueva York, 1956, pp. 96-111.
14. Wilensky, H. L.: "The Labor Vote: A Local Union's Impact on the Political Conduct of its Members", *Social Forces*, vol. 35, pp. 111-120, 1956.

XVI. ANÁLISIS DE LA VARIANCIA

EN EL capítulo XIII comparamos dos muestras investigando la significación de la diferencia entre las medias y las proporciones. Dichas pruebas eran adecuadas al tratamiento de situaciones en las que una de las dos variables mutuamente relacionadas era una escala nominal dicotómica. En el último capítulo vimos de qué modo podían compararse más de dos muestras por medio de la prueba de la χ -cuadrada. En el presente, por su parte, vamos a examinar un tipo muy importante de prueba, el análisis de la variancia, que puede utilizarse para verificar diferencias entre las medias de más de dos muestras. Así, pues, el análisis de la variancia representa una extensión de la prueba de la diferencia de las medias y puede emplearse siempre que estemos verificando una relación entre una escala nominal (o de orden superior) y una escala de intervalo. Veremos asimismo que, en determinadas circunstancias, las pruebas de análisis de variancia pueden extenderse a situaciones en las cuales hay una sola escala de intervalo y dos o más escalas nominales. Se examinarán, además, una prueba análoga no paramétrica y varias medidas de grado de asociación.

XVI.1. *Análisis simple de la variancia*

Si bien el análisis de la variancia puede considerarse como una extensión o generalización de la prueba de la diferencia de las medias, comporta con todo algunos principios fundamentalmente nuevos que requieren una explicación relativamente larga. De ahí que una breve exposición general resulte tal vez indicada, a fin de que el lector no se pierda en los detalles. Los supuestos del análisis de variancia son básicamente los mismos que los de la prueba de la diferencia de las medias, pero la prueba en sí misma es muy distinta. Habremos de suponer normalidad, muestras aleatorias independientes, poblaciones y desviaciones estándar iguales, y la hipótesis nula será que las medias de las poblaciones son iguales. La prueba misma supone el trabajar directamente con variancias y no con medias y errores estándar.

Supóngase que los datos del cuadro XVI.1 representan las tasas de criminalidad de tres tipos de ciudades distintos, a saber: centros industriales, comerciales, o políticos. Podemos calcular medias separadas para cada una de esas tres categorías o muestras, y podemos obtener también una media grande, prescindiendo de las clases y promediando los datos. En el presente ejemplo, las tres muestras son del mismo tamaño, pero esto no necesita ser siempre así.

Como quiera que se presume que todas las poblaciones tienen la misma desviación estándar, podemos formar dos apreciaciones independientemente de la variancia σ^2 común. Una de estas apreciaciones será directamente análoga a la estimación unificada que utilizamos en la prueba de la diferencia de las medias. Esta estimación será un promedio ponderado de las variancias *dentro* de las muestras separadas y será siempre insesgada, incluso si las medias de las muestras difieren considerablemente entre sí. Esto es así porque la variancia de cada muestra se calculará separadamente y sólo comportará las desviaciones respecto de la media de la muestra particular.

CUADRO XVI.1. *Datos para el análisis de variancia*

	Tasas de criminalidad			Total
	Centro industrial	Centro comercial	Centro político	
	4.3	5.1	12.5	
	2.8	6.2	3.1	
	12.3	1.8	1.6	
	16.3	9.5	6.2	
	5.9	4.1	3.8	
	7.7	3.6	7.1	
	9.1	11.2	11.4	
	10.2	3.3	1.9	
Sumas	68.6	44.8	47.6	161.0
Medias	8.58	5.60	5.95	6.71
Nº de casos	8	8	8	24

La segunda estimación de la variancia común comporta la variancia de las medias particulares de las muestras tratadas como datos individuales. En este caso, las desviaciones de las medias de las muestras respecto de la media grande se utilizarán para la estimación de σ^2 . Para los datos del cuadro XVI.1 obtendríamos la variación de las medias de las tres muestras, o sea 8.58, 5.60 y 5.95, respecto de la media total de 6.71. Esta estimación de σ^2 sólo será equilibrada si las medias de las poblaciones son de hecho iguales. Si las medias de las poblaciones son iguales, en efecto, puede esperarse que las de las muestras variarán una respecto de otra de acuerdo con el teorema del límite central, esto es, acercándose a una distribución normal a medida que aumenta el tamaño de la muestra, y podemos servirnos de esta ley y de las diferencias reales entre las medias de las muestras para apreciar la verdadera variancia. Por otra parte, si las medias de las poblaciones son en realidad diferentes, esperamos que las

medias de las muestras diferirán una respecto de otra más de lo que sería el caso si las medias de las poblaciones fueran las mismas. Por consiguiente, si la hipótesis nula es falsa, la segunda estimación de σ^2 será por lo general demasiado grande, y será una estimación sesgada.

La prueba empleada en el análisis de la variancia comporta una comparación de las dos estimaciones distintas de la variancia de la población. Sin embargo, en lugar de tomar la diferencia entre las dos estimaciones, tomamos la *razón* de la segunda a la primera. Si la hipótesis nula es correcta, las dos estimaciones serán insesgadas, y la razón habría de ser aproximadamente la unidad. En cambio, si las medias de la población difieren, la segunda estimación será por lo regular mayor que la primera, y la razón será mayor que la unidad. Como quiera que las variaciones de las muestras son siempre un factor, hemos de preguntarnos cuán grande sea la razón que estamos dispuestos a tolerar antes de poner a la hipótesis nula en duda. Afortunadamente, la razón F de las dos estimaciones tiene una distribución de muestreo conocida, a condición que las dos estimaciones de la variancia sean efectivamente independientes una de otra, y de ahí que pueda hacerse una prueba relativamente sencilla. Esto es lo que hacemos esencialmente en la prueba del análisis de la variancia. Veamos ahora en detalle el procedimiento que ello comporta.

Fragmentación de la variación total en partes componentes. Si bien nuestro objetivo último está en la formación de dos estimaciones distintas de la variancia, será menester introducir un nuevo concepto para explicar cómo dichas estimaciones se obtienen. Sirvamos del término *variación* (diferente del de variancia) para designar la suma de las desviaciones cuadradas con respecto a la media. En este caso, la variación total respecto de la media grande será para todas las muestras $\sum_{i=1}^N (X_i - \bar{X})^2$. Así,

pues, el término variación designa una suma de cuadrados, prescindiendo del número de casos implicado. Procedemos ahora a fragmentar esta variación total en dos partes componentes, cada una de las cuales se utilizará en el cálculo de las dos estimaciones.

Representemos nuestros datos en forma simbólica, como en el cuadro XVI.2. Los datos individuales están representados por $X_{11}, X_{21}, \dots, X_{ij}$; las medias de las muestras, por $\bar{X}_{.1}, \bar{X}_{.2}, \dots, \bar{X}_{.k}$, y la media grande por $\bar{X}_{..}$. Los puntos se emplean en los subíndices para distinguir las medias de las columnas de las medias de las hileras, que se emplearán cuando añadamos una segunda escala nominal. El símbolo general X_{ij} representa la marca del i -ésimo individuo en la columna j -ésima. La suma $\sum_{i=1}^{N_1} X_{i1}$ indica

que se han sumado las N_1 marcas de la primera columna, y lo mismo en relación con las columnas restantes.¹

Ahora practicamos algo de álgebra. Podemos escribir:

$$X_{ij} - \bar{X}_{..} = (X_{ij} - \bar{X}_{.j}) + (\bar{X}_{.j} - \bar{X}_{..})$$

o sea

$$\left(\begin{array}{c} \text{dato} \\ \text{individual} \end{array} - \begin{array}{c} \text{media} \\ \text{grande} \end{array} \right) = \left(\begin{array}{c} \text{dato} \\ \text{individual} \end{array} - \begin{array}{c} \text{media de} \\ \text{la clase} \end{array} \right) + \left(\begin{array}{c} \text{media de} \\ \text{la clase} \end{array} - \begin{array}{c} \text{media} \\ \text{grande} \end{array} \right)$$

en lo que hemos restado $\bar{X}_{.j}$ (la media de la columna j -ésima) de X_{ij} , para volverla a adicionar inmediatamente. Por lo tanto, hemos expresado la diferencia entre un dato individual singular y la media grande como suma de dos cantidades, a saber: 1) la diferencia entre su dato y la media de la categoría a la que pertenece, y 2) la diferencia entre la media de la clase y la de la me-

CUADRO XVI.2. Representación simbólica de los datos para el análisis de la variancia

	Categorías				Total
	A_1	A_2	\dots	A_k	
Marcas	X_{11}	X_{12}	\dots	X_{1k}	
	X_{21}	X_{22}	\dots	X_{2k}	
	X_{31}	X_{32}	\dots	X_{3k}	
	\vdots	\vdots	\dots	\vdots	
	$X_{N_1 1}$	$X_{N_2 2}$		$X_{N_k k}$	
Sumas	$\sum_{i=1}^{N_1} X_{i1}$	$\sum_{i=1}^{N_2} X_{i2}$	\dots	$\sum_{i=1}^{N_k} X_{ik}$	$\sum_i \sum_j X_{ij}$
Medias	$\bar{X}_{.1}$	$\bar{X}_{.2}$	\dots	$\bar{X}_{.k}$	$\bar{X}_{..}$
Nº de casos	N_1	N_2	\dots	N_k	N

¹ Como quiera que tenemos dos subíndices, i y j , importa distinguir entre \sum_i y \sum_j . En el último caso, los valores j se sumarían para cualquier i (fijo), y obtendríamos así la suma de los datos de la *hiler*a i -ésima.

dia grande. En el ejemplo numérico anterior podemos expresar la diferencia entre el dato del primer individuo de la primera clase y la media grande como:

$$4.3 - 6.71 = (4.3 - 8.58) + (8.58 - 6.71)$$

$$\text{o sea} \quad -2.41 = -4.28 + 1.87$$

Si elevamos al cuadrado ambos miembros de la ecuación, obtenemos:

$$(X_{ij} - \bar{X}_{..})^2 = (X_{ij} - \bar{X}_{.j})^2 + 2(X_{ij} - \bar{X}_{.j})(\bar{X}_{.j} - \bar{X}_{..}) + (\bar{X}_{.j} - \bar{X}_{..})^2$$

Sumando ambos lados obtenemos la suma de las desviaciones cuadradas de todos los individuos. Podemos sumar primero cada columna y añadir luego las cifras resultantes de cada clase. Al hacerlo así, el término central se anula. Para ver por qué esto es así, obsérvese que, al sumar cualquier columna particular, el valor de j será constante. Por lo tanto, para la columna j -ésima el factor $(\bar{X}_{.j} - \bar{X}_{..})$ será constante y puede tomarse fuera de la suma total. Así, por ejemplo, para la suma de los datos de la columna j -ésima el término central se convierte en

$$2(\bar{X}_{.j} - \bar{X}_{..}) \sum_i (X_{ij} - \bar{X}_{.j})$$

Pero, como quiera que para cada columna las desviaciones respecto de la media de la columna han de ser cero, vemos inmediatamente que el término central ha de desaparecer para todas y cada una de las columnas. Obtenemos, pues:

$$\sum_i \sum_j (X_{ij} - \bar{X}_{..})^2 = \sum_i \sum_j (X_{ij} - \bar{X}_{.j})^2 + \sum_i \sum_j (\bar{X}_{.j} - \bar{X}_{..})^2 \quad (\text{XVI.1})$$

$$\begin{array}{lcl} \text{Suma total de} & \text{suma de cuadra-} & \text{suma de cuadra-} \\ \text{los cuadrados} & \text{dos (dentro)} & \text{dos (entre)} \end{array} = +$$

Al proceder así, obtenemos una doble suma total que escribimos como $\sum_i \sum_j$, indicando que hemos sumado tanto las hileras como las columnas.

Hemos dividido la variación total en dos partes. La primera es una suma de las desviaciones cuadradas de los datos individuales respecto de las medias de sus clases respectivas. Esta se designa como suma *dentro* de los cuadrados y se empleará para obtener nuestra primera estimación de la variancia común σ^2 .

Obsérvese que esta suma de cuadrados se obtiene esencialmente en la misma forma en que se formó la estimación unificada en la prueba de la diferencia de las medias. Si escribimos la suma interior de cuadrados como:

$$\sum_{i=1}^{N_1} (X_{i1} - \bar{X}_{.1})^2 + \sum_{i=1}^{N_2} (X_{i2} - \bar{X}_{.2})^2 + \dots + \sum_{i=1}^{N_k} (X_{ik} - \bar{X}_{.k})^2$$

vemos que el primer término es exactamente $N_1 s_1^2$, en donde las desviaciones se han tomado respecto de la media de la categoría, y en forma análoga en relación con los otros términos. Por lo tanto:

$$\text{SC interior} = N_1 s_1^2 + N_2 s_2^2 + \dots + N_k s_k^2$$

Si dividimos entre los grados apropiados de libertad, que resultarán ser $N - k$, obtenemos una estimación unificada, basada en todas las k categorías. La segunda suma de cuadrados, o suma *entre* columnas, comporta las desviaciones de las medias de las categorías respecto de la media grande, siendo por consiguiente una medida de la variación entre las muestras. La segunda estimación de la variancia se basará en esta suma de cuadrados entre columnas.

Las sumas dentro y entre cuadrados se designan a menudo como variaciones *explicadas* e *inexplicadas* respectivamente. Resulta tal vez más fácil ver por qué la variación interior se designe como inexplicada, ya que se refiere a la variación que no se tiene en cuenta en la variable de la categoría. Si dentro de la categoría A hay todavía alguna otra variabilidad respecto de la media de la categoría, esta variabilidad no puede ciertamente explicarse por la categoría. Por otra parte, si las medias de las categorías difieren considerablemente entre sí, una fracción relativamente grande de la variación total puede atribuirse a diferencias entre varias categorías. Así, pues, es la magnitud de la variabilidad dentro de las categorías, comparada con las diferencias entre ellas, la que determina hasta qué grado las dos variables están asociadas. Categorías homogéneas que difieran considerablemente entre sí explican un alto grado de variación.² En el caso extremo, si tuviéramos categorías perfectamente homogéneas, la suma dentro de los cuadros sería cero, y toda la variabilidad podría atribuirse a la variable de la categoría. Así, por ejemplo, si todas las ciudades industriales tuvieran exactamente la misma

² Lo que sin embargo no implica causalidad, por supuesto. La palabra "explicado", tal como se la emplea en la bibliografía estadística, se traduce mejor como: "asociado con", no debiendo en forma alguna interpretarse en el sentido de suponer necesariamente que se haya localizado una variable explicativa en el sentido causal o teórico.

tasa de criminalidad y difirieran de la de los centros comerciales, de tasas también totalmente homogéneas, etcétera, entonces podría decirse que el tipo de ciudad explicaba toda la variación en materia de tasas de criminalidad. O sea que, sabiendo de cuál tipo de ciudad se trata, estaríamos en condiciones de anticipar dicha tasa exactamente.

Con objeto de obtener apreciaciones de estas dos sumas distintas de cuadrados, basta dividir entre los grados apropiados de libertad. Ahora bien, los grados de libertad asociados a la suma total de los cuadrados es $N - 1$, ya que, según vimos, $\hat{\sigma}^2$ es la estimación insesgada de σ^2 , habiéndose perdido un grado de libertad debido al cálculo de la media general $\bar{X}_{..}$. Veamos ahora la suma de los cuadrados entre columnas. Esta cantidad representa la suma de las desviaciones cuadradas de las medias de la muestra k respecto de la media general. En efecto, la media de cada categoría se trata como caso particular. Por lo tanto, están implicados $k - 1$ grados de libertad, ya que un grado se ha perdido, debido al hecho que el promedio ponderado de $\bar{X}_{.j}$ ha de ser $\bar{X}_{..}$. En el caso de la apreciación de la clase interior, se perderá un grado de libertad en cada columna a causa del cálculo de la $\bar{X}_{.j}$. Por lo tanto, en conjunto habrá $N - k$ grados de libertad asociados a la apreciación interior. Obsérvese que los grados de libertad se suman, lo mismo que las sumas de cuadrados. Así, pues:

$$N - 1 = (N - k) + (k - 1)$$

dif. total = dif. dentro + dif. entre columnas

En esta forma, nuestras dos estimaciones de la variancia común se convierten en:

$$\text{estimación dentro} = \frac{\sum_i \sum_j (X_{ij} - \bar{X}_{.j})^2}{N - k} \quad (\text{XVI.2})$$

$$\text{estimación entre columnas} = \frac{\sum_i \sum_j (\bar{X}_{.j} - \bar{X}_{..})^2}{k - 1} \quad (\text{XVI.3})$$

Llegados a este punto, es posible que se le haya ocurrido al lector que, si incluimos la estimación usual basada en la suma total de los cuadrados, tenemos en realidad tres estimaciones distintas de la variancia total. ¿Por qué, pues, no comparar aquella con cualquiera de las otras dos, ya que dicha estimación total bien podría constituir una estimación mejor que cualquiera de éstas? Se recordará, sin embargo, que la prueba F requiere que las es-

timaciones comparadas sean independientes unas de otras. Y la estimación basada en la suma total de los cuadrados no es independiente de las otras, siendo ésta la razón de que no se la pueda utilizar en la prueba de la F . Por lo regular, las sumas de cuadrados dentro y entre columnas no son independientes una de otra. Pero ocurre que la distribución normal posee la propiedad de que dichas cantidades sean independientes, a pesar de que las mismas $\bar{X}_{.j}$ figuren en ambas expresiones. Ésta es la razón en cuya virtud hemos de suponer que todas las poblaciones son normales. Se recordará que también en el caso de la distribución t se requería normalidad, debido a la necesidad de que el numerador fuera independiente del denominador. Como lo veremos en seguida, la distribución t es un caso particular de la distribución F .

Problema. Sirvámonos de los datos hipotéticos anteriores, que representan tasas de criminalidad para tres tipos de ciudades. Nos interesa saber si existen diferencias significativas entre las medidas de los tres tipos de ciudades en cuestión.

1. Supuestos.

Nivel de medición: Tasas de criminalidad, escala de intervalo. Tipo de ciudad como escala nominal.

Modelo: Muestreo independiente aleatorio.

Poblaciones normales para cada tipo de ciudad.

Las variancias de las poblaciones son iguales.

$$(\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 = \sigma^2)$$

Hipótesis: Las medias de las poblaciones son iguales.

$$(\mu_{.1} = \mu_{.2} = \dots = \mu_{.k})$$

Lo mismo que en el caso de la prueba de la diferencia de las medias, hay que suponer que las muestras se han seleccionado independientemente una de otra. En otros términos: las ciudades no están asociadas en forma alguna. Como quiera que se supone que las poblaciones de los tres tipos de ciudades son normales, con medias y variancias iguales, estamos suponiendo en realidad que son idénticas. Por lo tanto, las tres muestras pueden considerarse como si se hubieran tomado al azar de una misma población. Por lo regular, el investigador está interesado en el supuesto de medias iguales. En el presente ejemplo, anticipará probablemente diferencias en las tasas de criminalidad de los tres tipos de ciudades, y establecerá la hipótesis nula de que no existe entre ellos diferencia alguna. Conviene observar que no se requieren muestras grandes, debido al supuesto de normalidad. Sin embargo, es obvio que si en cada categoría sólo hubiera un caso, no habría variabilidad en el interior de las categorías, con lo que la prueba no sería posible.

La prueba de la F en sí misma no verifica el supuesto de va-

riancias iguales u *homoscedasticidad* (como se designa el supuesto en lenguaje técnico). En situaciones en las que las variancias de las muestras parecen diferir mucho entre sí, puede practicarse una prueba independiente para la igualdad de las variancias (véase [1], pp. 141 a 144). Si los resultados de una prueba de esta clase indican que hay desviaciones más bien extremas de la homogeneidad de la variancia, entonces no debería emplearse el análisis de ésta. Sin embargo, pueden con todo tolerarse desviaciones moderadas de la homogeneidad. Semejantes desviaciones pueden reducirse a menudo considerablemente mediante transformación de las variables.³ Si una categoría particular es o mucho más o mucho menos homogénea que las otras, puede resultar indicado descartarla del análisis de la variancia. En términos generales, las desviaciones moderadas respecto de la normalidad y de la igualdad de las variancias pueden tolerarse sin necesidad de recurrir al uso de las alternativas no paramétricas (véase [1], pp. 220 a 223).

2. *Nivel de significación y región crítica.* Sirvámolos de un nivel de .05. Si la hipótesis nula es en realidad incorrecta, entonces, si tomamos siempre la razón de la estimación dentro a la entre columnas, podemos esperar encontrar que el valor de F sea mayor que la unidad. Por lo tanto, nos serviremos de la cola mayor de la distribución F como región crítica. Si resulta que F es menor que la unidad, no tendrá objeto alguno buscar en la tabla el valor de la probabilidad, ya que, para descartar la hipótesis nula, se necesitarán valores de F superiores a la unidad. Una F menor que la unidad indicaría un grado mayor de heterogeneidad dentro de las categorías de lo que se esperaría al azar. El lector ha de recordar una vez más que, aunque sólo nos sirvamos de una sola cola de la distribución F , esto no significa en modo alguno que anticipemos cuál de las medias de las categorías será mayor.

3. *Distribución de muestreo.* La distribución de muestreo de F está dada en el cuadro J del Apéndice 2. El empleo de este cuadro se describe más abajo.

4. *Cálculo de la estadística de la prueba.* Con objeto de obtener un valor de F , razón de las estimaciones entre y dentro de columnas, será necesario calcular primero los totales entre y dentro de cuadrados. Como quiera que la variación total es igual a la suma de las otras dos, sólo necesitaremos calcular dos de los valores en cuestión, ya que el tercero se obtendrá por suma o diferencia de éstos. Se recordará que la suma dentro de los cuadrados comporta una operación de unificación. Esto repre-

³ Ocurre, por ejemplo, a veces que las categorías que tienen las medias más grandes son también las menos homogéneas. En tales casos, si se toma como escala de intervalo el logaritmo de la variable original, el efecto será el de igualar las variancias. Para un examen más detallado del empleo de la transformación logarítmica, véase la sec. XVIII.2.

senta considerablemente más trabajo que el que se requiere para las otras dos sumas de cuadrados y, por lo tanto, obtenemos la suma dentro de éstos restando la suma entre columnas de la suma total de los mismos.

La fórmula de cálculo para la suma total de los cuadrados se obtiene en la misma forma que la de la variancia [véase la ecuación (VI.6)]. Así pues:

$$\begin{aligned} \text{Suma total de los cuadrados} &= \sum_i \sum_j (X_{ij} - \bar{X}_{..})^2 = \\ &= \sum_i \sum_j X_{ij}^2 - \frac{(\sum_i \sum_j X_{ij})^2}{N} \end{aligned} \quad (\text{XVI.4})$$

Esta es la misma fórmula que empleamos al calcular las desviaciones estándar, sólo que ahora es necesario servirse de un doble signo de suma total.

La fórmula de cálculo de las variaciones entre columnas se presenta a primera vista como formidable, pero, si se mira más de cerca, encuéntrase que comporta un procedimiento relativamente sencillo. Es como sigue:

$$\begin{aligned} \text{Suma de cuadrados entre columnas} &= \sum_j \frac{(\sum_i X_{ij})^2}{N_j} - \frac{(\sum_i \sum_j X_{ij})^2}{N} \end{aligned} \quad (\text{XVI.5})$$

$$= \left[\frac{(\sum_i X_{i1})^2}{N_1} + \frac{(\sum_i X_{i2})^2}{N_2} + \dots + \frac{(\sum_i X_{ik})^2}{N_k} \right] - \frac{(\sum_i \sum_j X_{ij})^2}{N}$$

Obsérvese que el segundo término de la expresión anterior es el mismo factor que se sustrajo de $\sum_i \sum_j X_{ij}^2$ para obtener la suma total de cuadrados. El primer término, en cambio, es susceptible de desorientar al lector. Analizando esta expresión, vemos que calculamos primero la suma de cada columna y luego la elevamos al cuadrado para obtener $(\sum_i X_{ij})^2$. Dividimos luego dicha expresión entre el número de casos de la columna, que no necesita ser siempre el mismo. Tenemos así para la columna j -ésima: $(\sum_i X_{ij})^2/N_j$. Finalmente, hacemos lo mismo con cada columna y sumamos los resultados.

Los cálculos del problema numérico que se dan a continuación ayudarán a aclarar el procedimiento. Las sumas total y entre columna de cuadrados se calculan como sigue:

$$\sum_{i,j} X_{ij}^2 = (4.3)^2 + (2.8)^2 + \dots + (1.9)^2 = 1\,453.58$$

$$\frac{(\sum_{i,j} X_{ij})^2}{N} = \frac{(161.0)^2}{24} = 1\,080.042$$

$$SC \text{ totales} = 1\,453.58 - 1\,080.042 = 373.538$$

$$SC \text{ entre columnas} = \frac{(68.6)^2}{8} + \frac{(44.8)^2}{8} + \frac{(47.6)^2}{8} - 1\,080.042$$

$$= 1\,122.345 - 1\,080.042 = 42.303$$

Para obtener la suma de cuadrados dentro sustraemos simplemente la segunda expresión de la primera obteniendo:

$$SC \text{ dentro} = SC \text{ totales} - SC \text{ entre columnas}$$

o

$$331.235 = 373.538 - 42.303$$

Las apreciaciones de la variancia común pueden calcularse ahora dividiendo entre los grados apropiados de libertad. Finalmente, la F se calcula dividiendo la estimación entre columnas entre la estimación interior. Estos cálculos se resumen en el cuadro XVI.3.

CUADRO XVI.3. Cálculos para el análisis de la variancia

	Sumas de cuadrados	Grados de libertad	Estimación de la variancia	F
Total	373.538	$N - 1 = 23$		
Entre columnas	42.303	$k - 1 = 2$	21.152	
Dentro de columnas	331.235	$N - k = 21$	15.773	1.34

5. *Decisión.* Para decidir si descartamos o no la hipótesis nula, hemos de averiguar si el valor de F queda o no en la región crítica. Se observará que se dan tres cuadros distintos de F , que corresponden a los niveles de significación del .05, .01 y .001 respectivamente. Esta información no puede condensarse en un solo cuadro, porque hay que asociar con cada F dos grados de libertad, uno para el numerador y otro para el denominador. Los grados de libertad asociados al numerador, o sea la estimación entre columnas, se encuentran buscando horizontalmente arriba del cuadro, en tanto que los del denominador, o estimación dentro, se obtienen leyendo el cuadro de arriba abajo. Obsérvese que todos los valores de F dados en el cuadro son ≥ 1.0 , lo que indica

que el cuadro se ha establecido directamente para pruebas de una cola. En otros términos: el numerador es siempre la mayor de las dos estimaciones. En nuestro problema obtuvimos una F con 2 y 21 grados de libertad (se escribe $F_{2,21}$) igual a 1.34. Sirviéndonos del cuadro del nivel de significación del .05, y buscando los grados apropiados de libertad, encontramos la cifra de 3.47. Sabemos, pues, que, si los supuestos fueran correctos, obtendríamos un valor de F igual o mayor que éste menos del 5 por ciento de las veces. Como quiera que el valor efectivamente obtenido para F es menos que 3.47, no descartamos la hipótesis nula al nivel del .05. Decidimos que no se dispone de pruebas suficientes para concluir que los tipos de ciudades difieren realmente uno respecto de otro en cuanto a las tasas de criminalidad.

XVI.2. Comparación de medias específicas

Se habrá observado que el problema anterior pudo haberse tratado sirviéndonos de la prueba de la diferencia de las medias que comporta la distribución t . Pudieron haberse hecho tres comparaciones distintas, por pares, entre las ciudades industriales y comerciales, industriales y políticas, y comerciales y políticas. En contraste con esto, el análisis de la variancia brinda una *prueba sola* acerca de si los tres tipos de ciudades difieren o no significativamente entre sí o, en otros términos, si todos ellos pudieron proceder de la misma población. La ventaja del análisis de la variancia está en que puede emplearse una prueba sola en lugar de muchas. Si hubiera habido cuatro categorías, se habrían requerido $4(3)/2$, o sean 6 pruebas de diferencia de las medias. Con 6 categorías se necesitarían 15 pruebas, y con 10 categorías 45. Supóngase que se necesitaban 15 pruebas y que solamente 4 de ellas resultaban significativas, ¿qué concluiríamos? Sería difícil decirlo.

Hay una salida fácil que a primera vista parece ser un procedimiento razonable. ¿Por qué no efectuar simplemente una prueba de diferencia de medias con las dos categorías que presentan respectivamente las medias mayor y menor? Porque si éstas son significativamente distintas, podemos concluir que las categorías difieren efectivamente entre sí. Hemos de recordar, sin embargo, que (suponiendo muestras del mismo tamaño) en esta forma seleccionaríamos la prueba única que presentara mayores probabilidades de dar significado, prescindiendo de las demás. Como quiera que podemos esperar que al nivel del .05 una prueba sobre veinte dé significado incluso si todas las medidas de las poblaciones son iguales, es evidente que cargaríamos así los dados en favor del rechazo. En otros términos: el nivel de significación realmente empleado no sería del .05, sino tal vez el del .5 o .7, ya que estamos obteniendo la probabilidad

de conseguir *por lo menos un éxito* (significación al nivel del .05) en un gran número de pruebas.

Sin embargo, no debe deducirse de ello que el análisis de la variancia sea siempre preferible a una serie de pruebas de diferencia de medias. Estas últimas, en efecto, si se emplean cautamente, pueden suministrar considerable información. Así, por ejemplo, el análisis de la variancia puede conducir a resultados significativos sobre todo debido al hecho de que una de las categorías se aleje mucho de las restantes. De modo que si dicha categoría se hubiera excluido, la conclusión pudo haber sido totalmente distinta. En cambio, una serie de pruebas de diferencia de medias podría indicar el hecho en cuestión con mayor claridad. Si antes de empezar la prueba se sospecha, en particular, que una o varias categorías podrán acaso diferir mucho de las otras, entonces cierto número de pruebas de diferencia de medias de una sola cola podrá resultar más adecuado. Es posible también, en ocasiones, anticipar el orden en que quedarán las medias de las categorías. Supóngase, por ejemplo, que se hubiera predicho que las tasas de criminalidad serían las mayores en las ciudades industriales y mínimas en las políticas. En tal caso pudieran haberse utilizado dos pruebas de diferencia de las medias de una sola cola, o sea: una de ellas anticipando una diferencia entre las ciudades industriales y las comerciales, y otra anticipando una diferencia entre estas últimas y los centros gubernamentales. En términos generales, parece ser que cuanto mayor conocimiento tengamos para predecir las magnitudes relativas de las diferencias y sus direcciones, o éstas, tanto más probable resulta que las pruebas distintas de la diferencia de las medias sean adecuadas. El análisis de la variancia, en cambio, parece ser más útil al nivel de exploración.

Finalmente, puede observarse la relación entre las distribuciones t y F . Si sólo hubiera habido dos tipos de ciudades, podría también haberse hecho una prueba de análisis de variancia, comparando luego los resultados con los de una prueba t de diferencia de las medias. En este caso, los grados de libertad asociados al numerador de F habrían sido $2 - 1$, o sea 1. En tanto que los grados de libertad del denominador habrían sido $N - 2$, los mismos que para t en la prueba de la diferencia de las medias. Hay que recordar, también, que cuando suponemos $\sigma_1 = \sigma_2$, los denominadores tanto de t como de F comportan estimaciones unificadas de la variancia. Resulta que la distribución t puede considerarse como caso particular de la distribución F . Si calculáramos los valores de t^2 con $N - 2$ grados de libertad, encontraríamos que son exactamente los mismos que los de una F de 1 y $N - 2$ grados de libertad, como puede comprobarse comparando los cuadros F y t . En otros términos, t es la raíz cuadrada de una F que tenga un grado de libertad asociado a su

numerador. Esto significa, por supuesto, que se llegará exactamente a las mismas conclusiones en el caso de dos muestras, independientemente de si nos servimos de la prueba de análisis de variancia o de la de diferencia de las medias. En este sentido, el análisis de la variancia es en realidad una extensión de la prueba de la diferencia de las medias.

* *Comparaciones ortogonales.* En muchas ocasiones en que son comparadas más de dos categorías resulta conveniente hacer un cierto número de comparaciones específicas previamente planeadas, basadas en un interés teórico, y orientadas a comprobar los procedimientos de prueba. Supongamos por ejemplo que en un experimento aparecen cinco grupos, uno de los cuales es de control, en tanto los restantes están sujetos a diferentes tipos de manipulación experimental. Puede ocurrir que los grupos segundo y tercero cuenten con dirigentes autoritarios que se han visto sometidos a grados de frustración, mediano en el del segundo y elevado en el del tercero. También los grupos cuarto y quinto pueden haberse visto sujetos a grados moderados y extremos de frustración, pero han desarrollado experiencias de dirección democrática. Podemos desear comparar el grupo testigo con cada uno de los cuatro grupos experimentales, pero a la vez podemos proponernos comparar los dos grupos autoritarios con los dos democráticos, o los dos grupos sometidos a una frustración moderada con los otros dos en los que la frustración era extremada. ¿Son legítimas todas estas comparaciones, en el sentido de que no nos vayan a brindar información redundante? Dicho de otra manera: si conocemos el resultado de una comparación, ¿no podrá ocurrir que tal resultado haya de arrojar luz sobre alguna de las demás? Necesitamos un sistema que nos permita decidir si las comparaciones son *ortogonales* o si son realmente independientes.⁴

Podemos hacer uso de nuevo de la idea de las funciones lineales, mediante un procedimiento que viene a ser una ampliación directa de la prueba de la diferencia de medias. Si deseamos comparar el grupo de control (grupo I), con los grupos experimentales, se nos ocurriría naturalmente restar la media de las medias de los cuatro grupos experimentales, de la media del grupo de control. De manera análoga, si deseamos comparar los grupos autoritarios con los democráticos, restaríamos naturalmente la media de los grupos IV y V (democráticos) de la de los grupos II y III. Si damos la misma ponderación a todos los

⁴ La idea de ortogonalidad se deriva de una interpretación geométrica de las asociaciones estadísticas, y se refiere a aquellas situaciones en las que la relación puede ser representada mediante ejes perpendiculares u ortogonales. De interés para nosotros es que si también suponemos homocedasticidad y normalidad en la distribución de la variable dependiente, puede demostrarse que la ortogonalidad implica la independencia estadística.

grupos (con independencia del tamaño relativo de las muestras), ello supondría comparar las medias de las dos medias, o $(\frac{1}{2})(\bar{X}_2 + \bar{X}_3) - (\frac{1}{2})(\bar{X}_4 + \bar{X}_5)$, siendo la hipótesis nula la de que $(\frac{1}{2})(\mu_2 + \mu_3) - (\frac{1}{2})(\mu_4 + \mu_5) = 0$.

Definamos de modo más general una función de ψ_i para la i -ésima comparación que deseamos hacer, como sigue:

$$\psi_i = c_{i1}\mu_1 + c_{i2}\mu_2 + \dots + c_{ik}\mu_k = \sum_{j=1}^k c_{ij}\mu_j$$

En donde c_{ij} son ponderaciones muy simples, dependientes de la comparación que se lleva a cabo. Si imponemos la restricción de que la suma de las ponderaciones debe ser igual a cero, es decir, $\sum_{j=1}^k c_{ij} = 0$, se simplificará grandemente el análisis sin restringir de ninguna manera las comparaciones a realizar. Así, si nuestra primera comparación se refiere al grupo de control contra la media de los cuatro grupos restantes, podemos tomar $c_{11} = 1$, con los restantes c_{1j} iguales todos a $-\frac{1}{4}$. Si una particular comparación deja simplemente fuera una de las categorías (por ejemplo el grupo de control), haremos que el c_{ij} para aquella categoría sea igual a cero. De esta manera tendremos, en el caso de las tres comparaciones que estamos considerando:

	I	II	III	IV	V
ψ_1 : control contra los demás (I vs. II, III, IV y V)	1	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$
ψ_2 : autoritarios contra democráticos (II y III vs. IV y V)	0	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$
ψ_3 : frustración moderada contra extrema (II y IV vs. III y V)	0	$\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$

Si las variancias de población σ_i^2 son aproximadamente iguales, las poblaciones aproximadamente normales, y todas las muestras del mismo tamaño, las comparaciones separadas serán mutuamente independientes (como asimismo ortogonales), siempre que se produzca la siguiente relación entre los coeficientes:

$$\sum_{j=1}^k c_{hj}c_{ij} = 0 \quad \text{para todas las } h \neq i$$

En particular comenzaremos por examinar el primer par de comparaciones ($h = 1, i = 2$). En nuestro caso tendremos:

$$c_{11}c_{21} + c_{12}c_{22} + c_{13}c_{23} + c_{14}c_{24} + c_{15}c_{25}$$

$$= 1(0) + (-\frac{1}{4})(\frac{1}{2}) + (-\frac{1}{4})(\frac{1}{2}) + (-\frac{1}{4})(-\frac{1}{2}) + (-\frac{1}{4})(-\frac{1}{2}) = 0$$

viendo que la condición se aplica. Pasamos a continuación a las comparaciones primera y tercera y finalmente a las segunda y tercera, observando de nuevo que la suma requerida de los productos es igual a cero. Así:

$$1(0) + (-\frac{1}{4})(\frac{1}{2}) + (-\frac{1}{4})(-\frac{1}{2}) + (-\frac{1}{4})(\frac{1}{2}) + (-\frac{1}{4})(-\frac{1}{2}) = 0$$

y

$$0(0) + (\frac{1}{2})(\frac{1}{2}) + (\frac{1}{2})(-\frac{1}{2}) + (-\frac{1}{2})(\frac{1}{2}) + (-\frac{1}{2})(-\frac{1}{2}) = 0$$

Podemos demostrar en general que si hay k categorías, resultarán cuando más $k - 1$ comparaciones mutuamente ortogonales. Asimismo, si los tamaños de las muestras son distintos, resultará necesario ponderar con los tamaños N_j de la categoría de muestra, siendo el mejor criterio para lograr la ortogonalidad:

$$\sum_{j=1}^k \frac{c_{hj}c_{ij}}{N_j} = 0$$

En nuestro ejemplo hemos utilizado solamente tres comparaciones mutuamente ortogonales, en tanto que $k - 1$, o cuatro, son posibles. En la mayoría de los casos no tendrá por supuesto sentido teórico el utilizar todas las comparaciones ortogonales posibles; sin embargo, es instructivo determinar cuál sería la cuarta. Obsérvese que ya hemos comparado el grupo de control con todos los grupos experimentales, y por ello no es de esperar que una comparación de dicho grupo de control con cualquiera de los subgrupos (por ejemplo el de los grupos autoritarios), resulte ortogonal con la primera comparación. Puede comprobarse esto fácilmente aplicando el criterio de prueba. Obsérvese que hemos comparado el grupo II (junto al III o el IV) con el grupo V (en combinación a su vez con los grupos III y IV). Podríamos así esperar que si los grupos II y V son pareados contra los III y IV, la comparación resultante fuese ortogonal con las restantes comparaciones, como en efecto así ocurre. A menos que de manera específica se buscase una interacción, tal comparación particular carecería probablemente de sentido teórico, ya que requeriría promediar las puntuaciones del grupo autoritario con frustraciones medias con las del grupo democrático con elevada frustración.

Obsérvese que al comprobar la ortogonalidad o independencia mutua entre comparaciones no hemos dicho nada en relación con el tamaño real de la muestra, excepto en el caso de las muestras tamaño N_j . El criterio de prueba implica solamente

las ponderaciones c_{ij} y no las medias de las muestras o variancias. Las decisiones relacionadas con las comparaciones deben ser hechas, en efecto, antes de realizar la recogida de datos. Se puede entonces buscar la significación estadística de cada comparación, como se indica más abajo. Esta prueba incluye la distribución t de manera exactamente análoga a lo que ocurre con la prueba de la diferencia de medias, la que es por supuesto la comparación más simple posible, en la que $c_{11} = 1$, y $c_{12} = -1$. El numerador de t será una estimación de la función lineal ψ_i obtenida sustituyendo las contrapartes de la población con las medias de las muestras. Así, si hacemos:

$$\hat{\psi}_i = c_{i1}\bar{X}_1 + c_{i2}\bar{X}_2 + \dots + c_{ik}\bar{X}_k$$

tendremos el numerador para la i -ésima comparación. En el caso, por ejemplo, de nuestra primera comparación entre el grupo de control y todos los demás, habríamos tenido

$$\hat{\psi}_1 = \bar{X}_1 - (\frac{1}{4})(\bar{X}_2 + \bar{X}_3 + \bar{X}_4 + \bar{X}_5)$$

tal y como el sentido común lo habría sugerido.

Para nuestro denominador de t deseamos usar un estimador resumido basado en todas las muestras, incluso en los casos en que la comparación no abarque la totalidad de dichas muestras. Recordando nuestro resultado para la variancia de una combinación lineal, sabemos que

$$\text{var } \hat{\psi}_i = c_{i1}^2 \text{var } \bar{X}_1 + c_{i2}^2 \text{var } \bar{X}_2 + \dots + c_{ik}^2 \text{var } \bar{X}_k$$

Si suponemos normalidad y variancias iguales $\sigma_i^2 = \sigma^2$ la expresión se convierte en

$$\text{var } \hat{\psi}_i = c_{i1}^2 \frac{\sigma^2}{N_1} + c_{i2}^2 \frac{\sigma^2}{N_2} + \dots + c_{ik}^2 \frac{\sigma^2}{N_k} = \sigma^2 \sum_{j=1}^k \frac{c_{ij}^2}{N_j}$$

la que, al colocar un estimado por σ^2 y obteniendo la raíz cuadrada positiva, se convierte en el denominador deseado de t , el que tendrá $N - k$ grados de libertad. Esta misma expresión había sido usada en el denominador de t en los casos de nuestras comparaciones segunda y tercera, en las que no figura el grupo de control. En el caso de la segunda comparación, por ejemplo, habríamos tenido

$$\hat{\sigma}_2^2 = \frac{N_1 s_1^2 + N_2 s_2^2 + \dots + N_5 s_5^2}{N - 5}$$

(dentro del grupo estimado de variancia)

$$\begin{aligned} \sum_{j=1}^5 \frac{c_{ij}^2}{N_j} &= \frac{0}{N_1} + \frac{(\frac{1}{2})^2}{N_2} + \frac{(\frac{1}{2})^2}{N_3} + \frac{(-\frac{1}{2})^2}{N_4} + \frac{(-\frac{1}{2})^2}{N_5} \\ &= \left(\frac{1}{4}\right) \left(\frac{1}{N_2} + \frac{1}{N_3} + \frac{1}{N_4} + \frac{1}{N_5}\right) \end{aligned}$$

$$\begin{aligned} \text{y por lo tanto } t &= \frac{(\frac{1}{2})(\bar{X}_2 + \bar{X}_3) - (\frac{1}{2})(\bar{X}_4 + \bar{X}_5)}{\hat{\sigma}(\frac{1}{2})\sqrt{1/N_2 + 1/N_3 + 1/N_4 + 1/N_5}} \\ &= \frac{(\bar{X}_2 + \bar{X}_3) - (\bar{X}_4 + \bar{X}_5)}{\hat{\sigma}\sqrt{1/N_2 + 1/N_3 + 1/N_4 + 1/N_5}} \end{aligned}$$

lo que es una extensión evidente de la prueba de la diferencia de medias. Obsérvese que el factor $(\frac{1}{2})$ se cancela en el numerador y en el denominador, lo que refleja el hecho de que las magnitudes absolutas de c_{ij} no importan, en tanto que $\sum c_{ij} = 0$.

Debe recalarse que el estimador resumido $\hat{\sigma}$ será precisamente el basado en la suma interior de los cuadrados (tal como se calcula en la prueba F) y en él estarán incluidas *todas* las categorías, en tanto que el numerador de t y la expresión bajo el radical en el denominador no abarcarán todas las categorías.

XVI.3. Análisis bimodal de la variancia

En determinadas circunstancias resulta posible extender el análisis de la variancia añadiendo otras variables de escala nominal. Semejante procedimiento es posible ante todo en experimentos controlados, en los que el investigador puede asignar individuos a varios grupos al azar, controlando así el número de casos de cada categoría. En las situaciones naturales, sin embargo, en las que no puede efectuarse semejante tipo de control, la extensión que se describe en la presente sección será menos útil. Algunas de las ideas básicas contenidas en lo que se ha denominado análisis de variancia en dos formas ayudarán a comprender algo del material que se presenta en los capítulos XIX y XX.

Si es posible introducir otra variable de escala nominal de tal modo que todas las combinaciones de subcategorías de las dos escalas nominales tengan el mismo número de casos, la extensión del análisis de la variancia es muy sencilla.⁵ Supóngase que

⁵ Si colocamos el mismo número de casos en cada categoría, y si construimos un cuadro de contingencia que relacione a las dos escalas nominales, entonces podremos ver que no hay relación entre ellos en la muestra. Esta falta de relación entre las variables de escala nominal es lo que nos permite separar las sumas de cuadrados de hileras y columnas sin ambigüedad.

las categorías de la segunda escala nominal estén representadas por hileras. Obtenemos ahora cierto número de subcasillas, con el mismo número de casos cada una. Con objeto de cumplir dicha condición, hemos de limitarnos, por supuesto, a poner en columna categorías de la misma magnitud. A los datos numéricos del cuadro XVI.1 añadimos la escala nominal "región", empleando sólo las dos regiones Nordeste y Sudeste. Supongamos que hay el mismo número de ciudades en cada casilla de las seis en total. Si ello no fuera así, habría que recurrir a un método aproximado (véase *infra*). Los datos numéricos se dan ahora en el cuadro XVI.4, con las sumas y las medias de las subcategorías indicadas en cada casilla.

CUADRO XVI.4. Datos para el análisis de variancia en dos formas

Regiones	Tipo de ciudad						Total
	Industrial		Comercial		Gubernamental		
Nordeste	4.3	5.9	5.1	3.6	3.1	3.8	$\sum_j X_{1j} = 44.9$ $\bar{X}_{1.} = 3.74$
	2.8	7.7	1.8	3.3	1.6	1.9	
	$\Sigma X = 20.7$		$\Sigma X = 13.8$		$\Sigma X = 10.4$		
	$\bar{X} = 5.18$		$\bar{X} = 3.45$		$\bar{X} = 2.60$		
Sudeste	12.3	9.1	6.2	4.1	6.2	11.4	$\sum_j X_{2j} = 116.1$ $\bar{X}_{2.} = 9.68$
	16.3	10.2	9.5	11.2	7.1	12.5	
	$\Sigma X = 47.9$		$\Sigma X = 31.0$		$\Sigma X = 37.2$		
	$\bar{X} = 11.98$		$\bar{X} = 7.75$		$\bar{X} = 9.30$		
Total	$\sum_i X_{i1} = 68.6$		$\sum_i X_{i2} = 44.8$		$\sum_i X_{i3} = 47.6$		$\sum_i \sum_j X_{ij} = 161.0$ $\bar{X}_{..} = 6.71$
	$\bar{X}_{.1} = 8.58$		$\bar{X}_{.2} = 5.60$		$\bar{X}_{.3} = 5.95$		

Si hay el mismo número de casos en cada subcasilla, resulta posible fragmentar las sumas de cuadrados del interior de las columnas, o inexplicadas, en diversos componentes. Podemos, por supuesto, efectuar un análisis de variancia a través de las hileras, prescindiendo de las columnas por completo. Las sumas de cuadrados al interior de las hileras y entre las mismas se obtendrían en tal caso exactamente en la misma forma en que se calcularon anteriormente las cifras al interior de las columnas y entre ellas. Desde el punto de vista matemático, resulta que si hay el mismo número de casos en cada subcasilla la suma de cuadrados entre las hileras puede considerarse como procedente por completo de la suma de cuadrados dentro o inexplicadas

cada (por las columnas) de las columnas. Así, pues, la variación total puede dividirse ahora en tres porciones, como sigue:

$$\begin{aligned} \text{SC totales} = & \text{SC dentro de las columnas} + \text{SC entre-hileras} + \\ & + \text{SC inexplicadas} \end{aligned} \quad (\text{XVI.6})$$

Hemos tomado la variación total, explicando todo lo que podíamos por medio de la primera escala nominal (tipo de ciudad). De lo que permanece inexplicado (la suma de cuadrados dentro de la columna), cierta porción puede explicarse mediante la segunda escala nominal (región). En cuanto al remanente, llamado a menudo término de error, constituye la proporción de la variación total dejada sin explicar por ambas variables. Tenemos ahora tres apreciaciones de la variancia común, en adición a la estimación basada en la suma total de los cuadrados, y éstas pueden emplearse para efectuar dos pruebas F distintas. El término de error puede emplearse en los denominadores de ambas pruebas F , ya que la estimación basada en la suma de cuadrados inexplicada será siempre insesgada e independiente de las otras dos. Los numeradores de las F serán las estimaciones basadas en las sumas de cuadrados entre columnas y entre hileras. Cada prueba será una prueba de la existencia de una relación entre la variable de escala de intervalo y una de las variables de escala nominal, controlando la otra escala nominal.

Si bien este tipo de operación de control se examinará con mayor detalle en el capítulo XIX, es menester decir aquí unas palabras al propósito, ya que el control sirviéndose de un análisis de variancia de dos formas comporta un principio algo diferente del que se examinó en conexión con los problemas de contingencia. El lector observará, en efecto, que hasta aquí nuestro procedimiento de control ha consistido literalmente en mantener constante la variable de control y examinar lo que acontece en el interior de cada categoría de la variable de control. Así, por ejemplo, hicimos una serie de pruebas de la χ^2 -cuadrada, una para cada una de dichas categorías. Aquí, en cambio, hacemos una sola prueba F en vez de varias, como se hizo en el caso de la prueba de la chi al cuadrado resumida. En efecto, tomamos su presencia en consideración *ajustando* valores de la escala de intervalo, de acuerdo con la categoría de la variable de control.

El lector observará en el cuadro XVI.4, por ejemplo, que la tasa media de criminalidad es de 3.74 para todas las ciudades del Nordeste, en tanto que la de las ciudades del Sudeste es de 9.68. Supóngase que fuéramos a pretender que todas las ciudades estuvieran en la misma región, y realizando un ajuste estadístico de los niveles de criminalidad agregando a todas las ciudades del Nordeste una cantidad fija (esto es, 2.97) y sustrayendo la misma cantidad de las ciudades del Sudeste, de modo que

ambas categorías tuvieran la misma media (o sea la media general de 6.71). Semejante operación de control equivale a plantear la cuestión hipotética de cuáles serían las tasas de criminalidad si todas ellas estuvieran expuestas a las mismas influencias regionales. En lugar de tratar realmente las regiones separadamente, nos servimos del expediente auxiliar consistente en ajustar las marcas de la tasa de criminalidad, tomando con ello en consideración la variable de control en cuestión. Lo que perdemos en rigor científico lo ganamos en eficiencia del esquema, ya que podemos servirnos así de una sola prueba que comporta el número total de los casos.

Al ajustar las tasas de criminalidad en esta forma, reducimos en realidad la variación total de las marcas. En efecto, sustraemos la porción de la variación debida a la región. Tomando las *marcas ajustadas*, podríamos comparar a continuación las estimaciones entre las columnas y dentro de las mismas, en la forma habitual. Afortunadamente, no es necesario, en realidad, obtener las marcas ajustadas. Si lo hiciéramos, en efecto, encontraríamos que los resultados serían idénticos a aquellos hallados sirviéndonos del análisis de variancia de dos formas. En otros términos: el tipo de análisis que vamos a describir equivale a la operación de ajuste que acabamos de examinar. En efecto, lo que hacemos es, primero, dejar que la variable de control actúe sobre la variable dependiente, sacando la porción de la variación total explicada por la variable de control en cuestión. Tomamos luego el remanente como otra variación "total nueva" y determinamos cuánto de este remanente puede explicarse por la otra variable independiente. Este "nuevo total" es equivalente a la variación total de las *marcas ajustadas*. En términos generales, podemos controlar variables adicionales en la misma forma. Al practicar ajustes para cada una de las variables de control, extraemos todo aquello de la variación que puede explicarse por dichas variables. Y examinamos luego el remanente, para ver cuánto puede explicarse por la otra variable independiente. En los capítulos siguientes haremos un uso considerable de este mismo tipo de operación de control.

Interacción. No estamos todavía preparados para un ejemplo numérico, ya que mediante la adición de una segunda escala nominal se introduce una complicación más. Siempre que haya por lo menos dos casos en cada subcasilla, debería hacerse una prueba adicional. Esto constituye una prueba de "interacción", o del efecto posible debido a las combinaciones peculiares de las dos variables de escala nominal. Con objeto de efectuar la prueba del análisis de variancia en dos formas anteriormente descritas, es necesario suponer la propiedad de aditividad. Enunciada formalmente, esta propiedad requiere que las diferencias medias de población entre columnas sean las mismas para cada hilera,

así como, inversamente, que las diferencias entre hileras sean las mismas para cada columna. La aditividad puede ilustrarse mediante las siguientes cifras que representan medias hipotéticas de población:

	A_1	A_2	A_3
B_1	5	10	20
B_2	10	15	25
B_3	25	30	40

Obsérvese que las diferencias entre la primera y la segunda columnas son de 5 para cada hilera; entre la segunda y la tercera, las diferencias son de 10 para cada hilera. Y asimismo, las diferencias entre la primera y la segunda hileras son de 5 todas ellas, en tanto que entre la segunda y la tercera hileras son todas de 15. Supóngase, sin embargo, que la media de la casilla central fuera 35 en lugar de 15. Entonces la aditividad no se verificaría. Pese a que normalmente A_3 da mayores marcas que A_2 , y B_3 mayores que B_2 , ocurre algo peculiar cuando A_2 y B_2 se ponen juntas, en cuanto resulta una media muy alta. El proceso es algo parecido al que tiene lugar cuando se combinan hidrógeno y oxígeno y se produce agua. El resultado no es lo que podría esperarse si cada elemento se examinara separadamente.

Hemos encontrado ya esta posibilidad de interacción en el caso de los cuadros de contingencia, cuando vimos que la relación entre dos variables puede diferir de acuerdo con el nivel de una tercera variable. Ilustremos la idea con algunos ejemplos. Supóngase que por lo regular las ciudades industriales tengan tasas de criminalidad más altas que los centros políticos, y que las ciudades del Sudeste las tengan más altas que las del Nordeste. Se concibe, en tal caso, que podríamos hallar ciudades industriales en el Sudeste que presentaran una tasa media de criminalidad inesperadamente baja. Podríamos entonces buscar alguna clase de interacción tal, entre la industria y los factores regionales, que produjera una tasa baja. Otro tipo de ejemplo es tal vez más ilustrativo todavía. Supóngase que se tenga que elegir entre tres tipos de métodos pedagógicos. Se invita a cuatro maestros a que empleen los tres métodos. Es posible que en términos generales el maestro A sea más competente que el B . Y en forma análoga, el primer método puede ser, acaso en conjunto, superior al segundo. Pero se concibe que el maestro A no se adapte bien al primer método y tenga resultados inferiores a los esperados. Así, hay interacción entre el maestro y el método.

* Antes de pasar al cálculo de las distintas cantidades resultará instructivo trazar un modelo lineal general, que resultará ser análogo a los modelos formulados en relación con el análisis de la regresión. En él expresamos una variable de escala de inter-

valos como función de otras diversas variables que pueden ser tomadas, bien sea como escalas de intervalos, o como atributos. Supongamos que la puntuación del individuo k -ésimo en la fila i -ésima y columna j -ésima se representa por X_{ijk} , aceptando que dicha puntuación esté compuesta por los siguientes integrantes: 1) uno "debido a" la media general de población, μ ; 2) otro debido a los efectos que son consecuencia de aparecer en una determinada fila i , a los que denominaremos efecto de fila, α_i ; 3) un efecto similar β_j debido a encontrarse en la columna j ; 4) un efecto de interacción γ_{ij} debido a la combinación peculiar de la fila i -ésima y la columna j -ésima, y 5) un efecto único, o término de error, ε_{ijk} producido por factores no considerados de manera explícita en la ecuación. Ésta resultaría así:

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

la que por supuesto se refiere a los parámetros de población que han de ser estimados con base en los datos de la muestra. Resulta que si todos los supuestos requeridos en el caso de un análisis de la variancia por dos métodos se dan reunidos (véase más adelante), podemos obtener estimadores no sesgados de los parámetros de la anterior ecuación, como sigue:

$$\begin{aligned}\hat{\mu} &= \bar{X}_{..} & \hat{\gamma}_{ij} &= \bar{X}_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..} \\ \hat{\alpha}_i &= \bar{X}_{i.} - \bar{X}_{..} & &= \bar{X}_{ij} - (\hat{\alpha}_i + \hat{\beta}_j + \hat{\mu}) \\ \hat{\beta}_j &= \bar{X}_{.j} - \bar{X}_{..} & \hat{\varepsilon}_{ijk} &= X_{ijk} - \bar{X}_{ij}\end{aligned}$$

* Cada una de estas estimaciones tiene un sentido intuitivo, salvo, tal vez, la del efecto de interacción γ_{ij} . Utilizamos la gran media de la muestra $\bar{X}_{..}$ para estimar μ y las desviaciones entre $\bar{X}_{..}$ y las medias de fila y columna, para calcular los efectos de fila y de columna, α_i y β_j , respectivamente. La desviación de X_{ijk} en relación con la media \bar{X}_{ij} de la muestra de la subcategoría, representa la variación inexplicada en la muestra, la que estima el término residual comparable ε_{ijk} . La estimación del componente de interacción γ_{ij} podrá entonces ser obtenido por sustracción. Hemos expresado en efecto cada individuo X_{ijk} en función de los siguientes componentes:

$$\begin{aligned}X_{ijk} &= \bar{X}_{..} + (\bar{X}_{i.} - \bar{X}_{..}) + (\bar{X}_{.j} - \bar{X}_{..}) \\ &+ (\bar{X}_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..}) + (X_{ijk} - \bar{X}_{ij}) \\ &+ (\text{efecto de interacción}) + (\text{término de error})\end{aligned}$$

Por ejemplo, en el caso de la segunda ciudad política en el Nordeste tendríamos:

$$\begin{aligned}1.60 &= 6.71 + (3.74 - 6.71) + (5.95 - 6.71) \\ &+ (2.60 - 3.74 - 5.95 + 6.71) + (1.60 - 2.60)\end{aligned}$$

* El procedimiento básico tanto en el caso de este modelo, como en forma más generalizada, consiste en realizar pruebas separadas para cada uno de los efectos componentes α_i , β_j y γ_{ij} , evaluando la contribución de cada uno de ellos en relación con el tamaño del término de error. Como por otra parte siempre es deseable utilizar un modelo tan sencillo como resulte posible, comenzaremos observando si tiene sentido la eliminación del componente de interacción γ_{ij} . Volvamos ahora al procedimiento que utilizaremos para el cálculo.

La prueba de la interacción puede efectuarse independientemente de las dos pruebas descritas anteriormente y comporta el mismo procedimiento básico que ellas. La suma de cuadrados inexplicada, o término de error, se descompone más todavía, restándole la porción que puede explicarse por la interacción. En esta forma, la suma total de cuadrados se descompone en:

$$\begin{aligned}\text{SC total} &= \text{SC entre columnas} + \text{SC entre hileras} \\ &+ \text{SC de interacción} + \text{SC de error} \quad (\text{XVI.7})\end{aligned}$$

Esto puede efectuarse tomando cada combinación de las categorías A y B y tratándola como categoría de una variable sola combinada. En otros términos, tratamos el problema como si tuviéramos una sola escala nominal con las categorías, $A_1B_1, A_2B_1, \dots, A_kB_l$. Es obvio que si sólo hubiera un caso en cada subcasilla no podría haber variación alguna de subclase. Si no existe interacción en absoluto, deberíamos obtener exactamente el mismo error obtenido adicionando separadamente los efectos de las hileras y las columnas [como en la ecuación (XVI.6)]. Por otra parte, si se da una interacción significativa, el término de error será menor empleando este segundo método. Así, por ejemplo, el lector debería convencerse por sí mismo de que, si la casilla ij produjera efectos en discrepancia con las demás, dicha casilla será relativamente homogénea en comparación ya sea con la columna j o con la hilera i , y la suma dentro de cuadrados de las subclases será menor que el residuo obtenido restando la suma de las sumas de cuadrados entre columnas y entre hileras de la suma total de cuadrados.

La diferencia entre la cantidad de variación explicada sirviéndose de esas subcasillas y la cantidad explicada en el supuesto

de aditividad puede entonces atribuirse a interacción. Así tenemos:

$$SC \text{ total} = SC \text{ entre subclases} + SC \text{ dentro de las subclases}$$

en donde la suma de cuadrados entre subclases se ha descompuesto en tres componentes, a saber:

$$SC \text{ entre subclases} = SC \text{ entre columnas} + SC \text{ entre hileras} + SC \text{ de interacción}$$

Cálculos. Volviendo ahora una vez más al problema numérico que comporta tipos de ciudades, región y tasas de criminalidad, podemos empezar enumerando los supuestos requeridos.

1. Supuestos

Nivel de medición: Dos escalas nominales, una escala de intervalo;

Modelo: Muestras independientes aleatorias;

Todas las poblaciones de las subcasillas, las hileras y las columnas son normales;

Las variancias de las poblaciones de las subcasillas son iguales.

Hipótesis: 1. Las medias de las columnas de la población son iguales.

2. Las medias de las hileras de la población son iguales.

3. Adicionalidad de la población (sin interacción).

Tenemos ahora tres hipótesis distintas que pueden verificarse independientemente. La prueba de interacción ha de efectuarse primero, dependiendo las pruebas de las demás de aquélla. Si la hipótesis (3) no se rechaza, el procedimiento usual consiste en suponer aditividad en el modelo, poniendo las sumas de cuadrados debidas a interacción (en la muestra) en el término de error y sirviéndose de este término de error mayor para la verificación de las hipótesis (1) y (2). Pero si la hipótesis de falta de interacción se rechaza, entonces el procedimiento a emplear en las otras dos pruebas dependerá del carácter de los datos (véase *infra*). Obsérvese que, con objeto de verificar la interacción, hemos de suponer ahora normalidad e igualdad de variancias para cada una de las subcasillas. Los casos en las distintas subcasillas han de seleccionarse independientemente y no pueden aparecerse.

2. Nivel de significado. .05.

3. Distribución de muestreo. F.

4. **Cálculo de la estadística de la prueba.** Hemos obtenido ya las sumas de cuadrados total y entre columnas. La suma de cuadrados entre hileras se calcula exactamente del mismo modo que la de entre columnas. Así, pues:

$$SC \text{ entre hileras} = \frac{44.9^2}{12} + \frac{116.1^2}{12} - 1\,080.042$$

$$= 1\,291.268 - 1\,080.042 = 211.226$$

Con objeto de obtener la suma de cuadrados de interacción, nos servimos de las sumas de cada subclase. La suma de cuadrados entre subclases es:

$$SC \text{ entre subclases} = \frac{20.7^2}{4} + \frac{47.9^2}{4} + \dots + \frac{37.2^2}{4} - 1\,080.042$$

$$= 1\,341.585 - 1\,080.042 = 261.543$$

Obtenemos el término de error empleado en la verificación de la interacción restando la suma de cuadrados entre subclases del total. O sea:

$$SC \text{ del error} = 373.538 - 261.543 = 111.995$$

La cantidad debida efectivamente a interacción es la suma de cuadrados entre columnas menos la suma de las cantidades debidas a las hileras y las columnas separadamente. Por lo tanto:

$$SC \text{ de la interacción} = 261.543 - (42.303 + 211.226) = 8.014$$

Los resultados pueden resumirse como en el cuadro XVI.5.

CUADRO XVI.5. *Cálculos para el análisis de variancia, de dos formas con prueba de interacción*

	Sumas de cuadrados	Grados de libertad	Estimación de la variancia	F
Total	373.538	$N - 1 = 23$		
Entre subclases	261.543	$kl - 1 = 5$		
Entre columnas	42.303	$k - 1 = 2$	21.152	
Entre hileras	211.226	$l - 1 = 1$	211.226	
Interacción	8.014	$(k - 1)(l - 1) = 2$	4.007	
Error (dentro de las subclases)	111.995	$N - kl = 18$	6.222	0.644

Los grados de libertad se determinan por los medios usuales. Con l hileras y k columnas habrá $l - 1$ grados de libertad asociados con la suma de cuadrados entre hileras. Para obtener los grados de libertad del término de interacción, tomamos el número de subcasillas menos uno ($kl - 1$) y restamos de esta cantidad los grados de libertad asociados a las sumas de cuadrados entre hileras ($l - 1$) y entre columnas ($k - 1$). Una regla práctica más sencilla consiste en tomar el *producto* de los grados de libertad asociados a las sumas de cuadrados entre columnas y entre hileras. Así, pues, si multiplicamos los grados de libertad entre columnas y entre hileras obtenemos $(k - 1)(l - 1) = 2$ grados de libertad. Éste es el mismo resultado que obtendríamos tomando los grados de libertad entre subclases ($= 5$) y restando de ellos los grados de libertad de las sumas de cuadrados entre hileras y entre columnas ($= 1 + 2$). Esto puede expresarse algebraicamente con la siguiente identidad:

$$(kl - 1) - (k - 1 + l - 1) = (k - 1)(l - 1)$$

Los restantes grados de libertad, que deberían ser iguales al número total de casos menos 1 grado de libertad para cada subclase, pueden luego asociarse al término de error.

5. *Decisión.* La prueba de interacción da una F que es menor que la unidad. No tenemos, pues, motivo para rechazar la hipótesis nula de que no se da interacción. Esto significa que la pequeña cantidad adicional explicada por interacción al interior de estas *muestras* puede explicarse fácilmente por las fluctuaciones de la selección. En este caso propenderíamos probablemente a aceptar el supuesto de aditividad, pese a que nos encontramos en el extremo indebido de la prueba y que, en consecuencia, deberíamos preocuparnos en primer término por el riesgo de error de tipo II. Añadimos de paso que si hubiéramos dispuesto de cuadros, podríamos haber utilizado un nivel de significación de .3, por ejemplo, si realmente hubiéramos tenido interés en conservar el supuesto de aditividad. Habiendo decidido que no existe interacción, podemos ahora poner la suma de cuadrados debida a la interacción (de la muestra) junto con el término de error, y servirnos de este término de error mayor como base para la estimación del error de la variancia. Al hacerlo obtenemos el cuadro XVI.6, en el que el término de error de 120.009 representa la suma de los términos de interacción y de error del cuadro XVI.5.

Del cuadro correspondiente se desprende que para una F con 2 y 20 grados de libertad necesitamos una F de 3.49 o mayor para obtener significación al nivel de .05. Vemos asimismo que una F de 35.204 con 1 y 20 grados de libertad es altamente significativa, ya que para obtener significación al nivel de .001 se requeriría una F de sólo 14.82. Así, pues, existe poca duda de que se

CUADRO XVI.6. Cálculos para el análisis de variancia en dos formas, con la interacción añadida dentro del término de error

	Sumas de cuadrados	Grados de libertad	Estimación de la variancia	F	Nivel de significación
Total	373.538	23			
Entre columnas	42.303	2	21.152	3.525	$p < .05$
Entre hileras	211.226	1	211.226	35.204	$p < .001$
Error	120.009	20	6.000		

da una relación entre la región y la tasa de criminalidad. Obsérvese que cuando controlamos en relación con la región dejando que esta variable explique todo lo que puede acerca de las tasas de criminalidad, y dejando luego que el tipo de ciudad explique todo lo que puede a propósito del resto, obtenemos una relación significativa entre el tipo de la ciudad y las tasas de criminalidad. Se recordará que la relación sin el control relativo a la región no daba significación.

Cabe observar que si la interacción no es significativa, ganamos casi siempre más de lo que perdemos al adjuntar la interacción con el término de error, sirviéndonos de este término de error combinado en el denominador de F . En efecto, pese a que la suma de los cuadrados de error resultará en esta forma ligeramente aumentada, habrá también más grados de libertad asociados al término mayor del error. Y como quiera que el término de interacción será relativamente pequeño, el efecto neto será por lo regular el de obtener un denominador de F más pequeño. Habrá también, por supuesto, un mayor número de grados de libertad asociados con F y, por lo tanto, se requerirá para obtener significación un valor más pequeño de F .

Hemos de preguntar ahora qué habríamos hecho si la interacción hubiera sido significativa. La respuesta a dicha cuestión no tiene nada de sencillo, pero podemos, con todo, formular algunas cuantas sugerencias. El lector que se interese por un tratamiento más completo deberá consultar un texto como el de Hays [7], Kirk [9], o el Anderson y Bancroft [1].

Si la interacción es significativa, a veces será posible hallar una o dos filas o columnas, o aun unas cuantas subcasillas, que son las que producen la interacción. Si hubiésemos utilizado, por ejemplo, cinco regiones, podríamos haber observado que el Sudeste difiere, de manera fundamental, del resto de las regiones. De ser así podríamos haber excluido del análisis dicha región en especial, averiguando si había interacción entre las categorías restantes, aunque reconociendo la naturaleza *ex post facto* de tal procedimiento. En otras muchas aplicaciones no será tan sencillo localizar las filas, columnas o casillas aisladas responsables de

la interacción, en cuyo caso nos enfrentamos a un desafío teórico cuando deseamos formular una explicación general razonable del esquema logrado. En realidad, el localizar un efecto mayor de interacción puede resultar ser el hallazgo aislado de mayor importancia en el estudio. Aunque la conceptualización de modelos matemáticos en los que se halla implicada la interacción escapa al propósito de este libro, deberá observarse que cabe formular ciertas alternativas relativamente simples a los modelos aditivos lineales, tales como los modelos multiplicativos. (Véase Blalock [2]).

Además de centrar la atención en la interacción misma, puede tenerse interés en determinar si una u otra de las variables de la escala nominal se relaciona o no con la escala de intervalo. ¿Cuáles pruebas pueden efectuarse de tales relaciones? La cuestión se reduce a lo siguiente: "¿Qué estimación de la variancia debería emplearse en el denominador de F , en la estimación del error o en la estimación basada en el término de interacción?" La respuesta a esta pregunta parece depender de la naturaleza de las dos variables de la escala nominal y, en particular, de si las categorías empleadas representan todas las categorías de la población o no son más, por el contrario, que una mera selección de categorías. En los problemas sociales, en los que por lo regular no asignamos los individuos a las categorías al azar, suele darse con frecuencia el caso de que estas categorías representen todas las categorías posibles del esquema de clasificación. Así, por ejemplo, si dividimos todas las ciudades en tres tipos y no excluimos ninguna al proceder en esta forma, confiamos incluir algunas ciudades, por lo menos, de cada tipo. Y en forma análoga, si clasificamos a personas como varones o mujeres, o como protestantes, católicos o judíos, confiamos por lo regular haber incluido algunos representantes de todas (o casi todas) las categorías. Por otra parte, nuestras categorías podrían comportar ellas mismas una selección de todos los tipos. Así, por ejemplo, podríamos haber seleccionado a metodistas, cuáqueros y a testigos de Jehová como tres grupos religiosos que representan un número mucho mayor de ellos. Tal vez cada una de dichas denominaciones sea representativa de cierto tipo de religión. Examinemos cada una de estas situaciones por turno.

En la primera de ellas, nuestras categorías de *ambas* variables representan todos o casi todos los tipos posibles. No se da ciertamente error alguno en la selección de las categorías, como podría haberlo si sólo nos hubiéramos servido a título de comparación de tres denominaciones religiosas. En la mayoría de estos problemas, nuestro interés se centrará probablemente en el grado de homogeneidad de cada tipo, en relación con la magnitud de las diferencias entre los tipos. La segunda variable de la escala nominal puede considerarse en primer lugar como una

variable perturbadora que necesita controlarse. La interacción puede acaso constituir simplemente un resultado secundario interesante del análisis. En este caso será razonable comparar una estimación basada en la suma de cuadrados entre con la apreciación basada en la suma de cuadrados no explicada. Esta última estimación es una estimación dentro de las subclases y comporta la variación que permanece todavía inexplicada por la acción conjunta de la variable independiente mayor (digamos el tipo de la ciudad) y la variable de control. Dejamos que la variable de control actúe primero, y dejamos luego que la variable independiente mayor explique lo que puede del resto. Cierta cantidad adicional es explicada asimismo por la interacción de las dos variables. Cada una de estas sumas de cuadrados "explicadas" puede compararse con la suma de cuadrados "no explicada", o término de error. A continuación tomaríamos esta estimación del error como denominador en cada una de nuestras pruebas separadas de F . Al verificar en relación con la significación de una diferencia entre columnas, tomaríamos, por lo tanto, la estimación de entre columnas dividida por el término de error, y en forma análoga por lo que se refiere a las hileras. En nuestro problema numérico, si la interacción hubiera sido significativa, estas razones de F habrían sido respectivamente de 21.152/6.222 y 211.226/6.222.

Surgen otras consideraciones cuando las categorías de una u otra variable (o de ambas) de la escala nominal sólo comportan una pequeña selección de las categorías posibles. Si la interacción resulta significativa y es mayor, por lo tanto, que la estimación del error, añádase siempre la cuestión de saber si esto no se habría producido de haber sido distintas las categorías. Si *tanto* la variable de fila como la de columna abarcan una muestra de categorías, nos referimos a tal modelo denominándolo modelo de *efectos aleatorios*, por contraste con el modelo de *efectos fijos*, para el cual ninguna de las variables comprende un muestreo de categorías. Personalmente no he visto nunca una ilustración razonable de tal modelo de efectos aleatorios, aunque los modelos *mixtos* en los que figuran uno (o más) factores no muestreados y un factor muestreado, son razonablemente comunes. El más habitual de los modelos comunes en las aplicaciones a la ciencia social se presenta en los casos en que son personas (educadores, experimentadores, entrevistadores, operadores de equipo, etcétera) las que figuran como uno de los factores. En los experimentos en las aulas, por ejemplo, puede ser necesario considerar el "efecto del maestro" entre un cómputo de tal vez cinco educadores. En un laboratorio podrá ocurrir que el investigador haya contado con tres experimentadores. Aunque instruidas para conducirse de manera análoga, tales personas introducen inevitablemente en la situación algunos valores idiosincrásicos.

En una investigación puede el analista necesitar separar los "efectos del entrevistador" de entre las demás variables. En todos estos ejemplos se reconocerá que las personas que en ellos figuran constituyen una fracción muy reducida del número potencial en relación con el cual desea hacerse la generalización, y que la interacción entre las personas y el factor de mayor interés puede resultar especialmente perturbadora.

Estas ideas intuitivas pueden ser objeto de una fundamentación más rigurosa (véase Hays [7], capítulo XIII). Bastará indicar aquí el procedimiento preferido. Supongamos en primer lugar que tenemos interés en comprobar los efectos del factor *no muestreado* o fijo. Si la interacción ha sido significativa, ello implica por supuesto que el cálculo de la variancia, basado en el término de la interacción, debe haber sido mayor que la estimación del "error" (produciendo así una $F > 1.0$). Como se da la circunstancia de que el segundo factor ha sido muestreado, y que un segundo muestreo podría haber producido una estimación muy diferente de la interacción, el procedimiento más conservador consistiría en utilizar la interacción estimada (la mayor de las dos cantidades) como denominador para la razón de F en la prueba de la significancia del factor fijo o no muestreado. En efecto: la interacción es considerada como un error. En nuestro ejemplo numérico supongamos que considerásemos la región como un factor muestreado, ya que hemos seleccionado tan sólo dos regiones de entre tal vez cinco o seis. Si el efecto de interacción hubiese sido significativo y por tanto no incluido en el término de error, habríamos utilizado la razón $21.152/4.007$ al comprobar la significancia de los efectos de la ciudad en los niveles de delincuencia.

Si estamos además interesados en probar los efectos del factor muestreado (por ejemplo: personas o región), deberemos sin embargo continuar usando la estimación del error, de preferencia a la estimación de la interacción, en el denominador de F . La justificación intuitiva consiste en que el otro factor *no* está siendo muestreado, y por ello no puede ocurrir que un error de muestreo en dicho factor constituya una fuente de error en nuestro cálculo de los efectos del factor muestreado sobre la variable dependiente. Así, si la interacción hubiese resultado significativa en nuestro ejemplo, habríamos utilizado la razón $211.226/6.222$ al comprobar los efectos de la región sobre los índices de criminalidad. (El hecho de que el denominador, 6.222, es mayor que el de 4.007 usado en relación con los efectos de tipo ciudad, refleja el hecho de que la F , usada en este ejemplo para comprobar la interacción, resultó ser menor que la unidad, en tanto que una interacción significativa hubiera requerido desde luego una F mayor que la unidad.) Para una justificación más amplia de este procedimiento véase Hays [7].

Resulta necesario adoptar una precaución más con respecto a la interpretación de interacciones significativas. En la bibliografía estadística se encuentran frecuentes referencias a los "efectos principales" de las variables de fila o columna, más los "efectos de interacción". Resulta posible interpretar estos efectos principales como los efectos promedios de una de las variables independientes sobre el margen de la otra u otras variables. Pero si el componente de interacción es relativamente grande, esta simple distinción entre efectos principales y efectos de interacción resultará difícil de traducir a valores sustantivos o teóricos, ya que cuando la interacción es grande no tiene sentido teórico el oscurecer las diferencias reales hablando de los efectos promedios de, por ejemplo, el tipo de ciudad. Debe, pues, entenderse que esta distinción entre efectos principales y efectos de interacción se limita al uso estadístico, lo mismo que ocurre con la relativa a las sumas, "explicadas" y "no explicadas", de cuadrados.

Es fácil a veces caer en la trampa de utilizar la terminología de una sustantiva y personal disciplina en lugar de la terminología estadística, y creer que hay distintos tipos de "efectos" que cuentan con una simple contrapartida en la propia sustantiva teoría. Tal vez la precaución más oportuna consista en comprender que en cuantas ocasiones se encuentran interacciones estadísticas de magnitud sustancial, ello significa que dos o más variables tienen efecto conjunto sobre alguna variable dependiente; efectos demasiado complejos para ser adecuadamente descritos mediante un simple modelo aditivo. La presencia de la interacción estadística constituye así una indicación de que las relaciones son más complejas de lo que pudo pensarse, pero la interacción por sí misma no debe ser tratada como si fuese algo aparte de los efectos "principales" de las variables que están siendo consideradas.

Extensión a tres o más escalas nominales. En teoría nada hay que nos impida extender el análisis de variancia a variables adicionales. En la práctica, sin embargo, es probable que nos veamos restringidos por el requisito de números iguales de casos en cada subcasilla, a menos que estemos en condiciones de controlar este factor por vía experimental. Si añadimos una tercera escala nominal, podemos dividir la suma total de cuadrados en interacción entre A , entre B , entre C y los términos de error, y podemos efectuar cierto número de pruebas de hipótesis separadas. Ahora, sin embargo, tendremos más de un tipo de interacción. En efecto, puede darse interacción entre las variables A y B , A y C , B y C , así como entre las tres variables operando juntas. Procedemos primero a una prueba en relación con la interacción de tres factores ($A \times B \times C$). Si ésta no resulta significativa, podemos tomarla en el término de error y verificar las tres interacciones de dos factores. Pueden efectuarse pruebas de la significancia de

A, B y C. La extensión a cuatro o más escalas nominales tendría lugar en la misma forma. En el caso de que el investigador esté en condiciones de controlar el número de casos de cada categoría mediante asignación al azar, se dispone de muchos otros esquemas experimentales, y el lector hará bien en consultar un texto de éstos. Muchos de estos diseños alternativos hacen posible el logro de una mayor eficiencia (mediante una reducción del tamaño de la muestra), al costo de una simplificación de supuestos acerca de algunos de los términos de la interacción. Si uno está dispuesto, por ejemplo, a suponer que ciertas interacciones son despreciables, puede "confundir" deliberadamente estos efectos principales al trazar un diseño "incompleto" más eficiente.

* *Análisis de variancia de dos formas con subclases desiguales.* Cuando el número de casos no es igual en cada subclase, como ocurrirá por lo regular en la investigación sociológica, el análisis de la variancia de dos formas ya no resulta tan sencillo. Si el número de casos es suficientemente grande, será siempre posible, por supuesto, controlar en relación con una segunda escala nominal efectuando análisis separados en cada categoría de la variable de control, como lo hicimos en el caso de los problemas de contingencia. Pero si para empezar, el número de casos es relativamente pequeño, pueden emplearse algunos métodos aproximados. Uno de éstos comporta el empleo de los logaritmos, pero es sencillo por lo demás (véanse [8], pp. 260 a 266).

Otro procedimiento, descrito por Walker y Lev [11], es mucho más sencillo desde el punto de vista conceptual. Este último método consiste en tratar las medias de las distintas subcasillas como si constituyeran casos simples. Pueden obtenerse las sumas de cuadrados y las apreciaciones de variancia de los términos entre hileras, entre columnas y de interacción, suponiendo esencialmente que no hay más que un caso en cada subcasilla: la media. La suma de los cuadrados del error se obtiene luego, al igual que en el análisis corriente de variancia de dos formas, restando la suma de cuadrados de la subclase "entre" de la suma de cuadrados total, sirviéndonos para ello del número total de casos, y no de las medias de cada subcasilla. La apreciación del error se obtiene dividiendo entre el error en el número de grados de libertad, como antes, y dividiendo entonces esta última cifra entre la media armónica del número de casos en cada subcasilla. Esta última operación es necesaria para que la estimación del error, basada en el número total de los casos, pueda compararse con las estimaciones basadas únicamente en las medias de las subcasillas tratadas como casos singulares. Las pruebas *F* pueden luego efectuarse en la forma habitual.

Si las subclases contienen un número desproporcionado de casos, como ocurre habitualmente en la investigación no experi-

mental, tal cosa significa que las variables de fila y de columna estarán interrelacionadas. En efecto, alguna variación que viene "explicada" por la variable de la columna puede asimismo ser "explicada" por la variable de la fila, dándose casos de ambigüedad acerca de a cuál de las variables habrá de darse el crédito por una variancia, cuando ésta es explicable de dos maneras. Hallaremos esta misma dificultad en relación con el análisis de regresión múltiple, y, de manera implícita, en el análisis de la covariancia.

Después de estudiar tanto la regresión múltiple como el análisis de la covariancia, examinaremos brevemente (en el capítulo xx), lo que se denomina "variable simulada", utilizable para manejar un gran número de situaciones, entre ellas aquella en la que se tienen dos (o más) escalas nominales variables independientes e interrelacionadas. Veremos sin embargo que este muy amplio procedimiento estadístico no nos permite superar las dificultades *teóricas* que surgen en aquellos casos en que las variables independientes están interrelacionadas. Tales problemas sólo pueden ser resueltos por medio de procedimientos de cálculo basados en ecuaciones simultáneas, cuestión que excede los límites de este libro. Debe observarse que una de las grandes ventajas de los diseños experimentales consiste en que éstos permiten la manipulación de variables independientes, de tal manera que sus efectos pueden ser separados sin ambigüedad, lo que hace posible evaluar los efectos principales de cada variable, siempre a condición de que la interacción no sea demasiado notable.

XVI.4. Alternativas no paramétricas del análisis de variancia

En el caso en que los supuestos requeridos para el análisis de variancia no se cumplan, se dispone de pruebas no paramétricas que pueden utilizarse como análisis de variancia de una o de dos formas. Examinaremos en primer lugar el análisis de un procedimiento de la variancia con categorías de Kruskal-Wallis, para ver a continuación la prueba de Friedman para muestras pareadas, utilizable en aquellas ocasiones en que la variable de fila constituye un grupo de variables pareadas y en las que hay un "caso" en cada fila.

Prueba de Kruskal-Wallis. La prueba tratada en esta sección fue desarrollada por Kruskal y Wallis y resulta indicada siempre que tengamos cierto número de muestras al azar independientes y un nivel de medición de escala ordinal. La eficacia de su fuerza es aproximadamente, en las muestras grandes, del 95 por ciento. La prueba es básicamente muy sencilla y comporta la comparación de las sumas de los órdenes de cada una de las categorías

de la variable de la escala nominal. Se calcula una estadística H con objeto de medir el grado en que las distintas sumas de órdenes difieren de aquello que se esperaría bajo la hipótesis cero. Si hay más de cinco casos en cada clase, la distribución de selección de H es aproximadamente la χ -cuadrada.⁶

Con fines de comparación, ilustremos el empleo de la prueba de Kruskal-Wallis con los mismos datos. En el cuadro XVI.7, las tasas de criminalidad de los tres tipos de ciudades se han ordenado de altas a bajas (los órdenes bajos indican tasas bajas).

CUADRO XVI.7. Datos y cálculos para el análisis de variancia con rangos de Kruskal-Wallis

Ciudad industrial		Ciudad comercial		Ciudad política	
Cuota	Orden	Cuota	Orden	Cuota	Orden
4.3	10	5.1	11	3.1	5
2.8	4	1.8	2	1.6	1
5.9	12	3.6	7	3.8	8
7.7	16	3.3	6	1.9	3
12.3	22	6.2	13.5	6.2	13.5
16.3	24	9.5	18	7.1	15
9.1	17	4.1	9	11.4	21
10.2	19	11.2	20	12.5	23
Sumas	$R_1 = 124$	$R_2 = 86.5$		$R_3 = 89.5$	

1. Supuestos.

Nivel de medición: escalas ordinal y nominal

Modelo: muestreo al azar independiente

Hipótesis: muestras sacadas de la misma población continua

2. Nivel de significación y región crítica. Tomemos el nivel del .05.

3. Distribución de muestreo. La distribución de muestreo de H será aproximadamente la χ -cuadrada con $k - 1$ grados de libertad, en donde k representa el número de categorías empleadas.

4. Cálculo de la estadística de la prueba. Calculamos H por medio de la fórmula

$$H = \frac{\left(\frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{N_i} \right) - 3(N+1)}{1 - \sum T_i / (N^3 - N)} \quad (\text{XVI.8})$$

en donde N_i y N representan respectivamente el número de ca-

⁶ En caso de tres clases y N muy pequeñas, véase [10], pp. 195-198.

ses de la i -ésima categoría y de la muestra total. El denominador de la fórmula representa una corrección por ligaduras, en la que

$$T_i = t_i^3 - t_i$$

siendo t_i el número de observaciones ligadas en relación con un rango determinado.

En este ejemplo particular hay sólo un par de marcas empata-das. Por lo tanto: $T_i = 2^3 - 2 = 6$. Tenemos, pues:

$$H = \frac{[12/24(25)](124^2/8 + 86.5^2/8 + 89.5^2/8) - 3(25)}{1 - 6/(24^3 - 24)} = 2.17$$

5. Decisión. Refiriéndonos a la tabla de la χ -cuadrada, vemos que con 2 grados de libertad necesitamos una χ -cuadrada de 5.991 o mayor para obtener significación al nivel del .05. Habiendo, pues, obtenido una H de 2.17, decidimos no rechazar la hipótesis nula a dicho nivel de significación. Vemos, de paso, que llegamos a la misma conclusión que anteriormente.

Análisis de Friedman de dos métodos con categorías. Los datos ordinales no permiten en apariencia manejar el concepto de interacción, a no ser de manera muy burda y poco satisfactoria. Sin embargo, si se está dispuesto a suponer que la interacción carece de importancia, y se desea controlar para una o más variables utilizando lo que equivale a un procedimiento de pareado, puede procederse como sigue. Se emparejan los individuos (en este caso las ciudades) de acuerdo con el criterio que se desee aplicar. Uno de tales criterios puede ser el regional, otro el tamaño, un tercero la edad de las ciudades, etcétera. Se asigna a continuación un miembro de cada grupo a una situación experimental, tomando como número de "casos" el número de grupos de individuos pareados. Está bien claro que este procedimiento representa una ampliación de las pruebas mediante pares normalizados, ya estudiadas. En algunos casos pueden darse observaciones repetidas sobre cada individuo; en otros puede no haber resultado posible aplicar realmente el azar a la asignación a los grupos de tratamiento o experimentales, en cuyo caso nuestras interpretaciones deberán ser mucho más cautelosas. En el caso que estudiamos resultará evidentemente imposible distribuir al azar las ciudades en las categorías industrial, comercial o política.

Consideremos a continuación cada grupo de individuos pareados como una respuesta independiente. Dentro de cada uno de los grupos asignamos categorías 1, 2, 3, ..., k , de acuerdo con las puntuaciones de la variable dependiente. Hacemos lo mismo en cada caso y sumamos las filas, obteniendo una suma de filas T_i ,

para la columna j -ésima. Si la variable experimental (columna) no tiene efecto alguno, podremos esperar que las varias T_j resulten aproximadamente iguales. Estaríamos en efecto asignando las filas dentro de cada sector en forma totalmente al azar, y no esperaríamos que el total de puntuaciones de los sectores en cualquiera de las columnas resultase desusadamente grande o pequeño. Habrá, empero, de ordinario, diferencias menores de muestras entre las T_j , y por ello desearemos obtener una medida de las diferencias entre las T_j que cuenten con una distribución conocida de la muestra.

Si computamos la estadística

$$S = \sum_{j=1}^k (T_j - \bar{T})^2$$

en la que k es el número de categorías y \bar{T} es la media de las T_j , la distribución de la muestra de S puede ser calculada exactamente en el caso de muestras pequeñas, y aproximadamente en el de muestras grandes. En Bradley [3] y Siegel [10] se dan cuadros para las distribuciones exactas. Cuando $k \geq 4$ y $N \geq 10$, en donde N representa el número de grupos de individuos pareados, podemos usar una aproximación de χ^2 al cuadrado, como sigue:

$$\chi^2 = \frac{12S}{Nk(k+1)} = \frac{12}{Nk(k+1)} \sum_{j=1}^k T_j^2 - 3N(k+1)$$

en la que el grado de libertad para χ^2 al cuadrado es $k-1$, y en donde el lado de la extrema derecha resultará más conveniente para fines del cálculo. Suponemos de nuevo una distribución básica continua de puntuaciones verdaderas, de modo que los empates resulten tan sólo de la crudeza de las mediciones. Podemos asignar a las puntuaciones empatadas las medias de las filas que hubieran recibido en caso de no producirse empates o, más conservadoramente, podemos romper los empates, minimizando así el valor obtenido de χ^2 al cuadrado. Procedamos adelante con un ejemplo.

Continuando con la misma ilustración, por conveniencia de la comparabilidad, conservamos el supuesto de que hemos obtenido por lo menos un nivel ordinal de medición para los grados de delincuencia, y la hipótesis nula de que las muestras han sido obtenidas de la misma población continua. Esta hipótesis equivale, en efecto, al aserto de que, en el contexto de un experimento real, la variable experimental no tendría efecto alguno. Suponemos ahora, sin embargo, que las muestras están agrupadas, en este caso en tríos de ciudades, una industrial, otra comercial y otra política. Habrá ocho réplicas, de manera que $k=3$ y $N=8$.

Anteriormente, al hacer un análisis por dos métodos de la variancia, utilizamos sólo dos series, una de ellas para el Sudeste y otra para el Nordeste. Aquí contamos con ocho series, lo que permite emparejamientos individuales donde ello sea posible. Podemos por supuesto asignar arbitrariamente cada ciudad del Sudeste a cualquiera de las cuatro series de más abajo, pero el hacerlo supondría un diseño menos eficiente que el logrado usando controles más refinados en el proceso de pareado. Para ser concretos, supongamos que hemos utilizado cuatro clases de diferente tamaño para cada una de las dos regiones, de modo que las ciudades hayan resultado emparejadas simultáneamente por tamaño y por región. Supongamos que las ciudades han sido dispuestas como en el cuadro XVI.8.

CUADRO XVI.8. Datos y cálculos para la prueba de Friedman

Grupo	Ciudad industrial		Ciudad comercial		Ciudad política	
	Tasa	Rango	Tasa	Rango	Tasa	Rango
A	4.3	2	5.1	3	3.1	1
B	2.8	3	1.8	2	1.6	1
C	5.9	3	3.6	1	3.8	2
D	7.7	3	3.3	2	1.9	1
E	12.3	3	6.2	1.5 (1)	6.2	1.5 (2)
F	16.3	3	9.5	2	7.1	1
G	9.1	2	4.1	1	11.4	3
H	10.2	1	11.2	2	12.5	3
T_j		20		14.5 (14)		13.5 (14) $\bar{T} = 16$

Los rangos no coinciden por supuesto con los del cuadro XVI.7, ya que hemos tratado cada grupo como una réplica separada, con los rangos llegando sólo en cada caso hasta $k=3$. Obsérvese que hay sólo un empate dentro del grupo E, habiendo asignado un rango promedio de 1.5. El procedimiento más conservador habría consistido en asignar rango 1 a la ciudad comercial y rango 2 a la ciudad política, ya que para los demás grupos $T_2 > T_3$. Los resultados para el procedimiento más conservador aparecen entre paréntesis. Aun cuando estamos manejando un número muy reducido de casos y de columnas, utilizaremos, con fines de ilustración, la aproximación de χ^2 al cuadrado. Tenemos:

$$\chi^2 = \frac{12}{8(3)(4)} [20^2 + 14.5^2 + 13.5^2] - 3(8)(4) = 3.06$$

lo que para $d.f. = k-1 = 2$ no resulta significativo ni aun al nivel

de .10. Si hubiésemos empleado el método más conservador obteniendo $T_2 = T_3 = 14$, habríamos logrado una j al cuadrado de 3.00.

Bradley [3] hace notar que la eficacia de fuerza de la prueba de Friedman no sólo depende del tamaño de la muestra, sino del número de categorías usado. En el caso de muestras grandes la eficacia de la prueba en relación con el de la prueba F (suponiendo que todos los supuestos de esta última estuviesen justificados), es aproximadamente igual a

$$\frac{3}{\pi} \left(\frac{k}{k+1} \right)$$

De manera que para $k = 2$ la eficiencia de una muestra grande sería aproximadamente de $2/\pi = .64$, y para $k = 5$ resultaría aproximadamente de $5/2\pi = .80$. Bradley observa que a medida que k disminuye, disminuye también la ventaja de usar categorías a través de las columnas. En el caso límite en que $k = 2$, podemos asignar sólo las dos categorías 1 y 2, resultando esta prueba equivalente a la prueba de signo, teniendo por supuesto la misma escasa fuerza eficaz.

Si la variable dependiente (en este caso los niveles de la criminalidad), es medida de manera tan burda que sólo pueden asignarse los dos valores de *éxito* y *fracaso*, será posible hacer uso de una prueba no paramétrica muy similar, conocida como prueba Q de Cochran. El procedimiento que en ésta se sigue consiste en asignar unos (1) y ceros (0) a las equis (X) (tal vez según se encuentren por arriba o por abajo de la media global), utilizando una distribución exacta o una aproximación de j al cuadrado, como se hizo en el caso de la prueba de Friedman. La prueba de Cochran es discutida por Hays [7] y Bradley [3], y resulta también apropiada para el uso con muestras pareadas.

XVI.5. Medidas de asociación: correlación intraclase

Las pruebas de análisis de la variancia sólo nos permiten decidir si existe o no alguna relación entre dos variables. Como ya vimos, es relativamente fácil obtener significación estadística aun con una relación muy superficial, a condición que se tenga un número suficientemente grande de casos. Habiendo decidido que sí existe relación, sujetos sin duda al riesgo de error de tipo I, procedemos a continuación a medir la fuerza o grado de la misma. Puede obtenerse alguna indicación acerca de la magnitud de la relación, comparando simplemente las medias de las diversas categorías. Si estas medias difieren mucho, es probable que la relación sea fuerte; pero si las diferencias son pequeñas, en cam-

bio, podemos estar en condiciones de no atribuirles mucha significación práctica, aun en el caso en que hayamos obtenido significación estadística. Sin embargo, la mera comparación de las medias de las categorías puede resultar equívoca, a menos de observar también el grado de homogeneidad en el interior de cada grupo. Por lo regular, aunque tal vez no siempre, nuestro interés se centra en la magnitud relativa de las diferencias entre las medias, *en comparación con* las diferencias en el interior de las categorías. En otros términos: deseamos obtener una medida del grado en que las categorías son homogéneas en comparación con la variabilidad total en la variable de la escala de intervalo. Si las categorías son perfectamente homogéneas, la asociación entre las dos variables será completa, y sabiendo la categoría a la que un individuo pertenece, podemos predecir su marca exactamente.

Se han desarrollado varias medidas básicamente intercambiables de asociación, que se sirven de las sumas de cuadrados total, "entre" y "dentro", o bien de las apreciaciones de la variancia basadas en dichas sumas de cuadrados. La *razón de correlación* E^2 , la más simple tal vez de dichas medidas, comporta tomar la razón de la suma de cuadrados explicada, con respecto a la total. Así, pues:

$$E^2 = \frac{\text{SC explicada}}{\text{SC total}} = \frac{\text{SC "entre"}}{\text{SC total}} \quad (\text{XVI.9})$$

Según veremos en el próximo capítulo, la interpretación de la razón de correlación es directamente análoga a la de la correlación producto-momento corriente, salvo por su falta de signo, y nos serviremos de dicha medida para verificar la no linealidad de la relación entre dos escalas de intervalo.

La razón de correlación, sin embargo, es ligeramente sesgada. El lector recordará que la desviación estándar de la muestra, o variancia, tiende a subestimar la desviación estándar o variancia de la población, siendo el grado de sesgo relativamente importante en el caso de muestras pequeñas. De ahí que en el denominador nos sirviéramos de $N - 1$, en lugar de N , con objeto de obtener una estimación insesgada. Y en forma análoga, cuando el número de casos de cada categoría se hace relativamente pequeño, el valor esperado de la variabilidad en el interior de cada muestra tenderá, en comparación con la desviación estándar s , a ser menor que la de la población. Con objeto de corregir en relación con un sesgo correspondiente en la razón de correlación, podemos obtener lo que se designa como la *razón de correlación insesgada*, sirviéndonos de los grados de libertad adecuados y operando directamente con las estimaciones de la variancia y no con las sumas de cuadrados.

La fórmula de la razón de correlación insesgada ϵ^2 resulta ser la siguiente:

$$\epsilon^2 = 1 - \frac{V_w}{V_t} \quad (\text{XVI.10})$$

en donde V_w y V_t figuran en lugar de las *estimaciones* interior y total respectivamente. Si bien no hemos tenido necesidad hasta el presente de calcular la estimación total, su valor puede con todo obtenerse fácilmente dividiendo la suma de cuadrados total por $N - 1$. En el ejemplo numérico del que nos hemos servido, los valores de E y ϵ son respectivamente (véase el cuadro XVI.3, p. 342).

$$E^2 = \frac{42.303}{373.538} = .113 \quad E = .34$$

$$\epsilon^2 = 1 - \frac{15.773}{373.538/23} = .029 \quad \epsilon = .17$$

Obsérvese que el valor de ϵ es más pequeño que el de E .

Una medida de asociación algo más corriente es la del *coeficiente de correlación intraclase*. Esta medida deriva su nombre del hecho de que básicamente comporta una correlación de momento-producto entre todos los pares posibles de casos *dentro* de las categorías de la variable de la escala nominal.⁷ Al igual que las demás medidas examinadas en esta sección, el coeficiente de correlación dentro de las clases, r_i , puede considerarse también como medida del grado de homogeneidad de las clases en relación con la variabilidad total en la escala de intervalo. Su fórmula es como sigue:

$$r_i = \frac{V_b - V_w}{V_b + (\bar{n} - 1)V_w} \quad (\text{XVI.11})$$

en donde V_b y V_w son las estimaciones entre clases ($b = \text{between}$) y dentro de las clases ($w = \text{within}$) respectivamente, y \bar{n} representa un número de casos promedios en cada clase. Una fórmula alternativa para averiguar r_i en función de F es la siguiente:

$$r_i = \frac{F - 1}{F + (\bar{n} - 1)}$$

Si el número de casos en cada clase es el mismo, no existe problema, por supuesto, en cuanto al valor de \bar{n} . En el caso de cla-

⁷ Después de leído el capítulo XVII el lector podrá eventualmente consultar [5] para darse cuenta del carácter preciso de la relación entre estas dos medidas.

ses desiguales, en cambio, puede emplearse una simple media aritmética para obtener el valor en cuestión. Haggard [5] recomienda una clase algo distinta de valor promedio que habrá que utilizar siempre que el número de casos varíe considerablemente de una categoría a otra. Su fórmula para el cálculo de \bar{n} es:

$$\bar{n} = \frac{1}{k - 1} \left(\sum_{i=1}^k N_i - \frac{\sum_{i=1}^k N_i^2}{\sum_{i=1}^k N_i} \right) \quad (\text{XVI.12})$$

en donde N_i representa el número de casos de la categoría i -ésima y k el número de categorías. En nuestro ejemplo numérico todas las categorías son de la misma magnitud y, por consiguiente, $\bar{n} = 8$.

$$\text{Así, pues, } r_i = \frac{21.152 - 15.773}{21.152 + 7(15.773)} = \frac{5.379}{131.563} = .041$$

Pueden observarse algunas propiedades del coeficiente de correlación intraclase. Si las categorías son todas ellas perfectamente homogéneas, no habrá variación dentro de las clases (es decir, $V_w = 0$), y el valor de r_i será de + 1.0. En el caso extremo opuesto, supóngase, por el contrario, que toda la variación tiene lugar dentro de las clases y que las medias de las categorías son exactamente iguales. En este caso, V_b desaparecerá, y el límite inferior será:

$$\frac{-V_w}{(\bar{n} - 1)V_w} = -\frac{1}{\bar{n} - 1}$$

Así, pues, el límite inferior no es -1.0, excepto en el caso especial en que se tenga un promedio de 2 casos en cada clase. Normalmente, por lo tanto, el límite inferior será menor que la unidad en valor absoluto. De hecho, esto raramente nos preocupa, ya que pocas veces encontramos categorías que sean sustancialmente menos homogéneas de lo que se esperaría por azar. Cuando las apreciaciones "entre" y "dentro" son exactamente iguales, o sea cuando el valor de F es igual a la unidad, entonces r_i será cero. Así, pues, $r_i = 0$, cuando las categorías son exactamente tan homogéneas como se esperaría por azar, si no hubiera relación alguna entre las dos variables. Por lo regular, los valores de r_i se situarán entre 0 y 1.0. Por desgracia, no parece existir interpretación sencilla alguna de los valores de r_i entre dichos límites.

* La noción de correlación intraclase puede generalizarse fácilmente para abarcar el análisis de variancia de dos formas. En aquellas situaciones en las que nos serviríamos del término de

error en el denominador de F , podemos obtener una medida del grado de correlación entre la variable de las columnas y la escala de intervalo, con control en relación con la variable de las hileras, tomando como V_b la estimación entre columnas y sustituyendo V_w por el término del error. Y en forma análoga, podríamos tomar la estimación entre hileras como V_b , con lo que obtendríamos una medida del grado de asociación entre la escala de intervalo y la variable de las hileras, después de restada la variación debida a la variable de las columnas. Como veremos en el capítulo XIX, este procedimiento es directamente análogo a lo que hacemos al obtener correlaciones "parciales" entre dos escalas de intervalo, con control en relación con una tercera escala de intervalo.

GLOSARIO

Razón de correlación
Variación explicada e inexplorada
Homoscedasticidad
Interacción
Correlación intraclase
Comparaciones ortogonales.

EJERCICIOS

1. Como quiera que la prueba F puede utilizarse para probar la hipótesis nula de que tenemos dos estimaciones independientes de la misma variancia, podemos servirnos de la misma para verificar el supuesto de que $\sigma_1 = \sigma_2$ en problemas de diferencia de las medias. Ya que por lo regular no será posible anticipar cuál valor de s^2 será el mayor, tomamos la razón del mayor al menor y duplicamos el valor de la probabilidad dado en el cuadro F . Teniendo presentes estos hechos, tómense los datos del ejercicio 1, capítulo XIII, y verifíquese la hipótesis de que $\sigma_1 = \sigma_2$. Respuesta, $F = 1.75$, no rechazo al nivel .10.

2. Supóngase que los datos expuestos más abajo representan los ingresos de los presidentes de los consejos de administración de diversos tipos de organizaciones locales. Se han seleccionado al objeto al azar cinco organizaciones de cada tipo, tanto de localidades grandes como pequeñas, obteniendo en consecuencia números iguales de casos en cada subclase.

a) Empléese el análisis de variancia en dos formas con objeto de verificar la existencia de una relación entre el tipo de organización y los ingresos de los presidentes de consejos de administración, dejando de lado la extensión de la ciudad. ¿Cuáles son los valores de E y ϵ ? Respuesta, $F = 4.97$; $E = .52$; $\epsilon = .47$.

*b) Utilizando el análisis de variancia, de dos tipos ¿qué puede decirse acerca de la relación entre el tipo de organización y el ingreso, controlando en relación con la extensión de la localidad? ¿Cómo se comparan estos resultados con los de a)? Respuesta, para la interacción, $F = 3.52$, rechazo al nivel de .05.

c) Calcúlese la relación intraclase de los apartados a) y *b).

Tamaño de la localidad	Tipo de organización		
	Religioso	Social	Civil
Grande	\$ 13 000	\$ 15 000	\$ 20 800
	11 500	10 600	18 100
	17 300	12 300	14 600
	19 100	11 400	22 300
	16 700	10 800	16 500
Pequeño	15 000	9 300	14 400
	12 300	10 400	10 800
	13 900	12 900	9 700
	14 300	11 000	12 300
	11 700	9 100	13 100

3. Transfórmense los datos del ejercicio 2 relativos al ingreso en rango y, utilizando la prueba de Kruskal-Wallis, invéstiguese si existe o no relación entre el tipo de organización y el ingreso:

a) Prescindiendo del tamaño de la localidad. Respuesta, $H = 9.2$; rechazo al nivel .05.

b) Controlando en relación con el tamaño.

4. Utilice los datos del ejercicio 2, y suponga que las comunidades han sido organizadas en tríos de acuerdo con su tamaño, habiendo diez de aquéllos. Las organizaciones en la hilera superior (con ingresos de \$ 13 000, \$ 15 000 y \$ 20 800), representan las localidades mayores; las de la segunda hilera, las siguientes en tamaño, y así sucesivamente. Utilícese la prueba de Friedman para hallar la relación entre tipo de organización e ingresos del presidente (a nivel .05).

*5. El análisis de variancia puede practicarse lo mismo con datos agrupados que sin agrupar. Con objeto de reducir la confusión, lo más sencillo será servirse de las ecuaciones (XVI.4) y (XVI.5), sin modificar, pero recordando, sin embargo, que en el caso de datos agrupados tratamos las marcas como si estuvieran concentradas en los puntos medios de los intervalos. Teniendo en cuenta estas indicaciones, efectúese un análisis único de variancia con los datos del ejercicio 2, capítulo XIII. A título de control de los cálculos, ¿cómo se comparan entre sí los valores de F y t ?

6. Utilizando los datos del ejercicio 2 anterior:

a) Ignorando el tamaño de la localidad, búsquese la significación de la diferencia entre el ingreso medio de los presidentes de las organizaciones religiosas y las de los correspondientes a las organizaciones sociales y civiles combinadas.

b) ¿Qué comparación resultaría ortogonal con la hecha en a)?

c) Supóngase que tenemos seis tipos de organizaciones (religiosas en localidad grande, religiosas en localidad pequeña, social en

localidad grande, etcétera). ¿Cuántas comparaciones mutuamente ortogonales serían posibles? Hállese un grupo específico del anterior número de comparaciones que sean mutuamente ortogonales, comprobando que así es el caso.

BIBLIOGRAFÍA

1. Anderson, R. L., y T. A. Bancroft: *Statistical Theory in Research*, McGraw-Hill Book Company, Nueva York, 1952, caps. 17 y 18.
2. Blalock, H. M.: "Theory Building and the Statistical Concept of Interaction", *American Sociological Review*, vol. 30, pp. 374-380, 1965.
3. Bradley, J. V.: *Distribution-free Statistical Test*, Prentice-Hall, Inc., Englewood Cliffs, N. J., 1968, cap. 5.
4. Dixon, W. J., y F. J. Massey: *Introduction to Statistical Analysis*, 2ª ed., McGraw-Hill Book Company, Nueva York, 1957, cap. 10.
5. Haggard, E. A.: *Intraclass Correlation and the Analysis of Variance*, The Dryden Press, Inc., Nueva York, 1958, caps. 1-5.
6. Hagood, M. J., y D. O. Price: *Statistics for Sociologists*, Henry Holt and Company, Inc., Nueva York, cap. 22.
7. Hays, W. L.: *Statistics*, Holt, Rinehart and Winston, Inc. Nueva York, 1963, caps. 11-14.
8. Johnson, P. O.: *Statistical Methods in Research*, Prentice-Hall, Inc. Englewood Cliffs, N. J., 1949, caps. 10 y 11.
9. Kirk, R. E.: *Experimental Design: Procedures for the Behavioral Sciences*, Brooks/Cole Publishing Company, Belmont, Cal., 1968, cap. 3.
10. Siegel, S.: *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill Book Company, Nueva York, 1956, pp. 166-172, 184-193.
11. Walker, H. M., y J. Lev: *Statistical Inference*, Henry Holt and Company, Inc., Nueva York, 1953, cap. 14.

XVII. CORRELACIÓN Y REGRESIÓN

EN EL presente capítulo y en el siguiente examinaremos la relación entre dos escalas de intervalo. La extensión a tres o más variables de escala de intervalo se verá en el capítulo XIX, al tratar de la correlación múltiple y parcial. De momento, consideramos situaciones en las que tenemos dos medidas de escala de intervalo por cada individuo. Así, por ejemplo, podemos conocer el número de años de enseñanza completados y el ingreso anual de los varones adultos de una localidad determinada. O puede interesarnos relacionar el porcentaje de mano de obra empleado en la industria con el crecimiento demográfico de una población.

En algunos problemas de esta índole nos interesamos a menudo no sólo en las pruebas de significación y las medidas de grados de relación, sino que podemos también querer describir la *naturaleza* de la relación entre las dos variables, de modo que, conociendo una de ellas, podamos anticipar la otra. Así, por ejemplo, podemos querer predecir el ingreso futuro de una persona sobre la base de su instrucción, o la tasa de crecimiento de una ciudad a partir del porcentaje de su mano de obra empleada en la industria. Cuando el interés se centra ante todo en la tarea exploradora de encontrar *cuáles* variables se relacionan con una variable determinada, nos interesamos por lo regular principalmente por las medidas de grados o fuerza de las relaciones, tales como los coeficientes de correlación. Por otra parte, una vez halladas las variables significativas, propendemos a dirigir nuestra atención al análisis de regresión, en el que intentamos predecir el valor exacto de una variable a partir de la otra.

Si bien el lector ya está familiarizado con las pruebas de significación y las medidas de asociación, recomiéndase, con todo, empezar nuestro examen estudiando el problema de la predicción. Esto se debe a que la noción de regresión es a la vez anterior lógicamente y más importante teóricamente que la de correlación. La razón de ello se irá viendo más clara a medida que vayamos avanzando. Después de haber examinado el problema de la predicción, dirigiremos nuestra atención a la medición de la fuerza de la relación. En el capítulo XVIII, que de hecho representa la continuación del presente, examinaremos diversas pruebas de significación, así como la correlación del orden de lugares, que pueden emplearse para relacionar dos escalas ordinales.

XVII.1. Regresión lineal y mínimos cuadrados

En cierto sentido, el objetivo último de todas las ciencias es el de la predicción. Esto no implica, por supuesto, que sólo secun-

dariamente estemos interesados en comprender o suministrar explicaciones causales de por qué dos o más variables se relacionan como lo hacen. Tal vez sea más acertado decir que la comprensión constituye el objetivo final y que, en la medida en que la comprensión se va perfeccionando, la predicción se hace cada vez más precisa. Es posible que si la comprensión fuera completa la predicción perfecta sería también posible siempre que se conociera asimismo cierta información factual necesaria. Por ejemplo: si uno conoce las leyes del movimiento de los planetas, el campo gravitatorio dentro del sistema solar, y la posición y la velocidad de Venus en determinado momento, podría predecir su movimiento futuro. Sin embargo, independientemente de las implicaciones filosóficas de semejante punto de vista determinista, lo cierto es que la predicción constituye el objetivo de toda ciencia.

En sociología y en otras ciencias sociales, los enunciados predictivos se formulan a menudo, por necesidad, en forma relativamente burda. Por lo regular esto se debe a que no hemos alcanzado el nivel de medición de la escala de intervalo. Así, por ejemplo, podríamos predecir que cuanto más elevada sea la posición de una persona en el grupo, tanto mayor será su conformación a las normas de éste. Semejante enunciado no necesita implicar causalidad en una sola forma, sino que afirma simplemente que la posición y la conformidad se relacionan de modo positivo. Estableciendo una analogía con una terminología matemática que no es estrictamente correcta, decimos que la posición es una *función* de la conformidad, o que la conformidad es una función de la posición, eludiendo la cuestión de la causalidad. Obsérvese, sin embargo, que hemos dicho muy poco acerca de la *forma* de esta relación, aparte de indicar que es positiva. Y a menos que tengamos un nivel de medición de escala de intervalo para ambas variables, resulta efectivamente muy difícil decir mucho más.

Supóngase, sin embargo, que tenemos dos escalas de intervalo. Se hace entonces posible describir más exactamente de qué modo una de las variables varía con la otra. Así, por ejemplo podríamos estar en condiciones de decir que, por cada año de instrucción recibida, el ingreso aumentará en \$1 000. Si esto fuera efectivamente así, tendríamos en realidad una relación muy simple, o sea una relación lineal o en línea recta. Sin embargo, la mayoría de las relaciones no son ni con mucho tan sencillas, pese a que, según veremos, resulta a menudo posible obtener una aproximación muy buena de la verdadera relación suponiendo linealidad. La forma más elegante y sencilla de expresar una relación entre dos (o más) variables es por medio de una ecuación matemática. Así, por ejemplo, el lector estará familiarizado con ciertas leyes físicas que enuncian una relación entre la pre-

sión, el volumen y la temperatura ($PV/T = k$), o que indican una relación entre la razón de aceleración de un cuerpo al caer, la distancia recorrida y la duración del tiempo en que ha estado cayendo. Podemos también representar cada una de estas ecuaciones matemáticas como alguna clase de curva geométrica. Afortunadamente, en sociología solemos por lo regular operar con ecuaciones muy simples y con las curvas más simples posibles (rectas).

Cuando añadimos más variables, no podemos representar tan fácilmente las ecuaciones como figuras geométricas, ya que nos salimos de las dimensiones, de lo cual, sin embargo, no necesitamos preocuparnos por el momento.

Supóngase que hay una variable dependiente Y que ha de predecirse a partir de una variable independiente X . En algunos problemas, X precederá obviamente a Y en el tiempo. Por ejemplo: por lo regular una persona completa su instrucción antes de obtener un ingreso. En tales casos, semejante manera de representar las cosas resulta muy adecuada, pese a que hemos de poner cuidado en no implicar una relación necesaria o causal, o que X es la única variable que influye sobre el valor de Y . Si la dirección de la causa es ambigua, o si se piensa que cada variable es causa de la otra, necesitaremos, si es que deseamos suministrar una *explicación* teórica de la relación, usar un método de ecuaciones simultáneas que escape a este texto. (Véanse [1], [2] y [6]. Si nuestro objetivo es una simple estimación o una predicción a plazo breve de Y a partir de X , no se presentarán tales ambigüedades, aunque deba señalarse una vez más que no hay nada en las operaciones estadísticas que nos impida realizar operaciones matemáticas teóricamente carentes de sentido. En éste y en los capítulos sucesivos supondremos que la variable Y , seleccionada como dependiente en sentido matemático, es asimismo causalmente dependiente, de manera que la interpretación teórica puede resultar relativamente directa.

Ya vimos que si X y Y son estadísticamente independientes, no podemos predecir Y a partir de X o, más exactamente, el conocimiento de X no mejora en nada nuestra predicción de Y . Presumiblemente, pues, cuando las variables no son estadísticamente independientes, el conocimiento de X sí nos ayuda a predecir Y . Cuanto más fuerte sea la dependencia, tanto más precisa será nuestra predicción. Más adelante mediremos la fuerza de esta relación por medio de coeficientes de correlación. Nos concentramos de momento en la cuestión acerca de *cómo* predecimos Y a partir de X . Así, por ejemplo, podemos querer estimar el ingreso futuro de un individuo, sabiendo que ha completado tres años de escuela secundaria. Sin este conocimiento relativo a la instrucción, nuestra mejor estimación (suponiendo que no hay inflación) sería la del ingreso medio de todos los varones adul-

tos. En cambio, el hecho de conocer su instrucción debería permitirnos obtener una predicción mejor.

La ecuación de regresión. Representémonos el problema de la siguiente manera. Nos imaginamos que para cada valor fijo de la variable independiente X (instrucción) tenemos una distribución de Y (ingresos). En otros términos: para cada nivel educacional

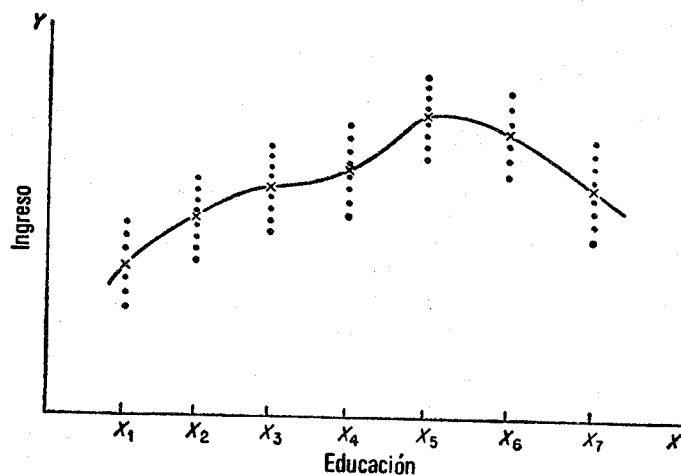


FIG. XVII.1. Forma general de la regresión de Y sobre X , o curso de las medias de los valores de Y para valores fijos de X .

habrá cierta distribución de ingresos en la población. No todas las personas que han terminado la escuela secundaria tendrán exactamente los mismos ingresos, por supuesto, pero dichos ingresos estarán con todo distribuidos alrededor de alguna media. Y habrá distribuciones de ingresos similares para los egresados de la escuela primaria, los de la universidad, los posgraduados, etcétera. Cada una de estas distintas distribuciones de ingresos (para X determinadas) tendrá una media, y podemos hacer una gráfica de la posición de dichas medias sirviéndonos del sistema familiar de las coordenadas rectangulares. Designamos el *curso resultante de estas medias* de las Y para X fijas como *ecuación de regresión de Y a X* . Semejante ecuación de regresión puede verse ilustrada en la figura XVII.1.

Estas ecuaciones de regresión son las "leyes" de la ciencia. En algunos casos hay muy poca dispersión alrededor de la ecuación de regresión. En tales casos, pueden hacerse predicciones muy precisas, y las desviaciones respecto de la ley se consideran a menudo como error de medición o como resultado de influencias menores no controladas. La "ley" puede formularse así como si existiera una perfecta relación entre Y y X . En el caso ideal,

se consideraría que todos los puntos caen exactamente en la curva, y la relación se abstraería como una función matemática perfecta en la que no hay más que una sola Y para cada X . En las ciencias sociales no podemos ser ni con mucho tan exigentes. En efecto, esperamos una variabilidad considerable alrededor de la ecuación de regresión, y preferimos pensar en términos de medias y de variancias de una distribución de Y para cada X . Sin embargo, el procedimiento es en principio el mismo en todas las ciencias, pese a que las leyes de las ciencias sociales no sean tan precisas como las de la física.

En la figura XVII.1 hemos indicado el carácter general de las ecuaciones de regresión, que comportan los cursos de las medias de los valores de Y para determinados valores de X . Vamos a tener que proceder ahora a algunos supuestos simplificadores, con objeto de poder tratar el problema estadísticamente. Si bien la idea de regresión es perfectamente general, la mayoría de la labor estadística sólo se ha realizado con los más simples de los modelos. En particular, vamos a suponer de momento: 1) que la forma de la ecuación de regresión es lineal, 2) que las distribuciones de los valores de Y para cada X son normales, y 3) que las variancias de las distribuciones de Y son las mismas para cada valor de X . Podemos ahora hacer un examen de estos diversos supuestos uno por uno, prestando la mayor atención al primero de ellos.

Si la regresión de Y a X es lineal, o sea una relación en línea recta, podemos escribir una ecuación como sigue:

$$Y = \alpha + \beta X \quad (\text{XVII.1})$$

en la que α y β son constantes. La ecuación (XVII.1) indica que la relación entre X y Y es exacta, pero en breve hemos de introducir en la ecuación un término de error. Una forma alternativa de escribir la ecuación es la siguiente: $E(Y|X) = \alpha + \beta X$; en la que $E(Y|X)$ pone de relieve que estamos preocupados con el valor esperado de Y , el que depende de X . Hemos utilizado letras griegas, ya que de momento tratamos de la población total. En una ecuación de esta clase, tanto α como β tienen interpretaciones geométricas definidas. Si ponemos X igual a cero, vemos que $Y = \alpha$. Por consiguiente, α representa el punto en donde la línea de la regresión corta el eje de las Y (o sea, allí donde $X=0$).

La inclinación de la línea de la regresión está dada por β , ya que esta constante indica la magnitud del cambio de Y para una unidad de cambio en X . El hecho de que la relación sea lineal significa que todo cambio de X , digamos en 5 unidades, produce siempre el mismo cambio en Y (esto es, 5β unidades, independientemente de la posición sobre el eje de X (véase fig. XVII.2). El lector ha de convencerse por sí mismo que si $\beta = 1$ y si las uni-

dades de X y Y están indicadas por distancias iguales a lo largo de los respectivos ejes, la línea de regresión formará un ángulo de 45 grados con el eje de las X . Una β mayor que la unidad indica una pendiente más rápida. Cuanto más rápida sea la pendiente, tanto mayor es el cambio de Y para un cambio dado de X . Y en forma análoga, si β es menor que la unidad pero mayor que cero, se requerirá un cambio mayor de X para producir un cambio

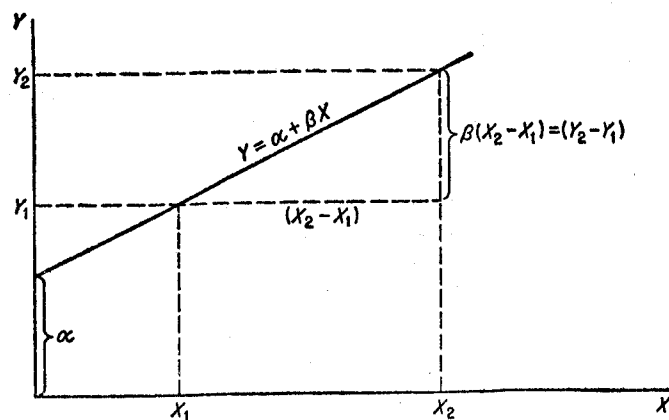


FIG. XVII.2. La ecuación lineal de regresión, mostrando interpretaciones geométricas de α y β .

dado en Y . En el caso límite, en que la línea es horizontal, β se hace cero, y los cambios de X no producen cambios de Y . En otros términos, si $\beta = 0$, no existe relación lineal entre X y Y . El conocimiento de X no nos ayuda a predecir Y , si se supone un modelo lineal.¹ Si β es negativa, sabemos que se da una relación negativa entre las dos variables, y que mientras X crece, Y decrece.

Una línea recta puede determinarse siempre por completo si conocemos ya sea dos puntos de la línea o un punto y la pendiente. Por lo tanto, no hay más que una sola línea de ecuación $Y = \alpha + \beta X$, a condición, por supuesto, que se considere a α y β como cantidades fijas (pero generales). Si α y β están dadas, podemos trazar la recta tomando simplemente dos puntos de la misma. Sabemos que cuando $X = 0$, $Y = \alpha$. Por consiguiente, el punto $(0, \alpha)$ se sitúa en la recta. Y así también, cuando $Y = 0$, tenemos $0 = \alpha + \beta X$ o $X = -\alpha/\beta$. Este punto $(-\alpha/\beta, 0)$ es, por supuesto, el punto en donde la línea corta el eje de las X . Si no

¹ Según veremos más adelante, la independencia estadística asegura que β sea cero, pero no se sigue necesariamente de ahí que si β es cero tengamos independencia.

conviene servirse de dichos dos puntos, pueden determinarse otros dos puntos cualesquiera por el mismo procedimiento.²

Supuestos acerca de X y el término de perturbación. Hasta ahora no hemos tratado en forma explícita el hecho de que, puesto que habrá dispersión alrededor de la ecuación de regresión, habremos de representar el valor real de Y para cada individuo mediante una ecuación que contenga un término de perturbación o de error que es único para cada individuo. Si suponemos que Y_i y X_i se refieren a las puntuaciones correspondientes al i -ésimo individuo, podremos representar la relación (lineal), como sigue:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

en la que ε_i representa el término de perturbación, cuyo comportamiento necesitamos estudiar. Podemos concebir este término como si contuviera el error de medición en Y (pero no en X), y como resultante de todas las varias causas de Y que no han sido llevadas a la ecuación de una manera explícita. Si la mayor parte de estas causas omitidas tienen individualmente un efecto menor, y si además están operando casi independientemente entre ellas, será razonable suponer que el valor esperado correspondiente al factor de perturbación $E(\varepsilon_i)$ será igual a cero, y que ε_i estará distribuido en forma aproximadamente normal. Lo que resulta muy importante es el hecho de que el factor de perturbación será estadísticamente independiente de X . Resulta que al usar mínimos cuadrados para estudiar los coeficientes de regresión α y β , es necesario suponer que $E(\varepsilon) = 0$, y que X_i y ε_i no están relacionados. La suposición de normalidad, más la suposición de homoscedasticidad, de que σ^2_ε es constante a través de todos los niveles de X será necesaria en las pruebas de significancia y para la determinación de los límites de confianza.

El supuesto fundamental que subraya el uso del análisis de regresión es el de que X sea independiente del factor de error. En aplicaciones experimentales nos encontramos con frecuencia en la posibilidad de elegir niveles fijos de X (como, por ejemplo, cuando mantenemos constantes de temperatura a intervalos de 50 grados). En tales casos, puesto que el nivel de X está bajo nuestro control y se presume que no es manipulado en forma que varíe sistemáticamente con el factor de perturbación, será raro preocuparse con este supuesto concreto. Un momento de reflexión nos convencería, sin embargo, de que en muchas situaciones experimentales incluso este supuesto es inocente, ya que al manipular X uno puede inadvertidamente afectar otros factores que se quedaron fuera de la ecuación y contenidos por lo tanto en el factor de perturbación.

En la investigación no experimental se toma tanto a las X como

² Véase un ejemplo numérico en la página 392.

a las Y como observadas y no como manipuladas, siendo por lo tanto X y Y variables aleatorias, o lo que se denomina variables *estocásticas*, las que tienen una distribución de probabilidad. En algunos casos la distribución de X será aproximadamente normal, aunque esto no es necesario en el caso del análisis de regresión. Lo que *resulta* esencial, sin embargo, es el formular algunos supuestos acerca de la distribución conjunta de X_i y el factor de perturbación ε_i . Si tuviéramos *a priori* razones sólidas para especificar alguna distribución particular, esto resultaría suficiente, pero en la práctica se carece siempre de tal información. Con mucha frecuencia suponemos que X_i y ε_i son estadísticamente independientes, supuesto que resultará justificado si las causas de Y omitidas son, 1) numerosas, aisladamente sin importancia, y no muy interrelacionadas, o 2) sin relación con X en situaciones en las que predominan uno o dos de los factores omitidos. Si uno no está dispuesto a hacer tal suposición en algún caso particular, deberá tratar de identificar los mayores factores perturbadores que hayan sido omitidos, introduciéndolos explícitamente en la ecuación como variables adicionales. En el capítulo XIX examinaremos la regresión múltiple, en la que han sido incluidos tales factores causales adicionales.

Una de las ventajas de la teoría estadística del análisis de regresión consiste en que está lo suficientemente desarrollada como para que tales supuestos acerca del comportamiento de los factores de perturbación resulten explícitos. Resultará bien claro que lo que hemos dicho acerca del comportamiento de las variables omitidas se aplica igualmente bien a *todos* los procedimientos que hasta aquí hemos examinado. Si se encuentra, por ejemplo, una diferencia estadísticamente significativa en medias o proporciones, y si se desea atribuir una explicación causal a la variable independiente (por ejemplo, sexo) en esta relación, habrá que suponer también que los factores omitidos no están sistemáticamente relacionados con la escala nominal dicotomizada (por ejemplo sexo). No es posible soslayar supuestos acerca de variables omitidas cambiando simplemente el tipo del análisis y confiando en que así desaparecerá el problema.

Ya se indicó más arriba que para las pruebas de significación hemos de suponer que las Y están distribuidas normalmente alrededor de cada valor de X . Para las X estocásticas convendrá también suponer que para cada valor fijo de Y las X están asimismo distribuidas normalmente. Decimos que la distribución conjunta de X y Y es una distribución *normal bivariable*, lo que significa que hay dos variables, cada una de las cuales está distribuida alrededor de la otra en forma normal. Semejante distribución normal bivariable tiene una ecuación matemática definida y puede representarse como una superficie tridimensional, como en la figura XVII.3. La altura de la superficie en un punto

dado (X, Y) es proporcional al número de casos en el mismo. Así, pues, se requiere un diagrama tridimensional para representar la distribución conjunta entre X y Y , del mismo modo que necesitábamos dos dimensiones para representar la distribución de frecuencia de la X sola. La forma exacta de esta figura, que se

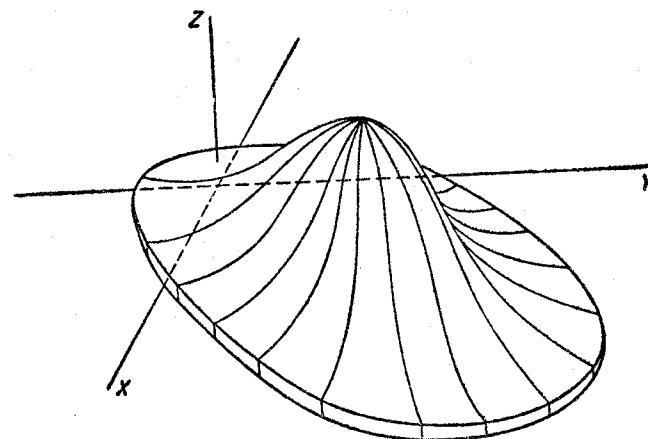


FIG. XVII.3. La distribución normal bivariable. (Con autorización de A. M. Mood, *Introduction to the Theory of Statistics*, McGraw-Hill Book Company, Inc., Nueva York, 1950, fig. 41, p. 165.)

parece mucho a un casco de bombero, dependerá de cuán cerca-
namente estén relacionadas las variables entre sí.

Si ambas variables se han expresado en términos de unidades de desviación estándar, entonces, cuanto más relacionadas estén las variables tanto más angosto será el casco. En el caso extremo, en el que Y puede predecirse exactamente a partir de X y, por consiguiente, todos los puntos están exactamente en la ecuación de regresión, las desviaciones estándar de las Y para cada X serían cero, y el casco no tendría grueso alguno. Por otra parte, si no existiera relación alguna entre X y Y , la base del casco sería más aproximadamente circular. Cualquier plano perpendicular al plano XY cortaría la superficie en una curva normal. En tanto que un plano paralelo al plano XY cortará el casco en una elipse. La distribución normal bivariable posee la propiedad de que la regresión de Y a X sea lineal. Por lo tanto, si tenemos una distribución normal bivariable, sabemos que, si trazamos las medias de las Y para cada X , el resultado será una recta. No se sigue de ahí, sin embargo, que si la regresión es lineal, la distribución conjunta sea necesariamente normal bivariable.

En el caso de las pruebas de significancia necesitaremos tam-

bién suponer que las desviaciones estándar de las Y para cada X son las mismas, independientemente del valor de X . Este supuesto se examinará en conexión con el tema de la correlación, ya que ésta es esencialmente una medida de dispersión alrededor de la línea de regresión. De momento basta, con todo, señalar que si la distribución conjunta es normal bivariable, las desvia-

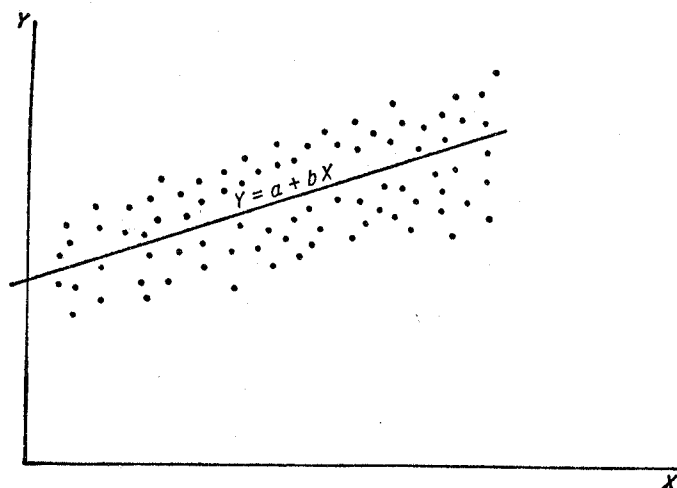


FIG. XVII.4. Diagrama de dispersión y recta de mínimos cuadrados.

ciones estándar de las Y para cada X serán de hecho todas idénticas. Esta propiedad de variancias iguales se designa como *homoscedasticidad* y es análoga al supuesto hecho en el análisis de variancia de que $\sigma_1 = \sigma_2 = \dots = \sigma_k$.

Mínimos cuadrados lineales. El modelo de regresión que hemos estado examinando es más bien sencillo en sus conceptos, pero no es por desgracia directamente útil en su forma teórica. Es raro, en efecto, que tengamos suficientes casos para examinar la distribución de las Y para valores fijos sucesivos de X . Con mayor frecuencia encontramos que hay relativamente pocos casos en los que las X sean idénticas o aproximadamente tales. Si hacemos una gráfica de la distribución de los casos alrededor de los ejes de las X y las Y en la forma convencional, encontramos por lo regular una dispersión de puntos como la que se indica en la figura XVII.4. Y si hacemos una gráfica de la distribución de los puntos en esta forma, obtenemos lo que se designa como *esquedograma* o diagrama de dispersión. El estudiante ha de acostumbrarse a dibujar un diagrama de dispersión antes de proceder al análisis ulterior. La mera inspección del

diagrama en cuestión, en efecto, puede acaso indicar que no tiene objeto seguir adelante. Así, por ejemplo, si los puntos aparecen en el diagrama como si estuvieran distribuidos al azar, resulta claro que no existe relación, o sólo una relación muy débil, entre las dos variables.

Una vez fijadas las marcas en un diagrama de dispersión, podemos querer acercarnos a dichos puntos por medio de alguna clase de curva que sea la más adecuada. Una de las maneras de hacerlo es trazar una curva (en el presente caso una recta) por inspección. Sin embargo, existen para ello métodos más precisos. Uno de éstos es el método de los mínimos cuadrados, que se examinará en la presente sección. Nuestro objetivo es ahora algo distinto del objetivo del análisis de regresión, en el que trazábamos el curso de la media de las Y . Aquí, en efecto, queremos aproximarnos a cierto número de puntos por medio de una curva de mejor adaptación.

Con objeto de servirnos de la teoría de los mínimos cuadrados, hemos de postular la forma de la curva a utilizar en la adaptación de los datos. En el caso del análisis de regresión, la forma de la curva se hallaría propiamente determinada por el curso de las medias, suponiendo que se dispone de datos relativos a la población entera. Vamos a tomar una vez más la curva más simple posible, la recta, como curva de nuestros mínimos cuadrados. Esto significa que hemos de adaptar los datos a una recta de mejor ajuste, conforme al criterio de los mínimos cuadrados, obteniendo una ecuación de la forma:

$$Y = a + bX \quad (\text{XVII.2})$$

Resultará así que la a y la b obtenidas con este método son las apreciaciones insesgadas más eficaces de los parámetros de la población, α y β , si la ecuación de regresión es efectivamente una recta y si suponemos: 1) Muestreo al azar, 2) Que $E(\epsilon_i) = 0$, y 3) Que X_i y ϵ_i son estadísticamente independientes.

Nuestro criterio de los mínimos cuadrados comporta hallar la única recta que posee la propiedad de que la suma de los cuadrados de las desviaciones de los valores reales de Y respecto de dicha recta sea mínima. Así, por ejemplo, si trazamos líneas verticales de los puntos a la línea de los mínimos cuadrados, y si elevamos al cuadrado dichas distancias y las sumamos, la suma resultante será menor que la suma correspondiente de cuadrados a cualquier otra recta posible (véase la figura XVII.5). Obsérvese que son las distancias verticales, y no las perpendiculares o las horizontales las que aquí se consideran. Sería posible minimizar la suma de los cuadrados de las distancias perpendiculares (designada como suma ortogonal de los mínimos cuadrados), pero las ecuaciones de ello resultantes no son ni

con mucho tan prácticas. Y si se emplearan las distancias horizontales, la recta de mínimos cuadrados resultante podría utilizarse para apreciar la regresión de X a Y . El lector ha de convencerse por sí mismo que minimizar la suma de cuadrados de las distancias verticales no minimiza necesariamente la suma de cuadrados de las distancias horizontales. Así, pues,

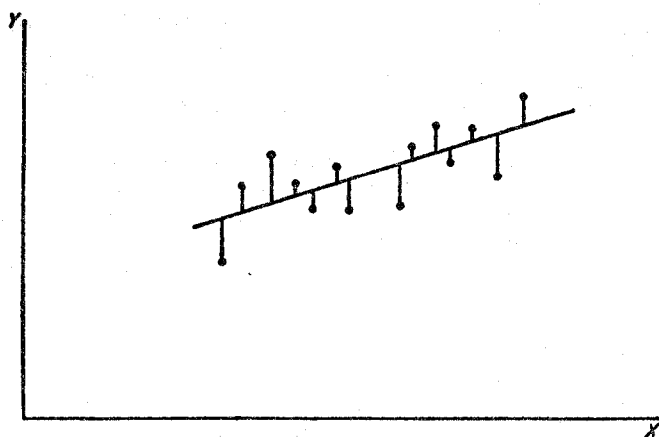


FIG. XVII.5. Ecuación de mínimos cuadrados, que minimiza las sumas de los cuadrados de las distancias verticales y estima la regresión de Y sobre X .

podemos obtener varias líneas de mínimos cuadrados distintas. Pero éstas sólo coincidirán si todos los puntos quedan exactamente en una sola línea. Resulta asimismo que, al minimizar la suma de los cuadrados de las distancias verticales, encontramos de hecho la recta que posee la propiedad de que la suma de las distancias verticales positivas y negativas sea cero y la desviación estándar de los puntos respecto de aquélla sea mínima. Este concepto de la desviación estándar de las Y se examinará con mayor detalle más adelante.

Con objeto de obtener la línea de mínimos cuadrados, pues, necesitamos calcular la a y la b que determinan la línea provista de la propiedad deseada. Esta clase de problemas puede resolverse fácilmente por medio del cálculo y conduce a las siguientes fórmulas de cálculo de a y b .³

³ Para los estudiantes familiarizados con el cálculo elemental vamos a delinear la naturaleza de la derivación. Comenzaremos con la ecuación $Y_i = a + bX_i + e_i$, en la que e_i es un término residual que puede ser utilizado para estimar el residual e_i de la ecuación de regresión. Deseamos minimizar la suma de los cuadrados de estos residuales, es decir: la cantidad $\sum e_i^2 = \sum (Y_i - a - bX_i)^2$ con respecto a las dos cantidades a y b ,

$$a = \frac{\sum_{i=1}^N Y_i - b \sum_{i=1}^N X_i}{N} = \bar{Y} - b\bar{X} \quad (\text{XVII.3})$$

$$b = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^N (X_i - \bar{X})^2} = \frac{\sum_{i=1}^N x_i y_i}{\sum_{i=1}^N x_i^2} \quad (\text{XVII.4})$$

en donde $x_i = X_i - \bar{X}$ y $y_i = Y_i - \bar{Y}$. Obsérvese que en estas ecuaciones a y b son las incógnitas, hallándose las otras cantidades determinadas a partir de los datos. Una vez que se haya obtenido b , a puede calcularse fácilmente a partir de la primera de las dos fórmulas. Podemos, pues, centrar nuestra atención en el cálculo de b .

El numerador de b comporta la expresión $\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$ que se designa como *covariación* de X y Y . Esta cantidad es directamente análoga a las sumas de cuadrados tanto de X como de Y , excepto que, en lugar de elevar al cuadrado $(X - \bar{X})$ o $(Y - \bar{Y})$, tomamos el producto de estos dos términos. Obtenemos en esta forma una medida de cómo X y Y varían juntas, y de ahí el nombre de *covariación*. Si dividimos esta expresión entre N , obtenemos, por analogía, lo que se designa como *covariancia*. Veremos inmediatamente que b puede ponerse igual a la razón de la covariancia a la variancia en X .

Si examinamos más de cerca la covariación de X y Y , vemos que, a diferencia de una suma de cuadrados, la covariación puede tomar valores tanto positivos como negativos. Si X y Y se relacionan positivamente, entonces valores grandes de X se asociarán por lo regular con valores grandes de Y . Así, pues, si $X > \bar{X}$, será por lo regular cierto que $Y > \bar{Y}$. Y asimismo, en el caso de una relación positiva, si $X < \bar{X}$, tendremos generalmente $Y < \bar{Y}$. Por consiguiente, el producto de $(X - \bar{X})$ y $(Y - \bar{Y})$ será normalmente positivo, y la suma de estos productos será asimismo positiva. Y en forma análoga, si X y Y se relacionan negativamente, esperaríamos que, si $X > \bar{X}$, entonces Y será menor que \bar{Y} , y la suma de productos resultante será negativa. Si no existe relación, entonces aproximadamente la mitad de los productos serán positivos y la otra mitad negativos, ya que X y Y variarán indepen-

a las que aquí se trata como desconocidas. Tomamos derivativos parciales con respecto a a y b ; las hacemos igual a cero, y resolvemos las dos ecuaciones resultantes (a las que se denomina *ecuaciones normales*) para a y b . Este mismo procedimiento es de aplicación al caso multivariado.

dientemente. En este caso, b será cero, o vecino de cero. Por lo tanto, cuanto mayor sea el valor numérico de la relación, independientemente de la dirección, tanto mayor será el valor numérico de la covariación. Como habremos de ver en breve, la covariación figura también en el numerador del coeficiente de correlación, que es nuestra medida del grado de asociación. En el caso de b , tomamos la covariación y la dividimos entre la suma de los cuadrados en X , con objeto de obtener nuestra estimación de la pendiente de la ecuación de regresión.

Es más conveniente servirse para la covariación de una fórmula que es directamente análoga a la fórmula de cálculo de la suma de los cuadrados y puede derivarse en forma similar. Podemos escribir la fórmula de cálculo de b como sigue:

$$b = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2} \quad (\text{XVII.5})$$

En la ecuación (XVII.5), tanto el numerador como el denominador se han multiplicado por N , con objeto de redondear los errores debidos a la división y con objeto de facilitar el cálculo con una calculadora.⁴

Problema. Supóngase que tenemos los datos del cuadro XVII.1, en donde X representa el porcentaje de negros en las grandes ciudades del Medio Oeste, y Y indica la diferencia entre las medianas de los ingresos de los blancos y los negros, como medida de discriminación económica.⁵

CUADRO XVII.1. Datos para un problema de correlación

Porcentaje de negros X	Diferencia de ingresos Y	Porcentaje de negros X	Diferencia de ingresos Y
2.13	\$ 809	4.62	\$ 859
2.52	763	5.19	228
11.86	612	6.43	897
2.55	492	6.70	867
2.87	679	1.53	513
4.23	635	1.87	335
		10.38	868

⁴ En esta y las fórmulas posteriores hemos prescindido de los subíndices, ya que se opera siempre la suma total de los casos, del cuadro N.

⁵ Aunque la palabra "negro" puede resultar ofensiva para algunos lectores, resulta necesario mantener esta terminología al referirse a los datos del censo, como contraste con otros datos hipotéticos o los obtenidos de otras fuentes.

A partir de los datos podemos calcular cinco sumas que, junto con N , son todo lo que necesitamos para tratar los problemas de regresión y correlación. Todas estas sumas menos una se emplearán en los cálculos de a y b . Los cálculos pueden resumirse como sigue:

$$\begin{aligned} N &= 13 & \Sigma Y &= 8\,557 \\ \Sigma X &= 62.88 & \Sigma Y^2 &= 6\,192\,505 \\ \Sigma X^2 &= 432.2768 & \Sigma XY &= 43\,943.32 \end{aligned}$$

Aquí la única cantidad nueva es ΣXY . Si ponemos estos valores en las fórmulas de a y b , tenemos ahora:

$$\begin{aligned} b &= \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2} \\ &= \frac{13(43\,943.32) - (62.88)(8\,557)}{13(432.2768) - (62.88)^2} = \frac{33\,199.0}{1\,665.7} = 19.931 \end{aligned}$$

$$\begin{aligned} \text{y} \quad a &= \frac{\Sigma Y - b\Sigma X}{N} \\ &= \frac{8\,557 - (19.931)(62.88)}{13} = 561.83 \end{aligned}$$

Por lo tanto, la ecuación lineal resultante es:

$$Y_p = a + bX = 561.83 + 19.931X$$

en donde hemos utilizado Y_p para indicar que los valores de Y se han estimado a partir de una ecuación de mínimos cuadrados. Como ya se indicó anteriormente, las a y b obtenidas por este método son las estimaciones insesgadas más eficaces de α y β , o sea los coeficientes de regresión reales a condición de que el factor de perturbación ε_i en la ecuación $Y_i = \alpha + \beta X_i + \varepsilon_i$ tenga un valor esperado de cero no relacionado con X , y siempre, por otra parte, de que tengamos una muestra al azar de la población que estudiamos. Por consiguiente, la línea de mínimos cuadrados será la mejor apreciación de la verdadera regresión, si la ecuación de regresión es efectivamente lineal.

La ecuación de los mínimos cuadrados posee asimismo la propiedad de pasar por el punto (\bar{X}, \bar{Y}) , que representa las medias de X y de Y . Esto puede verse en la ecuación (XVII.3). Ya que

$$a = \bar{Y} - b\bar{X}$$

tenemos:

$$\bar{Y} = a + b\bar{X}$$

lo que indica que estos valores de X y Y satisfacen la ecuación. Por consiguiente, el punto (\bar{X}, \bar{Y}) queda exactamente sobre la línea.

En el problema anterior, si sabemos el valor de X (porcentaje de negros) para cualquier ciudad dada del Medio Oeste, nuestra mejor estimación del valor de Y sería aquel valor de Y que co-

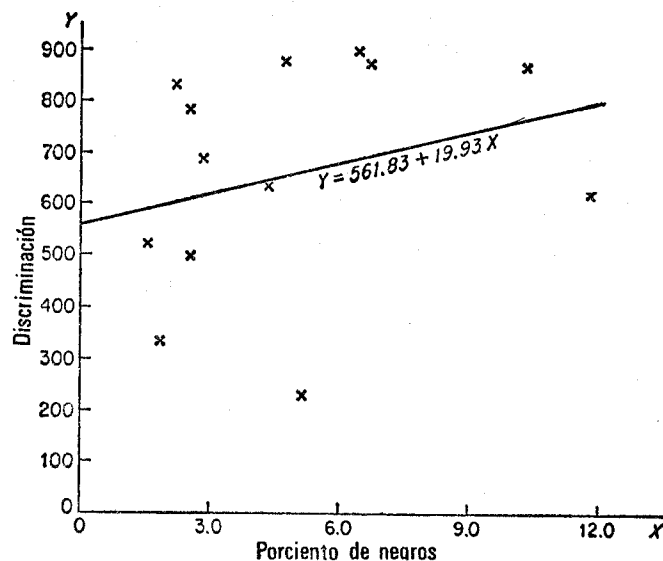


FIG. XVII.6. Diagrama de dispersión y recta de mínimos cuadrados para los datos del cuadro XVII.1.

responde en la ecuación de los mínimos cuadrados a la X dada. Como quiera que las marcas de discriminación indican diferencias (en dólares) entre los ingresos (en medianas) de los blancos y los negros, vemos que un aumento del 1 por ciento de los negros corresponde a una diferencia de \$ 19.93 en dichos ingresos. En la figura XVII.6 se han trazado un diagrama de dispersión y la ecuación de los mínimos cuadrados. Con objeto de ilustrar el empleo de semejante ecuación de predicción, si supiéramos que había un 8 por ciento de negros en una ciudad determinada, la diferencia estimativa del ingreso mediano sería:

$$Y_p = a + b(8) = 561.83 + (19.931)(8) = \$ 721.28$$

Vemos en la figura que se habría obtenido aproximadamente el mismo resultado con la gráfica. Observemos de paso que, haciendo $X = 8$ y resolviendo en relación con Y , hemos localizado un

segundo punto de la línea, que puede utilizarse a continuación con objeto de trazar la línea en el diagrama de dispersión.

XVII.2. Correlación

Supongamos a partir de ahora que X es estocástica, y no sometida por tanto al control del investigador. No sólo deseamos conocer la *forma* o la naturaleza de la relación entre X y Y , de modo que una de las variables pueda predecirse a partir de la otra, sino que es necesario al propio tiempo conocer el *grado* o fuerza de la relación. Es obvio que si la relación es muy débil, no tiene objeto tratar de predecir Y a partir de X . Los sociólogos tienen a menudo interés ante todo en descubrir *cuáles* de un gran número de variables se relacionan más de cerca con una variable dependiente determinada. En los estudios de exploración de esta clase, el análisis de regresión reviste importancia secundaria. A medida que una ciencia va madurando y que se descubren variables importantes, la atención puede centrarse en métodos de predicción exacta. Algunos estadígrafos son del parecer que en conjunto se ha prestado demasiada atención a la correlación y casi ninguna al análisis de regresión. Que esto sea así o que no lo sea depende, por supuesto, del estado del conocimiento en la ciencia considerada.

El coeficiente de correlación r , que vamos a examinar en esta sección, fue introducido por Karl Pearson y se designa a menudo como correlación momento-producto, con objeto de distinguirla de otras medidas de asociación. Este coeficiente mide la cantidad de dispersión alrededor de la ecuación *lineal* de los mínimos cuadrados. Hay un coeficiente correspondiente de población rho (ρ), que mide la bondad del ajuste a la verdadera ecuación de

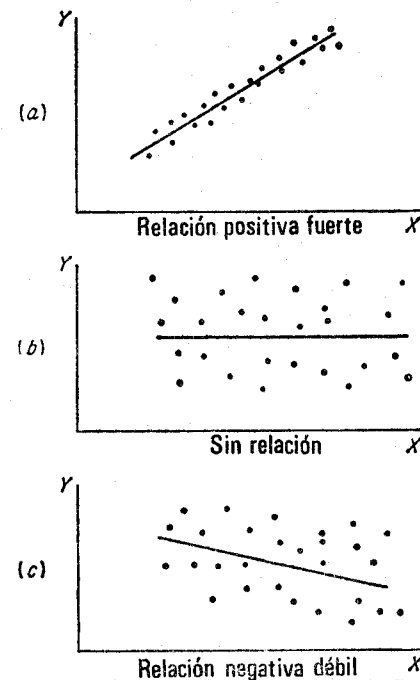


FIG. XVII.7. Diagrama de dispersión que muestra las diferentes fuerzas y direcciones de las relaciones entre X y Y .

regresión. Obtenemos una estimación r de dicho parámetro midiendo las desviaciones respecto de la línea calculada por medio de los mínimos cuadrados.

Como quiera que la ecuación de regresión representa el curso de las medias de las Y para unas X dadas, sería también posible medir la dispersión respecto de esa línea tomando una desviación estándar de la misma.⁶ Sin embargo, los investigadores de la mayoría de los campos de aplicación se han acostumbrado al coeficiente de correlación; es probable, con todo, que el coeficiente de correlación se mantenga. Posee la ventaja de ser de fácil interpretación, y su recorrido va de -1.0 a 1.0 , hecho que resulta atractivo para la mayoría de los prácticos. Según veremos, en efecto, la relación entre el coeficiente de correlación y la desviación estándar respecto de la línea de los mínimos cuadrados es muy sencilla, hecho que puede utilizarse para proporcionar una interpretación de r .

Se acaba de indicar que r tiene un límite superior de 1.0 . Si todos los puntos se hallan exactamente sobre la recta, r será 1.0 o -1.0 , según que la relación sea positiva o negativa. Y si los puntos están dispersados al azar, r será cero. Cuanto mejor sea el ajuste, tanto mayor será la magnitud de r . Es lo que se indica en la figura XVII.7.

Obsérvese que r es una medida de relación *lineal*, ya que es una medida de la bondad de ajuste de la línea de los mínimos cuadrados. El lector no debe caer en el error de suponer que si $r = 0$ (o si $\rho = 0$) no existe relación alguna. En efecto, si no hay relación, síguese que r será aproximadamente cero y habrá una dispersión de puntos al azar. Sin embargo, puede haber una relación perfectamente curvilínea y, con todo, ser r cero, indicando que no se da *recta* alguna que satisfaga los datos. Éste es el caso en la figura XVII.8, por ejemplo. Por lo tanto, si el investigador encuentra una correlación de cero, habrá de precaverse contra la deducción de que no existe relación entre las variables. Por lo regular, la inspección del diagrama de dispersión indicará si hay o no relación de hecho, o si la relación es suficientemente no lineal para producir una correlación de cero. En la mayoría de los problemas sociológicos, las relaciones pueden aproximarse razonablemente por medio de rectas. Sin embargo, esto no significa que no se deba estar bastante alerta contra excepciones eventuales.

Hasta el presente no hemos definido todavía el coeficiente de correlación, pero podemos hacerlo fácilmente en los términos de la fórmula:

⁶ La naturaleza exacta de semejante medida se examinará más adelante. De momento podemos señalar simplemente que representa una extensión del concepto de la desviación estándar, en la que la media de las Y ya no se toma como fija, sino que se considera función de X .

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{[\Sigma(X - \bar{X})^2][\Sigma(Y - \bar{Y})^2]}} = \frac{\Sigma xy}{\sqrt{(\Sigma x^2)(\Sigma y^2)}} \quad (\text{XVII.6})$$

U oralmente: el coeficiente de correlación es la razón de la covariación a la raíz cuadrada del producto de la variación de X y la variación de Y . Dividiendo el numerador y el denominador entre N y poniendo esta cantidad como N^2 bajo el radical, vemos

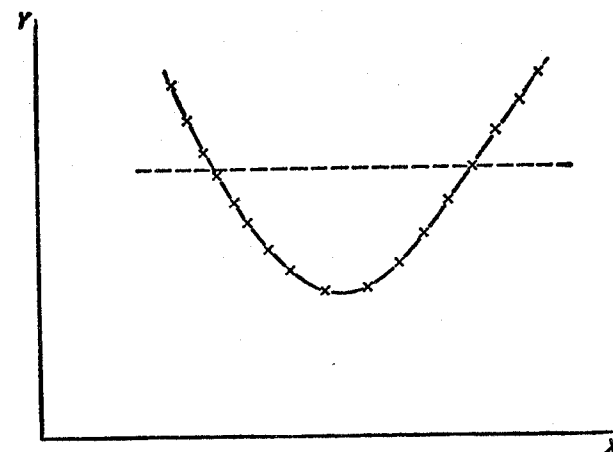


FIG. XVII.8. Diagrama de dispersión de una relación no lineal perfecta, en que $r = 0$.

que r puede también definirse como la razón de la covariancia al producto de las desviaciones estándar de X y Y . La covariancia es la medida de la variación conjunta de X y Y , pero su magnitud depende de la cantidad total de variabilidad de las dos variables. Como quiera que el valor numérico de la covariancia puede ser considerablemente mayor que la unidad, no resulta conveniente emplearlo directamente como medida de asociación. En lugar de ello, estandarizamos dividiendo entre el producto de las dos desviaciones estándar, con lo que obtenemos una medida que varía entre -1.0 y 1.0 .

Ya vimos que la covariancia será cero siempre que X y Y no estén relacionadas. Puede demostrarse también fácilmente que el límite superior de r es la unidad. Tomemos, por ejemplo, el caso en que b es positiva y todos los puntos se encuentran exactamente sobre la recta. En tal caso, para cada Y podemos escribir $Y = a + bX$. Y como quiera que el punto (\bar{X}, \bar{Y}) se encuentra tam-

bién sobre la recta, tenemos $\bar{Y} = a + b\bar{X}$. Por consiguiente, para todos los puntos sobre la recta tenemos:

$$Y - \bar{Y} = (a + bX) - (a + b\bar{X}) = b(X - \bar{X})$$

De donde: $\Sigma(X - \bar{X})(Y - \bar{Y}) = b\Sigma(X - \bar{X})^2$

y $\Sigma(Y - \bar{Y})^2 = b^2\Sigma(X - \bar{X})^2$

La inspección del numerador y el denominador de r indica ahora que, en estas condiciones, $r = 1.0$. Y en forma análoga, puede demostrarse que si todos los puntos se encuentran exactamente sobre una línea de pendiente negativa, la r resultante será -1.0 .

Conviene observar asimismo la relación entre el coeficiente de correlación y las pendientes de las dos ecuaciones de los mínimos cuadrados. Si hacemos que b_{yx} sea la pendiente de la ecuación de mínimos cuadrados estimando la regresión de Y sobre X , y dejamos que b_{xy} indique la pendiente de la estimación de la regresión de X sobre Y , tenemos, por simetría, que:

$$b_{xy} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\Sigma(Y - \bar{Y})^2}$$

en donde

$$X = a_{xy} + b_{xy}Y$$

Así, pues, r tiene el mismo numerador que las dos b . Si éstas son cero, síguese que r ha de ser también cero y viceversa.

Para sumas de cuadrados en X y Y dadas, el valor de b_{yx} (o de b_{xy}) será proporcional a r . Esto parecería conducir a la conclusión de que la fuerza de la relación sea proporcional a la pendiente de la línea de los mínimos cuadrados. Sin embargo, esto sólo será así si el denominador permanece fijo. Así, pues, b es una función no sólo de la fuerza de la relación, sino también de las desviaciones estándar.⁷ Si hay bastante variabilidad en X , en relación con Y , el valor de b será relativamente pequeño, indicando que se requiere un gran cambio de X para producir un cambio moderado de Y . Como lo veremos después, los valores numéricos de las b dependen, por consiguiente, de la magnitud de las unidades de medida.

El valor de r se ha estandarizado de modo que sea hasta cierto punto independiente de las magnitudes relativas de las desviaciones estándar en X y Y . Sería en efecto desdichado que no fuera así, ya que difícilmente deseamos una medida que variara

⁷ Excepto en los casos en que ello pudiera dar lugar a confusión, seguiremos sirviéndonos de b sin subíndice para representar b_{yx} .

según que escogiéramos como unidad monetaria dólares o centavos. Se observará en las fórmulas de r y las b que r^2 puede expresarse en términos de estas últimas. Así, pues:

$$r^2 = b_{yx}b_{xy} = \frac{[\Sigma xy]^2}{\Sigma x^2 \Sigma y^2} \quad (\text{XVII.7})$$

El lector hará bien en verificar que cuando r es 1.0 (o -1.0), $b_{yx} = 1/b_{xy}$, lo que significa que las dos ecuaciones de mínimos cuadrados coinciden. Por lo regular, a medida que r se acerca a cero, el ángulo entre las dos líneas se va haciendo cada vez mayor, hasta que, $r = 0$, las líneas se hacen perpendiculares.

Finalmente, podemos introducir una fórmula de cálculo para r que comporta las cinco sumas previamente obtenidas en conexión con los cálculos de a y b . La fórmula es:

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}} \quad (\text{XVII.8})$$

El numerador, por supuesto, ha sido ya calculado, lo mismo que una parte del denominador. Así, pues, la correlación entre el porcentaje de negros y el índice de discriminación es:

$$\begin{aligned} r &= \frac{13(43\ 943.32) - (62.88)(8\ 557)}{\sqrt{[13(432.2768) - (62.88)^2][13(6\ 192\ 505) - (8\ 557)^2]}} \\ &= \frac{33\ 199}{110\ 120} = .301 \end{aligned}$$

Conviene observar que se pueden adicionar valores tanto a X como a Y , o sustraerlos, sin afectar el valor del coeficiente de correlación. De forma análoga, r no se verá afectado por un cambio de escala en cualquiera de las variables. Esto equivale a decir, de hecho, que la correlación entre el ingreso y la educación es la misma, ya sea que se mida el ingreso en dólares o en centavos. Sin embargo, aunque el coeficiente de correlación sea invariante en transformaciones de esta clase, la ecuación de los mínimos cuadrados, en cambio, no lo es. En efecto, la adición o sustracción de valores afecta el valor numérico de a . Y un cambio de escala afecta la pendiente de la línea. Así, por ejemplo, si cada X se divide entre 10 manteniendo a la Y fija, la b resultante se verá multiplicada por 10. El lector hará bien en verificar que estas propiedades se mantienen, examinando las fórmulas de r , a y b . Estos hechos pueden utilizarse con objeto de simpli-

ficar los cálculos. Así, por ejemplo, si X comporta un número muy grande o un decimal muy pequeño, un cambio de escala puede reducir el riesgo de errores de cálculo. O bien, si la variable X consta de valores tales como 1 207, 1 409, 1 949 y 1 568, se recomendará probablemente sustraer 1 000 de cada marca. Algunas rutinas de cálculo requieren que todos los valores sean positivos.

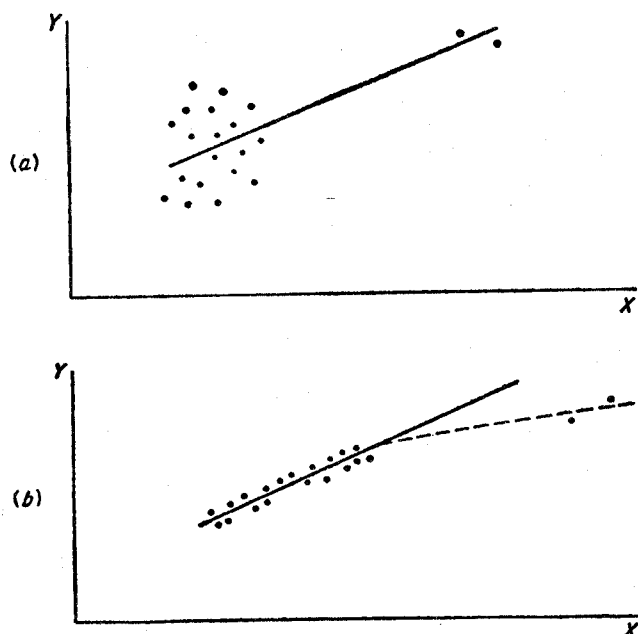


FIG. XVII.9. Diagramas de dispersión que muestran los efectos posibles de valores extremos de X .

Por lo tanto, al calcular r puede resultar necesario añadir a cada valor un número ligeramente superior a la marca negativa mayor.

Hay que tener presente, en este punto, otro hecho relativo a la correlación. Y es que, como quiera que esta medida comporta variancias y covariancias a la vez, se ve sumamente afectada por unos pocos valores extremos de cualquiera de las dos variables. Por otra parte, la magnitud de r depende del grado de variabilidad general de la variable independiente. Es lo que ilustra la figura XVII.9. En la figura XVII.9a, el efecto de uno o dos valores extremos produce una correlación moderadamente alta cuando no se da ninguna en los casos restantes. En la figura XVII.9b, tenemos una relación lineal moderadamente elevada, excepto en cuanto al hecho de que los casos extremos no quedan en línea recta con los demás. En este último caso tenemos probablen-

te un ejemplo de relación no lineal. El diagrama de dispersión resultará siempre útil para indicar la naturaleza de la situación en un problema determinado. Veamos ahora lo que puede hacerse cuando se presenta una u otra de estas situaciones.

La figura XVII.9a ilustra el punto anteriormente señalado de que la magnitud del coeficiente de correlación depende del mar-

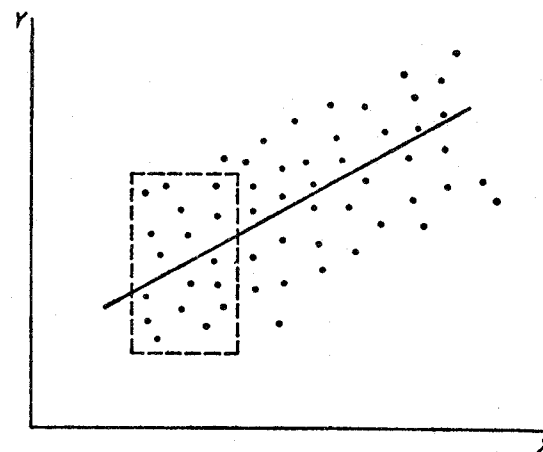


FIG. XVII.10. Diagrama de dispersión que no muestra relación alguna dentro de un recorrido limitado de variación de X , pero con relación positiva sobre el recorrido total.

gen de variabilidad de ambas variables. Si hubiera habido un número mayor de casos extremos, la distribución resultante habría podido ser como en la figura XVII.10. En este caso, la correlación conjunta podría ser alta, pero en el interior de cualquier recorrido limitado de las X la correlación puede ser vecina de cero. Esto indica de hecho que hay insuficiente variabilidad de X en el interior de dicho recorrido limitado para contrarrestar los efectos de las numerosas variables incontroladas. En realidad, X está siendo mantenida casi constante. Por consiguiente, si el diagrama de dispersión resulta ser semejante al de la figura XVII.9a, habría que tratar de extender el recorrido de variabilidad de X hallando más casos extremos.

Si la extensión del recorrido de variabilidad no resulta prácticamente posible, o si el interés del investigador se centra ante todo en casos menos extremos, será tal vez más razonable prescindir totalmente en el análisis de los casos extremos. Así, por ejemplo, supongamos que X es el tamaño de las ciudades y que la ciudad de Nueva York figura en la muestra. A menos que haya un gran número de ciudades de tamaño correspondiente, y no las hay, puede resultar necesario limitar la atención a ciudades de

menos de 500 000 habitantes. En algunos casos podrá parecer indicado calcular r tanto con los casos extremos como sin ellos. Es obvio que la decisión dependerá de la naturaleza del problema y del interés del sociólogo. El lector ha de percatarse bien del hecho de que una o dos marcas extremas pueden eventualmente ejercer un efecto muy pronunciado sobre el tamaño de r , hecho que en alguna forma debe tenerse siempre en cuenta. De ahí que el recorrido de variabilidad debiera consignarse juntamente con los coeficientes de correlación. Esto constituye otra ilustración del punto importante relativo a que una simple medida de resumen, por muy superior que sea respecto de otras, puede ser a menudo desorientadora.

Si los datos se presentan como en la figura XVII.9b, sospecharíamos, por supuesto, que no existe linealidad. Aquí también, pues, habría que obtener, de ser posible, más casos extremos. Si éstos son sólo uno o dos, resultará tal vez preferible excluirlos del análisis. Las situaciones de esta índole ilustran el hecho de que, al interior de cierto recorrido una relación de variación puede ser aproximadamente lineal, resultando en cambio inapropiada si se extiende el modelo lineal. De ahí, pues, que se imponga prudencia en cuanto a generalizar más allá de los límites de los datos. Un enunciado por el estilo de "dentro los límites de —y— la relación resulta ser aproximadamente lineal" será más apropiado.

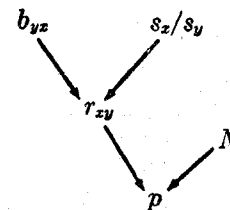
Comparación entre correlaciones y declives. Las observaciones anteriores acerca de la sensibilidad de los coeficientes de correlación ante las diferencias en la cantidad de variación de X , relativa a la dispersión producida por factores extraños, apunta uno de los problemas fundamentales con *cualquier* medida del grado de asociación. Nuestra atención debe estar centrada en la naturaleza de la ley que relaciona X y Y , de si la relación es o no es lineal, y, si lo es, en la magnitud del declive. Al comparar los resultados de dos estudios o de varias submuestras, debemos reconocer que es perfectamente posible obtener diferencias sustanciales entre los coeficientes de correlación, aun cuando se apliquen las mismas leyes (medidas por los declives). Es decir, que las r pueden diferir aunque no los declives, lo que puede ser debido únicamente a diferencias en la cantidad de variación en la variable independiente X , o a diferencias en la amplitud con que han sido sometidos a control otros factores extraños que producen variaciones aleatorias en Y . Como veremos al tratar del análisis de la covarianza, al buscar la interacción estamos en efecto buscando una diferencia entre declives, y no correlaciones. En el capítulo siguiente estudiaremos brevemente las pruebas para encontrar diferencias entre correlaciones, pero el lector debe estar prevenido acerca del peligro de que tales diferencias, una vez encontradas, puedan ser fácilmente mal interpretadas.

Puede ser útil concebir un coeficiente de correlación r_{xy} como función de dos tipos variables, con el declive b_{yx} y un factor s_x/s_y abarcando la razón de las dos desviaciones estándar que se aplican a la muestra o submuestra particular que nos ocupa. Así:

$$r_{xy} = b_{yx} (s_x/s_y)$$

El valor numérico de b_{yx} es, por supuesto, determinado no sólo por la ley que une a X con Y , sino también por la elección que el investigador hace entre las unidades de medida. El factor s_x/s_y es también una función de tales unidades, las que son por supuesto conocidas antes que los datos de la población o la muestra. Pero la razón s_x/s_y será también única para cada muestra (y σ_x/σ_y para cada población), y se utiliza para obtener la medida estandarizada r_{xy} . Un coeficiente de correlación tiene la ventaja de ser estandarizado, independizándolo así de la elección que se haga de unidades de medida, pero lamentablemente tiene que ser estandarizado en función de algo que resulta ser una cantidad no invariable en relación con muestras o poblaciones. Esta circunstancia debe ser claramente entendida, debiendo ser señalados *siempre* los declives no estandarizados, de modo que las réplicas no resulten desorientadoras a este respecto.

Planteando el asunto en forma algo diferente, podemos reconocer que en la inferencia y estimación estadísticas se da una jerarquía de metas científicas. Probamos buscando primero la significancia, para decidir si se ha encontrado una relación que no pueda ser fácilmente explicable por mecanismos casuales. Observamos a este respecto que el nivel de probabilidad o significación es función del grado de relación y del tamaño de la muestra. Si ésta es muy grande podremos obtener un pequeño nivel de probabilidad, incluso con una relación muy débil y tal vez sin importancia práctica. Pero habiendo encontrado al menos una relación moderadamente fuerte, se nos plantea de nuevo una tarea más importante, a saber: la de estimar la naturaleza de tal relación, medida por un coeficiente de regresión en el caso lineal. Cuando las correlaciones son moderadamente fuertes, en lugar de comparar estas r directamente, estimamos los declives, y los comparamos en nuestras pruebas de interacción. El proceso puede presentarse diagramáticamente así:



donde la dirección de las flechas representa el "curso causal" (por ejemplo: probabilidades influidas por magnitudes de relaciones y tamaños de muestras), lo que va frecuentemente en dirección opuesta a la que siguen los pasos del procedimiento empleado en un análisis estadístico. El diagrama indica que p es una función de dos variables, una de las cuales (el tamaño de la muestra) no es de interés inherente, y que la correlación r_{xy} es asimismo una función de dos factores, uno de los cuales (s_x/s_y), no es de interés. Nuestro objetivo consiste en llevar el análisis hacia arriba en el diagrama hasta la estimación de los coeficientes de regresión, en lugar de detenernos en los niveles de probabilidad, o formulando declaraciones en relación con los coeficientes de correlación.

Resulta que en cuantas ocasiones se manejan medidas ordinales de asociación, tales como las que se verán en el capítulo siguiente, desaparece la distinción entre declives y medidas de asociación. En el caso de dicotomías, sin embargo, puede demostrarse que si se sigue la regla de colocar la variable independiente al través de la parte alta del cuadro, y se computan las proporciones (o porcentajes) de modo que sumen 1.00 (o 100) hacia abajo, comparando a continuación de izquierda a derecha, la diferencia de proporciones resultante puede ser considerada como un caso especial del declive b_{yx} , en tanto que ϕ pasa a ser un caso especial de r_{xy} . Si se computan las proporciones en la otra dirección, la diferencia de proporciones pasa a ser un caso especial de b_{xy} , de modo que tendremos una justificación más para seguir la regla empírica previamente sugerida. Pueden obtenerse estos resultados por el simple procedimiento de asignar puntuaciones de 0 y 1 tanto a X como a Y , utilizando a continuación las fórmulas básicas para el cálculo de r_{xy} y b_{yx} .

* *Cálculos a partir de datos agrupados.* Si el número de casos es grande o si no se dispone de una calculadora moderna, el cálculo de los coeficientes de correlación puede resultar extremadamente laborioso. En tal caso será tal vez más indicado servirse de datos agrupados, aun a riesgo de introducir eventualmente algunas imprecisiones. En principio, estos cálculos de datos agrupados no son más que aplicaciones abreviadas de los procedimientos empleados para obtener la media y la desviación estándar. Tenemos ahora dos variables que han de clasificarse cruzadamente como en el cuadro XVII.2. Hemos de anticipar una media para cada variable, tomando desviaciones graduales de cada una de las medias y sirviéndonos de factores de corrección en cada caso. Además, necesitaremos un término de producto cruzado equivalente a Σxy . Como que las desviaciones tanto de X como de Y se tomarán de las medias *adivinadas* respectivas, necesitamos servirnos de un factor de corrección a sustraer del término del producto cruzado apreciado. Podemos modificar así las fórmulas

de cálculo de r y b de modo que se tenga en cuenta que nos hemos servido de medias adivinadas en lugar de las correctas.

Se recordará que una de las fórmulas de s sirviéndose de datos agrupados era (dejando de lado los subíndices):

$$s = \frac{i}{N} \sqrt{N \Sigma f d'^2 - (\Sigma f d')^2}$$

Como quiera que tenemos ahora dos variables, X y Y , nos serviremos de subíndices con objeto de distinguir las frecuencias y las desviaciones graduales de X (esto es, f_x y d'_x) de las de Y (o sea, f_y y d'_y). Al calcular el término del producto cruzado, necesitamos obtener también las frecuencias f_{xy} de cada subcasilla. Estas últimas serán por lo regular más pequeñas que f_x o f_y . Así, pues, si bien hay 24 casos en la categoría de 40.0 a 49.9 para la variable X y 30 casos en la categoría de 15.0 a 19.9 de Y , sólo hay 6 casos en la subcasilla correspondiente a ambas categorías. El lector ha de convencerse por sí mismo de que la fórmula de cálculo de r (ecuación XVII.8) puede modificarse como sigue:

$$r = \frac{N \Sigma f_{xy} d'_x d'_y - (\Sigma f_x d'_x)(\Sigma f_y d'_y)}{\sqrt{[N \Sigma f_x d'^2_x - (\Sigma f_x d'_x)^2][N \Sigma f_y d'^2_y - (\Sigma f_y d'_y)^2]}} \quad (\text{XVII.9})$$

Y en forma análoga, la fórmula de b se convierte en:

$$b = \frac{N \Sigma f_{xy} d'_x d'_y - (\Sigma f_x d'_x)(\Sigma f_y d'_y)}{N \Sigma f_x d'^2_x - (\Sigma f_x d'_x)^2} \frac{i_y}{i_x} \quad (\text{XVII.10})$$

en donde i_y e i_x representan las amplitudes de intervalos de Y y X respectivamente. El valor de a puede calcularse ahora a partir de la ecuación:

$$a = \frac{\Sigma Y - b \Sigma X}{N} = \bar{Y} - b \bar{X}$$

en donde \bar{X} y \bar{Y} pueden obtenerse sirviéndonos de la fórmula usual de los datos agrupados.

Calculemos ahora los valores en esos coeficientes en relación con los datos de 150 distritos del Sur consignados en el cuadro XVII.2. Tomaremos como variable dependiente Y , o sea el porcentaje de mujeres de la clase trabajadora, siendo la variable independiente el porcentaje de la población clasificada como granjas rurales. Convendrá servirse de una fórmula de cálculo como la que se da en el cuadro XVII.3. En ésta, los límites de

las clases y los puntos medios se indican horizontalmente en la parte superior (para Y) y de arriba abajo, a mano izquierda, para X. Obsérvese el área cerrada en el interior del cuadro. Se verá que hay tres números en cada subcasilla. En cada casilla, el número de arriba representa el número de casos de la subcasilla, tal como se da en el cuadro XVII.2. Los números restantes de la

CUADRO XVII.2. Datos clasificados cruzados para obtener correlaciones de datos agrupados

Porcentaje de granjas rurales, X	Porcentaje de mujeres de la clase trabajadora, Y								
	10.0-14.9	15.0-19.9	20.0-24.9	25.0-29.9	30.0-34.9	35.0-39.9	40.0-44.9	Totales	
0.0-9.9	0	0	0	1	8	4	0	13	
10.0-19.9	1	2	0	2	4	1	3	13	
20.0-29.9	2	5	1	2	3	3	0	16	
30.0-39.9	2	0	5	5	7	3	0	22	
40.0-49.9	4	6	6	7	1	0	0	24	
50.0-59.9	3	10	9	6	2	0	0	30	
60.0-69.9	2	4	3	7	4	0	0	20	
70.0-79.9	2	3	4	1	0	0	0	10	
80.0-89.9	1	0	1	0	0	0	0	2	
Totales	17	30	29	31	29	11	3	150	

FUENTE: Censo de los Estados Unidos de 1950.

subcasilla se emplean para calcular el término del producto cruzado. La cifra central de cada subcasilla representa el producto de las desviaciones graduales $d'_x d'_y$. Así, por ejemplo, en la subcasilla más baja de la izquierda (correspondiente a las categorías de 80.0 a 89.9 y de 10.0 a 14.9), la cifra -12 es el producto de 4 por -3. En otros términos: la categoría de 80.0 a 89.9 se halla 4 desviaciones graduales *por encima* de la media anticipada de X, y la categoría de 10.0 a 14.9 se encuentra 3 desviaciones graduales *por debajo* de la media anticipada de Y. Finalmente, el número inferior en cada subcasilla representa el producto de los dos números que tiene arriba y puede por consiguiente representarse simbólicamente como $f_{xy} d'_x d'_y$. Por lo tanto, la suma de estas cifras inferiores de todas las subcasillas nos da el término del producto cruzado, sin corrección de los errores introducidos sirviéndose de medias estimadas. Esta suma se empleará en el primer término del numerador de r ; es numéricamente igual a -200, y se ha dispuesto en el ángulo inferior derecho del cuadro.

Las cantidades restantes necesitadas en el cálculo de r y b pueden obtenerse en la forma usual. Las cuatro últimas columnas

CUADRO XVII.3. Cálculos de la correlación de datos agrupados *

Límites de clase	Y	10.0-14.9	15.0-19.9	20.0-24.9	25.0-29.9	30.0-34.9	35.0-39.9	40.0-44.9	f_x	d'_x	$f_x d'_x$	$f_x (d'_x)^2$
X	Puntos medios	12.45	17.45	22.45	27.45	32.45	37.45	42.45				
0.0-9.9	4.95					1	8	4				
						0	-4	-8	13	-4	-52	208
						0	-32	-32				
10.0-19.9	14.95	1	2		2	4	1	3				
		+9	+6		0	-3	-6	-9	13	-3	-39	117
		9	12		0	-12	-6	-27				
20.0-29.9	24.95	2	5	1	2	3	3					
		+6	+4	+2	0	-2	-4		16	-2	-32	64
		12	20	2	0	-6	-12					
30.0-39.9	34.95	2		5	5	7	3					
		+3		+1	0	-1	-2		22	-1	-22	22
		6		5	0	-7	-6					
40.0-49.9	44.95	4	6	6	7	1						
		0	0	0	0	0			24	0	0	0
		0	0	0	0	0						
50.0-59.9	54.95	3	10	9	6	2						
		-3	-2	-1	0	+1			30	1	30	30
		-9	-20	-9	0	2						
60.0-69.9	64.95	2	4	3	7	4						
		-6	-4	-2	0	+2			20	2	40	80
		-12	-16	-6	0	8						
70.0-79.9	74.95	2	3	4	1							
		-9	-6	-3	0				10	3	30	90
		-18	-18	-12	0							
80.0-89.9	84.95	1		1								
		-12		-4					2	4	8	32
		-12		-4								
f_y		17	30	29	31	29	11	3	N = 150		-37	643
d'_y		-3	-2	-1	0	1	2	3				
$f_y d'_y$		-51	-60	-29	0	29	22	9	-80		$\Sigma f_{xy} d'_x d'_y = -200$	
$f_y (d'_y)^2$		153	120	29	0	29	44	27	402			

* Esta forma de cálculo se ha tomado, con ligeras adaptaciones, de [1], cuadro XIX.4 de la p. 476, con la amable autorización del editor.

del cuadro se emplean para obtener $f_x, d'_x, f_x d'_x$ y $f_x (d'_x)^2$, las sumas de las dos últimas de estas cantidades utilizándose directamente en la fórmula de r . Obsérvese que al calcular los valores numéricos de estas cuatro columnas prescindimos por completo de los valores de Y . Así, pues, si dejamos totalmente de lado el área encerrada, tenemos exactamente la misma clase de tabla de la que nos servimos al calcular la media y la desviación estándar de datos agrupados. Y en forma análoga, las cuatro hileras inferiores pueden emplearse para obtener sumas correspondientes en relación con la variable Y . Todas las cantidades necesitadas en las fórmulas de r y b pueden ponerse ahora en las casillas inferiores de la derecha de la tabla mayor.

Obtenemos ahora los valores de r y b como sigue:

$$r = \frac{150(-200) - (-37)(-80)}{\sqrt{[150(643) - (-37)^2][150(402) - (-80)^2]}} = \frac{-32\,960}{71\,590} = -.460$$

$$b = \frac{150(-200) - (-37)(-80)}{150(643) - (-37)^2} \cdot \frac{5.0}{10.0} = \frac{-32\,960}{95\,081} \cdot \frac{1}{2} = -.1733$$

Como quiera que los valores de \bar{X} y \bar{Y} son 42.48 y 24.78, respectivamente, obtenemos:

$$a = \bar{Y} - b\bar{X} = 24.78 - (-.1733)(42.48) = 32.14$$

y la ecuación de los mínimos cuadrados puede escribirse como:

$$Y_p = 32.14 - .1733X$$

Interpretación del coeficiente de correlación. Con objeto de obtener una interpretación de r que tenga sentido cuando r no es ni cero ni 1.0, volvamos al concepto de variabilidad a propósito de la ecuación de regresión. Hemos definido la variancia respecto de la media de Y como:

$$\sigma_y^2 = \frac{\sum(Y - \mu_y)^2}{M}$$

en donde M representa la magnitud de la población (frente al tamaño de la muestra N) y donde nos servimos de los subíndices para recalcar el hecho de que tenemos ahora dos variables que han de distinguirse. Así, pues, el concepto corriente de la variancia comporta desviaciones respecto de una medida *fija* de tendencia central, o sea la media conjunta. Pero podemos obtener

también la media de las Y para una X fija, y estamos suponiendo que estos valores varían con X de manera que produzcan una regresión lineal. Podemos generalizar en esta forma el concepto de la media, obteniendo una especie de media condicional de Y para una X dada, que podemos simbolizar como $\mu_{y|x}$ o como $E(Y|X)$.

Si generalizamos el concepto de variancia en forma similar, podemos obtener una medida de dispersión respecto de la ecuación de regresión tal como:

$$\sigma_{y|x}^2 = \frac{\sum(Y - \mu_{y|x})^2}{M} \quad (\text{XVII.11})$$

en donde el símbolo $\sigma_{y|x}^2$ se emplea para señalar el hecho de que la magnitud de la variabilidad respecto de la ecuación de regresión, lo mismo que la media de Y , depende del valor de X . En otros términos: para cada X se dan tanto una media de las Y como una variancia respecto de dicha media. La cantidad de dispersión alrededor de la línea no necesita ser siempre la misma para cada X , pese a que vamos a *suponer* la propiedad de homocedasticidad o de variancias iguales.

Tenemos ahora dos medidas de variabilidad para Y . La primera mide la dispersión alrededor del valor de Y , la gran media μ_y , que sería el mejor valor anticipado de Y si no se conociera X . En otros términos: si se nos pidiera anticipar Y no conociendo X , la mejor anticipación sería μ_y (o \bar{Y} , si sólo se dispusiera de los datos de la muestra). En cambio, si conociéramos X , anticiparíamos el valor correspondiente de Y que se sitúa en la ecuación de regresión. A menos que no existiera relación entre X y Y , el conocimiento de X nos ayudará a predecir el valor de Y . Si la relación fuera perfecta, podríamos predecir Y exactamente, ya que todos los puntos quedarían exactamente sobre la línea. Por lo regular, no estaremos en condiciones de hacerlo así, pero, como quiera que estamos suponiendo una distribución normal de las Y y una desviación estándar $\sigma_{y|x}$ fija, podemos emitir enunciados de probabilidad acerca de los riesgos y de la magnitud del error. Y lo que es más importante todavía desde el punto de vista de nuestros propósitos, podemos comparar las dos desviaciones estándar (o variancias) y obtener una medida acerca de en qué proporción se ha mejorado la anticipación por el conocimiento de X . Al proceder en esta forma, podemos servirnos de procedimientos con los que estamos ya familiarizados a partir del análisis de la variancia.

En dicho análisis, en efecto, tomamos la variación total o suma de cuadrados y descompusimos dicha cantidad en porciones explicadas e inexplícadas. Vamos a servirnos ahora exacta-

mente del mismo procedimiento, obteniendo casi a manera de producto accesorio los valores de $\sigma_{y|x}^2$ y r^2 . Con lo que estaremos en condiciones de dar una interpretación lógica del coeficiente de correlación. Primero, podemos expresar las desviaciones de cada Y respecto de \bar{Y} como suma de dos cantidades $(Y - Y_p) + (Y_p - \bar{Y})$ (véase la figura XVII.11). La primera de estas cantida-

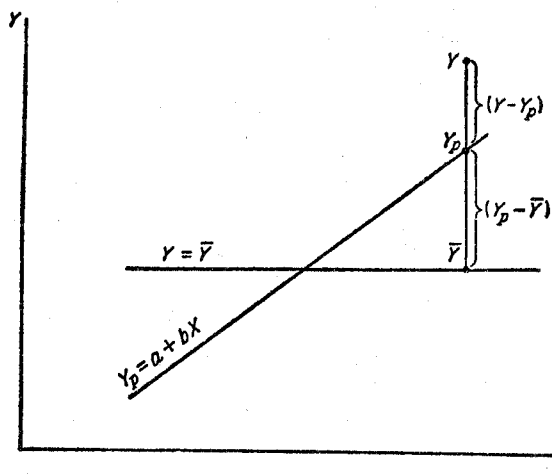


FIG. XVII.11. Representación geométrica que muestra las desviaciones respecto de la media \bar{Y} como una suma de desviaciones respecto de la recta de mínimos cuadrados y desviaciones de la recta de mínimos cuadrados respecto de la \bar{Y} .

des representa la desviación del valor de Y respecto de la línea de los mínimos cuadrados e indica la cantidad de error que se comete cuando se emplea Y_p para predecir Y . La segunda expresión, en cambio, indica la desviación de la línea de mínimos cuadrados (para una X dada) respecto de \bar{Y} . En la mayoría de los casos, esta cantidad representará el monto en que se reduce el error al conocer Y_p . Si elevamos al cuadrado ahora ambos miembros de la ecuación y sumamos luego todos los casos, obtenemos:

$$\Sigma(Y - \bar{Y})^2 = \Sigma(Y - Y_p)^2 + 2\Sigma(Y - Y_p)(Y_p - \bar{Y}) + \Sigma(Y_p - \bar{Y})^2$$

Afortunadamente, el término central vuelve a desaparecer, y nos quedamos con:

$$\Sigma(Y - \bar{Y})^2 = \Sigma(Y - Y_p)^2 + \Sigma(Y_p - \bar{Y})^2 \quad (\text{XVII.12})$$

$$\text{SC total} = \text{SC inexplicada} + \text{SC explicada}$$

La primera cantidad de la derecha de la ecuación representa la suma de los cuadrados de las desviaciones de los valores reales de Y respecto de la línea de los mínimos cuadrados. Esta cantidad es inexplicada, ya que indica la magnitud del error en la predicción. Y la cantidad restante indica lo que hemos ganado al servirnos de Y_p con preferencia a \bar{Y} , pudiendo designarse como la suma de cuadrados explicada. Por *explicada* no entendemos, por supuesto, una explicación causal, sino simplemente una asociación entre las dos variables. Consideremos ahora más de cerca cada una de estas cantidades.

Si tomamos una suma de cuadrados inexplicada y dividimos entre el número total de casos, obtenemos la variancia de la muestra $s_{y|x}^2$ respecto de la línea de los mínimos cuadrados. O sea:

$$s_{y|x}^2 = \frac{\Sigma(Y - Y_p)^2}{N} \quad (\text{XVII.13})$$

Si deseamos obtener una estimación insesgada de la variancia de la población $\sigma_{y|x}^2$ respecto de la regresión real, hemos de dividir no entre N sino entre los grados apropiados de libertad. En este caso hemos perdido 2 grados de libertad al calcular a y b como estimaciones de α y β . Por consiguiente, si deseamos estimar $\sigma_{y|x}^2$ nos serviremos de:

$$\hat{\sigma}_{y|x}^2 = \frac{\Sigma(Y - Y_p)^2}{N - 2} \quad (\text{XVII.14})$$

En esta forma, la suma de cuadrados inexplicada puede convertirse fácilmente en una estimación de la variancia respecto de la ecuación de regresión. El lector hará bien en convencerse por sí mismo de que lo que hemos hecho es directamente paralelo a nuestro tratamiento anterior del análisis de la variancia. La variabilidad respecto de la ecuación de mínimos cuadrados ha sustituido la noción de variabilidad en el *interior* de las categorías de X .

Volviendo ahora a la suma de cuadrados explicada $\Sigma(Y_p - \bar{Y})^2$, podemos mostrar fácilmente que esta cantidad es equivalente a $r^2[\Sigma(Y - \bar{Y})^2]$, o $r^2\Sigma y^2$. Como quiera que $Y_p = a + bX$ y $\bar{Y} = a + b\bar{X}$, tenemos:

$$(Y_p - \bar{Y}) = b(X - \bar{X})$$

Por consiguiente:

$$\begin{aligned}
\Sigma(Y_p - \bar{Y})^2 &= b^2 \Sigma(X - \bar{X})^2 = b^2 \Sigma x^2 \\
&= \frac{(\Sigma xy)^2}{(\Sigma x^2)^2} (\Sigma x^2) = \frac{(\Sigma xy)^2}{\Sigma x^2} \\
&= \frac{(\Sigma xy)^2}{\Sigma x^2 \Sigma y^2} (\Sigma y^2) = r^2 \Sigma y^2 \\
&= r^2 \Sigma(Y - \bar{Y})^2
\end{aligned}$$

Hemos demostrado así que:

$$r^2 = \frac{\Sigma(Y_p - \bar{Y})^2}{\Sigma(Y - \bar{Y})^2} = \frac{\text{SC explicada}}{\text{SC total}}$$

Por medio de un razonamiento similar pudimos haber demostrado que r^2 representa la razón de la variación explicada en X a la variación total en X . Por lo tanto, el cuadrado del coeficiente de correlación puede interpretarse como la proporción de variación total en una de las variables explicada por la otra. La cantidad de $\sqrt{1-r^2}$, designada a menudo como *coeficiente de alienación*, representa la raíz cuadrada de la proporción de la suma total de cuadrados que permanece sin explicar por la variable independiente.

Cabe observar que no se da interpretación directa y simple alguna de la r misma. De hecho, es posible dejarse desorientar por los valores de r , ya que estos valores serán numéricamente mayores que los de r^2 (a menos que r sea 0 o ± 1.0). Así, por ejemplo, podría parecer que una r de .5 sea la mitad de buena que una correlación perfecta, en tanto que vemos que, en este caso, sólo explicamos un 25 por ciento de la variación. Una correlación de .7 indica que algo menos de la mitad de la variación resulta explicada. Vemos asimismo que correlaciones de .3 o menores significan que sólo una fracción muy pequeña de la variación es explicada. El cuadro XVII.4 indica las relaciones entre las diversas cantidades.

Como quiera que $1-r^2$ representa la proporción de variación inexplicada, tenemos:

$$(1-r^2)[\Sigma(Y - \bar{Y})^2] = \Sigma(Y - Y_p)^2$$

Por consiguiente:

$$(1-r^2) \frac{\Sigma(Y - \bar{Y})^2}{N} = \frac{\Sigma(Y - Y_p)^2}{N}$$

o bien:

$$(1-r^2)s_y^2 = s_{y|x}^2$$

De donde:

$$s_{y|x} = \sqrt{1-r^2} s_y$$

Este resultado nos proporciona una indicación acerca de en qué medida podemos reducir la desviación estándar conociendo X .

CUADRO XVII.4. Relaciones numéricas entre r , r^2 , $1-r^2$ y $\sqrt{1-r^2}$

r	r^2	$1-r^2$	$\sqrt{1-r^2}$
.90	.81	.19	.44
.80	.64	.36	.60
.70	.49	.51	.71
.60	.36	.64	.80
.50	.25	.75	.87
.40	.16	.84	.92
.30	.09	.91	.95
.20	.04	.96	.98
.10	.01	.99	.995

(Véase la última columna del cuadro XVII.4.) Si r es cero, las dos desviaciones estándar son iguales. Este hecho es obvio, por supuesto, si nos percatamos de que la línea de los mínimos cuadrados será en tal caso una recta horizontal de ecuación $Y = \bar{Y}$. Si r^2 es igual a la unidad, $s_{y|x}$ será cero, por supuesto, ya que todos los puntos quedarán exactamente sobre la recta. Del cuadro XVII.4 se desprende que la magnitud de r ha de ser grande para que obtengamos una reducción sustancial de las desviaciones estándar. Para una r de .80, la desviación estándar respecto de la línea de los mínimos cuadrados es de .60 de la desviación estándar corriente; pero, con una r de .40, vemos que no hemos ganado mucho en cuanto a apreciar Y a partir de X .

GLOSARIO

Distribución normal bivariada
 Coeficiente de alienación
 Coeficiente de correlación
 Covariancia
 Intercepción
 Ecuación de los mínimos cuadrados
 Regresión de Y sobre X
 Declive.

EJERCICIOS

1. Los siguientes datos relativos a 29 ciudades de 100 mil o más habitantes de regiones fuera del Sur están tomados del estudio de R. C. Angell sobre la integración moral de las ciudades norteamericanas. El índice de integración moral se ha derivado combinando los índices de tasas de criminalidad con los de la labor de mejoramiento. La heterogeneidad se midió en términos de los números relativos de los no blancos y los blancos nacidos en el extranjero contenidos en la población. Y se calculó asimismo, a título de segunda variable independiente, un índice de movilidad, que mide los números relativos de las personas que se establecen o dejan la ciudad.

Ciudad	Índice de integración	Índice de heterogeneidad	Índice de movilidad
Rochester	19.0	20.6	15.0
Syracuse	17.0	15.6	20.2
Worcester	16.4	22.1	13.6
Erie	16.2	14.0	14.8
Milwaukee	15.8	17.4	17.6
Bridgeport	15.3	27.9	17.5
Buffalo	15.2	22.3	14.7
Dayton	14.3	23.7	23.8
Reading	14.2	10.6	19.4
Des Moines	14.1	12.7	31.9
Cleveland	14.0	39.7	18.6
Denver	13.9	13.0	34.5
Peoria	13.8	10.7	35.1
Wichita	13.6	11.9	42.7
Trenton	13.0	32.5	15.8
Grand Rapids	12.8	15.7	24.2
Toledo	12.7	19.2	21.6
San Diego	12.5	15.9	49.8
Baltimore	12.0	45.8	12.1
South Bend	11.8	17.9	27.4
Akron	11.3	20.4	22.1
Detroit	11.1	38.3	19.5
Tacoma	10.9	17.8	31.2
Flint	9.8	19.3	32.2
Spokane	9.6	12.3	38.9
Seattle	9.0	23.9	34.2
Indianapolis	8.8	29.2	23.1
Columbus	8.0	27.4	25.0
Portland (Ore.)	7.2	16.4	35.8

FUENTE: R. C. Angell, "The Moral Integration of American Cities" ("La integración moral de las ciudades norteamericanas"), *American Journal of Sociology*, vol. 57, 2ª parte, p. 17, julio de 1951, con la amable autorización del autor y el editor. (Copyright 1951 de la Universidad de Chicago).

- Trácese un diagrama de dispersión que relacione la integración moral con la heterogeneidad.
 - Calcúlense r , a y b para las mismas variables, y trácese en el diagrama de dispersión la línea de mínimos cuadrados, tomando la integración moral como Y . Respuesta, $r = -.156$; $a = 13.9$; $b = -.049$.
 - ¿De cuánto es la desviación estándar respecto de la línea de los mínimos cuadrados comparada con la desviación estándar respecto de Y ?
2. Con objeto de resolver los ejercicios del capítulo XIX, se necesitará obtener las correlaciones entre la integración moral y la movilidad, así como entre la heterogeneidad y la movilidad. Calcúlense las dos r . Respuesta, $r = -.456$; $r = -.513$.
3. Agrúpense los índices de integración moral y heterogeneidad en intervalos y calcúlense r , a y b sirviéndose de las fórmulas de datos agrupados. Compárense los resultados con los datos sin agrupar.

BIBLIOGRAFÍA

- Blalock, H. M.: *Causal Inferences in Nonexperimental Research*. University of North Carolina Press, Chapel Hill, 1964, caps. 2 y 3.
- Christ, Carl: *Econometric Models and Methods*, John Wiley & Sons, Inc., Nueva York, 1966, Parte III.
- Croxtton, F. E., y D. J. Cowden: *Applied General Statistics*, 3ª ed., Prentice-Hall, Inc.: Englewood Cliffs, N. J. 1967, caps. 19 y 20.
- Hagood, M. J., y D. O. Price: *Statistics for Sociologists*, Henry Holt and Company Inc., Nueva York, 1952, cap. 23.
- Hays, W. L.: *Statistics*, Holt, Rinehart and Winston. Inc., Nueva York, 1963, cap. 15.
- Johnston, J.: *Econometric Methods*, McGraw-Hill Book Company, Nueva York, 1963, Parte II.
- McCullough, C., y L. Van Atta: *Introduction to Descriptive Statistics and Correlation*, McGraw-Hill Book Company, Nueva York, 1965, caps. 5-8.
- Mueller, J. H., K. Schuessler, y H. L. Costner: *Statistical Reasoning in Sociology*, 2ª ed., Houghton Mifflin Company, Boston, 1970, cap. 11.
- Wallis, W. A., y H. V. Roberts: *Statistics: a New Approach*, The Free Press of Glencoe, Ill., Chicago, 1956, cap. 17.
- Weinberg, G. H., y J. A. Schumaker: *Statistics: An intuitive Approach*, Wadsworth Publishing Company, Inc., Belmont, Cal., 1962, caps. 16-18.

XVIII. CORRELACIÓN Y REGRESIÓN [conclusión]

EN EL presente capítulo proseguimos el examen de la correlación y la regresión. Se tratarán primero algunas pruebas de significación, a continuación de lo cual pasaremos a las relaciones no lineales, tema que se examinará asimismo brevemente en el capítulo XIX. A continuación estudiaremos los efectos de los errores de medición en las pendientes y las correlaciones. Finalmente, se examinará el tema de la correlación grado-orden.

XVIII.1. Prueba de significación e intervalos de confianza

Prueba de significación de r y b . Como quiera que r y los coeficientes de mínimos cuadrados a y b sólo describen los datos de las muestras, nuestro interés se centra por lo regular en los parámetros correspondientes de las poblaciones, ρ , α y β . En particular, desearíamos probar la hipótesis nula de que no hay relación (lineal) alguna en la población, o podemos querer obtener intervalos de confianza para ρ o para los coeficientes de regresión. Examinaremos primero la prueba de la hipótesis nula en el sentido de que no se da relación en la población. Según veremos, si podemos suponer una distribución normal de Y acerca de X y homocedasticidad, podemos también servirnos del análisis de la variancia para verificar la hipótesis de que $\rho = \beta = 0$.

Sirvámonos del hecho de que, toda vez que r y b (y, por consiguiente, también ρ y β) tienen los mismos numeradores, una verificación de la hipótesis de que $\rho = 0$ lo es asimismo de la hipótesis $\beta = 0$ y viceversa. En otros términos: si no se da asociación lineal en la población, la pendiente de la ecuación de regresión será cero y, por tanto, la línea será horizontal. Recordando que la ecuación de regresión representa el camino de las medias de las Y para valores fijos de X , vemos inmediatamente que siempre que $\beta = 0$, las medias de las Y han de ser las mismas para todos los valores de X (véase figura XVIII.1). Esto implica, por supuesto, que la ecuación de regresión sea realmente de forma lineal. En particular, si dividiéramos el eje de las X en cierto número de categorías, encontraríamos que las medias de las categorías de la población son exactamente iguales. Así, pues, podemos traducir la hipótesis de que $\beta = \rho = 0$ en el enunciado de que las medias de Y serán iguales para cada una de las categorías de X . Si nos imaginamos una población infinita, como habrá que hacerlo para satisfacer el supuesto de normalidad, podemos concebir el eje de las X como dividido en un número indefinido de categorías, cada una de las cuales tenga medias idénticas en Y . En esta forma, nuestra hipótesis cero se con-

vierte en $\mu_{y1} = \mu_{y2} = \mu_{y3} = \dots$, en donde nos servimos del subíndice doble para recalcar que son las medias de las Y las que nos interesan y que tenemos un número indefinidamente grande de categorías X .

El curso del razonamiento anterior sugiere obviamente una extensión de la prueba de análisis de variancia para abarcar un

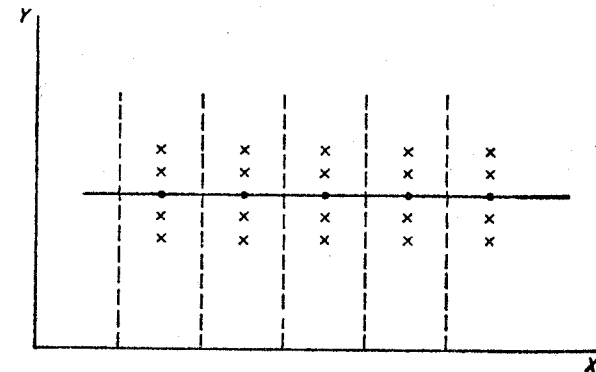


FIG. XVIII.1. Representación geométrica del hecho de que la hipótesis de $\beta = 0$ es equivalente a la hipótesis $\mu_1 = \mu_2 = \dots = \mu_k$.

número indefinidamente grande de categorías de la variable de escala nominal (ahora X). Recordemos los supuestos requeridos en el análisis de variancia. Además de la hipótesis nula y del supuesto de que los casos se han muestreado aleatoria e independientemente de cada una de las categorías, hemos de suponer también poblaciones normales y variancias iguales dentro de cada categoría. A condición, pues, de que podamos suponer también muestreo aleatorio, vemos que todos estos supuestos pueden cumplirse si suponemos que la distribución conjunta de X y Y sea normal bivariable. El lector recordará que este último supuesto nos asegura simultáneamente una ecuación de regresión lineal, normalidad de las Y para cada valor fijo de X e iguales variancias para todos los valores de esta variable. De hecho, pues, los supuestos de muestreo al azar y de normalidad bivariable nos capacitan para servirnos del análisis de variancia con objeto de verificar la hipótesis de que $\rho = \beta = 0$, aun cuando no se requiere la normalidad de las X en tanto las ϵ_i tengan una distribución aproximadamente normal.

Anteriormente encontramos que era necesario obtener las sumas totales de cuadrados y la de entre clases y restarlas, con objeto de obtener la suma de cuadrados dentro. Sin embargo, al verificar la hipótesis de que $\rho = 0$, el proceso se simplifica con-

siderablemente. Ya vimos, en efecto, que la proporción de la suma de cuadrados total de la Y explicada por X nos es dada por r^2 . Y en forma análoga, la proporción que dejamos inexplicada por X será $1 - r^2$. Como quiera que la suma total de cuadrados puede simbolizarse con Σy^2 , las sumas de cuadrados explicada e inexplicada se convierten en $r^2 \Sigma y^2$ y $(1 - r^2) \Sigma y^2$ respectivamente.

Los grados de libertad asociados a la suma total de cuadrados son, por supuesto, $N - 1$. Al calcular la suma inexplicada de cua-

CUADRO XVIII.1. Prueba de análisis de variancia de la hipótesis $\rho = 0$

	Suma de cuadrados	Grados de libertad	Apreciaciones de la variancia	F
Total	Σy^2	$N - 1$		
Explicada	$r^2 \Sigma y^2$	1	$\frac{r^2 \Sigma y^2}{1}$	$\frac{r^2 (N - 2)}{(1 - r^2)}$
Inexplicada	$(1 - r^2) \Sigma y^2$	$N - 2$	$\frac{(1 - r^2) \Sigma y^2}{N - 2}$	

drados, tomamos la suma de las desviaciones al cuadrado respecto de la línea de mínimos cuadrados, y no respecto de la gran media de las Y. Pero, con objeto de obtener la línea de los mínimos cuadrados, hemos de servirnos de los dos coeficientes a y b . Por consiguiente, hemos perdido 2 grados de libertad, o sea uno más de los que perdimos al tomar las desviaciones respecto del valor particular de \bar{Y} . Podemos, pues, asociar $N - 2$ grados con la suma inexplicada de cuadrados y, restando, vemos que hay que asociar un grado de libertad a la suma de cuadrados explicada.

Los resultados pueden resumirse ahora como en el cuadro XVIII.1. La ventaja de insertar símbolos en lugar de números en tabla está en que vemos inmediatamente que la cantidad Σy^2 desaparece cuando formamos la razón de las apreciaciones explicadas a las inexplicadas. En otros términos: la suma total de cuadrados se elimina, y podemos escribir una fórmula de F en términos de las proporciones de las sumas de cuadrados explicada e inexplicada. De este modo, la fórmula de F sólo comporta las cantidades r^2 y $1 - r^2$, junto con los grados de libertad de $N - 2$ y 1. Podemos, por consiguiente, servirnos de la fórmula:

$$F_{1, N-2} = \frac{r^2}{1 - r^2} (N - 2) \quad (\text{XVIII.1})$$

sin tener que ocuparnos en construir una tabla de análisis de

variancia, como fue el caso en el capítulo anterior. Como los cuadros para F sólo admiten pruebas a los niveles de .05, .01 y .001, puede resultar preferible tomar la raíz cuadrada positiva de (XVIII.1) y utilizar la distribución t , con $N - 2$ grados de libertad.

Podemos ilustrar el empleo de esta prueba de análisis de variancia para la significancia de r con los datos del cuadro XVII.1. Obtuvimos allí una correlación de $r = .301$ entre el porcentaje de negros y nuestro índice de discriminación. Al verificar en relación con el significado de r hacemos en realidad la importante pregunta: "¿Con qué probabilidad obtendríamos una r de .301 o mayor (en valor absoluto) si no hubiera efectivamente asociación lineal alguna en la población?" Con objeto de efectuar la prueba F , calculamos simplemente r^2 y $1 - r^2$ y nos servimos de la ecuación XVIII.1. Así, pues, ya que r se basaba en 13 casos, tenemos:

$$F_{1,11} = \frac{(.301)^2}{[1 - (.301)^2]} 11 = \frac{.0906}{.9094} 11 = 1.10$$

Refiriéndonos a la tabla F , vemos que para 1 y 11 grados de libertad necesitamos una F de 4.84 o mayor para descartar al nivel de .05 suponiendo que la dirección no hubiese sido establecida con anticipación. Decidimos, por consiguiente, no descartar la hipótesis nula de que $\rho = 0$. Aparentemente podríamos haber obtenido una r de .301 o mayor, simplemente por casualidad, aun si no se diera asociación alguna en la población.

Una vez más, es necesario insistir en la diferencia entre una prueba de significación y una medida del grado de relación. Si hubiéramos obtenido una r de .301 con un tamaño de muestra de 50, habríamos tenido:

$$F_{1,48} = \frac{.0906}{.9094} 48 = 4.78$$

o sea un valor significativo al nivel de .05. En ambos casos hemos explicado aproximadamente el 9 por ciento de la variación total de la muestra, pero en el último de ellos tenemos más confianza, aunque ligeramente, de que se da una relación en la población.

Intervalos de confianza. Siempre que pueda presuponerse o apreciarse aproximadamente una población normal bivariable, es posible construir intervalos de confianza para ρ y β , así como la línea de regresión. El error estándar de r nos está dado por la fórmula.

$$\sigma_r = \frac{1 - \rho^2}{\sqrt{N - 1}}$$

Por desgracia, la distribución de muestreo de r no será por lo regular simétrica, excepto en el caso especial en que $\rho = 0$. En efecto, la distribución de selección se distorsiona más y más a medida que el valor absoluto de ρ se aproxima a la unidad. Además, observamos que, para poder servirnos de la fórmula anterior del error estándar de r , necesitaríamos conocer o poder apreciar el valor de ρ . Estas dos complicaciones hacen que sea difícil obtener intervalos de confianza para ρ en forma abreviada.

Al calcular un intervalo de confianza respecto de r , convertimos primero r en una nueva estadística z que tiene una distribución de muestreo aproximadamente normal. Ponemos luego un intervalo de confianza alrededor de z en la forma habitual. Finalmente, una vez anotados los límites superior e inferior de confianza de z , reconvertimos estos valores particulares de z en r , con lo que obtenemos los límites de confianza de esta última.

Transformamos r en z por medio de la fórmula:

$$z = 1.151 \log \frac{1+r}{1-r}$$

en donde z puede tomar valores de cero al infinito. Conviene llamar la atención del lector acerca del hecho de que el valor z calculado mediante la fórmula anterior no tiene en absoluto conexión alguna con los valores de Z que utilizamos con la curva normal estándar. Los valores de z pueden obtenerse directamente del cuadro K, Apéndice 2, en lugar de servirse de los logaritmos. Los dos primeros dígitos de r se buscan de arriba abajo en el margen izquierdo, en tanto que el tercero se localiza horizontalmente en la parte superior. Los valores de z correspondientes están dados en el cuerpo del cuadro. Así, por ejemplo, una z de 0.3228 corresponde a una r de .312; una z de 1.3892 corresponde a una r de .883. Al servirnos del cuadro K, prescindimos del signo de r , asignando a z el signo correspondiente una vez hallado su valor numérico. Obsérvese que los valores de z sólo son ligeramente mayores que r cuando $|r| \leq .40$, pero a medida que r crece, z empieza a tomar valores mayores que la unidad.

Podemos servirnos ahora de la transformación de z en un problema de intervalo de confianza. La distribución de selección de z es aproximada a la normal, aun para N pequeñas y desviaciones moderadas de la normalidad bivariada. Su error estándar nos está dado por:

$$\sigma_z = \frac{1}{\sqrt{N-3}} \quad (\text{XVIII.2})$$

Y esto no sólo permite servirse de la tabla normal, sino que he-

mos eliminado además la necesidad de haber estimado ρ , ya que el error estándar de z sólo depende de N . Tomando como ejemplo numérico la correlación de .301 entre el porcentaje de negros y la discriminación, hallamos que el valor correspondiente de z es de 0.3106. Como quiera que no había más que 13 casos, tenemos:

$$\sigma_z = \frac{1}{\sqrt{13-3}} = \frac{1}{\sqrt{10}} = 0.3162$$

Supóngase que deseamos obtener para ρ un intervalo de confianza del 95 por ciento. Primero calculamos dicho intervalo en términos de valores de z . Así, pues, tomaríamos:

$$z \pm 1.96\sigma_z = 0.3106 \pm 1.96(0.3162) \\ = 0.3106 \pm 0.6198$$

Por consiguiente, el intervalo de confianza alrededor de z va de -.3092 a +.9304. Obsérvese que para obtener el límite inferior tuvimos que restar un número mayor, numéricamente, que 0.3106. Esto da un resultado negativo, lo cual significa a su vez que el valor de r correspondiente a dicho límite inferior ha de tomarse también como negativo. Buscando los valores de r correspondientes a los dos límites de confianza de z , obtenemos los valores de -.300 y .731 para los límites inferior y superior respectivamente.

Obsérvese que el intervalo no es totalmente simétrico en relación con el valor de .301 obtenido para r . En este caso, el límite superior está algo más cerca de r que el límite inferior. Si hubiéramos hallado una r de .80, el intervalo resultante habría estado todavía más distorsionado en la misma dirección. Puede comprenderse intuitivamente que esto sea así si tenemos presente que, siempre que empezamos a acercarnos al límite superior de la unidad, ponemos también una restricción al límite superior del intervalo de confianza. En esta forma, resultaría imposible, por ejemplo, obtener un intervalo de confianza de .86 a .16. Si ocurre que r sea negativa, la dirección de la distorsión será opuesta, por supuesto, a la anterior. El intervalo solamente llegará a ser simétrico en relación con r cuando ésta sea igual a cero.

Podemos interpretar este intervalo de confianza en la forma habitual. Nuestro procedimiento es tal que a la larga podemos esperar obtener intervalos que incluyan el valor (fijo) de ρ el 95 por ciento de las veces. Podemos también utilizar tales intervalos de confianza como verificaciones implícitas de hipótesis. En el problema anterior, en efecto, ya hemos observado que el

límite inferior del intervalo es negativo. Y como quiera que cero está incluido en el intervalo, sabemos inmediatamente que no descartaríamos la hipótesis nula de que $\rho = 0$. Y si quisiéramos verificar algún otro valor supuesto de ρ , procederíamos igual. Si por ejemplo hubiéramos anticipado que $\rho = .80$, habríamos descartado al nivel de .05, ya que este valor cae fuera del límite superior de .731.

Sería conveniente también calcular intervalos de confianza a propósito de otras medidas de relación. Por desgracia, se conoce demasiado poco acerca de las distribuciones de muestreo de la mayoría de las medidas de asociación en materia de problemas de contingencia para poder construir intervalos de confianza en relación con ellas. Haggard [11] sugiere un método para computar intervalos de confianza acerca de r_i o correlación interclase, y Goodman y Kruskal [10] discuten la distribución de muestras de varias medidas nominales y ordinales.

Ocasionalmente se quiere poder poner un intervalo de confianza con referencia a b , o se puede tener necesidad de encontrar un cinturón a cuyo interior pueda esperarse que la verdadera ecuación de regresión se encuentre. En ambos casos podemos servirnos de la distribución t en forma relativamente directa. La apreciación del error estándar de b está dada por:

$$\hat{\sigma}_b = \frac{\hat{\sigma}_{y|x}}{\sqrt{\sum_{i=1}^N (X_i - \bar{X})^2}} \quad (\text{XVIII.3})$$

en donde se recordará que:

$$\hat{\sigma}_{y|x} = \sqrt{\sum_{i=1}^N \frac{(Y_i - \bar{Y}_p)^2}{N-2}}$$

Con fines de cálculo puede demostrarse algebraicamente que:

$$\hat{\sigma}_{y|x} = \sqrt{\frac{\sum_{i=1}^N (Y_i - \bar{Y})^2 - b \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{N-2}} \quad (\text{XVIII.4})$$

Podemos servirnos ahora de los cálculos numéricos obtenidos ya para los datos de discriminación del cuadro XVII.1, con lo que obtenemos:

$$\hat{\sigma}_{y|x} = \sqrt{\frac{560\,024 - 19.931(2\,553.77)}{11}} = \sqrt{46\,284} = 215.1$$

y

$$\hat{\sigma}_b = \frac{215.1}{\sqrt{128.131}} = \frac{215.1}{11.32} = 19.00$$

Si deseamos calcular el intervalo de confianza del 99 por ciento, recurrimos directamente a la tabla t y nos servimos de $N-2$ u 11 grados de libertad. Obtenemos en esta forma:

$$b \pm (3.106)(19.00) = 19.931 \pm 59.014$$

* Al apreciar la ecuación de regresión, vemos que nuestra mejor apreciación singular (de "punto") es la línea de los mínimos cuadrados. Como quiera que la cantidad que estamos apreciando ahora ya no es un valor singular, sino una línea entera, nuestra apreciación del intervalo ya tampoco será un intervalo, sino una banda a ambos lados de la línea de mínimos cuadrados. De buenas a primeras podría esperarse que dicha banda consistiera en dos líneas paralelas a la de los mínimos cuadrados. Sin embargo, semejante banda implicaría que conocemos la verdadera pendiente y que la única fuente de error está en la apreciación de a . Hemos de recordar que se aprecian ahora dos cantidades (α y β), y, por lo tanto, tenemos dos fuentes de error. El lector ha de percatarse por sí mismo de que toda vez que la pendiente puede habese apreciado asimismo incorrectamente, cuanto más nos vamos alejando del punto (\bar{X}, \bar{Y}) , tanto mayor resulta la imprecisión. La banda de confianza adopta la forma general de la figura XVIII.2.

* Para trazar esta banda de confianza, será necesario calcular el error estándar de Y_p para varios valores de X . La apreciación del error estándar nos está dada por la fórmula:

$$\hat{\sigma}_{Y_p} = \hat{\sigma}_{y|x} \sqrt{\frac{1}{N} + \frac{(X - \bar{X})^2}{\sum_{i=1}^N (X_i - \bar{X})^2}} \quad (\text{XVIII.5})$$

en donde el valor particular de X a utilizar en $(X - \bar{X})^2$ puede ponerse en cualquier lugar del eje de las X . Obsérvese, de paso, que cuanto más lejos X queda de \bar{X} , tanto mayor es el valor numérico del error estándar. Supóngase que deseamos obtener el error estándar estimado cuando $X = 10.0$. Como quiera que $X = 4.837$, obtenemos:

$$\hat{\sigma}_{Y_p} = 215.1 \sqrt{\frac{1}{13} + \frac{(10.0 - 4.837)^2}{128.131}} = 215.1 \sqrt{.28496} = 114.86$$

* Sirviéndonos nuevamente de la tabla t y de un intervalo del 99 por ciento respecto de Y_p calculado para este valor fijo de X , obtendríamos:

$$Y_p \pm (3.106)(114.86) = Y_p \pm 356.8$$

Una vez que hayamos obtenido otros intervalos semejantes de Y_p para otros valores particulares de X , podemos trazar la gráfica

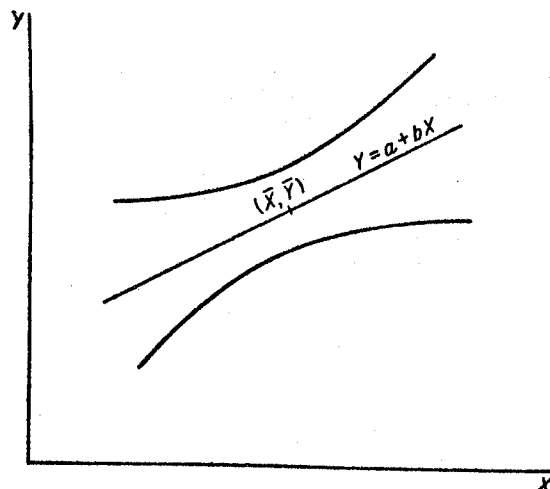


FIG. XVIII.2. Banda de confianza con respecto de la recta de mínimos cuadrados.

fica de la banda entera. Inútil es decir que el procedimiento en cuestión se haría muy fastidioso si se deseara obtener la banda entera y no se contara con calculadoras.

Probando la diferencia entre dos correlaciones. Como se indicó antes, tiene habitualmente más sentido teórico el comparar dos o más declives que el comparar correlaciones; tal comparación entre declives ocupará nuestra atención en el capítulo xx sobre análisis de covarianza. Sin embargo, ocurre con frecuencia que se han obtenido varias correlaciones y se desea establecer que una de ellas es significativamente más alta que las demás. Mientras nos contentamos en describir relaciones dentro de nuestra muestra particular, podemos comparar simplemente las magnitudes relativas de las dos r y registrar la magnitud de la diferencia. Sin embargo, si deseamos generalizar a una población mayor, plantéase la cuestión de si la diferencia obtenida pueda o no deberse acaso al azar. Supóngase, por ejemplo, que se han obte-

nido una r de .50 y otra de .30. Puede desearse verificar la hipótesis nula de que las dos correlaciones de las poblaciones son idénticas, esto es, $\rho_1 = \rho_2$.

Cabe imaginar dos situaciones distintas en las que podrían hacerse verificaciones de esta clase. Primero, pueden acaso tenerse *dos muestras independientes* y desearse comparar los grados de relación entre X y Y y dentro de cada una de las muestras. Así, por ejemplo, la relación entre el porcentaje de negros y la discriminación puede acaso no ser la misma en los estados del Sur que en los del Norte. Podría en este caso establecerse la hipótesis de investigación de que ρ_{xy} es más alta en el Sur que en el Norte, verificando la hipótesis nula de que las dos correlaciones son iguales. Un segundo tipo de situación, fácil de confundir con el primero, puede presentarse cuando se dispone de *una sola muestra*. Puede haber en este caso una sola variable dependiente (por ejemplo, la discriminación) y dos variables independientes (por ejemplo, el porcentaje de negros y el porcentaje de mano de obra empleada en la industria). Puede acaso desearse establecer que una de estas variables independientes está más directamente relacionada con la variable dependiente que la otra. Si designamos la segunda variable independiente como Z , podemos tener interés en verificar la hipótesis nula de que $\rho_{xy} = \rho_{zy}$. Veamos primeramente cómo tratamos el primer tipo de situación, para pasar luego a la prueba de una sola muestra.

Si las dos correlaciones se basan en muestras independientes, podemos convertir cada una de las r en z y servirnos de la fórmula del error estándar de la diferencia entre las z , que es análoga a la del error estándar de una diferencia entre medias y se presenta como sigue:

$$\sigma_{z_1 - z_2} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}} \quad (\text{XVIII.6})$$

Podemos a continuación ya sea establecer un intervalo de confianza relativo a $(z_1 - z_2)$ o buscar el valor de:

$$Z = \frac{(z_1 - z_2) - 0}{\sigma_{z_1 - z_2}}$$

en la tabla normal. El cero figura en la fórmula anterior debido al hecho de que nuestra hipótesis nula adopta la forma $\rho_1 = \rho_2$.

Supóngase que para 17 ciudades del Sur la correlación entre el porcentaje de negros y la discriminación resulta ser de .567, frente a la de .301 de las ciudades del Norte. Así, pues:

$$\begin{array}{ll} r_1 = .301 & r_2 = .567 \\ z_1 = 0.3106 & z_2 = 0.6431 \end{array}$$

y

$$\sigma_{z_1-z_2} = \sqrt{1/10 + 1/14} = \sqrt{.1000 + .0714} = .414$$

Por lo tanto:

$$Z = \frac{.3106 - .6431}{.414} = \frac{-.3325}{.414} = -.803$$

y vemos que esta diferencia de las r no es significativa al nivel de .05. Así, pues, pese a que la correlación sea mayor por lo que se refiere a las ciudades del Sur, esta diferencia puede deberse simplemente al azar.

En el segundo tipo de situación mencionado, no disponemos de dos muestras independientes y no podemos, por consiguiente, servirnos de la misma fórmula del error estándar de $z_1 - z_2$. Se dispone de un método para tratar este tipo de problema, a condición que sólo tengamos interés en generalizar a subpoblaciones de todas las muestras posibles para los que X y Z (las dos variables independientes) tienen las mismas combinaciones de valores que las de la muestra particular que hemos obtenido. En la mayoría de los casos prácticos puede prescindirse impunemente de esta restricción, a menos que exista alguna razón para suponer que el margen de variación es mucho mayor en la población que en la muestra estudiada, en cuyo caso deberemos de todos modos guardarnos de generalizar en un sentido o en otro.

Si verificamos la hipótesis nula de que $\rho_{xy} = \rho_{zy}$, formamos t de la manera siguiente:

$$t = (r_{xy} - r_{zy}) \sqrt{\frac{(N-3)(1+r_{xz})}{2(1-r_{xy}^2 - r_{xz}^2 - r_{zy}^2 + 2r_{xy}r_{xz}r_{zy})}} \quad (\text{XVIII.7})$$

Podemos buscar luego el valor de t en el cuadro, sirviéndonos de $N-3$ grados de libertad. En nuestro ejemplo numérico, supóngase que la correlación entre X y Z para las ciudades del Norte resulta ser de .172 y que la correlación entre Y y Z es de .749. Tendríamos en esta forma:

$$t = (.301 - .749) \sqrt{\frac{10(1 + .172)}{2[1 - .301^2 - .172^2 - .749^2 + 2(.301)(.172)(.749)]}} = -1.72.$$

Como tenemos 10 grados de libertad, vemos que no podemos descartar la hipótesis nula de que no hay diferencia entre las

correlaciones de las poblaciones de cada una de las variables independientes con discriminación.

XVIII.2. Correlación no lineal y regresión

Hasta aquí hemos venido suponiendo que la ecuación de regresión era de forma lineal. En muchos problemas sociológicos prácticos, el modelo lineal, aunque tal vez no exacto, da con todo una aproximación bastante cercana a la forma verdadera de la ecuación, de modo que no necesitamos ocuparnos de modelos alternativos más complicados. Esto es así, en particular, en relación con los estudios de exploración en los que el grado de adaptación no es excesivamente exacto. Hay casos, sin embargo, en los que la inspección del diagrama de dispersión podrá indicar claramente una relación no lineal, o en los que nuestra teoría ha anticipado una relación de esta clase. Siempre que se dé una relación no lineal semejante, el coeficiente momento-producto dará obviamente una subestimación del grado verdadero de relación, ya que este coeficiente sólo mide el grado de adaptación de la mejor recta singular. Ya vimos que con una curva en forma de U es posible tener una fuerte relación con una r de aproximadamente cero, y se advirtió al lector que era, por lo tanto, incorrecto sacar la conclusión de que dos variables son independientes simplemente porque r sea cero. Si el diagrama de dispersión indica una distribución de puntos más o menos al azar, podemos concluir que no existe relación, pero hemos de estar al acecho al propio tiempo de las relaciones no lineales. Ésta es, por supuesto, una razón más en favor de que el lector debe acostumbrarse a trazar siempre diagramas de dispersión antes de seguir adelante con el análisis.

El tema general de la correlación y la regresión no lineales es demasiado complejo para poder tratarlo adecuadamente en este texto. La razón de la complejidad del análisis no lineal está en que, una vez que progreseemos más allá de la ecuación de la recta, hay numerosos tipos de ecuaciones que representan las distintas formas posibles susceptibles de ser adoptadas por las relaciones no lineales. Sólo las más simples de estas ecuaciones pueden tratarse aquí. Afortunadamente, estas ecuaciones relativamente sencillas suelen ser por lo regular adecuadas para la solución de las clases de relaciones que se plantean en la investigación sociológica. Un tipo general de función no lineal puede representarse en términos de polinomios de grado enésimo, que tienen ecuaciones de la forma:

$$Y = a + bX + cX^2 + dX^3 + \dots + kX^n$$

El examen de las relaciones no lineales de este tipo general lo

dejaremos hasta el próximo capítulo, o sea hasta el momento de emprender el estudio de los problemas de regresión múltiple. En efecto, una vez comprendidos estos problemas de regresión, dispondremos de un método relativamente simple para el tratamiento de aquellos tipos de relaciones no lineales que se dejan describir adecuadamente por medio de polinomios.

Algún otro tipo de relaciones no lineales relativamente sencillo puede tratarse a menudo mediante una transformación de variables que permite el empleo del modelo lineal familiar. Este proceso puede ilustrarse con el caso de las funciones logarítmicas representadas por ecuaciones del tipo:

$$Y = a + b \log X$$

que presentan la forma general de la figura XVIII.3. En una ecuación de este tipo, en efecto, Y es en realidad una función lineal no de la X misma, sino de su logaritmo. Esto sugiere que si podemos transformar cada una de las marcas de X en una nueva variable $Z = \log X$, podemos escribir Y como función lineal de Z . Así, por ejemplo:

$$Y = a + b \log X = a + bZ$$

Podemos calcular ahora la correlación entre Y y Z (o sea de Y y de $\log X$) en la forma habitual. Si damos a conocer la distribución de las marcas a los ejes de las Y y las Z , el resultado habrá de ser aproximadamente de forma lineal. Si queremos, podemos comparar el grado de relación entre Y y Z con el que existe entre Y y X . Si r_{yz} es significativamente mayor que r_{xy} , entonces el modelo logarítmico da una mejor aproximación que el modelo lineal entre X y Y .

Los modelos logarítmicos del tipo anterior se presentan a menudo en casos en que la variable independiente X asume un gran margen de valores, pero en los que, una vez alcanzado cierto valor, los aumentos ulteriores producen cada vez menos efecto sobre la variable dependiente. La magnitud de una ciudad es una variable que presenta con frecuencia esta clase de efecto. Es posible, por tanto, que las ciudades de más de 500 mil habitantes presenten todas ellas marcas de Y muy parecidas. Pero, si se incluye en la muestra a la ciudad de Nueva York, por ejemplo, el valor de X para esta ciudad será tan superior al de las demás ciudades, que el efecto neto consistirá en inclinar la relación en forma muy parecida a la de la figura XVIII.3. En tal caso podrá resultar preferible relacionar Y con $\log X$, ya que el hecho de tomar el logaritmo de la magnitud urbana producirá el efecto de agrupar las marcas extremadamente grandes y de disminuir el "efecto de curvatura" de estas ciudades mayores.

En cierto número de casos el investigador no tendrá tal vez interés en hallar la forma exacta de la ecuación de predicción que mejor se adapte a sus datos. Acaso sólo trate, por ejemplo, de demostrar que la relación es de forma no lineal, o de obtener una medida para el grado de relación, independientemente de su forma. Cuando pueda efectuarse una transformación sencilla

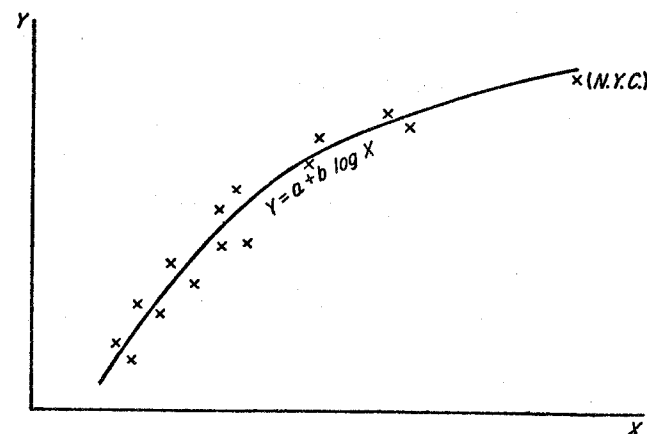


FIG. XVIII.3. Ecuación logarítmica de mínimos cuadrados de la forma $Y = a + b \log X$.

como la logarítmica, será indudablemente ventajoso servirse de dicho procedimiento. Pero aun así, el investigador querrá acaso verificar si la medida que ha obtenido constituye o no una buena aproximación del resultado que habría hallado si se hubiera encontrado la mejor adaptación posible. Con objeto de tratar los problemas de esta índole, podemos servirnos de los principios básicos del análisis de variancia y de algunas de las medidas de los grados de asociación desarrolladas en el capítulo sobre análisis de variancia.

El lector recordará que para obtener la suma de cuadrados "dentro" en el análisis de variancia de una forma tomamos la suma de las desviaciones al cuadrado de cada una de las medias de las categorías. Supongamos ahora que las X se han subdividido en cierto número de categorías y que la suma de los cuadrados en Y se analizaban en la forma habitual. Sabemos que para toda categoría dada de X la suma de los cuadrados alrededor de la media de la categoría producirá un resultado numérico inferior al de la suma de los cuadrados alrededor de cualquier otro número. Síguese, en particular, que la suma interior de cuadrados será menor que la suma de las desviaciones cuadradas respecto de aquellos puntos de la línea de mínimos cua-

drados que caen en los puntos medios de los intervalos (véase la figura XVIII.4).

Si ocurre que la ecuación sea de forma lineal, podemos esperar que \bar{Y}_j caerá aproximadamente en la línea de los mínimos cuadrados, de modo que cambiará poco que las desviaciones se tomen respecto de las medias de las categorías o respecto de la lí-

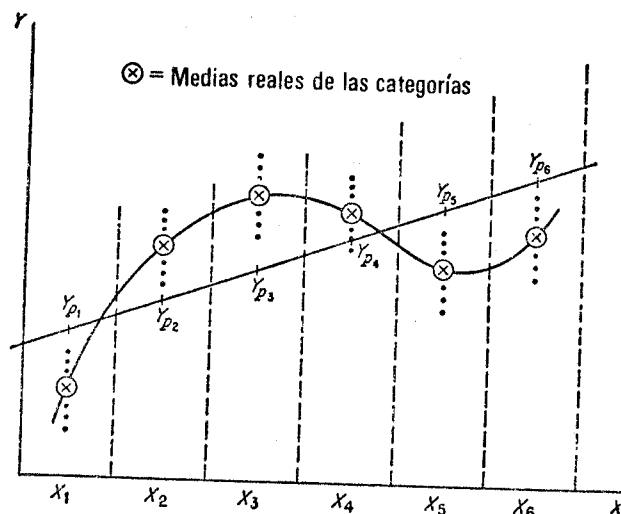


FIG. XVIII.4. Comparación de las desviaciones respecto de la recta de mínimos cuadrados con las desviaciones respecto de las medias de las categorías.

nea en cuestión. Por otra parte, si la ecuación es realmente no lineal, entonces, para algunas, al menos, de las categorías, la suma de los cuadrados referidos a la media de la categoría será considerablemente más pequeña que la de los cuadrados referidos a la línea de los mínimos cuadrados. En otros términos: la suma interior o inexplicada de cuadrados será mínima empleando las medias de las categorías y, por consiguiente, la suma de cuadrados entre categorías, o explicada, será máxima. Así, pues, la proporción de variación explicada por las categorías, medida por la razón de correlación E^2 , será mayor que la proporción explicada por la línea de mínimos cuadrados, a menos que la verdadera relación sea efectivamente lineal.

Podemos sacar utilidad de este hecho practicando una prueba de no linealidad. Si formamos la cantidad $E^2 - r^2$, obtenemos la proporción de variación explicada en el supuesto de una forma cualquiera de relación no explicada por una relación lineal. Es obvio que para obtener E^2 permitimos que la relación adopte

cualquier forma posible, ya que sólo hemos tomado desviaciones respecto de las medias de las categorías, prescindiendo de dónde estas medias acontezcan encontrarse. Nos estamos preguntando fundamentalmente en cuánto podemos mejorar nuestra posibilidad de predecir valores de Y no restringiéndonos al modelo lineal. Si la mejora es mayor de lo que esperaríamos del azar

CUADRO XVIII.2. Prueba de análisis de variancia para el caso de no linealidad

	Sumas de cuadrados	Grados de libertad	Estimaciones de la variancia	F
Total	Σy^2	$N-1$		
Explicada por el modelo lineal	$r^2 \Sigma y^2$	1		
Adicional, explicada por el modelo no lineal	$(E^2 - r^2) \Sigma y^2$	$k-2$	$\frac{(E^2 - r^2) \Sigma y^2}{k-2}$	$\frac{(E^2 - r^2)(N-k)}{(1 - E^2)(k-2)}$
Inexplicada	$(1 - E^2) \Sigma y^2$	$N-k$	$\frac{(1 - E^2) \Sigma y^2}{N-k}$	

en el supuesto de que la ecuación de regresión sea efectivamente lineal, entonces podemos concluir que la relación es no lineal.

La prueba de análisis de variancia que emplearemos para verificar la no linealidad asume una forma con la que no tardaremos en familiarizarnos. Hallamos primero la cantidad de variación que puede explicarse sirviéndonos del modelo lineal. Algebraicamente, esta cantidad puede representarse como $r^2 \Sigma y^2$. De la variación dejada sin explicar por el modelo lineal, $(1 - r^2) \Sigma y^2$, vemos a continuación qué tanto pueda explicarse por medio del modelo general. Como quiera que $E^2 \Sigma y^2$ nos da la suma de cuadrados que puede explicarse por Y cuando no pesa sobre la forma de la relación restricción alguna, la cantidad $(E^2 - r^2) \Sigma y^2$ representa el incremento explicado debido a la no linealidad. Suponiendo que no haya errores de redondeo, esta cantidad habrá de ser siempre positiva.¹ Y como quiera que la cantidad $(1 - E^2) \Sigma y^2$ nos da la suma de cuadrados que no resulta explicada ni siquiera por el modelo mejor adaptado, podemos efectuar una prueba F tal como se indica en el cuadro XVIII.2. Como de costumbre, el denominador de F es el término de error y, como

¹ Siempre que N sea pequeña y sólo pueda, por tanto, usarse un corto número de categorías, resulta poco realista el supuesto de que las puntuaciones de X están agrupadas en los puntos medios de cada intervalo. Esto puede llevar a agrupar los errores, dando un valor a E^2 menor que r^2 .

quiera que estamos verificando en relación con desviaciones respecto de la linealidad, tomamos como numerador una apreciación de la variancia basada en $(E^2 - r^2)\Sigma y^2$, o sea la cantidad explicada por el modelo general mejor, que no ha sido explicada todavía por el modelo lineal. Los grados de libertad asociados al numerador pueden obtenerse por sustracción.

Una vez más observamos que la suma total de cuadrados se elimina, dejándonos con la siguiente fórmula para F :

$$F_{k-2, N-k} = \frac{(E^2 - r^2)(N - k)}{(1 - E^2)(k - 2)} \quad (\text{XVIII.8})$$

en donde k representa el número de categorías en las que se ha descompuesto X .

Ilustremos la prueba de no linealidad con los datos que se agruparon en el cuadro XVII.2. Según puede comprobarse fácilmente, las sumas total y entre categorías de cuadrados en Y son como sigue:

$$\begin{aligned} \text{SC total} &= 101\,115.38 - 92\,132.04 = 8\,983.34 \\ \text{SC entre categorías} &= 94\,792.59 - 92\,132.04 = 2\,660.55 \end{aligned}$$

en donde hemos tratado todas las marcas de Y como si se encontraran en los puntos medios de sus respectivos intervalos y en donde nos hemos servido de los procedimientos para los datos agrupados (véase sec. VI.4). Por lo tanto:

$$E^2 = \frac{\text{SC entre cuadrados}}{\text{SC total}} = \frac{2\,660.55}{8\,983.34} = .2962$$

Toda vez que anteriormente encontramos una r de $-.460$ suponiendo una relación lineal, obtenemos:

$$F_{7,141} = \frac{.2962 - (-.460)^2}{1 - .2962} \cdot \frac{150 - 9}{9 - 2} = \frac{.0846}{.7038} \cdot \frac{141}{7} = \frac{11.929}{4.927} = 2.42$$

y vemos que al nivel de .05 podemos descartar la hipótesis nula de una relación lineal entre el porcentaje de personas clasificadas como trabajadoras de granjas rurales y el porcentaje de mujeres que trabajan en la industria.

Si una relación resulta ser no lineal en cuanto a la forma, es muy posible que r no sea significativa estadísticamente, en tanto que E sí lo será. Por supuesto, la significación de E puede comprobarse por medio de un análisis directo de variancia, tomando la razón de las estimaciones explicada e inexplicada de la variancia. Son, pues, así tres las pruebas que pueden efectuarse,

a saber: 1) la de la significación de r ; 2) la de la significación de las desviaciones respecto de la linealidad ($E^2 - r^2$), y 3) la de la significación de E .

Si se encuentra una relación no lineal y se desea una estimación del grado de relación en la población, es preferible servirse de la razón de correlación insesgada ϵ , examinada en el capítulo XVI y dada por la fórmula:

$$\epsilon^2 = 1 - \frac{V_w}{V_t}$$

ya que el valor numérico de E es función del número de categorías empleadas y probablemente sobrestimaré ligeramente por lo regular la relación en la población. Si ya se ha calculado E , el valor de ϵ puede también calcularse a partir de la fórmula:

$$\epsilon^2 = \frac{E^2(N - 1) - (k - 1)}{N - k} \quad (\text{XVIII.9})$$

XVIII.3. Efectos de los errores de medición

Si hay mediciones de error en X o Y , bien sean al azar o sistemáticas, puede esperarse una alteración en nuestros resultados. Esto se aplica por supuesto a todas las pruebas y mediciones que hemos examinado hasta ahora, incluso los procedimientos no paramétricos. En realidad, uno de los tipos de errores de medición más comunes en sociología, ciencia política y la mayoría de las restantes ciencias sociales, parecería ser consecuencia del uso de dicotomías más bien burdas, tales como *alto* y *bajo* o *presente* y *ausente*. No se comprenden bien las consecuencias que se derivan de los errores de medición, pero la mayor parte del trabajo sistemático sobre el tema se ha llevado a cabo en las escalas de intervalo y en los problemas que implican análisis de correlación y regresión. El tema es por desgracia demasiado técnico para ser tratado en el presente texto, pero resultará conveniente pronunciar por lo menos algunas palabras precautorias.

Si hay una medición de error sistemática, o no aleatoria, cualquier tipo de distorsión resulta posible, siendo así necesario explicar cuáles son las fuentes del error no aleatorio y la forma en que actúan. Si se comparan por ejemplo las medias de tres muestras, y el error de medición es tal que coloque las medias de las muestras segunda y tercera cercanas a la correspondiente a la primera, no se logrará significación estadística cuando, con base en mediciones más exactas, pueda rechazarse fácilmente la hipótesis nula. Pero si los errores de medición son estrictamente al azar, resultará posible tener una mayor claridad acerca de los

efectos de tales errores. En general, las medidas de asociación resultarán atenuadas por los errores aleatorios de medición en cualquier variable. Por ejemplo, en el análisis de las situaciones de variancia, las mediciones aleatorias de error en la escala de intervalos aumentarán las variaciones *dentro* de las categorías, pero no afectarán sistemáticamente las variaciones entre las categorías, lo que hará bajar tanto el valor de F como la correlación interclases.

En el caso de dos escalas de intervalo los errores aleatorios de medición en cualquier variable reducirán la magnitud del coeficiente de correlación. En algunos textos elementales de estadística se examinan los procedimientos correctivos de atenuación, pero se hace basándose en supuestos especiales, inapropiados para uso en la investigación sociológica. (Véase [3].) En general, cuando se cuenta con dos o más medidas de cada variable, resulta posible obtener estimaciones corregidas bajo grupos variables de supuestos. (Véanse [2], [6] y [14].)

Si hay errores aleatorios de medición en Y pero no en X , podemos concebir la situación como una contribución que alcanza sólo al factor de error en la ecuación $Y_i = \alpha + \beta X_i + \varepsilon_i$, pudiendo demostrarse que no habrá efecto sistemático en la estimación $b_{y'}$ del declive, salvo que el error estándar en tal estimación se verá incrementado debido al aumento del error en la variancia. Pero si hay también error aleatorio de medición en X —lo que es muy posible en toda investigación realista—, la estimación $b_{y'}$ del declive se verá asimismo atenuada. En el caso de muestras grandes puede aplicarse una fórmula aproximada para determinar el valor esperado del declive $b_{y'}$:

$$E(b_{y'}) = \beta \frac{\sigma_x^2}{\sigma_{x'}^2} = \beta \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}$$

en la que X' representa el valor medido de X , tal como se le representa en la ecuación $X' = X + u$, en donde se supone a u como un componente estrictamente aleatorio, con valor esperado igual a cero, y sin que haya correlación entre u y X . La razón de la atenuación estriba en que la variancia del valor medido X' será mayor que la variancia verdadera de X , según la fórmula:

$$\sigma_{x'}^2 = \sigma_x^2 + \sigma_u^2$$

Vemos así que la atenuación en la estimación de un desnivel es función de la variancia del error de medición, *relativa* a la variancia en X .

Este hecho tiene consecuencias importantes en la práctica. Significa que en cuantos casos haya error aleatorio de medición en

una variable independiente, no podemos contar con iguales declives estimados, incluso en el caso de que los declives verdaderos lo sean. Si varias poblaciones (o muestras) difieren con respecto a la cantidad de variación en X , incluso con las mismas variancias de error de medición, las atenuaciones de los declives diferirán. Vale la pena tener esto presente cuando se llevan a cabo comparaciones de los resultados de diferentes estudios. La dificultad señalada se aplica también a todas las medidas de asociación, y no puede ser considerado como un defecto privativo del análisis de regresión.

XVIII.4. Escalas ordinales: correlación de rangos

Nos hemos ocupado ahora de medidas de asociación que pueden utilizarse para relacionar dos escalas nominales (ϕ^2 , τ_b , etcétera), una escala nominal y una de intervalo (correlación intraclass), y dos escalas de intervalo (r). Las tres medidas que vamos a examinar en esta sección, o sean la r_s de Spearman y la tau y la gamma de Kendall, pueden emplearse para relacionar entre sí dos escalas ordinales. A condición que las dos variables pueden alinearse, cualquiera de estas últimas medidas puede emplearse para dar correlaciones que son algo parecidas a las del momento producto.

Las medidas ordinales examinadas en esta sección resultan apropiadas cuando la relación entre X y Y es la que se denomina *monotónica en aumento* o bien *monotónica en disminución*. La idea de linealidad es desde luego inapropiada en el caso de las escalas ordinales, como lo es también la idea de una distancia entre valores de X (o de Y). Podemos, sin embargo, hablar de relaciones que se encuentran en aumento (o disminución) constante. Una función de aumento monotónico es aquella que o bien *aumenta* siempre o permanece constante, a medida que X aumenta. En otras palabras: cuando X aumenta, Y no disminuye. Una función lineal constituye un caso especial de una función monotónica de aumento (o disminución), pero también lo es una función logarítmica tal como $Y = a + b \log X$. Reconocemos dos clases de relación no lineal, a saber: las que son monotónicas y las que no lo son. El último tipo de relación no lineal tendrá por supuesto una o más curvaturas o inversiones de dirección, como ejemplifica una parábola o ecuación de tercer grado.

Con frecuencia encontramos proposiciones teóricas de la forma "cuanto mayor la X , mayor la Y (o menor la Y)". Estas afirmaciones quieren decir que la relación entre X y Y es monotónica, pero no especifican en qué forma. Las medidas ordinales resultan apropiadas cuando se trata de proposiciones de esta naturaleza. Sería por supuesto preferible refinar nuestras teorías, de modo que se especificase si existe linealidad o alguna clase

particular de no linealidad (por ejemplo, logarítmica), pero si la medición no ha superado el nivel ordinal, resultará imposible distinguir empíricamente entre alternativas lineales o no lineales. (Véase [22].)

La r_s de Spearman. El principio que se halla en la base de la medida de Spearman es muy simple. Comparamos la ordenación de dos grupos de marcas tomando las diferencias de los rangos, cuadrándolas y luego adicionándolas, y tratando finalmente dicha medida de modo que su valor sea $+1.0$, siempre que los órdenes estén perfectamente de acuerdo, -1.0 si los órdenes discrepan totalmente, y cero si no se da relación alguna. Si simbolizamos la diferencia entre dos lugares cualesquiera como D_i , hallamos el valor de $\sum_{i=1}^N D_i^2$ y calculamos r_s por medio de la fórmula:

$$r_s = 1 - \frac{6 \sum_{i=1}^N D_i^2}{N(N^2 - 1)} \quad (\text{XVIII.10})$$

Esta fórmula para r_s se obtiene tomando la fórmula para una correlación momento-producto y aplicándola a rangos y no a puntuaciones brutas, pudiendo así interpretar la medida de Spearman como la correlación momento-producto entre los rangos de X y los de Y .

Ilustrémosla con algunos datos reunidos por el autor. Los miembros de un campamento de trabajo fueron ordenados de superior a inferior desde los puntos de vista de la popularidad, medida por las amistades y de la participación en las discusiones de grupo. Para ambas variables el orden de clasificación de uno significa una marca elevada. Los órdenes empatados se calculan atribuyendo a cada marca empatada la media aritmética de la puntuación que habría recibido si no hubiera empates. Los valores de D_i se calculan a continuación, tal como se indica en el cuadro XVIII.3. Si el número de empates es pequeño, como en el presente caso, no necesitamos introducir modificación en la fórmula de r_s . Pero si el número de empates es considerable, entonces puede calcularse un factor de corrección (véase [19], pp. 215-220). Obtenemos, pues:

$$r_s = 1 - \frac{6(207.50)}{16(255)} = 1 - .305 = .695$$

Obsérvese que si las clasificaciones concuerdan perfectamente, $\sum_{i=1}^N D_i^2$ será cero, y el valor de r_s será la unidad. Si bien la ins-

pección directa de la fórmula no nos da inmediatamente los valores de r_s para la independencia y la asociación perfectamente negativa, resulta que para la asociación negativa perfecta el valor del segundo término será de -2.0 y, por lo tanto, r_s será -1.0 .

CUADRO XVIII.3. Cálculo del coeficiente de Spearman de la correlación de rango

Personas	Orden de popularidad	Orden de participación	D_i	D_i^2
Ann	1	5.5	4.5	20.25
Bill	2.5	5.5	3.0	9.00
Jim	2.5	1	-1.5	2.25
Hans	4	2	-2.0	4.00
Marcia	5	3	-2.0	4.00
Joan	6	9.5	3.5	12.25
Ruth	7	5.5	-1.5	2.25
Doris	8	13.5	5.5	30.25
Barbara	9	9.5	0.5	0.25
Cynthia	10	16	6.0	36.00
Ullie	11.5	5.5	-6.0	36.00
Plo	11.5	11.5	0.0	0.00
Nancy	13.5	8	-5.5	30.25
Mart	13.5	15	1.5	2.25
Ntan	15	11.5	-3.5	12.25
Narah	16	13.5	-2.5	6.25
Total			0.0	207.50

Para la no asociación, el segundo factor será exactamente la unidad.

Si $N = 10$, la distribución de selección de r_s es aproximadamente normal, con una desviación estándar de $1/\sqrt{N-1}$. Por lo tanto, en el ejemplo que estamos examinando, el error estándar será de $1/\sqrt{15}$. Como prueba de la hipótesis nula de que no se da relación en la población, podemos calcular Z como sigue:

$$Z = \frac{r_s - 0}{1/\sqrt{N-1}} = .695 \sqrt{15} = 2.69$$

Al mirarnos de la tabla normal vemos que la relación es significativa al nivel de .01.

La tau de Kendall. Al calcular la r_s de Spearman nos servimos de los cuadrados de las diferencias en los rangos. La tau de Kendall, en cambio, que también varía entre -1.0 y 1.0 , se basa en una operación algo distinta. En efecto, calculamos primero

una estadística S buscando todos los pares posibles de casos y observando si las puntuaciones están o no en el mismo orden. Así, por ejemplo, supongamos que teníamos las siguientes combinaciones de lugares:

	a	b	c	d
A	1	2	3	4
B	2	3	1	4

Como quiera que las marcas de A se han dado en orden ascendente, podemos calcular S examinando las clasificaciones de B una por una. Fijándonos en el primer valor de la hilera B (individuo a), vemos que la marca de B está en el orden apropiado para los pares (a,b) y (a,d) . En otros términos: el individuo a ocupa un lugar inferior a b y d en ambas variables A y B . Por otra parte, la marca de B discrepa (con respecto a la marca de A) para el par (a,c) , ya que a ocupa un lugar inferior a c en cuanto a A , pero inversamente en cuanto a B .

Sirvámolos de +1 cada vez que un par determinado se halla ordenado igualmente para A y B (lo que se denomina par "concordante") y de -1 cada vez que se halla ordenado al revés (lo que se denomina par "discordante"). El valor de S se obtiene sumando dichos +1 y -1 para todos los pares posibles. Por lo tanto, S es igual al número de pares concordantes C , menos el número de pares discordantes D . Por lo tanto, la contribución de los pares (a,b) , (a,c) y (a,d) es: $+1 -1 +1 = (2 - 1) = 1$. Con objeto de tener en cuenta los demás pares, recorreremos la tabla de izquierda a derecha. Vemos así que la contribución de los pares (b,c) y (b,d) es de $-1 + 1$, o sea cero. Finalmente, la contribución del par (c,d) es de +1. Obsérvese que de hecho podemos obtener el valor total de S disponiendo primero A en el orden apropiado y examinando luego sucesivamente los lugares de la hilera B , contando cada vez el número de lugares de la derecha que están en el orden apropiado y sustrayendo los que están en el orden contrario. De este modo, en este sencillo ejemplo obtenemos:

$$S = C - D = (2 - 1) + (1 - 1) + (1 - 0) = 2$$

Si ahora dividimos S entre el valor máximo posible que podría tener, esto es: $(N - 1) + (N - 2) + \dots + 2 + 1 = N(N - 1)/2$, obtenemos un coeficiente que puede variar de -1 a +1. Definimos así el coeficiente tau_k (según Kendall [16]), adecuado cuando no hay empates, como sigue:²

² Este coeficiente, derivado de los datos de la muestra, se denomina a veces t , en tanto que tau se reserva para la contrapartida de la pobla-

$$\tau_k = \frac{S}{\frac{1}{2}N(N - 1)} = \frac{C - D}{\frac{1}{2}N(N - 1)} \quad (\text{XVIII.11})$$

Ha obvio que si hay discrepancia perfecta entre los dos sistemas de ordenación (esto es, si B estuviera ordenado como 4, 3, 2, 1), el valor de S será $-\frac{1}{2}N(N - 1)$, y τ será -1.0. Y asimismo, si las dos variables no tienen relación alguna entre sí, las contribuciones a S positivas y negativas se invalidarán, y τ será cero.

Con objeto de ilustrar el caso de los órdenes empatados, sirvámolos nuevamente del ejemplo del campamento de trabajo. Dispongamos a los individuos en orden horizontal y reemplacemos los nombres por letras. Nuestra disposición se presenta en esta forma:

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p
A	1	2.5	2.5	4	5	6	7	8	9	10	11.5	11.5	13.5	13.5	15	16
B	5.5	5.5	1	2	3	9.5	5.5	13.5	9.5	16	5.5	11.5	8	15	11.5	13.5

Hemos de seguir la regla de que siempre que algún par comporte un empate, ya sea en la marca A o B , su contribución a S será cero. Mirando primero todos los pares que pueden formarse con a , vemos que los pares (a,b) , (a,g) y (a,k) no contribuirán con nada a S , ya que las marcas de B para todos dichos individuos están ligadas en 5.5. Por lo tanto, la contribución de todos los demás pares será:

$$(a,d) (a,e) (a,f) (a,h) (a,i) (a,j) (a,l) (a,m) (a,n) (a,o) (a,p) \\ 1 \quad 1 \quad -1 \quad +1 \quad +1 \quad +1 \quad +1 \quad +1 \quad +1 \quad +1 \quad +1 = 9 - 3 = 6$$

A continuación comparamos las marcas de b con cada una de las marcas a su derecha. Obsérvese, sin embargo, que b y c están ligados con respecto a A . Como quiera, por lo tanto, que b y c pudieron haberse dado lo mismo en el orden inverso, hemos de eliminar el par (b,c) . Y en forma análoga, los pares (b,g) y (b,k) están ligados en B y, por consiguiente, no harán contribución alguna a S . En esta forma, para los pares de b , obtenemos una suma de $9 - 2$, o sea 7. Recorriendo la tabla de izquierda a derecha obtenemos finalmente:

$$S = C - D = (9 - 3) + (9 - 2) + (13 - 0) + (12 - 0) + (11 - 0) \\ + (6 - 3) + (8 - 0) + (2 - 5) + (5 - 2) + (0 - 6) \\ + (4 - 0) + (2 - 1) + (2 - 0) + (0 - 2) + (1 - 0) \\ 60$$

Como seguiremos, sin embargo, el uso más convencional. La tau de Kendall no debe confundirse con las tau_k y tau_b de Goodman y Kruskal, las que son apropiadas para datos nominales.

Con objeto de corregir en relación con los empates, hemos de practicar ahora un ajuste en el denominador de tau. Semejante ajuste tiene el efecto de producir un aumento del valor numérico de tau, si bien dicho aumento será ligero, a menos que el número de empates sea muy grande. La fórmula de tau (la que Kendall designó como τ_b) puede generalizarse como sigue:

$$\tau_b = \frac{S}{\sqrt{\frac{1}{2}N(N-1) - T} \sqrt{\frac{1}{2}N(N-1) - U}} \quad (\text{XVIII.12})$$

en donde $T = \frac{1}{2}\sum t_i(t_i - 1)$, siendo t_i el número de empates en cada grupo de empates en A, y $U = \frac{1}{2}\sum u_i(u_i - 1)$, siendo u_i el número de empates en cada grupo de empates en B. En el ejemplo anterior tenemos tres empates, de dos cada uno, en la variable A (popularidad). Por lo tanto:

$$T = \frac{1}{2}[2(1) + 2(1) + 2(1)] = 3$$

Y en forma análoga, hay tres empates, de dos cada uno, y una marca con cuatro empates en la variable B (participación). Por consiguiente:

$$U = \frac{1}{2}[2(1) + 2(1) + 2(1) + 4(3)] = 9$$

De donde:

$$\tau_b = \frac{60}{\sqrt{[8(15) - 3][8(15) - 9]}} = \frac{60}{\sqrt{(117)(111)}} = \frac{60}{114.0} = .526$$

Prueba de significación para tau. Kendall [16] ha demostrado que para tamaños de muestras de 10 o más, la distribución de muestreo de S bajo la hipótesis nula será aproximadamente normal, con media de cero y variancia dada por:

$$\sigma_s^2 = \frac{1}{18}N(N-1)(2N+5) \quad (\text{XVIII.13})$$

Hablando estrictamente, la fórmula anterior es aplicable sólo cuando no hay empates, pero puede ser usada cuando el número de éstos es relativamente pequeño. Si se da un gran número de empates, un factor de corrección bastante voluminoso habrá de ser aplicado.

Para probar la significancia de tau con los datos del campo de trabajo, comenzamos por computar σ_s^2 como sigue:

$$\sigma_s^2 = \frac{1}{18}(16)(15)(37) = 493.3$$

Obteniendo la raíz cuadrada tenemos:

$$\sigma_s = 22.21$$

valor que puede ser usado en el denominador de Z al probar la hipótesis nula de que A y B no están relacionados. Así

$$Z = \frac{S - 0}{\sigma_s} = \frac{60.0}{22.21} = 2.70$$

y vemos que un valor de tau de .526 es significativo al nivel de .01.

Medidas ordinales para datos agrupados: τ_o , γ , d_{yx} y d_{xy} . Una de las ventajas de tau respecto de r_s es que aquélla puede utilizarse fácilmente cuando se da un número grande de empates. Pese a que el cálculo de rutina que se acaba de describir resultaría sumamente fastidioso en tales casos, podemos simplificar mucho el procedimiento cuando ambas categorías se han agrupado en categorías algo toscas. Así, por ejemplo, puede haberse colocado a personas en cinco clases sociales, considerándolas como empatadas con respecto a la posición. Si la segunda variable se ha categorizado en la misma forma, podemos servirnos de una fórmula de tau modificada, aprovechando con ello la información de que los datos han sido efectivamente ordenados, y no simplemente puestos en categorías.

Podemos calcular $S = C - D$ mediante un procedimiento que no describe más abajo. Sirviéndonos de las fórmulas que se acaban de dar, encontraremos que el límite superior de tau, sólo será la unidad cuando el número de hileras y de columnas sea el mismo. Con objeto de corregir para el caso en que $r \neq c$, formamos la razón:

$$\tau_c = \frac{S}{\frac{1}{2}N^2[(m-1)/m]} \quad (\text{XVIII.14})$$

donde

$$m = \text{Min}(r, c)$$

Aquí seguimos a Kendall en el empleo del símbolo τ_o , con objeto de distinguir la ecuación (XVIII.14) de las fórmulas precedentes. Veamos ahora cómo se calcula τ_o .

Los datos del cuadro XVIII.4 representan los lugares asignados a 217 estudiantes de introducción a la sociología en la Universidad de Michigan. La variable B comporta el interés general del estudiante en cuanto a adoptar las formas "apropiadas" o "correctas" de comportamiento en los medios convencionales. En tanto que la variable A comporta el deseo de formar parte de organizaciones únicamente con objeto de mejorar la posición social.

Toda vez que la medición de ambas variables fue más bien tosca, se decidió dividir cada una de ellas en cuatro categorías: interés alto, moderadamente alto, moderadamente bajo y bajo. De este modo, si bien cada variable comporta una escala ordinal con un

CUADRO XVIII.4. Datos comparados para el cálculo de la tau de Kendall a partir de datos agrupados

Grado del deseo de formar parte de organizaciones (A)	Interés en la conducta adecuada (B)				Total
	Alto	Moderadamente alto	Moderadamente bajo	Bajo	
Alto	18	19	12	8	57
Moderadamente alto	16	16	12	10	54
Moderadamente bajo	11	14	18	16	59
Bajo	5	5	15	22	47
Total	50	54	57	56	217

gran número de empates, los resultados pueden con todo reunirse en forma de una tabla de contingencia.

Al calcular S será conveniente obtener separadamente C y D , ya que dichas cantidades serán utilizadas también para otras mediciones discutidas en esta sección. Observamos en primer término que las marcas de A se han ordenado nuevamente de altas a bajas, con la diferencia de que ahora tenemos 57 individuos "empatados" en cuanto a las marcas altas, 54 en cuanto a las moderadamente altas, 59 en cuanto a las moderadamente bajas y 47 en cuanto a las bajas. Considerando primero a los de marcas altas en cuanto a A , vemos que 18 las tienen también altas en B ; 19 moderadamente altas, etcétera. Para obtener las contribuciones a C y D (y por lo tanto a S) observamos que, como quiera que todos los individuos de la categoría alta de A están empatados, ninguno de estos pares contribuirá a C o D . Y en forma análoga, ninguno de los pares de la misma columna contribuirá a C o D , debido al hecho de que todos ellos están empatados con respecto a B . Si nos fijamos en una casilla determinada cualquiera, todas las marcas que se hallan por debajo y a la derecha de la misma contribuirán al número de pares C concordantes, en tanto que todas las que se encuentran por debajo y a la izquierda contribuirán a D . Así, por ejemplo, cada uno de los 18 individuos de la casilla producirá pares concordantes con cada una de las marcas

$$16 + 14 + 5 + 12 + 18 + 15 + 10 + 16 + 22$$

que quedan por debajo y a la derecha de dicha casilla. En total, pues, la contribución de la casilla en cuestión a C será de:

$$18(16 + 14 + 5 + 12 + 18 + 15 + 10 + 16 + 22) = 18(128)$$

A continuación nos fijamos en los 16 casos inmediatamente debajo del ángulo izquierdo superior. Cada uno de estos individuos tiene también marcas altas de B . Con objeto de contar los pares de contribuciones a C , volvemos a adicionar las cantidades que figuran debajo y a la derecha. Multiplicando luego por el número de casos tenemos:

$$16(14 + 5 + 18 + 15 + 16 + 22) = 16(90)$$

Al pasar a las columnas segunda y siguientes, empezamos a encontrar contribuciones a C y D , ya que las columnas de la izquierda tienen marcas superiores de B . Así, para la primera casilla de la segunda columna obtenemos como contribución a C :

$$19(12 + 18 + 15 + 10 + 16 + 22) = 19(93)$$

y como contribución a D la cantidad $19(16 + 11 + 5) = 19(32)$. Recorriendo la tabla hacia abajo y hacia la derecha en forma semejante, podemos obtener S hasta cierto punto con facilidad, como sigue:

$$C = 18(128) + 16(90) + 11(42) + 19(93) + 16(71) + 14(37) + 12(48) + 12(38) + 18(22) = 9055$$

$$D = 19(32) + 16(16) + 14(5) + 12(67) + 12(35) + 18(10) + 8(112) + 10(68) + 16(25) = 4314$$

$$\text{Por tanto:} \quad S = 9055 - 4314 = 4741$$

Así pues:

$$\tau_o = \frac{4741}{\frac{1}{2}(217)^2[(4-1)/4]} = .268$$

Obsérvese que el denominador de τ_o depende sólo del número de hileras y columnas, y no de las distribuciones marginales, las que por supuesto determinan el número de empates. Esto hace que τ_o sea difícil de interpretar, y, en este sentido, menos satisfactoria que τ_b .⁸ Hay también otras varias medidas que di-

⁸ Puede demostrarse que en el caso $k \times k$, en el que todos los totales marginales son exactamente N/k , τ_b y τ_o serán iguales. De otra forma, en el caso $k \times k$, τ_o será generalmente menor que τ_b en valor numérico, aun cuando pueda ser mayor que τ_b en el caso $r \times c$.

fieren en relación con el manejo de los empates en el denominador. La más conocida de dichas medidas es gamma (γ), la que excluye por completo los empates en el denominador, y puede además ser aplicada a datos no agrupados. La fórmula para gamma es la siguiente:

$$\gamma = \frac{C - D}{C + D}$$

En el ejemplo que estamos considerando obtenemos:

$$\gamma = \frac{9055 - 4314}{9055 + 4314} = .354$$

Se indicó en el capítulo xv que la Q de Yule, igual a $(ad - bc)/(ad + bc)$ es un caso especial de gamma. Podemos por ello esperar que gamma se conduzca esencialmente igual en los casos en que las distribuciones marginales son muy desiguales, debiendo observarse las mismas precauciones que se aplicaron a Q . Como tanto gamma como τ_a y τ_b tienen todas los mismos numeradores y puesto que el denominador de gamma excluye todos los empates, puede verse fácilmente que $|\gamma| \geq |\tau_b| \geq |\tau_a|$. En general, hasta el grado en que los totales marginales para A y B son muy diferentes, gamma puede exceder a τ_b por una cantidad apreciable. Por ejemplo, en el caso del siguiente cuadro hipotético:

A	B			Total
	Alta	Media	Baja	
Alta	100	80	0	180
Media	0	20	80	100
Baja	0	0	20	20
Total	100	100	100	300

observamos que no hay pares discordantes, de modo que $\gamma = 1.0$. Sin embargo, $\tau_b = .77$ y $\tau_c = .68$. El que uno desee o no referirse a la anterior asociación considerándola "perfecta", dependerá de los supuestos en relación con la causa de que las distribuciones marginales no sean idénticas.

Además de las taus y gamma, tenemos dos medidas asimétricas, d_{yx} y d_{xy} , ideadas por Sommers [20] y definidas como sigue:

$$d_{yx} = \frac{C - D}{C + D + T_y}$$

y

$$d_{xy} = \frac{C - D}{C + D + T_x}$$

en donde T_x es el número de pares que están empatados en X pero no en Y , y T_y es el número de pares empatados en Y pero no en X . Si hacemos que T_{xy} se refiera al número de pares empatados tanto en X como en Y , y volviendo a la ecuación (XVIII.12) para τ_b , veremos que $T = T_x + T_{xy}$, y $U = T_y + T_{xy}$, y por tanto, ya que el número total de pares $\frac{1}{2}N(N-1) = C + D + T_x + T_y + T_{xy}$, tendremos $C + D + T_y = \frac{1}{2}N(N-1) - (T_x + T_{xy}) = \frac{1}{2}N(N-1) - T$. De manera análoga, el denominador de d_{xy} es $C + D + T_x = \frac{1}{2}N(N-1) - U$. Así, el producto $d_{yx}d_{xy} = \tau_b^2$. En este sentido puede pensarse en las medidas asimétricas como *análogos declives*. Sin embargo, como su asimetría es función del número de empates, los que habitualmente dependen de los procedimientos de clasificación, la analogía con los declives b_{yx} y b_{xy} es, en el mejor de los casos, muy tenue.

Costner [5] ha señalado que puede darse a gamma una interpretación de reducción proporcional en el error semejante a la dada a las τ_b o λ_b de Goodman y Kruskal. Supongamos que deseamos predecir el orden de un par de casos con respecto a B . Si prescindimos de empates, nuestra probabilidad de incurrir en error, no conociendo nada más, sería de .5. Pero si conocemos el orden con respecto a A , resulta que el valor absoluto de gamma es igual al número de errores esperados conociendo A , menos el número esperado no conociendo A , dividido entre el número esperado no conociendo A .

Tenemos así disponible un número de medidas ordinales que difieren sólo en relación con el tratamiento de los empates en el denominador. Por desgracia, no tenemos de ordinario reglas claras de decisión para elegir entre dichas medidas, ya que las razones para los empates permanecen frecuentemente en la oscuridad. Wilson [23] ha demostrado que la propiedad de gamma, de reducción proporcional en el error, desaparece si se admite que los errores pueden cometerse cuando se predice un orden con respecto a B si, en realidad, el par está empatado con respecto a B .⁴ Parece como si este problema del manejo de los empates no tuviera solución sencilla. Tal vez la mejor regla empírica consista en hacer uso de tantas categorías de cada variable como sea posible, reduciendo así el número de empates, a la vez que las diferencias entre las distintas medidas.

⁴ Wilson [23] hace observar que tales empates no están excluidos del análisis de los modelos de regresión. Así, si dos casos se encuentran sumamente próximos en relación con sus puntuaciones de X , predeciríamos que sus puntuaciones de Y también lo estarían. En este sentido, si hay un par empatado con respecto a X , podemos esperar que lo esté también con respecto a Y , y cometeríamos un error si así no fuera. ¿Cuál es la importan-

Kruskal [17] ha demostrado que la medida de la r_s de Spearman puede ser interpretada en función de tríos de observaciones en lugar de pares, preguntándose cuál es la probabilidad de que, por lo menos, una de las tres observaciones sea concordante con las otras dos *a la vez*. Tal interpretación tiene una mucho menor atracción intuitiva que las interpretaciones mediante pares, aparte del hecho de que son mayores nuestros conocimientos acerca de los errores de muestreo de tau y gamma. Por estas razones prefiere Kruskal la tau a la r_s . Sin embargo, si la distribución básica de las dos variables es realmente bivariada normal, el valor absoluto de r_s será mayor que el de tau, y su comportamiento puede resultar mucho más semejante al de la correlación momento-producto. Trabajos previos no publicados muestran que el comportamiento de las r_s parciales (después de corregidos los empates) es muy singular al de las correlaciones parciales cuando las relaciones verdaderas son normales multivariadas (véase la definición en el próximo capítulo), por lo que sigue sin aclararse cuál de las medidas es preferible. Ante tal situación, el investigador deberá aplicar varias medidas diferentes para comprobar si se comportan de manera semejante al aplicarse a los datos que se examinan.

Finalmente, debemos tomar nota de un argumento de Wilson [22], quien afirma que *ninguna* medida ordinal que implique la idea de pares (o tríos) puede tener propiedades plenamente deseables. El punto básico de Wilson está en que el razonamiento teórico se funda normalmente en leyes que son apropiadas para un caso *único*, como cuando especificamos por ejemplo que el cambio de una unidad en X debe producir en Y el cambio de b_{yx} unidades. Con base en tales teorías, no tiene sentido pensar en función de pares ordenados, los que por necesidad nos fuerzan a realizar comparaciones a través de los casos. Si, por ejemplo, la propia teoría especifica que un cambio en el porcentaje de negros producirá un cambio en los niveles de discriminación, uno se está refiriendo tal vez a una "ley" que opera en el interior de una simple localidad (u otras unidades de observación). No se aplica directamente a comparaciones a través de pares de observaciones. Por supuesto que, en tanto uno defina su tarea como una simple generalización de poblaciones fijas, no se planteará este tipo de dificultad conceptual. El lector deberá consultar a Wilson si desea un análisis más completo. Está bien claro que el

cia de este "error" al predecir incorrectamente los empates, comparada con la del error de hacer predicciones equivocadas en los casos no empatados? Como puede verse, toda esta cuestión de la exclusión de empates, procedimiento que tiende a favorecer a gamma en relación con las demás medidas, no resulta cosa sencilla. Por ello, cuanto mayor sea el número de empates debidos a la crudeza de la medición, tanto más ambigua será la elección entre las medidas y mayor la sensibilidad de los resultados de tal elección.

empleo de medidas ordinales trae consigo cierto número de dificultades que hasta el momento no han sido resueltas adecuadamente.

EJERCICIOS

1. En los ejercicios 1 y 2 del capítulo XVII se calcularon tres coeficientes de correlación.
 - a) Para cada uno de dichos coeficientes, empléese el análisis de variancia para verificar la hipótesis nula de que $\rho = 0$. Respuesta, $F = .67$; $F = 7.09$; $F = 9.6$.
 - b) Colóquense intervalos de confianza del 99.9 por ciento con respecto a las tres r .
 - c) Verifíquese la relación entre la integración moral y la heterogeneidad en el caso de no linealidad.
 - d) Conviértanse los mismos datos en órdenes y obténganse la tau de Kendall y la r_s de Spearman para las tres correlaciones.
 - e) Verifíquese cada uno de estos coeficientes de rango ordenados en cuanto a significación.
2. En el ejercicio 3 del capítulo XVII se agruparon los índices de integración moral y de heterogeneidad. Cálculense para estos datos agrupados la tau, y la gamma de Kendall y compárese el resultado con el que se acaba de obtener antes en el ejercicio 1d de esta sección.

BIBLIOGRAFÍA

1. Anderson, T. R., y M. Zelditch: *A Basic Course in Statistics*, 2ª ed., Holt, Rinehart and Winston, Inc., Nueva York, 1968, caps. 7 y 8.
2. Blalock, H. M.: "Estimating Measurement Error Using Multiple Indicators and Several Points in Time", *American Sociological Review*, vol. 35, pp. 101-111, 1970.
3. Bohrnstedt, G. W.: "Observations on the Measurement of Change", en Edgar Borgatta (ed.), *Sociological Methodology 1969*, Jossey-Bass Inc., Publishers, San Francisco, 1969, cap. 4.
4. Christ, Carl: *Econometric Models and Methods*, John Wiley & Sons, Inc., Nueva York, 1966, Parte III.
5. Costner, H. L.: "Criteria for Measures of Association", *American Sociological Review*, vol. 30, pp. 341-353, 1965.
6. Costner, H. L.: "Theory, Deduction and Rules of Correspondence", *American Journal of Sociology*, vol. 75, pp. 245-263, 1969.
7. Croxton, F. E., y D. J. Cowden: *Applied General Statistics* 3ª ed., Prentice-Hall, Inc., Englewood Cliffs, N. J., 1967, cap. 20.
8. Goodman, L. A., y W. H. Kruskal: "Measures of Association for Cross Classifications", *Journal of the American Statistical Association*, vol. 49, pp. 732-764, 1954.
9. Goodman, L. A., y W. H. Kruskal: "Measures of Association for Cross Classifications, II: Further Discussion and References", *Journal of the American Statistical Association*, vol. 54, pp. 123-163, 1959.
10. Goodman, L. A., y W. H. Kruskal: "Measures of Association for

- Cross Classifications, III: Approximate Sampling Theory", *Journal of the American Statistical Association*, vol. 58, pp. 310-364, 1963.
11. Haggard, E. A.: *Intraclass Correlation and the Analysis of Variance*, The Dryden Press, Inc., Nueva York, 1958, pp. 22-26.
 12. Hagood, M. J., y D. O. Price: *Statistics for Sociologists*, Henry Holt and Company, Inc., Nueva York, 1952, cap. 23.
 13. Hays, W. L.: *Statistics*, Holt, Rinehart and Winston, Inc., Nueva York, 1963, cap. 16.
 14. Heise, D. R.: "Separating Reliability and Stability in Test-Retest Correlation", *American Sociological Review*, vol. 34, pp. 93-101, 1969.
 15. Johnston, J.: *Econometric Methods*, McGraw-Hill Book Company, Nueva York, 1963, Parte II.
 16. Kendall, M. G.: *Rank Correlation Methods*, Hafner Publishing Company, Inc., Nueva York, 1955, caps. 1, 3 y 4.
 17. Kruskal, W. H.: "Ordinal Measures of Association", *Journal of the American Statistical Association*, vol. 53, pp. 814-861, 1958.
 18. Mueller, J. H., K. Schuessler, y H. L. Costner: *Statistical Reasoning in Sociology*, 2ª ed., Houghton Mifflin Company, Boston, 1970, cap. 10.
 19. Siegel, Sidney: *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill Book Company, Nueva York, 1956, cap. 9.
 20. Somers, R. H.: "A New Asymmetric Measure of Association for Ordinal Variables", *American Sociological Review*, vol. 27, pp. 799-811, 1962.
 21. Wallis, W. A., y H. V. Roberts: *Statistics: A New Approach*, The Free Press of Glencoe, Ill., Chicago, 1956, cap. 17.
 22. Wilson, T. P.: "A Critique of Ordinal Variables", *Social Forces*, vol. 49, pp. 432-444, 1971.
 23. Wilson, T. P.: "A Proportional-Reduction-in-Error Interpretation for Kendall's tau-b", *Social Forces*, vol. 47, pp. 340-342, 1969.

XIX. CORRELACIÓN MÚLTIPLE Y PARCIAL

EN LOS dos últimos capítulos nos hemos ocupado de la relación entre dos escalas de intervalo, entre una variable dependiente y una sola variable independiente. Los análisis de correlación y regresión pueden extenderse fácilmente para comprender cualquier número de escalas de intervalo, una de las cuales puede tomarse como dependiente, y las demás como independientes. El problema se puede concebir como un problema de predicción en el que tratamos de predecir una variable dependiente Y a partir de las variables X_1, X_2, \dots, X_k . Habremos de servirnos de nuevo de un modelo muy sencillo, que será directamente análogo a la regresión lineal, excepto en cuanto al hecho de que habrá más de dos dimensiones.

El concepto de correlación se generalizará en dos formas. Emplearemos el término de *correlación parcial* para designar la correlación entre dos variables cualesquiera cuando los efectos de otras variables se han controlado. El de *correlación múltiple*, en cambio, servirá para indicar qué tanto de la variación total de la variable dependiente puede explicarse por todas las variables independientes actuando conjuntamente. Veremos que los materiales examinados en el presente capítulo comportan en su mayor parte extensiones directas de razonamientos presentados anteriormente. Una vez que hayamos ampliado las nociones de correlación y regresión, estaremos en condiciones, en el capítulo siguiente, de emprender el análisis de covariancia, que comporta una combinación de las técnicas de regresión con el análisis de la variancia.

XIX.1. Regresión múltiple y mínimos cuadrados

En la regresión múltiple tratamos de predecir una sola variable dependiente a partir de cualquier número de variables independientes. Si se da un gran número de variables de escala de intervalo que deban relacionarse entre sí, será posible, por supuesto, predecir cualquier variable particular a partir de cualquier combinación de las demás. Por lo regular resultará claro del contexto cuáles variables han de considerarse como independientes y cuáles como dependientes.¹ Así, por ejemplo, puede querer predecirse el éxito en la universidad a partir de una serie de marcas de aptitud y del éxito en la escuela secundaria. O puede resultar posible predecir la tasa de crecimiento de una ciudad

¹ Cuando se crea que existe una causación recíproca, o retroalimentación, de la variable "dependiente" hacia alguna de las demás, deberán emplearse ecuaciones simultáneas en lugar de mínimos cuadrados. Véanse [4] y [12].

conociendo factores como la magnitud actual, los porcentajes de mano de obra en las diversas ocupaciones, o la magnitud y la distancia del gran centro urbano más próximo.

En el análisis de regresión múltiple definimos la ecuación de regresión como el curso de la media de la variable dependiente Y para todas las combinaciones de X_1, X_2, \dots, X_k . En otros términos: para cada combinación de X fijas habrá una distribución de las Y . Cada distribución tendrá una media $\mu_{Y|X_1, X_2, \dots, X_k}$ y una desviación estándar $\sigma_{Y|X_1, X_2, \dots, X_k}$, y habremos de suponer una vez más que todas estas distribuciones son normales y que las desviaciones estándar son iguales (homoscedasticidad). El recorrido de las medias ya no seguirá siendo una curva en el espacio bidimensional, sino que será, antes bien, una especie de hipersuperficie en un espacio de $(k+1)$ dimensiones. Es obvio que ya no estaremos en condiciones de representar un curso semejante, excepto en el caso en que sólo tengamos dos variables independientes X_1 y X_2 .

En el capítulo anterior supusimos una ecuación de regresión lineal de la forma $Y = \alpha + \beta X$. Y habremos de volver a suponer una forma sencilla de la ecuación de regresión. Supongamos, pues, que el curso de las medias de Y adopta la forma:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k \quad (\text{XIX.1})$$

en donde $\alpha, \beta_1, \beta_2, \dots, \beta_k$ son constantes. Esta es la ecuación más sencilla posible de regresión múltiple, y es directamente análoga a la regresión lineal en el caso de dos variables. En efecto, si todas las β , excepto una, son cero, el problema se reduce al caso bidimensional.

Si podemos suponer una población "normal multivariable" en la que cada variable esté distribuida normalmente alrededor de todas las demás, entonces podemos satisfacer los tres supuestos requeridos. En otros términos, una distribución normal multivariable nos asegura que las ecuaciones de regresión serán de la forma anterior, que las distribuciones de las Y para X determinadas serán todas normales, y que las variancias serán también iguales. Esto constituye una generalización obvia de las propiedades de la distribución normal bivariable. Sobre decir que la distribución normal multivariable no puede representarse geoméricamente (pese a que tiene una ecuación algebraica perfectamente definida), toda vez que tuvimos ya necesidad de tres dimensiones para representar el caso bivariable.

Con objeto de proporcionar una mejor comprensión intuitiva de la naturaleza de las extensiones implicadas, será conveniente examinar el caso en que no hay más que dos variables independientes (véase la figura XIX.1). La ecuación de regresión $Y = \alpha + \beta_1 X_1 + \beta_2 X_2$ puede representarse en este caso por medio de un

plano en un espacio tridimensional. Si dejamos que $X_1 = X_2 = 0$, obtenemos $Y = \alpha$, lo que indica que el plano de regresión corta el eje de las Y a una altura α . Con objeto de obtener una interpretación de las β , tomamos las intersecciones del plano de regresión con planos perpendiculares a los ejes de X_1 y X_2 . Así, por ejemplo, si tomamos un plano perpendicular al eje de X_2 ,

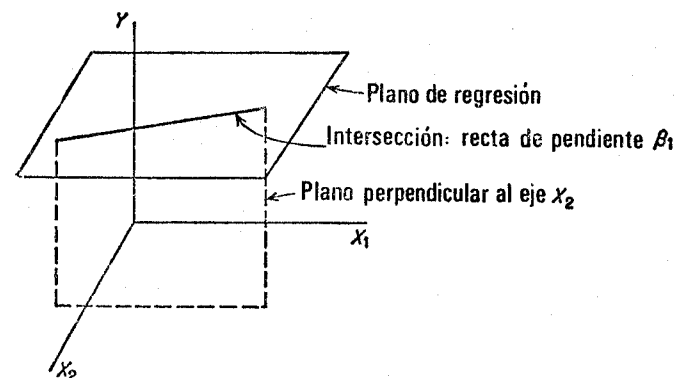


FIG. XIX.1. Interpretación geométrica de la regresión múltiple de Y sobre X_1 y X_2 .

mantenemos de hecho a X_2 constante, ya que todos los puntos situados en este plano tendrán el mismo valor de X_2 . Este plano corta el plano de regresión en una recta, y la pendiente de esta recta será β_1 . En otros términos, si mantenemos X_2 en un valor fijo, β_1 representa la pendiente de la línea de regresión de Y a X_1 . Y en forma análoga, el hecho de mantener constante a X_1 nos da un plano que interseca el plano de regresión en una línea de pendiente β_2 .

Conviene observar que las β empleadas en la regresión múltiple no serán por lo regular las mismas que las que se obtuvieron en el caso de dos variables. Designando el caso de dos variables como regresión *total*, vemos que la β empleada en la regresión total se obtiene *prescindiendo* de las demás variables independientes, y no manteniéndolas constantes. Las β obtenidas en las ecuaciones de regresión múltiple se designan como coeficientes *parciales*, porque comportan pendientes que se obtendrían eliminando o manteniendo constantes cada una de las demás variables independientes consideradas en la ecuación de regresión.

El concepto de los mínimos cuadrados puede ampliarse en una forma semejante. Como quiera que es casi siempre necesario apreciar una ecuación de regresión adaptando una a los datos empíricos, habremos de requerir una vez más que la ecuación de

- Cross Classifications, III: Aproximate Sampling Theory", *Journal of the American Statistical Association*, vol. 58, pp. 310-364, 1963.
11. Haggard, E. A.: *Intraclass Correlation and the Analysis of Variance*, The Dryden Press, Inc., Nueva York, 1958, pp. 22-26.
 12. Hagood, M. J., y D. O. Price: *Statistics for Sociologists*, Henry Holt and Company, Inc., Nueva York, 1952, cap. 23.
 13. Hays, W. L.: *Statistics*, Holt, Rinehart and Winston, Inc., Nueva York, 1963, cap. 16.
 14. Heise, D. R.: "Separating Reliability and Stability in Test-Retest Correlation", *American Sociological Review*, vol. 34, pp. 93-101, 1969.
 15. Johnston, J.: *Econometric Methods*, McGraw-Hill Book Company, Nueva York, 1963, Parte II.
 16. Kendall, M. G.: *Rank Correlation Methods*, Hafner Publishing Company, Inc., Nueva York, 1955, caps. 1, 3 y 4.
 17. Kruskal, W. H.: "Ordinal Measures of Association", *Journal of the American Statistical Association*, vol. 53, pp. 814-861, 1958.
 18. Mueller, J. H., K. Schuessler, y H. L. Costner: *Statistical Reasoning in Sociology*, 2ª ed., Houghton Mifflin Company, Boston, 1970, cap. 10.
 19. Siegel, Sidney: *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill Book Company, Nueva York, 1956, cap. 9.
 20. Somers, R. H.: "A New Asymmetric Measure of Association for Ordinal Variables", *American Sociological Review*, vol. 27, pp. 799-811, 1962.
 21. Wallis, W. A., y H. V. Roberts: *Statistics: A New Approach*, The Free Press of Glencoe, Ill., Chicago, 1956, cap. 17.
 22. Wilson, T. P.: "A Critique of Ordinal Variables", *Social Forces*, vol. 49, pp. 432-444, 1971.
 23. Wilson, T. P.: "A Proportional-Reduction-in-Error Interpretation for Kendall's tau-b", *Social Forces*, vol. 47, pp. 340-342, 1969.

XIX. CORRELACIÓN MÚLTIPLE Y PARCIAL

EN LOS DOS últimos capítulos nos hemos ocupado de la relación entre dos escalas de intervalo, entre una variable dependiente y una sola variable independiente. Los análisis de correlación y regresión pueden extenderse fácilmente para comprender cualquier número de escalas de intervalo, una de las cuales puede tomarse como dependiente, y las demás como independientes. El problema se puede concebir como un problema de predicción en el que tratamos de predecir una variable dependiente Y a partir de las variables X_1, X_2, \dots, X_k . Habremos de servirnos de nuevo de un modelo muy sencillo, que será directamente análogo a la regresión lineal, excepto en cuanto al hecho de que habrá más de dos dimensiones.

El concepto de correlación se generalizará en dos formas. Emplearemos el término de *correlación parcial* para designar la correlación entre dos variables cualesquiera cuando los efectos de otras variables se han controlado. El de *correlación múltiple*, en cambio, servirá para indicar qué tanto de la variación total de la variable dependiente puede explicarse por todas las variables independientes actuando conjuntamente. Veremos que los materiales examinados en el presente capítulo comportan en su mayor parte extensiones directas de razonamientos presentados anteriormente. Una vez que hayamos ampliado las nociones de correlación y regresión, estaremos en condiciones, en el capítulo siguiente, de emprender el análisis de covariancia, que comporta una combinación de las técnicas de regresión con el análisis de la variancia.

XIX.1. Regresión múltiple y mínimos cuadrados

En la regresión múltiple tratamos de predecir una sola variable dependiente a partir de cualquier número de variables independientes. Si se da un gran número de variables de escala de intervalo que deban relacionarse entre sí, será posible, por supuesto, predecir cualquier variable particular a partir de cualquier combinación de las demás. Por lo regular resultará claro del contexto cuáles variables han de considerarse como independientes y cuáles como dependientes.¹ Así, por ejemplo, puede querer predecirse el éxito en la universidad a partir de una serie de marcas de aptitud y del éxito en la escuela secundaria. O puede resultar posible predecir la tasa de crecimiento de una ciudad

¹ Cuando se crea que existe una causación recíproca, o retroalimentación, de la variable "dependiente" hacia alguna de las demás, deberán emplearse ecuaciones simultáneas en lugar de mínimos cuadrados. Véanse [4] y [12].

conociendo factores como la magnitud actual, los porcentajes de mano de obra en las diversas ocupaciones, o la magnitud y la distancia del gran centro urbano más próximo.

En el análisis de regresión múltiple definimos la ecuación de regresión como el curso de la media de la variable dependiente Y para todas las combinaciones de X_1, X_2, \dots, X_k . En otros términos: para cada combinación de X fijas habrá una distribución de las Y . Cada distribución tendrá una media $\mu_{Y|X_1, X_2, \dots, X_k}$ y una desviación estándar $\sigma_{Y|X_1, X_2, \dots, X_k}$, y habremos de suponer una vez más que todas estas distribuciones son normales y que las desviaciones estándar son iguales (homoscedasticidad). El recorrido de las medias ya no seguirá siendo una curva en el espacio bidimensional, sino que será, antes bien, una especie de hipersuperficie en un espacio de $(k+1)$ dimensiones. Es obvio que ya no estaremos en condiciones de representar un curso semejante, excepto en el caso en que sólo tengamos dos variables independientes X_1 y X_2 .

En el capítulo anterior supusimos una ecuación de regresión lineal de la forma $Y = \alpha + \beta X$. Y habremos de volver a suponer una forma sencilla de la ecuación de regresión. Supongamos, pues, que el curso de las medias de Y adopta la forma:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k \quad (\text{XIX.1})$$

en donde $\alpha, \beta_1, \beta_2, \dots, \beta_k$ son constantes. Esta es la ecuación más sencilla posible de regresión múltiple, y es directamente análoga a la regresión lineal en el caso de dos variables. En efecto, si todas las β , excepto una, son cero, el problema se reduce al caso bidimensional.

Si podemos suponer una población "normal multivariable" en la que cada variable esté distribuida normalmente alrededor de todas las demás, entonces podemos satisfacer los tres supuestos requeridos. En otros términos, una distribución normal multivariable nos asegura que las ecuaciones de regresión serán de la forma anterior, que las distribuciones de las Y para X determinadas serán todas normales, y que las variancias serán también iguales. Esto constituye una generalización obvia de las propiedades de la distribución normal bivariable. Sobra decir que la distribución normal multivariable no puede representarse geoméricamente (pese a que tiene una ecuación algebraica perfectamente definida), toda vez que tuvimos ya necesidad de tres dimensiones para representar el caso bivariable.

Con objeto de proporcionar una mejor comprensión intuitiva de la naturaleza de las extensiones implicadas, será conveniente examinar el caso en que no hay más que dos variables independientes (véase la figura XIX.1). La ecuación de regresión $Y = \alpha + \beta_1 X_1 + \beta_2 X_2$ puede representarse en este caso por medio de un

plano en un espacio tridimensional. Si dejamos que $X_1 = X_2 = 0$, obtenemos $Y = \alpha$, lo que indica que el plano de regresión corta el eje de las Y a una altura α . Con objeto de obtener una interpretación de las β , tomamos las intersecciones del plano de regresión con planos perpendiculares a los ejes de X_1 y X_2 . Así, por ejemplo, si tomamos un plano perpendicular al eje de X_2 ,

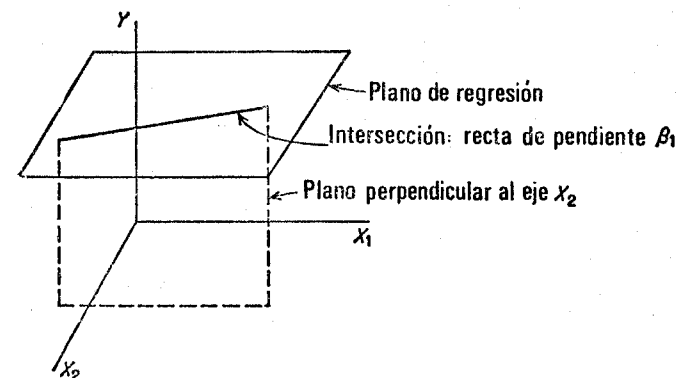


FIG. XIX.1. Interpretación geométrica de la regresión múltiple de Y sobre X_1 y X_2 .

mantenemos de hecho a X_2 constante, ya que todos los puntos situados en este plano tendrán el mismo valor de X_2 . Este plano corta el plano de regresión en una recta, y la pendiente de esta recta será β_1 . En otros términos, si mantenemos X_2 en un valor fijo, β_1 representa la pendiente de la línea de regresión de Y a X_1 . Y en forma análoga, el hecho de mantener constante a X_1 nos da un plano que intersecta el plano de regresión en una línea de pendiente β_2 .

Conviene observar que las β empleadas en la regresión múltiple no serán por lo regular las mismas que las que se obtuvieron en el caso de dos variables. Designando el caso de dos variables como regresión *total*, vemos que la β empleada en la regresión total se obtiene *prescindiendo* de las demás variables independientes, y no manteniéndolas constantes. Las β obtenidas en las ecuaciones de regresión múltiple se designan como coeficientes *parciales*, porque comportan pendientes que se obtendrían eliminando o manteniendo constantes cada una de las demás variables independientes consideradas en la ecuación de regresión.

El concepto de los mínimos cuadrados puede ampliarse en una forma semejante. Como quiera que es casi siempre necesario apreciar una ecuación de regresión adaptando una a los datos empíricos, habremos de requerir una vez más que la ecuación de

estimación revista una forma particular y se sirva del criterio de los mínimos cuadrados para conseguir el "mejor" ajuste. Nos serviremos de una ecuación de mínimos cuadrados de la forma:

$$Y_p = a + b_1X_1 + b_2X_2 + \dots + b_kX_k \quad (\text{XIX.2})$$

y volverá a resultar que, a condición que la ecuación de regresión sea efectivamente de la misma forma, la ecuación de los mínimos cuadrados representa la mejor estimación de la ecuación de regresión. En otros términos, si nos servimos de a para estimar α , y de b_i para estimar β_i , estas estimaciones serán insesgadas y, al propio tiempo, de eficiencia máxima. Por consiguiente, nuestra atención puede fijarse en el análisis de los mínimos cuadrados como método práctico de estimar una ecuación teórica que se aplica a la población. Si sólo hay dos variables independientes, ajustaremos una serie de puntos en el espacio tridimensional con un plano de mejor ajuste. En un espacio de $(k+1)$ dimensiones, por su parte, ajustaremos puntos con un hiperplano de k dimensiones, si es que semejante figura se puede concebir.

Tomando el caso tridimensional, reduciremos al mínimo la cantidad $\Sigma(Y - Y_p)^2$, que representa la suma de las desviaciones al cuadrado respecto del plano de mínimos cuadrados en la dimensión vertical de Y (véase la figura XIX.2). El resultado será un plano único de mejor ajuste, determinado por valores específicos de a , b_1 y b_2 . Según veremos, puede utilizarse luego un coeficiente de correlación múltiple para medir la bondad de ajuste de los puntos al plano de mínimos cuadrados. Sería también posible, por supuesto, medir el grado de ajuste mediante una desviación estándar referida al plano, y podríamos comparar esta desviación con la desviación estándar en relación con la \bar{Y} fija (representada ahora como plano perpendicular al eje de las Y). Algebráicamente, el caso más general es una ampliación di-

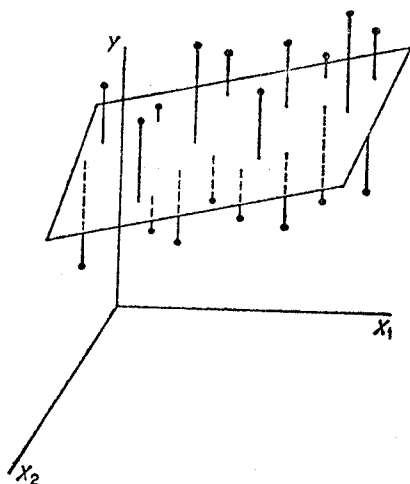


FIG. XIX.2. Plano de mínimos cuadrados, que reduce al mínimo las sumas de las desviaciones al cuadrado en la dimensión vertical Y .

recta del caso de tres variables. La cantidad $\Sigma(Y - Y_p)^2$ se minimiza, y habrá que calcular $(k+1)$ coeficientes, esto es, a , b_1 , b_2 , ..., b_k . El cálculo efectivo de estos coeficientes será posible examinarlo más adelante, cuando hayamos efectuado el estudio de la correlación parcial.

XIX.2. Correlación parcial

Podemos servirnos de este modelo de regresión múltiple para obtener medidas del grado de relación entre una variable dependiente Y y cualquiera de las variables independientes, controlando una o más de ellas. El término de *correlación parcial* se emplea para designar este tipo de procedimiento de control, el cual, según veremos, es básicamente muy similar al referente al análisis de la variancia por dos métodos. En la correlación parcial controlamos ajustando valores de las variables dependientes e independientes con objeto de tomar en cuenta las puntuaciones de las variables de control. Para comprender la naturaleza de la correlación parcial y el procedimiento de ajuste, limitaremos por ahora nuestra atención a los problemas más sencillos, en los

que figuran sólo tres variables, y supondremos modelos de regresión lineal entre las tres combinaciones de variables tomadas de dos en dos.

Supongamos que queremos medir el grado de relación entre una variable dependiente Y y una variable independiente X_1 , controlando en relación con otra variable independiente X_2 . Para servirnos de un ejemplo concreto, podemos tener interés en predecir la tasa de discriminación económica contra los negros, medida por las diferencias de ingreso entre los blancos y los negros, y el grado de urbanización, según resulta del porcentaje de un distrito designado como urbano. Se espera con seguridad que el porcentaje de negros en el distrito afectará asimismo la tasa de discrimi-

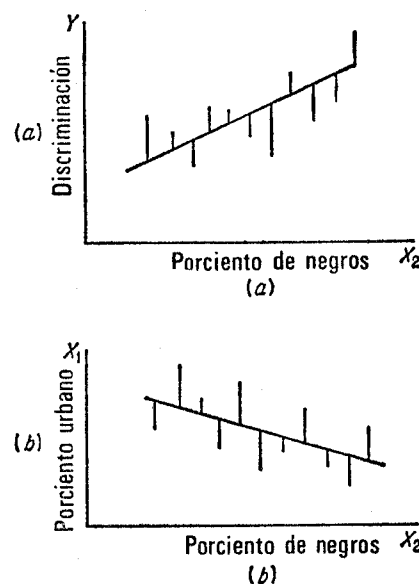


FIG. XIX.3. Rectas de mínimos cuadrados indicando los residuos entre: a) Y y X_2 , y b) entre X_1 y X_2 ...

nación, y se decide en consecuencia utilizar el porcentaje de negros como variable de control.

Supóngase que las líneas de mínimos cuadrados entre la discriminación Y y el porcentaje de negros X_2 y entre el porcentaje urbano X_1 y el porcentaje de negros son como las que se indican en la figura XIX.3. La relación entre la discriminación y el porcentaje de negros es positiva, lo que indica que tasas elevadas de discriminación se hallan asociadas a porcentajes elevados de la minoría en cuestión. Por otra parte, la relación entre el índice de urbanización y el porcentaje de negros es negativa. Sobre la base de la sola información, anticiparíamos una relación negativa entre las tasas de discriminación y la urbanización. En otros términos: las áreas urbanas podrían tener tasas bajas, debido simplemente al hecho de que en promedio cuentan con relativamente menos negros. Supóngase, sin embargo, que estuviéramos en condiciones, en alguna forma, de "forzar" todos los distritos a tener los mismos porcentajes de la minoría. Podríamos eliminar en esta forma el efecto perturbador de esta tercera variable. En realidad, por supuesto, no podemos hacer efectivamente todos los porcentajes de la minoría iguales, pero podemos por lo menos efectuar ajustes en relación con el hecho de que difieren. Como quiera que sabemos (o podemos apreciarla) la relación entre la variable de control y cada una de las otras dos variables, podemos predecir cómo se comportarían éstas respecto de cambios en la variable de control.² De hecho, las ecuaciones de mínimos cuadrados representadas en la figura XIX.3 constituyen nuestras ecuaciones de predicción y pueden emplearse en el proceso de ajuste.

Al relacionar la discriminación Y al porcentaje de negros X_2 , podemos concebir la variación de las tasas de discriminación como resultado de dos componentes, siendo la primera de ellas el porcentaje de negros y, la segunda, debiéndose a otros factores, uno de los cuales puede ser la urbanización. Como ya vimos, esta segunda componente puede representarse como *desviaciones* respecto de la ecuación de mínimos cuadrados que comporta Y y X_2 . En términos de X_2 , estas desviaciones o *residuos* representan error. Aun si X_2 se mantuviera constante, subsistirían. En estos residuos, por consiguiente, son en los que estamos en verdad interesados, ya que representan la cantidad de variación en la discriminación que subsiste una vez que el porcentaje de negros ha explicado todo lo que podía de la variación.

Y en forma análoga, nos interesaremos también en los residuos o desviaciones respecto de la ecuación empleada para predecir el porcentaje urbano a partir del porcentaje de negros. En otros

² Debe insistirse de nuevo en que la justificación para interpretar resultados de tal ajuste, hecho con lápiz y papel, implica el supuesto causal de que la variable de control puede afectar a las otras dos variables.

términos: dejamos que el porcentaje de negros explique la variación en las otras dos variables todo lo que pueda. Y si ahora ponemos los residuos en relación de unos con otros, obtenemos una medida de la relación entre Y y X_1 que es independiente de los efectos de X_2 . *La correlación parcial entre Y y X_1 controlando X_2 , puede definirse como la correlación entre los residuos de las regresiones de Y sobre X_2 y de X_1 sobre X_2 .* En cierto sentido, pues, la correlación parcial representa la correlación entre "errores" respecto de la variable de control.

El que tenga algún sentido controlar en relación con X_2 correlacionando residuos podrá parecer oscuro aún. Quizá la explicación sea más aceptable si examinamos más de cerca una relación hipotética entre dichos residuos. Supongamos, por ejemplo, que para el distrito A encontramos un gran residuo negativo al correlacionar Y con X_2 . Esto significa que el distrito A presenta considerablemente menos discriminación de lo que se esperaría conociendo solamente su porcentaje de minoría. El punto que representara dicho distrito particular se situaría en algún lugar por debajo de la línea de mínimos cuadrados. Supóngase, asimismo, que el residuo para este mismo distrito fuera positivo al correlacionar X_1 con X_2 . En tal caso sabemos que el distrito en cuestión está más urbanizado de lo que se esperaría conociendo solamente su porcentaje de minoría. Tenemos, por lo tanto, un distrito relativamente urbanizado con tasas bajas de discriminación, y sabemos, además, que dichos valores son altos y bajos respectivamente en comparación con otros distritos del mismo porcentaje de minoría. No podemos, por consiguiente, atribuir la relación negativa entre los residuos al hecho de que la cifra del porcentaje de negros acontezca ser alta o baja. Y en forma análoga, el distrito B puede tener grandes residuos positivos para Y , pero negativos para X_1 . Por consiguiente, este distrito tendría mayores tasas de discriminación de lo que se esperaba, pero estaría al propio tiempo menos urbanizado que otros distritos del mismo porcentaje de minoría. Es obvio que si muchos distritos son similares a A o B, obtendremos una correlación negativa, entre los residuos, indicando una correlación negativa entre la discriminación y la urbanización, ajustando en relación con el porcentaje de negros.

La correlación parcial da una sola medida que resume el grado de relación entre dos variables al controlar en relación con otra. Según veremos al examinar los procedimientos de cálculo, el razonamiento puede extenderse fácilmente a variables de control adicionales. Podemos concebir varias ecuaciones de regresión múltiple, una de las cuales comporte Y y todas las variables de control, y la otra relacionando X_1 con estas mismas variables. Pueden obtenerse los residuos de cada una de esas ecuaciones de regresión múltiple y relacionarlos luego. Ajustaremos en esta

forma en relación con todas las variables de control al mismo tiempo. El punto importante, aquí, es que sólo obtenemos una correlación parcial, en tanto que al controlar con las tablas de contingencia (con concesiones para la interacción) obteníamos una medida separada para cada una de las categorías de las variables de control.

En el capítulo xv vimos que el grado de relación entre dos variables podía variar de una categoría de la variable de control a otra. Así, por ejemplo, si el porcentaje de negros se hubiera categorizado, es perfectamente posible que hubiéramos obtenido una elevada correlación negativa entre la discriminación y la urbanización para distritos de porcentajes de minoría muy bajos, pero con una correlación positiva, de todos modos, en el extremo opuesto del continuo porcentaje de negros. Por lo tanto, el hecho de que en la correlación parcial hayamos obtenido una sola medida de resumen puede acaso oscurecer cierta información acerca de la interacción.

Resulta que el coeficiente de correlación parcial puede ser también interpretado como un *promedio ponderado* de los coeficientes de correlación que se hubieran obtenido si la variable de control hubiera sido dividida en muy pequeños intervalos y calculando correlaciones separadas dentro de cada una de estas categorías. La naturaleza exacta de este procedimiento de ponderación carece de importancia, ya que nunca se hace uso de él en la práctica. No tendría, por tanto, sentido pensar que las correlaciones parciales relacionan dos variables que "mantienen constante" a una tercera, ya que la fuerza de su relación puede variar de acuerdo con el valor particular en que se mantiene constante la variable de control.

En el caso de la distribución normal multivariable, sabemos que todas las ecuaciones de regresión tendrán la forma especial descrita por la ecuación (XIX.1). Pero la distribución normal multivariable posee además otra propiedad notable. Y es que la fuerza de la relación entre dos variables será la misma independientemente de los valores de las variables de control. En otros términos: si se seleccionara un gran número de categorías de una variable de control y se obtuvieran correlaciones dentro de cada una de dichas categorías, todas las correlaciones tendrían el mismo valor. Por lo tanto, la correlación parcial tendría el mismo valor que cada una de esas correlaciones dentro de las categorías. En este caso especial, tendría así cierto objeto pensar en términos del mantenimiento constante de la tercera variable de control. Sin embargo, como quiera que en el mejor de los casos sólo podemos aproximarnos a la distribución normal multivariable con datos reales, es más seguro pensar en la correlación parcial como promedio ponderado, o como si comportara un ajuste en relación con la variable de control.

Cálculo de los coeficientes de correlación parcial. El cálculo de las correlaciones parciales es sumamente sencillo, a menos que se desee controlar en relación con tres o más variables a la vez. Antes de presentar la fórmula de la correlación parcial, hemos de introducir un cambio de notación. Por desgracia, lo que constituye una notación conveniente para un objeto no lo es para otro, ni es el uso convencional totalmente concordante. Hemos venido representando la variable dependiente por Y y las variables independientes por X_1, X_2, \dots, X_k . En reconocimiento del hecho de que la elección de la variable dependiente es a menudo más o menos arbitraria y que, por consiguiente, podemos querer calcular correlaciones parciales entre varias combinaciones de variables, convendrá reenumerar simplemente las variables de 1 a $k+1$ y representar la correlación entre las variables 1 y 2, controlando en relación con 3 mediante $r_{12 \cdot 3}$. Y en forma análoga, la correlación entre las variables 2 y 3, controlando en relación con 1, por medio de $r_{23 \cdot 1}$.

Esta notación puede extenderse fácilmente a cualquier número de variables de control añadiendo más números a la derecha del punto central del subíndice. Así, por ejemplo, la relación entre las variables 5 y 7, con control de las variables 1, 2, 3, 4 y 6, nos vendría dada por $r_{57 \cdot 12346}$. El orden de las dos variables a la izquierda del punto no juega papel alguno, lo mismo que el de la derecha. Para distinguir entre parciales con números diferentes de control, designamos el número de controles como el *orden* de la correlación. Así, pues, un primer orden parcial tendrá un control; un segundo orden, dos controles, y así sucesivamente. En concordancia con esta terminología, la correlación sin controles se designa a menudo como correlación de orden cero. Según se ha indicado más arriba, el término *correlación total* se emplea también para designar una correlación entre dos variables sin controles.

Podemos dar ahora la fórmula del primer orden parcial $r_{ij \cdot k}$:

$$r_{ij \cdot k} = \frac{r_{ij} - (r_{ik})(r_{jk})}{\sqrt{1 - r_{ik}^2} \sqrt{1 - r_{jk}^2}} \quad (\text{XIX.3})$$

Obsérvese que la primera correlación del numerador es la correlación total entre las dos variables a relacionar (i y j). La variable de control figura en la segunda expresión del numerador, en donde se la relaciona con cada una de las otras variables, así como en ambos términos del denominador. Cualquier correlación parcial particular puede obtenerse a partir de esta fórmula general, sustituyendo i, j y k por los números apropiados. Así, por ejemplo:

$$r_{13.2} = \frac{r_{13} - (r_{12})(r_{23})}{\sqrt{1 - r_{12}^2} \sqrt{1 - r_{23}^2}}$$

En un estudio de 150 distritos del Sur [3], la correlación entre la discriminación en los ingresos y el porcentaje de negros fue de .536; aquella entre la discriminación en los ingresos y el porcentaje urbano fue de .139, y la correlación entre los porcentajes de negros y urbano fue de -.248. Si llamamos al índice de discriminación variable 1, al porcentaje de negros variable 2 y al porcentaje urbano variable 3, podemos obtener la correlación parcial entre la discriminación y el porcentaje urbano, controlado en relación con el porcentaje de negros. Tenemos así:

$$r_{13.2} = \frac{.139 - (.536)(-.248)}{\sqrt{1 - (.536)^2} \sqrt{1 - (-.248)^2}} = \frac{.2719}{.8178} = .332$$

Este resultado puede interpretarse como correlación entre la discriminación y el porcentaje urbano una vez que se ha dejado que el porcentaje de negros explique todo lo que puede de ambas variables.

Si bien no resultará inmediatamente evidente que la fórmula anterior pueda derivarse de la definición de la correlación parcial en términos de una correlación de residuos, la fórmula de cálculo, por lo menos, tiene un sentido. En efecto, en el numerador sustraemos esencialmente un factor de corrección de la correlación total. En cuanto al denominador, éste consta de dos factores de corrección, ninguno de los cuales puede ser mayor que la unidad, que toman en cuenta el hecho de que la variable de control explica cierta proporción de la variación de las otras variables. Si elevamos al cuadrado el coeficiente de correlación parcial, el número resultante representará la proporción de variación de la variable 1 (discriminación), dejada inexplicada por 2 (porcentaje de negros), pero que puede explicarse por los valores ajustados de X_3 (porcentaje urbano).

Examinemos la ecuación (XIX.3) con mayor atención, para ver cómo la correlación parcial se comporta en relación con las tres correlaciones totales. Con fines de simplificación, supongamos primero que r_{ij} es positiva. Si r_{ik} y r_{jk} tienen ambas el mismo signo (ya sea positivo o negativo), su producto será positivo, y el numerador será o bien un número positivo menor que r_{ij} , o será incluso cero o negativo. Por otra parte, el denominador será siempre menor que la unidad, a menos que $r_{ik} = r_{jk} = 0$. Por consiguiente, la fracción resultante puede ser casi cualquier número entre -1.0 y +1.0, según sea la magnitud de las tres correlaciones totales. Veremos más adelante exactamente lo que

podemos y lo que no podemos decir acerca del comportamiento de la parcial en estas circunstancias.

Supongamos ahora que las correlaciones con la variable de control son de signos opuestos. Obtenemos en tal caso un producto negativo a sustraer de un número positivo, y el resultado será un número positivo mayor. Esto significa que si empezamos con dos variables relacionadas positivamente y si podemos encontrar una variable de control relacionada negativamente con una de ellas pero positivamente con la otra, la parcial resultante será mayor que la correlación de orden cero. Si la correlación de la variable de control con una u otra de las otras variables acontece ser cero, el factor de corrección del numerador será cero. Pero si la variable de control se halla correlacionada ya sea positiva o negativamente con la variable restante, el denominador será menor que la unidad, y la correlación parcial volverá a ser mayor que la correlación total.

Si hubiéramos empezado con una correlación total negativa, una variable de control relacionada con cada una de las otras dos en la misma dirección (ya sea positiva o negativa) produciría una correlación negativa mayor. Sin embargo, si se relacionara con ellas en sentido opuesto, el resultado sería análogo al que se ha descrito en primer término (en donde la correlación total era positiva y el factor de corrección positivo asimismo). ¿Por qué? En cambio, si la variable de control no se relacionara con una de las otras variables, el resultado sería una correlación parcial con un valor absoluto mayor que la total. Y si la variable de control no se relacionara con *ninguna* de las otras variables, la correlación parcial sería exactamente igual, por supuesto, a la correlación total. Una vez que hayamos examinado la relación entre la correlación parcial y las interpretaciones causales, estaremos en condiciones de dar una justificación intuitiva del comportamiento de las correlaciones parciales en estas diversas condiciones.

Las fórmulas de las parciales de segundo orden o superior son directamente análogas a las de la parcial de primer orden. En efecto, vamos añadiendo simplemente variables de control sucesivas, empezando cada vez con la parcial de orden uno menos que el deseado. Así, por ejemplo, las fórmulas de $r_{ij.kl}$ y $r_{ij.klm}$ serán:

$$r_{ij.kl} = \frac{r_{ij.kl} - (r_{il.k})(r_{jl.k})}{\sqrt{1 - r_{il.k}^2} \sqrt{1 - r_{jl.k}^2}} \quad (\text{XIX.4})$$

y

$$r_{ij.klm} = \frac{r_{ij.klm} - (r_{im.kl})(r_{jm.kl})}{\sqrt{1 - r_{im.kl}^2} \sqrt{1 - r_{jm.kl}^2}} \quad (\text{XIX.5})$$

Obsérvese que en la ecuación (XIX.4) suponemos que ya hemos controlado en relación con la variable X_k . Por lo tanto, la k aparece a la derecha del punto en las tres parciales de primer orden. Y en forma análoga, en la ecuación (XIX.5) hemos controlado previamente en relación con X_k y X_i , y de aquí que estas cantidades figuren en cada una de las parciales de segundo orden.

Las parciales de cuarto y quinto orden podrían obtenerse en forma análoga, y resultará instructivo tratar de escribir las fórmulas de estas parciales de orden superior. De modo que la manera de calcular estas últimas es idéntica a la que empleamos en el caso del primer orden. Pero el trabajo que supone se hace prontamente aburrido. Así, por ejemplo, con objeto de obtener una parcial de tercer orden con este método, han de haberse obtenido previamente tres parciales de segundo orden, cada una de las cuales ha de haberse obtenido a su vez calculando parciales de primer orden a partir de correlaciones de orden cero. Si el lector tratara de expresar la fórmula de las parciales de tercer orden directamente en términos de las correlaciones de orden cero, se daría cuenta del trabajo que esto representa.

Afortunadamente, en la investigación sociológica rara vez resulta necesario ir mucho más allá de las parciales de segundo o tercer orden. Por lo regular, la adición de controles más allá del segundo o tercer control proporciona muy pocos conocimientos nuevos. Si se hace necesario servirse de parciales de orden superior, o de ecuaciones de regresión múltiple de cuatro o más variables, existen ciertas rutinas de cálculo que facilitan considerablemente la labor. Para tratar tales problemas el lector podrá referirse ya sea al método abreviado de Doolittle o al de Dwyer, de la raíz cuadrada (véanse [9] y [11]). De estos dos métodos, el primero tal vez sea más satisfactorio, por cuanto permite obtener directamente las parciales sucesivas $r_{12\cdot3}$, $r_{12\cdot34}$, $r_{12\cdot345}$, etcétera.

Correlación parcial de rangos ordenados. La teoría de las correlaciones parciales de rangos ordenados está menos bien desarrollada. Puede extenderse al caso de las parciales de primer orden la tau de Kendall, aunque la interpretación de la tau parcial no resulta tan aceptable intuitivamente como en el caso de la correlación de producto-momento. Si no hay empates, resulta que la fórmula de la tau parcial es idéntica a la que hemos estado empleando. (Véanse [13] y [19].) Así, por ejemplo:

$$\tau_{ij\cdot k} = \frac{\tau_{ij} - (\tau_{ik})(\tau_{jk})}{\sqrt{1 - \tau_{ik}^2} \sqrt{1 - \tau_{jk}^2}} \quad (\text{XIX.6})$$

En el caso que haya un gran número de empates podrá usarse un procedimiento alternativo, sugerido por Davis [7] para el caso

de gamma, pero su principio puede aplicarse a cualquiera de las medidas de tau o a d_{yx} y d_{xy} . Si controlamos para W , categorizaremos simplemente W , computando gammas (u otras medidas) dentro de las categorías de W , obteniendo un promedio ponderado de dichas gammas. Pero en lugar de ponderar según el número de casos en cada categoría, lo haremos según el número de pares afectados. De esta manera, en el caso de una gamma parcial, estamos considerando tan sólo aquellos pares que no están empatados, bien en X o en Y , pero que lo están con respecto a la categoría de la variable de control. Davis demuestra que tal promedio ponderado puede recibir una simple interpretación de reducción proporcional en el error. Quade [16], ofrece un procedimiento análogo de promedio ponderado para el caso de tau, facilitando asimismo una prueba de significancia para dicho parcial.

En la investigación exploratoria puede tener sentido el utilizar múltiples variables de control, bien por ampliación de la fórmula (XIX.6) o dividiendo las variables de control en múltiples subcategorías. Los cimientos teóricos de tales procedimientos no son, sin embargo, muy firmes, particularmente cuando se dan numerosos empates (véase [20]). Somers [19] ha observado que en el caso de las relaciones no monotónicas marcadas, el procedimiento que Davis sugiere puede ser engañoso. Como norma general, en vista de nuestra ignorancia acerca de las propiedades y comportamiento de las medidas ordinales parciales, puede resultar prudente utilizarlas con precaución, complementándolas con medidas momento-producto aun allí donde las escalas legítimas de intervalo no estén plenamente justificadas. En un terreno ideal debe, por supuesto, intentarse mejorar los procedimientos de medición, justificando así el uso de pruebas y medidas paramétricas más poderosas.

Como está implícito en nuestras anteriores consideraciones sobre los datos ordinales, una de las razones fundamentales por las que resulta difícil llegar a conclusiones definitivas en orden a la adecuación de medidas alternativas está en que tales respuestas parecen depender del concepto que uno tenga acerca de la "realidad básica" que los datos reflejan. Ya hemos observado esto en relación con la manipulación de los empates, y, en forma implícita, con el proceso de categorización. Una manera muy prometedora de atacar este difícil problema supone la construcción de una "realidad" cuyas propiedades sean conocidas, mediante el empleo de datos originados en la computadora, o de datos simulados. Pueden, por ejemplo, crearse variables con distribuciones de frecuencia normales, rectangulares o desviadas. Pueden usarse modelos lineales o no lineales, variar las magnitudes relativas de las variancias de error y formar grupos de datos multivariados con estructuras causales conocidas (por ejemplo, X y Y con relación espuria debida a Z o varias Z_i). Los datos podrían

a continuación ser agrupados de distintas maneras, utilizando diversos procedimientos, comparando las diferentes medidas ordinales en vista de su conformidad con el comportamiento deseado. Por ejemplo: ¿se reduce casi a cero la parcial entre X y Y cuando se controla para Z , allí donde los datos han sido creados de conformidad con el modelo $X \leftarrow Z \rightarrow Y$?

Reynolds [17] ha logrado algunos resultados esperanzadores utilizando variedad de modelos, tipos de distribución de frecuencias y modelos no lineales, y mediante la introducción de cierto número de complicaciones adicionales, habiendo encontrado que si se utilizan por lo menos cinco niveles de cada variable (aunque preferentemente deban ser hasta diez), pueden lograrse muy buenas aproximaciones al comportamiento de las parciales momento-producto, utilizando diferentes procedimientos de separación y cualesquiera de las medidas τ_b , τ_c , d_{yx} o r_s , corregida para empates. Si el número de estos últimos es apreciable, los valores numéricos de las asociaciones totales que utilicen τ_a (la que no corrige para empates) tienden a ser tan bajos que resulta difícil distinguir sus valores de los de las parciales. Si el total τ_a es de solamente .20, el error de muestreo puede ser suficiente para que resulte difícil determinar si hubo o no una reducción suficientemente grande en la parcial que permita apoyar la hipótesis de que la relación es espuria.

También ha encontrado Reynolds que la gamma no se comporta tan bien bajo el seccionamiento como las otras medidas, tal vez por causa de su sensibilidad extremada ante marginales desiguales. En los casos en que el modelo correcto implica una relación espuria entre X y Y debida a W , los controles sobre W no reducen la gamma parcial a cero. Los datos de Reynolds parecen también favorecer el empleo de los procedimientos de seccionamiento de promedios ponderados por comparación con el uso de la fórmula de seccionamiento de la ecuación (XIX.6), aun cuando debe tenerse presente la advertencia de Somers relativa a las variables de control no monotónicas. Por último, y esto es importante, Reynolds ha encontrado que el seccionamiento (usando promedios ponderados), con τ_b , τ_c y d_{yx} dio excelentes resultados en el caso de las relaciones monotónicas, pero no lineales, en tanto que los procedimientos momento-producto o paramétricos no los daban. En el último caso, si se conocen las puntuaciones reales, sería preferible trabajar con modelos explícitos no lineales y procedimientos paramétricos. En ausencia de tal conocimiento, el empleo de los procedimientos paramétricos con puntuaciones asignadas arbitrariamente (y conservando el orden) dio resultados engañosos.

Debe observarse, por último, que el problema de crear medidas de correlación múltiple, usando técnicas ordinales, no ha sido estudiado sistemáticamente. Morris [15] ha encontrado incluso

que tanto la gamma como la d_{yx} tienen la indeseable propiedad de que si se forman medidas de correlación múltiple usando procedimientos plenamente razonables, la agregación de más valores explicativos puede traducirse realmente en la *disminución* de los valores de dichas dos medidas. Sugiere una medida alternativa γ_k que es una generalización multivariada de la d_{yx} (no de la d_{yx}) de Somers, como medida asimétrica de asociación múltiple ordinal más apropiada.

XIX.3. Correlación parcial e interpretaciones causales

Ya se señaló que el análisis de correlación no se puede emplear directamente para establecer causalidad debido al hecho de que las correlaciones sólo miden la covariación, o sea el grado en que diversas variables cambian juntas. Sin embargo, uno de los objetivos básicos de toda ciencia está en establecer relaciones causales. Independientemente de las reservas filosóficas que se puedan sentir en cuanto a las nociones de causa y efecto, es sumamente difícil pensar teóricamente en cualesquiera otros términos. En el capítulo II se señaló que existe una brecha muy real entre el lenguaje teórico, que empleamos para pensar, y el lenguaje operativo, del que nos servimos para verificar las hipótesis. El problema espinoso de la causalidad no es más que otra indicación de la existencia de dicha brecha. Pensamos a menudo en términos de relaciones causales que comportan secuencias temporales *necesarias*. Así, por ejemplo, si A es causa de B , entonces B ha de seguir necesariamente a A , y si A está ausente, B ha de estarlo asimismo. Por supuesto, este concepto de la causalidad está excesivamente simplificado. Por lo pronto, no se han tenido en cuenta otras variables, y sólo tiene sentido hablar de causa y efecto cuando se pueden establecer ciertos supuestos a propósito de esos otros factores. Por otra parte, A y B pueden variar en grado, y no simplemente estando presentes o ausentes.

Empíricamente, por supuesto, nunca podemos probar que la conexión entre dos variables sea necesaria. Podemos averiguar, en cambio, el grado en que varíen juntas, y resulta asimismo posible, en ocasiones, registrar la secuencia temporal implicada. A partir de estos dos fragmentos de información podemos formular deducciones causales si queremos. Si nuestra teoría puede demostrar una conexión lógica entre dos variables, o predecir que B seguirá a A , no necesitamos atormentarnos demasiado por el hecho de efectuar el salto intelectual a la interpretación causal. Por otra parte, si no logramos hallar razón teórica alguna para enlazar directamente dos acontecimientos, solemos, por lo regular, sentirnos más vacilantes. Tenemos mayor propensión, por ejemplo, a considerar la relación como *espuria*. Por desgracia, nada hay en el análisis de correlación que nos ayude a decidir

al respecto, *a menos* que estemos dispuestos a admitir algunos supuestos a propósito de las variables particulares consideradas y a propósito de otras, que acaso puedan producir también sus efectos. Veamos cómo habrán de ser dichos supuestos.

Supóngase que estamos investigando la relación entre el consumo *per capita* de helados y las tasas de la delincuencia juvenil. Es probable que hallemos una relación negativa. Una de las interpretaciones causales posibles sería la de pensar que los helados son tan buenos para los niños que previenen la delincuencia. Otra podría ser la de que las tasas elevadas de delincuencia hacen que los niños pierdan su gusto por los dulces. Por supuesto, descartaríamos inmediatamente dichas interpretaciones por absurdas, pese a que otras no menos absurdas se hayan tomado en serio en algún momento u otro. Se razonaría probablemente en el sentido de que la relación hallada era *espuria*, por cuanto una tercera variable, el ingreso, por ejemplo, era causa de que las dos variables variaran de tal modo que resultara de ello una correlación negativa.

Una prueba del carácter espurio, válida además si se emplea adecuadamente, consiste en controlar en relación con el nivel del ingreso. Si la correlación parcial entre el consumo de helados y la delincuencia se reduce a cero, o a cerca de cero, podemos deducir que no se da relación causal entre las dos variables. ¿Podemos, efectivamente? Tomemos otro ejemplo muy parecido. Supóngase que encontramos una relación negativa entre el nivel del ingreso y la delincuencia, y decidimos controlar en relación con el porcentaje de hogares deshechos en el área considerada. Podemos hallar de nuevo que la parcial se reduce a cero. ¿Es por ello esta relación espuria? Esta vez ya no estamos tan seguros, pese a que no haya acaso absolutamente nada en la magnitud de las correlaciones o en el comportamiento de las parciales que difiera en modo alguno del primer caso. Con el propósito de atacar el problema básico que aquí se nos plantea, volvamos atrás y considerémoslo en forma un poco más sistemática.

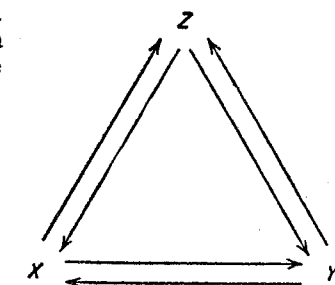


FIG. XIX.4. Las seis flechas causales posibles entre X, Y y Z.

Limitando de momento nuestra atención al caso de tres variables, observamos que se dan seis conexiones causales posibles entre éstas. Si designamos las variables como X, Y y Z e indicamos la dirección de la causalidad por medio de flechas, podemos trazar un diagrama de las conexiones posibles, como en la figu-

ra XIX.4. En todo problema determinado, por supuesto, algunas de esas flechas habrán de borrarse. Descartamos la posibilidad de la causalidad de doble sentido razonando en el sentido de que, si se seleccionan acontecimientos discretos, la secuencia temporal habrá de ser en un sentido o en otro, pero no en ambos a la vez.³ Así, por ejemplo, en lugar de sostener que el desempleo produce recesión económica y viceversa, digamos que el desempleo de Jo-

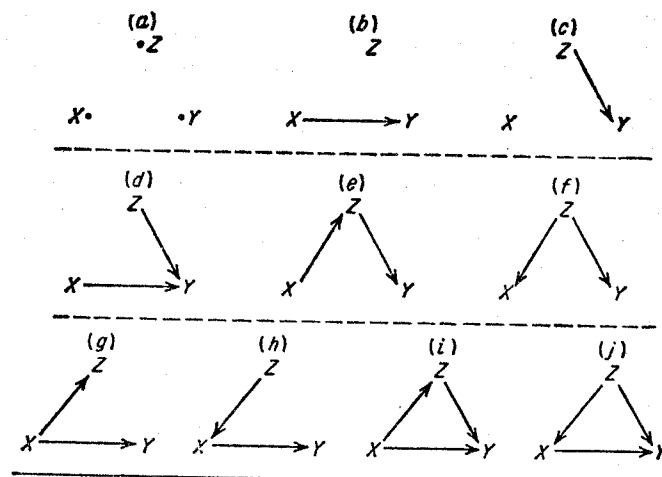


FIG. XIX.5. Relaciones causales posibles entre X, Y y Z, tomando a Y como variable dependiente y excluyendo la causalidad en dos direcciones.

nes es causa de que éste gaste menos dinero, lo cual deja a su vez sin empleo a Smith, etcétera. Nos quedamos entonces con sólo ciertas relaciones causales posibles, que se han indicado en la figura XIX.5. Con objeto de reducir el número de figuras de la figura XIX.5, se ha decidido arbitrariamente escoger a Y como variable dependiente, o sea como aquella que ha de ocurrir como última en cualquier secuencia temporal. De ahí que no se hayan trazado flechas de Y a X o a Z. De todas estas relaciones posibles, las tres primeras no revisten interés ni requieren comentario ulterior. Además, con objeto de simplificar la cosa, limitemos nuestra atención a aquellas figuras en las que sólo se han trazado dos flechas causales (d, e, f, g y h).

³ La mayoría de las situaciones empíricas son, por supuesto, mucho más complejas de lo que esta sencilla ilustración sugiere, requiriéndose técnicas más avanzadas, procedimiento que se aplica cuando los supuestos adecuados para las pruebas de mínimos cuadrados no se reúnen. Véanse [2], [4] y [12] para un examen más detallado de este problema.

¿Podemos distinguir entre estos varios modelos inspeccionando las magnitudes relativas de los coeficientes de correlación? La respuesta es afirmativa si estamos dispuestos a hacer dos clases de supuestos. Simon [18] ha demostrado matemáticamente lo que dichos supuestos deban ser. *Primero*, hemos de estar en condiciones de eliminar algunos de los modelos, postulando que por lo menos algunas de las relaciones posibles son *inconsistentes*. Esto ya se ha hecho hasta cierto punto, al eliminar todas las dobles flechas, así como al tomar a Y como variable dependiente, o sea suponiendo que no podía ser causa ni de X ni de Z . Habrán de hacerse además otros supuestos, pero éstos los dejamos para más adelante.

El *segundo* tipo de supuesto general que hemos de establecer se refiere a otras variables que podrían eventualmente actuar. Supongamos, siguiendo a Simon, que todas las demás variables que influyen sobre X no están relacionadas con todas las otras que afectan a Y y Z , etcétera. En otros términos: podemos admitir la existencia de otras variables incontroladas, pero hemos de suponer que la influencia que ejercen sobre X , Y y Z es esencialmente fortuita. Obsérvese que esto implica en realidad una combinación de supuestos más débil que la que suele comportar el modelo del experimento ideal, en el que se supone que todas las variables "relevantes" han sido controladas. Reconocemos la influencia perturbadora de otras variables en que no esperamos que las correlaciones sean perfectas. Por otra parte, hemos de suponer que operan de tal modo que no perturben el *patrón* de las relaciones entre X , Y y Z . Esta condición puede realizarse aproximadamente en la práctica si actúa un gran número de otras variables, ninguna de las cuales ejerce gran efecto sobre más de una de las variables consideradas.

Si existe una variable externa de efecto perturbador, deberá introducirse en el modelo como cuarta variable. Simon argumenta en el sentido de que esto es lo que debemos hacer siempre, y que el hecho de que no nos satisfaga la explicación causal en el caso de dos variables es la razón de que introduzcamos la noción de una relación espuria. Así, por ejemplo, si estuviéramos convencidos de que no existía variable tal alguna que perturbara la relación entre el consumo de helados y la delincuencia, y si pudiéramos excluir la posibilidad de que la delincuencia produce una baja de la venta de helados, entonces nada tendríamos que oponer a la explicación de que la flecha causal vaya en sentido opuesto. E introducimos el factor ingreso precisamente porque esperamos que esta última variable afecte a la relación entre las dos primeras. Y en forma análoga, añadiríamos al sistema una cuarta o quinta variable, pero hemos de estar dispuestos a detenernos en algún lugar. En este punto, si hemos de formular en principio alguna deducción causal, cualquiera que sea, hemos

de suponer que el sistema está *cerrado*, en el sentido que acabamos de describir.

Obsérvese que nos encontramos en la posición con la que estamos ya familiarizados de tener que adoptar algunos supuestos que no se dejan verificar empíricamente mediante el análisis estadístico. No será posible, por consiguiente, establecer el carácter correcto de un modelo causal particular cualquiera, pero podemos proceder por eliminación. Así, por ejemplo, uno de los modelos indicados en la figura XIX.5 podrá parecer eficaz; sin embargo, el modelo correcto podría comportar en realidad cuatro o más variables, y el cuadro presentarse en forma totalmente distinta. Con todo, habiendo adoptado los supuestos en cuestión, podemos servirnos del análisis matemático formulado por Simon, para llegar a ciertas relaciones anticipadas que deberían verificarse entre las correlaciones si el modelo particular es efectivamente correcto. Según veremos, exactamente las mismas relaciones empíricas se anticipan por algunos de los modelos anteriores, obligándonos a escoger sobre otras bases. Es aquí donde hemos de servirnos del primer tipo de supuesto examinado más arriba, o sea de que algunas relaciones causales *no* se realizan. Con todo, sin embargo, examinemos primero las predicciones matemáticas relativas a las interrelaciones entre coeficientes de correlación.

Si nos fijamos en la figura XIX.5g, vemos que las relaciones entre X y Y y entre X y Z son directas, en tanto que la relación entre Y y Z sólo es indirecta. Lo propio se aplica a la figura h . En estos dos casos, el sentido común sugeriría que, si todas las demás variables actuaran de modo esencialmente fortuito, esperaríamos encontrar que la correlación entre Y y Z es menor en magnitud que cualquiera de las otras dos. Y en forma análoga, en las figuras XIX.5e y f esperaríamos que la relación entre X y Y fuera la más pequeña de las tres, prescindiendo de los signos. Como lo revelan las matemáticas, podemos incluso pronunciarlos en forma más categórica. Es posible, en efecto, demostrar tanto para (g) como para (h), en las que la relación entre Y y Z es indirecta, que:

$$\rho_{YZ} = \rho_{XY}\rho_{XZ}$$

Nos hemos servido de las ρ para indicar que estas relaciones exactas sólo pueden esperarse en la población, y que los valores de las r de la muestra se apartarán por lo regular de esa relación estimada, a causa de las fluctuaciones de la muestra. Y en forma análoga puede demostrarse que para los casos (e) y (f) tendremos:

$$\rho_{XY} = \rho_{XZ}\rho_{YZ}$$

Toda vez que los valores absolutos de los coeficientes de correlación no pueden ser mayores que la unidad, está claro que en

el primer caso el valor numérico de ρ_{YZ} ha de ser menor que el de cualesquiera de los otros coeficientes, a menos que uno de estos valores acontezca ser la unidad. En este caso especial, por supuesto, una de las variables puede predecirse exactamente a partir de una de las otras, y tenemos así esencialmente un problema de sólo dos variables.

Fijándonos con mayor detenimiento en la primera de estas ecuaciones, que se aplica a las figuras XIX.5g y h, vemos inmediatamente que si esta ecuación se verificara, la correlación parcial (en la población) entre Y y Z, controlando en relación con X, desaparecería, ya que el numerador de la fórmula de la parcial sería en tal caso cero. Así, pues, si (g) o (h) se verificaran, el valor de $r_{YZ.X}$ debería ser cero o muy cerca de cero, habida cuenta de los errores de muestreo. Y en forma semejante, puede esperarse tanto para (e) como para (f) que la parcial entre X y Y, controlando respecto de Z, sea aproximadamente cero. ¿Qué indican estos hechos? Si limitamos nuestra atención a una comparación entre (e) y (f), ya que la relación entre (g) y (h) puede compararse directamente si se intercambian X y Z, vemos que en (f) interpretaríamos la relación entre X y Y como espuria, toda vez que Z actúa en el sentido de producir variación tanto en X como en Y. Esta situación es exactamente la misma que se describió en el ejemplo del consumo de helados X y las tasas de delincuencia Y. Debido a que sospechamos que la relación entre estas dos variables se deba a otra, o sea al nivel de ingreso Z, controlamos en relación con ésta para ver si la correlación entre X y Y se reduce a casi cero. Si (f) es de hecho el modelo correcto, acabamos de ver matemáticamente que tal será el caso.

Vimos también, sin embargo, que la parcial habría sido cero si el modelo correcto fuera el de la figura XIX.5e. En (e), en efecto, tenemos que Z actúa como variable interventora, en el sentido de que X causa Z, la cual a su vez causa Y. Pero, ¿tiene algún objeto controlar en relación con Z en estas condiciones? Probablemente no. Porque si X es efectivamente causa de Z, ¿cómo podemos concebir que mantengamos a Z constante mientras X sigue variando? No tiene sentido ciertamente pensar obtener residuales tomando aquella porción de la variación de X que es "debida a" Z cuando Z es un efecto de X. Puede, sin embargo, tener sentido el controlar para Z si lo que tratamos de demostrar es la ausencia de una conexión causal entre X y Y, excepto a través de la variable interventora Z. La manipulación de fórmulas estadísticas no constituye sustituto alguno del conocimiento de lo que se está haciendo. En este caso, saber lo que se está haciendo consiste en estar en condiciones de elegir entre los modelos (e) y (f), yendo más allá de la información estadística disponible y haciendo un supuesto acerca de la dirección de la flecha entre X y Z.

Hasta aquí hemos prescindido de la situación (d) de la figura XIX.5, en la que las flechas van a Y tanto de X como de Z, pero en la que no se da relación directa alguna entre X y Z. ¿Qué sucede en este caso si controlamos en relación con Z? Observamos en primer lugar que tiene objeto controlar aquí en relación con Z porque ésta se concibe como variable totalmente independiente que afecta también a Y. Desde el punto de vista de la relación entre X y Y, opera como influencia perturbadora. Es una variable "extraña" que produce esencialmente en Y variaciones fortuitas con respecto a las variaciones de X. Por lo tanto, esperaríamos que, controlando en relación con Z, aumentara la magnitud de la relación entre X y Y. Puede demostrarse matemáticamente que si establecemos los supuestos requeridos a propósito de otras variables, la correlación en la población entre X y Z será cero. Señalemos de paso que este hecho nos permitirá distinguir (d) empíricamente de cada una de las situaciones que hemos venido examinando. Ésta es, pues, la situación en la que la variable de control no se relaciona con una de las otras variables, y ya vimos que en tal caso la parcial será mayor en valor absoluto que la correlación total, lo que concuerda con el sentido común. Es asimismo la situación a la que nos enfrentamos en el análisis por dos métodos de la variancia, en la que la condición de subceldillas iguales suponía una independencia completa entre las variables de fila y de columna, y en la que también vimos que un control para una de las variables reducía la suma inexplicada de los cuadrados, sin reducir la variación explicada por la otra variable independiente.

Hay otro tipo de situación de control que no se ha examinado, pero que puede tratarse brevemente, ya que son pocos los casos, si los hay, en que podríamos vernos inducidos a servirnos de un control.

Supóngase, en efecto, en una de las situaciones (e) o (h), que iba a relacionar las variables dependientes que intervienen, controlando en relación con la variable independiente. En (h), por ejemplo, ¿qué sucedería si fuéramos a obtener la parcial entre X y Y controlando en relación con Z? Puede demostrarse algebraicamente que la parcial resultante sería menor en magnitud que la correlación total. Esto concuerda con la noción intuitiva de que, manteniendo constante la variable independiente, se reduce necesariamente la variación de la variable interferente, con lo que se debilita la relación con la variable dependiente. Una vez más, tendría poco objeto llevar a cabo semejante operación. Por lo regular, en efecto, nuestro interés se centrará en la cuestión de saber si existe o no una relación directa entre X y Y, y no en el problema de las causas antecedentes de X. Puede demostrarse, sin embargo, que si hubiéramos controlado inadvertidamente para Z en (h), no hubiéramos afectado sistemáticamente

el *declive* estimado b_{xy} , excepto en el sentido de que habríamos aumentado el valor del error de muestreo.

Las extensiones a cuatro o más variables son directas, con tal de que nos restrinjamos a una causación en sólo un sentido. Puede demostrarse que en los casos en que no hay lazo directo entre dos variables, se dará una parcial de orden más elevado entre estas variables, la que será aproximadamente igual a cero, excepto por los errores de muestreo. En general, debemos controlar para todas las variables antecedentes e interventoras, con objeto de hacer desaparecer la apropiada correlación parcial, pero habremos de tener cuidado, evitando controlar para variables que se supone son dependientes de las dos que están siendo consideradas. Por ejemplo, en el modelo

$$\begin{array}{ccc} X_1 & \rightarrow & X_2 \\ \downarrow & & \downarrow \\ X_3 & \rightarrow & X_4 \end{array}$$

será necesario controlar *tanto* para X_2 como para X_3 , con el fin de reducir a cero la parcial $r_{14.23}$. De manera análoga, el modelo predice para $r_{23.1} = 0$ (excepto por errores de muestreo), pero *no* deberemos esperar que $r_{23.14}$ sea igual a cero, ni tendría sentido alguno controlar en este caso para X_4 . (Véase [2] para más amplia discusión.)

Son de nuevo necesarias varias advertencias. Como en el caso de las tres variables, habrá siempre modelos alternativos que predigan exactamente las mismas intercorrelaciones empíricas, y habrá que confiar en el conocimiento de las secuencias temporales, o supuestos *a priori*, cuando haya que escoger entre tales alternativas. Por otra parte, la existencia de errores de medición aleatorios y no aleatorios invalidará las predicciones de cualquier modelo dado. Como observamos en el capítulo anterior, el error aleatorio de medición en una variable independiente atenuará las correlaciones entre ésta y otras variables. En el caso de regresión *múltiple*, y cuando las variables independientes estén altamente intercorrelacionadas, los errores aleatorios de medición, en algunas de ellas, tenderá a *aumentar* los efectos visibles de aquellas variables con las que estén más altamente intercorrelacionadas. Se ve de esta manera que los errores de medición en presencia de una alta intercorrelación entre variables independientes se prestan a conducirnos a deducciones erróneas.

Resultará claro de las observaciones anteriores que si uno suma variables a una ecuación de regresión podrá esperar que las correlaciones parciales cambien según sea la naturaleza de las intercorrelaciones entre las variables independientes. Esto es aplicable a los declives parciales y estandarizados que se examinan en la sección siguiente. Suponemos que el error, o término resi-

dual para la ecuación de regresión, no está relacionado con cada una de las variables independientes de la propia ecuación. En términos causales, esto hace suponer que los factores que son causa mayor de la variable dependiente no están sistemáticamente relacionados con las variables independientes.

Si somos capaces de localizar las variables que contribuyen a este factor de error y si las hacemos figurar de manera explícita en la ecuación, tales variables deberán *no* estar relacionadas con las variables independientes originales, aparte los errores de muestreo, no resultando afectados sistemáticamente los declives parciales. Las correlaciones parciales, por otra parte, aumentarán en su valor numérico, debido a que habrá sido eliminado algo de la variancia no explicada. Sin embargo, si las variables adicionales llevadas a la ecuación están relacionadas sistemáticamente con las variables independientes originales, podrá darse por seguro que todos los coeficientes resultarán afectados.

XIX.4. Mínimos cuadrados múltiples y los coeficientes beta

Nos hemos servido de las correlaciones parciales para indicar el grado de relación entre una variable dependiente y una variable independiente, controlando en relación con una o varias variables independientes más. Si tenemos un número grande de variables independientes, podemos obtener una indicación de su importancia relativa asociando la variable dependiente con cada una de las variables independientes sucesivamente y controlando en cada caso con las variables restantes. Anteriormente, en nuestro examen de la regresión múltiple y de los mínimos cuadrados, ya observamos también que las b y las β que figuran en nuestras ecuaciones y relacionan a Y con las variables independientes podrían interpretarse en cierto sentido como parciales. Se recordará que representan las pendientes de las ecuaciones de regresión o de los mínimos cuadrados en la dimensión de la variable independiente apropiada, esto es, con todas las demás variables independientes mantenidas constantes. Por lo tanto, cada coeficiente representa la cantidad de variación de Y que puede asociarse con un cambio determinado de las X , manteniendo fijas las demás variables independientes. Teniendo en cuenta esta similitud con los coeficientes de la correlación parcial, no debería sorprender que las fórmulas empleadas en el cálculo de esas b parciales resultaran muy semejantes a las que se emplearon en obtener las r parciales y que, además, esas pendientes pudieran emplearse para dar una indicación de la importancia relativa de cada una de las variables independientes en la determinación de la variación de Y .

Hemos de modificar nuevamente nuestra notación, con objeto de distinguir entre el gran número de combinaciones posibles de

las pendientes. Designando nuestras variables simplemente como 1, 2, 3, etcétera, nos servimos del símbolo $b_{12\cdot3}$ si anticipamos la variable uno a partir de las variables 2 y 3 con referencia al coeficiente de la segunda variable. El coeficiente $b_{12\cdot3}$ ha de distinguirse de $b_{21\cdot3}$, que emplearíamos si la segunda variable se tomara como variable dependiente. En ambos casos, el hecho de que el número tres se haya colocado a la derecha del punto indica que se ha controlado la tercera variable. Y en forma análoga, $b_{12\cdot34}$ se emplea para indicar el coeficiente de la segunda variable en una ecuación de predicción en la que la primera variable se toma como variable dependiente y que comporta dos variables de control. En este último caso, la ecuación de los mínimos cuadrados se daría en la siguiente forma:

$$X_1 = a_{1\cdot234} + b_{12\cdot34}X_2 + b_{13\cdot24}X_3 + b_{14\cdot23}X_4$$

en donde el subíndice de a indica que estamos anticipando en relación con la variable uno a partir de las variables 2, 3 y 4. La razón de que hayamos considerado conveniente apartarnos de la práctica consistente en designar la variable dependiente con Y está en servirnos de una combinación más sencilla de subíndices para seguir la traza de las distintas b .

Las fórmulas de cálculo de $a_{i\cdot jk}$ y $b_{ij\cdot k}$ son como sigue:

$$a_{i\cdot jk} = \bar{X}_i - b_{ij\cdot k} \bar{X}_j - b_{ik\cdot j} \bar{X}_k \quad (\text{XIX.7})$$

$$y \quad b_{ij\cdot k} = \frac{b_{ij} - (b_{ik})(b_{kj})}{1 - b_{jk}b_{kj}} \quad (\text{XIX.8})$$

Obsérvese que si bien el denominador de (XIX.8) difiere en cuanto a la forma del de la fórmula de $r_{ij\cdot k}$, el numerador, en cambio, es esencialmente similar en carácter.

En efecto, recordando que

$$r_{jk}^2 = b_{jk}b_{kj}$$

vemos que incluso los denominadores no son demasiado dispares en cuanto a la forma. Con todo, al emplear esta fórmula para obtener las b parciales, hay que poner cuidado en distinguir entre b_{jk} y b_{kj} , ya que los subíndices ya no pueden intercambiarse.

La extensión a parciales de orden superior es directa (véase [5]). Las ecuaciones de $a_{i\cdot jkl}$ y $b_{ij\cdot kl}$ pueden escribirse:

$$a_{i\cdot jkl} = \bar{X}_i - b_{ij\cdot kl} \bar{X}_j - b_{ik\cdot jl} \bar{X}_k - b_{il\cdot kj} \bar{X}_l \quad (\text{XIX.9})$$

$$y \quad b_{ij\cdot kl} = \frac{b_{ij\cdot k} - (b_{il\cdot k})(b_{kj\cdot l})}{1 - b_{jk\cdot l}(b_{lj\cdot k})} \quad (\text{XIX.10})$$

Igualmente cierto en el cálculo de correlaciones parciales de orden superior a medida que el número de variables aumenta, el empleo de estas fórmulas puede comportar acaso considerablemente más trabajo que el que requieren los métodos abreviados de Doolittle o de la raíz cuadrada de Dwyer. Normalmente será, por supuesto, más conveniente utilizar programas de computación, cuando se trate de obtener estos coeficientes.

Se puede interpretar un declive parcial como el cambio hipotético que ocurriría en la variable dependiente si una de las variables independientes hubiera de cambiar en una unidad y si las demás variables permanecieran constantes. Esto puede ser interpretado como una medida del efecto directo de la variable independiente sobre la variable dependiente; si un declive parcial es igual a cero, ello no implicaría un efecto directo. Pero no habiendo especificado las conexiones causales entre las propias variables independientes y teniendo en cuenta únicamente sus intercorrelaciones, no nos es posible decir nada en relación con el efecto total de cada variable. Si, por ejemplo, la primera variable independiente es una causa de la segunda, un cambio en la primera variable produciría un cambio también en la segunda, produciéndose efectos tanto directos como indirectos. De esta manera no podemos valorar la importancia relativa de cada variable, a menos que conozcamos más acerca de la estructura causal del sistema en su totalidad. Esto requeriría trabajar con todo un grupo de ecuaciones, una por cada variable que sea tomada como dependiente de cualesquiera de las otras. Por desgracia, los mínimos cuadrados ordinarios no son en general adecuados para tal sistema de ecuaciones (véanse [4] y [12]).

En tanto no estemos interesados en generalizar más allá de los límites de una sola población, en ocasiones es deseable obtener una medida asimétrica de los efectos directos de cada variable independiente, que no dependa de las unidades de medida utilizadas. Obtenemos así, en efecto, una medida del efecto directo real en el caso particular de la población que estudiamos, dado que algunas variables independientes varían más que otras. Una variable puede ser medida en dólares, otra en años. Carecería de sentido comparar la unidad de cambio en una con la unidad de cambio en la otra.

Si cada variable es estandarizada, dividiéndola por su desviación estándar, en la misma forma que se aplicó para obtener la curva normal estándar obtendremos declives ajustados, comparables de una variable a la siguiente. Medimos así los cambios en la variable dependiente en función de unidades de desviación estándar para cada una de las otras variables, lo que nos asegura

una misma variabilidad en cada una de estas variables. Estos declives parciales ajustados resultan así *b* estandarizadas, llamadas frecuentemente *ponderaciones beta*, siendo denominados *coeficientes de curso* en los más simples modelos causales, en los que está implicada una determinante de causa en una sola dirección (véase [14]).

Por desgracia, una vez más nos vemos envueltos en incongruencias de notación. En efecto, estas ponderaciones de beta no son las mismas que las de las β en la ecuación de regresión, que se refieren a características de la *población* y no han sido ajustadas en relación con las diferencias de variabilidad. Las ponderaciones de beta se obtienen de los datos de la *muestra* y son simples funciones de las *b* parciales. Las fórmulas generales de $\beta_{ij \cdot k}$ y $\beta_{ij \cdot kl}$ son:

$$\beta_{ij \cdot k} = b_{ij \cdot k} \frac{s_j}{s_i} \quad (\text{XIX.11})$$

y

$$\beta_{ij \cdot kl} = b_{ij \cdot kl} \frac{s_j}{s_i} \quad (\text{XIX.12})$$

Así, pues, la ponderación de beta puede obtenerse multiplicando la *b* comparable por la razón de la desviación estándar de la variable independiente (no controlada) a la de la variable dependiente.

La comparabilidad de las ponderaciones de beta y los coeficientes de correlación parcial puede verse en la fórmula:

$$\beta_{ij \cdot k} = \frac{r_{ij} - r_{ik}r_{jk}}{1 - r_{jk}^2} \quad (\text{XIX.13})$$

Las dos medidas sólo difieren en los denominadores. De hecho, vemos inmediatamente que:

$$r_{ij \cdot k}^2 = (\beta_{ij \cdot k})(\beta_{ji \cdot k})$$

ya que $\beta_{ji \cdot k}$ sólo difiere de $\beta_{ij \cdot k}$ en que el denominador de r_{jk}^2 será remplazado por r_{ik}^2 . Ya que las ponderaciones de beta y las correlaciones parciales representan tipos de medida de asociación algo distintos, no darán exactamente los mismos resultados, aunque por lo regular comprendan variables del mismo orden de importancia. En efecto, la correlación parcial es una medida de la *cantidad de variación explicada* por una de las variables independientes después que las otras han explicado todo lo que podían. Las ponderaciones de beta, en cambio, indican *cuánto cambio* se produce en la variable dependiente por un cambio estandarizado en una de las variables independientes al controlar en relación con las otras.

XIX.5. Correlación múltiple

Como quiera que nuestro interés puede acaso residir en el poder explicativo de cierto número de variables independientes tomadas juntas más que en la relación entre la variable dependiente y cada una de las variables independientes tomadas separadamente, preferiremos tal vez servirnos del *coeficiente de correlación múltiple*, que es una medida de la bondad de ajuste de la superficie de mínimos cuadrados a los datos. Al igual que el cuadrado del coeficiente de la correlación de orden cero indicaba el porcentaje de variación explicada por la recta de mejor ajuste, el cuadrado del coeficiente de correlación múltiple puede emplearse para dar el porcentaje de variación explicado por la ecuación de mejor ajuste de la forma:

$$Y_p = a + b_1X_1 + b_2X_2 + \dots + b_kX_k$$

Otra manera de concebir la correlación múltiple está en que representa la *correlación de orden cero* entre los *valores reales* obtenidos para la variable dependiente y los *valores anticipados a partir de la ecuación de mínimos cuadrados*. Si todos los puntos se encuentran exactamente en la superficie de mínimos cuadrados, los valores real y anticipado coincidirán, y la correlación múltiple será la unidad. Y cuanto mayor sea la dispersión alrededor de la ecuación de mínimos cuadrados tanto más baja será la correlación entre los valores real y predicho.

La fórmula de la correlación múltiple puede desarrollarse fácilmente sirviéndose del hecho de que el cuadrado del múltiple será igual al porcentaje de la variación explicada por todas las variables independientes. Conviene recalcar una vez más que se supone un modelo de tipo lineal. Al escribir la fórmula de la correlación múltiple, dejamos primero que una de las variables independientes explique todo lo que puede. Dejamos luego que la segunda variable independiente haga lo propio en relación con la porción de la variación no explicada por la primera. Sin embargo, con objeto de evitar duplicación, hemos de controlar en relación con esta primera variable independiente. Dejamos a continuación que la tercera explique todo lo que puede del resto, controlando ahora en relación con las dos primeras variables independientes. El procedimiento puede prolongarse de manera indefinida.

De momento, sin embargo, nos limitamos al caso de tres variables, que sólo comporta dos variables independientes. Si tomamos la primera variable como la variable dependiente, y designamos el coeficiente de correlación múltiple por $R_{1 \cdot 23}$, podremos escribir:

$$R_{1-23}^2 = r_{12}^2 + r_{13-2}^2 (1 - r_{12}^2) \quad (XIX.14)$$

$\left(\begin{array}{c} \text{Proporción} \\ \text{explicada} \\ \text{por 2 y 3} \end{array} \right) = \left(\begin{array}{c} \text{proporción} \\ \text{explicada} \\ \text{por 2} \end{array} \right) + \left(\begin{array}{c} \text{proporción} \\ \text{adicional} \\ \text{explicada} \\ \text{por 3} \end{array} \right) \left(\begin{array}{c} \text{proporción} \\ \text{no explica-} \\ \text{da por 2} \end{array} \right)$

Obsérvese que las correlaciones múltiples sólo tienen una cifra a la izquierda del punto, cifra que indica la variable dependiente. Los números de la derecha, en cambio, indican aquellas variables independientes que se están empleando para explicar la variación de la variable dependiente. Así, pues, la fórmula general (para tres variables) puede escribirse como sigue:

$$R_{i \cdot jk}^2 = r_{ij}^2 + r_{ik \cdot j}^2 (1 - r_{ij}^2) \\ = r_{ik}^2 + r_{ij \cdot k}^2 (1 - r_{ik}^2) \quad (XIX.15)$$

No importa, por supuesto, cuál de las dos variables independientes se emplee primero en la fórmula, a condición que dicha variable se controle en los términos siguientes de la ecuación.

Operamos con los cuadrados tanto de la correlación total como de las correlaciones parciales, ya que obtenemos los porcentajes de la variación explicada. Por lo tanto, no tenemos por qué preocuparnos por los signos de estas correlaciones. Y de hecho, la dirección de la múltiple carece de significado, ya que comporta correlaciones con cierto número de variables, algunas de las cuales son positivas y otras posiblemente negativas. Por convención, al designar el coeficiente de correlación múltiple, tomamos siempre la raíz cuadrada positiva de R^2 .

Si resolvemos la ecuación (XIX.14) en relación con la parcial r_{13-2}^2 , obtenemos:

$$r_{13-2}^2 = \frac{R_{1-23}^2 - r_{12}^2}{1 - r_{12}^2} \quad (XIX.16)$$

Esto nos permite ver la relación entre los coeficientes de las correlaciones múltiples y parciales bajo una perspectiva algo distinta. En el numerador hemos sustraído la proporción de la variación de 1 explicada por la 2 sola, de la proporción explicada por 2 y 3 actuando juntas (R_{1-23}^2). El resultado es el incremento explicado por 3, después de haber permitido actuar a 2. Si dicho incremento se divide entre la proporción de variación dejada sin explicar por 2, obtenemos la parcial entre 1 y 3 controlando en relación con 2. Esto concuerda con nuestra interpretación anterior del coeficiente de la correlación parcial.

De la ecuación (XIX.14) pueden derivarse diversas fórmulas alternativas pero equivalentes de R_{1-23}^2 . Sustrayendo ambos miembros de dicha ecuación de la unidad, obtenemos:

$$1 - R_{1-23}^2 = 1 - r_{12}^2 - r_{13-2}^2 (1 - r_{12}^2) \\ = (1 - r_{12}^2)(1 - r_{13-2}^2) \quad (XIX.17)$$

Esta ecuación indica que podemos escribir la proporción de variación dejada sin explicar por 2 y 3 juntas, como producto de la proporción inexplorada por 2 y de aquella inexplorada por 3, controlando en relación con 2.

La fórmula de la múltiple puede escribirse también totalmente en términos de correlaciones de orden cero. En efecto, sirviéndonos de la ecuación (XIX.3) de r_{13-2} en términos de coeficientes de orden cero y simplificando la expresión algebraica resultante, obtenemos:

$$R_{1-23}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$

o bien, en general:

$$R_{i \cdot jk}^2 = \frac{r_{ij}^2 + r_{ik}^2 - 2r_{ij}r_{ik}r_{jk}}{1 - r_{jk}^2} \quad (XIX.18)$$

En particular, si la correlación entre las dos variables independientes j y k acontece ser cero, obtenemos:

$$R_{i \cdot jk}^2 = r_{ij}^2 + r_{ik}^2$$

Pueden observarse ahora algunas relaciones entre la múltiple R y las diversas correlaciones totales. Es obvio que R no puede ser menor en magnitud que cualesquiera de las correlaciones totales, ya que es imposible explicar *menos* variación añadiendo más variables. Normalmente, por supuesto, la múltiple R será mayor que una cualquiera de las r totales. Su valor máximo en relación con los coeficientes totales suele producirse cuando las intercorrelaciones entre las variables independientes son todas cero. Como acabamos de ver, el cuadrado de la correlación múltiple será en este caso igual a la suma de los cuadrados de las demás correlaciones. Por otra parte, si las intercorrelaciones entre las variables independientes son muy grandes en magnitud,

la múltiple R no será por lo regular mucho mayor que la correlación total mayor con la variable dependiente. En otros términos: si deseamos explicar lo más posible de la variación de la variable dependiente, hemos de buscar variables independientes que tengan relativamente poca relación unas con otras, pero que tengan por lo menos correlaciones moderadamente altas con la variable dependiente. O expresado en otra forma: si tenemos dos variables independientes de interrelación alta, la segunda explicará esencialmente la misma variación que la primera, ya que las dos se traslaparán considerablemente. Y si no están correlacionadas, entonces cada una explicará una porción diferente de la variación total.

Hay otra razón importante para preferir las variables independientes que no estén altamente intercorrelacionadas. No sólo habrá menos superposiciones en la variancia explicada, y por ello menos ambigüedad en nuestra interpretación causal de sus supuestos efectos, sino que en la medida en que las variables independientes estén altamente intercorrelacionadas, tanto las correlaciones parciales como las estimaciones de declives se harán cada vez más sensibles a los errores de muestreo y medición. Esta dificultad se denomina *multicolinealidad* en la bibliografía económica (véanse [4] y [12]). El problema resulta especialmente serio cuando se utilizan bloques de variables independientes altamente intercorrelacionadas, y cuando dichos bloques difieren en cuanto al número de variables que contienen. (Véase [10]). Puede demostrarse, por ejemplo, que con muy pequeñas diferencias en las correlaciones totales con la variable dependiente se producen diferencias considerables en las correlaciones parciales y en la estimación de los declives, de tal manera que si se confía en las magnitudes relativas de estos coeficientes parciales, cabe esperar encontrar diferencias considerables de una muestra a la siguiente, o bien entre réplicas en las que se utilicen instrumentos de medición algo distintos. La conclusión es que en cuantas ocasiones las variables independientes estén altamente intercorrelacionadas, resultará necesario contar *tanto* con muestras grandes *como* con las mediciones exactas.

A título de ejemplo numérico del cálculo de la múltiple R , veamos cuánta variación en materia de discriminación puede explicarse por el porcentaje de negros y el porcentaje urbano. Sirviéndonos de la ecuación (XIX.14) obtenemos:

$$R_{1\cdot23}^2 = r_{12}^2 + r_{13\cdot2}^2(1 - r_{12}^2) = (.536)^2 + (.332)^2[1 - (.536)^2] \\ = .2873 + (.1102)(.7127) = .3658$$

De ahí: $R_{1\cdot23} = .605$

Por consiguiente, el porcentaje urbano explica muy poca variación por encima y por debajo de aquella explicada por el porcentaje de negros.

A título de control de nuestros cálculos, observamos que el mismo resultado deberá obtenerse si dejamos que actúe primero el porcentaje urbano.

Obtenemos en este caso:

$$r_{12\cdot3} = \frac{r_{12} - r_{13}(r_{23})}{\sqrt{1 - r_{13}^2} \sqrt{1 - r_{23}^2}} = \frac{.536 - (.139)(-.248)}{\sqrt{1 - (.139)^2} \sqrt{1 - (-.248)^2}} = .595$$

$$\text{Así pues, } R_{1\cdot23}^2 = r_{13}^2 + r_{12\cdot3}^2(1 - r_{13}^2) \\ = (.139)^2 + (.595)^2[1 - (.139)^2] = .3667$$

$$\text{y por lo tanto: } R_{1\cdot23} = .605$$

Las fórmulas del coeficiente de correlación múltiple pueden extenderse fácilmente asimismo a un número cualquiera de variables independientes. Al introducir una tercera variable independiente, designada como X_4 , no hacemos más que añadir a la fórmula de $R_{1\cdot23}^2$ un término que comporta el producto del cuadrado de la parcial entre 1 y 4, controlando en relación con 2 y 3, y la proporción de variación queda inexplicada por 2 y 3. Así, pues:

$$R_{1\cdot234}^2 = r_{12}^2 + r_{13\cdot2}^2(1 - r_{12}^2) + r_{14\cdot23}^2[1 - r_{12}^2 - r_{13\cdot2}^2(1 - r_{12}^2)] \\ = R_{1\cdot23}^2 + r_{14\cdot23}^2(1 - R_{1\cdot23}^2) \quad (\text{XIX.19})$$

Podemos, pues, ir añadiendo a la proporción de la variación explicada, siempre que controlemos en relación con todas las variables previamente empleadas y a condición que permitamos que la nueva parcial sólo actúe sobre aquella porción de variación dejada inexplicada por las variables que la han precedido. Obsérvese, de paso, el paralelo con lo que hicimos en el análisis de la variancia. Según veremos a continuación, podemos servirnos de este hecho en las pruebas de significación tanto de la correlación múltiple como de la parcial. Si procediéramos a añadir una quinta variable, obtendríamos:

$$R_{1\cdot2345}^2 = R_{1\cdot234}^2 + r_{15\cdot234}^2(1 - R_{1\cdot234}^2)$$

Podemos resolver de nuevo estas ecuaciones en relación con los

coeficientes parciales. Así, por ejemplo, tenemos (de XIX.19):

$$r_{14 \cdot 23}^2 = \frac{R_{1 \cdot 234}^2 - R_{1 \cdot 23}^2}{1 - R_{1 \cdot 23}^2} \quad (\text{XIX.20})$$

indicando que la parcial entre 1 y 4, controlando en relación con 2 y 3, puede interpretarse como la razón de la proporción de variación adicional explicada por 4, además de la explicada por 2 y 3, a la proporción de variación dejada sin explicar por estas dos últimas variables. Podemos también extender la ecuación (XIX.17) para comprender más variables. Así, por ejemplo:

$$1 - R_{1 \cdot 234}^2 = (1 - r_{12}^2)(1 - r_{13 \cdot 2}^2)(1 - r_{14 \cdot 23}^2)$$

y, en general,

$$1 - R_{1 \cdot 234 \dots k}^2 = (1 - r_{12}^2)(1 - r_{13 \cdot 2}^2) \dots (1 - r_{1k \cdot 234 \dots (k-1)}^2) \quad (\text{XIX.21})$$

El coeficiente parcial-múltiple. En ocasiones resulta deseable calcular una correlación múltiple entre una variable dependiente y algunas variables independientes, controlando en relación con una o varias de éstas. Supóngase, por ejemplo, que se está tratando de predecir el tamaño real de la familia a partir de cierto número de variables independientes. Es obvio que ciertas variables como la duración del matrimonio y la edad de la esposa en el momento de celebrarlo han de tomarse en consideración. Por otra parte, estas variables son tan obvias, que el hecho de conjuntarlas al coeficiente general múltiple podría oscurecer los efectos de las variables restantes. Así, pues, el interés podría fijarse en la variación del tamaño de la familia después que dichas variables teóricamente poco importantes han explicado de la variación todo lo que podían. Siguiendo a Croxton y Cowden [6], indicamos la parcial-múltiple entre la variable 1 (dependiente) y 2 y 3, controlando en relación con 4, por medio de $r_{1(23) \cdot 4}^2$. La fórmula se convierte en tal caso en:

$$r_{1(23) \cdot 4}^2 = \frac{R_{1 \cdot 234}^2 - r_{14}^2}{1 - r_{14}^2}$$

La fórmula anterior de la parcial-múltiple es una simple extensión de las fórmulas que hemos utilizado en las correlaciones múltiple y parcial. Dejamos primero que la variable de control 4 explique todo lo que puede. Observamos luego que $R_{1 \cdot 234}^2$ repre-

senta la proporción de variación explicada por las tres variables independientes tomadas juntas. La diferencia, pues, ha de deberse a las variables 2 y 3. De este modo, el numerador representa la proporción de variación explicada por 2 y 3, además de aquella explicada por 4. Pero, como quiera que sólo hemos de operar con la variación no explicada por la variable de control, dividimos entre la cantidad $1 - r_{14}^2$. Sirviéndonos del principio consistente en dejar actuar primero todas las variables de control, podemos escribir la fórmula general de la parcial-múltiple como:

$$r_{i(jk \dots n) \cdot tu \dots w}^2 = \frac{R_{i \cdot jk \dots w}^2 - R_{i \cdot tu \dots w}^2}{1 - R_{i \cdot tu \dots w}^2} \quad (\text{XIX.22})$$

Por ejemplo:

$$r_{3(56) \cdot 124}^2 = \frac{R_{3 \cdot 12456}^2 - R_{3 \cdot 124}^2}{1 - R_{3 \cdot 124}^2}$$

La parcial-múltiple no parece haberse utilizado con mucha frecuencia en la investigación sociológica, debido tal vez al hecho de que las personas del ramo están poco familiarizadas con ella. Sin embargo, como medida que permite tratar problemas de correlación múltiple y parcial simultáneamente, su empleo potencial parece ser grande. En la próxima sección examinaremos otro tipo de empleo de esta medida.

XIX.6. Regresión múltiple y no linealidad

Hasta aquí toda nuestra labor se ha basado en el supuesto de modelos lineales. En el capítulo anterior vimos una prueba de no linealidad, pero sólo pudimos decir muy poco a propósito de la forma de la relación no lineal, excepto en el caso de transformaciones logarítmicas. En otros términos: no hicimos más que verificar en relación con la existencia de una *desviación* respecto de la linealidad, pero no efectuamos prueba alguna por lo que se refiere a la forma de la curva. Si bien el problema general de la no linealidad rebasa el objetivo de este texto, podemos, con todo, examinar brevemente de qué modo las técnicas de la regresión múltiple y de los mínimos cuadrados se dejan modificar ligeramente para permitirnos tratar algunos tipos de problemas que comportan no linealidad.

Como ya se señaló en el capítulo anterior, el número de formas que la relación no lineal puede adoptar es sumamente grande. Consideremos ecuaciones del tipo:

$$Y = a + b_1X + b_2X^2 + b_3X^3 + \dots + b_kX^k \quad (\text{XIX.23})$$

Si todos los coeficientes b_2, b_3, \dots, b_k son cero, la ecuación se reduce a la forma lineal familiar. En otros términos: la recta puede considerarse como caso particular de este tipo general de curva, que se designa como *polinomial*. Y en forma análoga, si todos los coeficientes, excepto a, b_1 y b_2 , son cero, obtenemos una

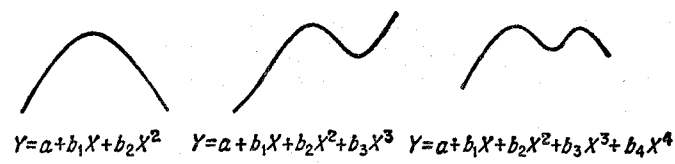


FIG. XIX.6. Formas de polinomios de segundo, tercero y cuarto grados.

polinomial de segundo grado. El grado de la polinomial se refiere al exponente más elevado de X que tenga un coeficiente no cero.

Las polinomiales tienen una propiedad muy importante, la que nos permite decir cuál es el grado de la ecuación que puede resultar más apropiada para nuestros datos. Obsérvese que una polinomial de primer grado es una línea recta sin desviaciones. Sucede que una ecuación de segundo grado contará con una desviación, describiendo de hecho la curva que llamamos parábola. Una polinomial de tercer grado tendrá dos desviaciones; la de cuarto grado, tres, y así sucesivamente. Si ignoramos ciertos casos degenerados en los que las "desviaciones" no se comportan adecuadamente, podremos dibujar las ecuaciones de segundo, tercero y cuarto grados, como se ve en la figura XIX.6. La dirección en que la parábola o curva de más alto grado "se abre", dependerá del signo de los coeficientes. Lo importante es observar que siempre habrá una desviación menos que lo que indica el grado de la ecuación.

Algunas veces obtenemos curvas empíricas que se parecen a una u otra de esas polinomiales, aunque raras veces, si es que alguna, necesitamos ir más allá de una ecuación de tercer grado. La parábola simple proporciona a menudo una adaptación razonablemente buena a los datos, sobre todo si nos damos cuenta de que nuestra curva puede ser perfectamente plana y que nuestros datos no necesitan extenderse lo suficiente para completar la flexión. Así, por ejemplo, los datos podrían ser similares a los que se indican en la figura XIX.7. Aquí, aunque no exista acaso razón teórica alguna para esperar que las marcas vuelvan a bajar una vez que hayamos avanzado cierta distancia a lo largo del eje de las X , la parábola puede constituir con todo una adaptación razonable, dentro de los límites de variación dados en el problema. Es, pues, perfectamente posible que una parábola de

mínimos cuadrados se adapte a los datos mucho mejor que la recta.

Supóngase que sea efectivamente así. ¿Cómo puede tratarse el problema? El lector se habrá ya dado cuenta, sin duda, de la semejanza entre la fórmula de la polinomial general y la de la ecuación de los mínimos cuadrados de más de una variable indepen-

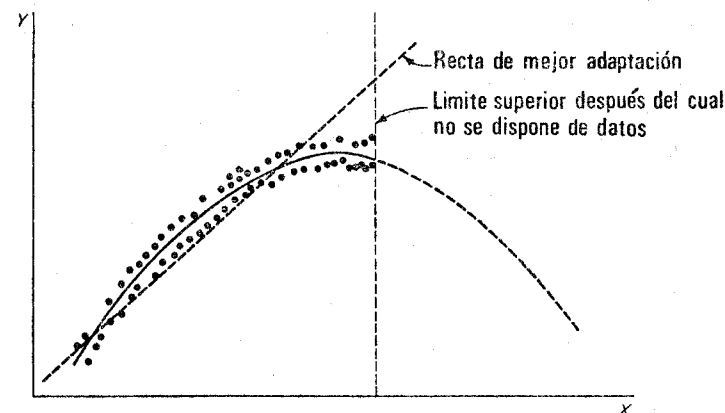


FIG. XIX.7. Datos hipotéticos con una parábola de mejor ajuste.

diente. La única diferencia, en efecto, está en que hemos escrito X^2 en lugar de X_2 , etcétera. Supóngase ahora que hubiéramos de representarnos X^2 como variable separada y distinta de X . Mientras nos servimos de símbolos abstractos esto es perfectamente posible, aunque, admitiéndolo, esta práctica no tendría mucho sentido en términos de una variable concreta. Con todo, las matemáticas del caso resultan ser idénticas. Así, por ejemplo, si sospechamos que la relación entre la discriminación y el porcentaje de negros pueda representarse acaso más adecuadamente por medio de una curva de segundo grado, tomamos el porcentaje de negros como una de las variables independientes X_1 y (el porcentaje de negros)² como segunda variable independiente X_2 . Por consiguiente, la ecuación de segundo grado:

$$Y = a + b_1X + b_2X^2$$

difícil de tratar por medio de los mínimos cuadrados, se reduce a la ecuación familiar:

$$Y = a + b_1X_1 + b_2X_2$$

Para obtener una medida de bondad de ajuste a la parábola, podemos servirnos ahora de la correlación múltiple entre Y y X_1

y X_2 . La diferencia entre el cuadrado de esta correlación múltiple y el cuadrado de la r total (suponiendo linealidad) nos dará una medida del grado en que hemos aumentado nuestra habilidad en cuanto a predecir la discriminación sirviéndonos, sin embargo, de una ecuación de segundo grado en lugar de una recta.

En principio, el procedimiento indicado puede extenderse de varios modos. Podrían emplearse ecuaciones de tercer grado y mayores con objeto de conseguir una adaptación algo mejor. Además, pueden añadirse al cuadro otras variables. Respecto de algunas de estas variables independientes, puede suponerse un modelo no lineal, y uno lineal respecto de otras. Al anticipar tasas de discriminación a partir de cierto número de variables independientes, podremos acaso encontrar que cabría obtener ecuaciones de predicción algo mejores suponiendo modelos no lineales para algunas de las variables. En particular, tal vez la relación entre la discriminación y el porcentaje de negros pueda ser de forma parabólica, en tanto que las variables independientes restantes se relacionan con la discriminación en forma lineal. Por lo tanto, la ecuación múltiple de los mínimos cuadrados adoptará la siguiente forma:

$$Y = a + (b_1X_1 + b_2X_2) + b_3X_3 + \dots + b_kX_k$$

en la que los dos términos al interior del paréntesis se necesitan para describir la relación (no lineal) entre la discriminación y el porcentaje de negros. En este caso también, la variable X_2 vuelve a representar el porcentaje de negros al cuadrado: (porcentaje de negros)². Se concibe que también alguna de las otras X de la ecuación pueda emplearse asimismo para indicar relaciones no lineales entre la discriminación y las demás variables.

En el ejemplo anterior, supóngase que deseábamos obtener la correlación parcial entre la discriminación y el porcentaje de negros controlando en relación con las variables restantes. Como quiera que X_1 y X_2 se han utilizado para referirse a la primera y la segunda potencias del porcentaje de negros, no tendría sentido referir Y a X_1 controlando en relación con todas las demás "variables", incluida X_2 . Antes bien, necesitamos obtener la correlación múltiple entre Y y tanto X_1 como X_2 , controlando en relación con X_3, X_4, \dots, X_k . Para lograr dicho propósito, podemos servirnos del coeficiente parcial-múltiple.

Manejo de la interacción como productos cruzados. En el análisis por dos métodos de la variancia, en el de la covariancia (véase capítulo xx), y en relación con las variables dependientes de escala nominal, concebíamos la interacción estadística como si implicara *cualquier* diferencia de la simple adición. Una alternativa obvia para un modelo aditivo la constituye una relación multiplicativa del tipo que podría ser sugerido mediante argu-

mentos verbales orientados a señalar que, al objeto de tener "presente" la Y , deberán tenerse "presentes" *tanto* la X_1 *como* la X_2 . Cuando se avanza más allá de las simples dicotomías, la idea, generalizada, nos dice que Y puede ser una función multiplicativa de X_1 y X_2 . La ecuación que sigue puede constituir una formulación general de tal relación.

$$Y = (\alpha_1 + \beta_1X_1)^{\gamma_1}(\alpha_2 + \beta_2X_2)^{\gamma_2}$$

en las que los exponentes de gamma pueden ser, o bien positivos, en cuyo caso estará implícita la multiplicación, o negativos, con división implicada. La función puede desde luego ser convertida en aditiva, haciendo una transformación logarítmica de todas las variables, pudiendo extender fácilmente el principio general a más de dos variables independientes.

Supongamos, como aproximación razonable, que ambos exponentes fuesen la unidad, lo que reduciría la ecuación a:

$$Y = (\alpha_1 + \beta_1X_1)(\alpha_2 + \beta_2X_2) = \alpha_1\alpha_2 + \alpha_2\beta_1X_1 + \alpha_1\beta_2X_2 + \beta_1\beta_2X_1X_2$$

Vemos inmediatamente que mediante la suma de un factor que abarca X_1X_2 podremos manejar este tipo de modelo simple multiplicativo, conservando el formato aditivo. Nos limitamos a denominar X_1X_2 como X_3 , construyendo en consecuencia nuestra medida de X_3 , y continuamos adelante. Deseamos, por ejemplo, medir el grado en que X_3 agrega a la variancia explicada, y podríamos probar la significancia de este factor adicional como se indica en la sección siguiente. Si hubiéramos comenzado con tres variables independientes, podríamos haber formado tres factores con los productos X_1X_2 , X_1X_3 y X_2X_3 para determinar las tres interacciones de primer orden, y un triple producto $X_1X_2X_3$ para manejar la interacción de orden superior.

Es necesario formular varias advertencias. En primer lugar, el uso de factores de productos cruzados está justificado con base en que la relación "verdadera" sea multiplicativa y no aditiva, en tanto que la "no aditividad" se refiera a *cualquier* tipo de separación de la aditividad. Tenemos así una medida de interacción algo más restrictiva que la que se obtuvo en relación con el análisis de la variancia, y es posible que otros factores de interacción hubieran funcionado mejor (por ejemplo: $X_1 \log X_2$, $X_1 \cos X_2$, o $e^{X_1} \log X_2$). Segundo: si tomamos $X_3 = X_1X_2$, debemos tener presente que X_3 es una función no lineal exacta de X_1 y X_2 , y por tanto las correlaciones momento-producto de X_3 tanto con X_1 como con X_2 serán de ordinario muy altas. Tendremos así entre manos un problema de multicolinealidad, y no podremos tener mucha fe en nuestras estimaciones de los coeficientes de

los factores X_i . Este problema resulta particularmente serio cuando se comienza con cinco o seis variables independientes y se desea tener en cuenta todas las posibles interacciones. Si las propias variables originales están altamente intercorrelacionadas, o bien forman parte de bloques, los factores de productos cruzados se relacionarán con tales bloques en formas peculiares (véase [1]). En tales casos puede resultar razonable medir hasta qué punto el grupo completo de factores de productos cruzados aumenta significativamente la variancia explicada, mediante el uso del coeficiente parcial-múltiple, o comparando los múltiples, con y sin los factores de los productos. La determinación de los efectos de determinados factores de los productos cruzados puede, sin embargo, resultar demasiado arriesgada, por razón de un gran volumen de errores de muestreo en los que pudiera haberse incurrido.

Hay evidentemente muchos más usos y más posibles extensiones de las técnicas de correlación y regresión múltiples, de los que pueden ser examinados en un texto general. Hemos visto, sin embargo, algunos de los principios básicos más elementales, los que permitirán consultar inteligentemente con los especialistas en caso de que se plantearan problemas más complicados.

XIX.7. Pruebas de significación e intervalos de confianza

En relación con la significación será necesario verificar, por supuesto, tanto el coeficiente múltiple como el parcial. La hipótesis nula y los supuestos serán similares a los que se establecieron en el caso de la correlación total. Una muestra aleatoria será supuesta como de costumbre. El supuesto de una distribución normal multivariable nos asegurará que cada variable está normalmente distribuida alrededor de las otras, que las variancias son iguales, y que la ecuación de regresión tendrá la forma indicada por la ecuación (XIX.1).⁴ Hechos estos supuestos, podemos servirnos de las pruebas de análisis de variancia para la significación de varios coeficientes parciales y múltiples. Veremos primero pruebas de significancia de correlaciones múltiples, ya que éstas son más sencillas desde el punto de vista de los conceptos que las de las correlaciones parciales.

Como quiera que el cuadrado de la correlación múltiple representa siempre la proporción del total de la variación explicada por las variables independientes actuando juntas, hemos dividido

⁴ Debe recalarse una vez más que no todas las X_i necesitan tener distribuciones normales, en tanto la variable dependiente esté normalmente distribuida alrededor de todas las combinaciones de niveles fijos de las variables independientes con la misma variancia σ^2 . Suponemos, con otras palabras, que el factor de perturbación ε_i se encuentra distribuido normalmente con la variancia constante.

de hecho esta variación total en dos porciones: las sumas explicada e inexplicada de cuadrados. Por lo tanto, el cuadro del análisis de variancia será siempre similar al cuadro XIX.1.

CUADRO XIX.1. Prueba de análisis de variancia para la significación de la correlación múltiple

	Sumas de cuadrados	Grados de libertad	Apreciación de la variancia	F	
Total	Σx_i^2	$N - 1$			
Explicada	$R^2 \Sigma x_i^2$	k	$\frac{R^2 \Sigma x_i^2}{k}$	$\frac{R^2}{1 - R^2}$	$\frac{N - k - 1}{k}$
Inexplicada	$(1 - R^2) \Sigma x_i^2$	$N - k - 1$	$\frac{(1 - R^2) \Sigma x_i^2}{N - k - 1}$		

En el cuadro XIX.1 hemos indicado la variable dependiente con X_1 , dejando que k represente el número de las variables independientes. Si R tiene, por ejemplo, una variable dependiente y tres variables independientes, habrá en la ecuación de regresión cuatro parámetros que hay que apreciar. Por consiguiente, sirviéndonos de la ecuación de los mínimos cuadrados para apreciar la variable dependiente, deberíamos perder 4 o $(k + 1)$ grados de libertad. Así, pues, los grados de libertad asociados al término de error serán por lo regular

$$N - (k + 1) = N - k - 1$$

Los grados de libertad asociados a la suma de cuadrados explicada puede obtenerse a continuación por sustracción. Toda vez que los grados de libertad para las sumas de cuadrados explicada e inexplicada resultarán ser siempre k y $N - k - 1$, respectivamente, podemos escribir una fórmula general de F . Obvérese que, al igual que en el caso de las correlaciones totales, el factor que representa la suma total de cuadrados se elimina. Obtenemos así una fórmula general para verificar la significación de una R múltiple, o sea:

$$F_{k, N-k-1} = \frac{R^2}{1 - R^2} \frac{N - k - 1}{k} \quad (\text{XIX.24})$$

No es necesario, por consiguiente, establecer la tabla del análisis de variancia en la forma convencional. Verificando la significación de la correlación múltiple que obtuvimos al explicar la discriminación a partir del porcentaje de negros y el porcentaje urbano (p. 476), obtenemos ahora:

$$F_{2,147} = \frac{.3658}{1 - .3658} \frac{150 - 3}{2} = \frac{.3658}{.6342} \frac{147}{2} = 42.39$$

que es significativa al nivel de .001.

Al verificar la significación de coeficientes parciales, operamos sobre la base del principio de dejar que las variables de control expliquen primero todo lo que pueden. Tomamos a continuación la porción de la suma total de cuadrados que queda *inexplicada* por la variable de control, y nos servimos de ella como nuevo total. Esta última cantidad se descompone luego en dos componentes, las porciones explicada e inexplicada, y una prueba F efectuada tomando la razón de las apreciaciones de la variancia basadas en estas dos últimas componentes. El procedimiento se ilustra en el cuadro XIX.2, en el que verificamos la significación de $r_{13.2}$ (o sea, $H_0: \rho_{13.2} = 0$).

CUADRO XIX.2. Prueba de análisis de variancia para la significación de la correlación parcial $r_{13.2}$

	Sumas de cuadrados	Grados de libertad	Estimación de la variancia	F
Total	Σx_i^2	$N - 1$		
Explicada por 2	$r_{12}^2 \Sigma x_i^2$	1		
Inexplicada por 2	$(1 - r_{12}^2) \Sigma x_i^2$	$N - 2$		
Explicada por 3	$r_{13.2}^2 (1 - r_{12}^2) \Sigma x_i^2$	1	$r_{13.2}^2 (1 - r_{12}^2) \Sigma x_i^2$	
Inexplica- da por 3	$(1 - r_{13.2}^2)(1 - r_{12}^2) \Sigma x_i^2$	$N - 3$	$\frac{(1 - r_{13.2}^2)(1 - r_{12}^2) \Sigma x_i^2}{N - 3}$	$\frac{r_{13.2}^2 (N - 3)}{1 - r_{13.2}^2}$

Obsérvese que los grados de libertad inexplicados decrecen en uno cada vez que se añade una nueva variable. Por otra parte, en la fórmula de F la expresión se simplifica de tal modo, que resulta innecesario escribir la tabla entera cada vez que deseamos efectuar una prueba. En el problema numérico del que nos hemos venido sirviendo (p. 456) el valor de F de la prueba de significancia de la relación entre la discriminación y el porcentaje urbano, controlándolo en relación con el porcentaje de negros, se convierte en:

$$F_{1,N-3} = \frac{r_{13.2}^2}{1 - r_{13.2}^2} (N - 3) \quad (\text{XIX.25})$$

$$= \frac{(.332)^2}{1 - (.332)^2} (147) = 18.21$$

Así pues, la parcial es significativa al nivel de .001.

Si al extender este procedimiento deseamos verificar la significación de $r_{14.23}$, podemos tomar como nuevo total la porción no explicada por 2 y 3 combinadas. Esta cantidad puede luego descomponerse en porciones explicada e inexplicada, practicándose la prueba de F lo mismo que anteriormente. Una vez más, todas las cantidades tanto del numerador como del denominador de F se eliminarán, excepto en cuanto a los factores que comportan las parciales. Toda vez que los grados de libertad asociados al numerador serán siempre la unidad y, como quiera que los del denominador serán $N - k - 1$, podemos escribir la fórmula general de la verificación de la parcial $r_{ij.mn...t}$ como sigue:

$$F_{1,N-k-1} = \frac{r_{ij.mn...t}^2}{1 - r_{ij.mn...t}^2} (N - k - 1) \quad (\text{XIX.26})$$

en donde el número total de variables es $k + 1$.

Obsérvese que al comparar las pruebas de la significación de las correlaciones múltiples y las parciales el término final de error que comporta la suma de cuadrados inexplicada por todas las variables deberá ser el mismo en ambas tablas, a condición, por supuesto, que se empleen las mismas variables dependientes e independientes. Ya demostramos que era así, toda vez que sabemos que:

$$1 - R_{1.23}^2 = (1 - r_{12}^2)(1 - r_{13.2}^2)$$

De los cuadros XIX.1 y XIX.2 puede verse que estas expresiones son las que figuran en las hileras inferiores de las tablas respectivas.

El procedimiento que acabamos de describir para verificar las correlaciones parciales puede utilizarse asimismo para verificar la significación de la parcial-múltiple. A estas alturas el lector estará ya en condiciones de verificar que, con objeto de hacer la prueba de significación de $r_{1(23).45}$ tomaremos la suma de cuadrados no explicada por 4 y 5, sirviéndonos luego del cuadrado de la parcial-múltiple para obtener la proporción de esta nueva suma de cuadrados, que resulta explicada por las variables 2 y 3.

Pueden calcularse asimismo intervalos de confianza para los coeficientes parcial y múltiple, mediante una ligera modificación del procedimiento de transformación de la z descrito en el capítulo anterior. Podemos convertir de nuevo los dos tipos de coefi-

cientes en z sirviéndonos de la tabla. El único cambio que se requiere es que el error estándar de z ya no nos venga dado por

$$\sigma_z = \frac{1}{\sqrt{N-3}}$$

En lugar de ello, en cambio, perdemos un grado más de libertad por cada variable añadida, de modo que el error estándar se convierte en general en:

$$\sigma_z = \frac{1}{\sqrt{N-k-2}} \quad (\text{XIX.27})$$

en donde k representa el número total de variables.

Obtenemos por consiguiente los intervalos de confianza del 95 por ciento para $R_{1.23}$ y $r_{13.2}$ de la manera siguiente:

$$1.96\sigma_z = 1.96 \frac{1}{\sqrt{146}} = .1622$$

	z	$z_l = z - 1.96\sigma_z$	$z_u = z + 1.96\sigma_z$	r_l	r_u
$R_{1.23} = .605$.7010	.5388	.8632	.492	.698
$r_{13.2} = .332$.3451	.1829	.5073	.181	.468

Así, pues, el intervalo de confianza del 95 por ciento alrededor de $R_{1.23}$ va de .492 a .698, en tanto que el de $r_{13.2}$ va de .181 a .468.

Antes de terminar el presente capítulo conviene observar un importante punto más. Cada vez que añadimos a la ecuación de los mínimos cuadrados otra variable, sólo perdemos un grado de libertad más. Podemos, por consiguiente, añadir variables, con una pérdida muy pequeña de eficacia, por lo que se refiere a las pruebas de significación. En ocasiones, la adición de más variables podrá bajar el nivel de significación, debido al hecho de que aquellas no contribuyen a explicar bastante variación adicional para compensar la pérdida en grados de libertad. No obstante, tenemos en la correlación múltiple y parcial un instrumento que, si se aplica adecuadamente, es mucho más potente que cualquiera de los métodos que examinamos anteriormente. Sin embargo, si el número de variables utilizadas empieza a aproximarse al de los casos, podemos esperar obtener unas correlaciones múltiples muy grandes, debido simplemente a que estamos en condiciones de sacar partido de las fluctuaciones fortuitas. Con 15 casos y 15 variables, será posible pasar una superficie de mínimos cuadrados

exactamente entre todos los puntos, incluso si suponemos un modelo de tipo lineal. Por consiguiente, la múltiple R será automáticamente la unidad. De ahí que, lo mismo que las demás técnicas estadísticas, las de regresión y correlación múltiple deban emplearse con precaución. A estas alturas ya no será probablemente necesario señalar que, excepto con fines de exploración, no deberán emplearse, a menos que los supuestos requeridos se cumplan, si no totalmente, por lo menos aproximadamente.

GLOSARIO

Ponderaciones de beta
Correlación múltiple
Correlación parcial-múltiple
Ecuación de regresión múltiple
Distribución normal multivariable
Correlación parcial
Ecuación polinomial

EJERCICIOS

1. Sirviéndose de los datos del ejercicio 1 del capítulo XVII.

- Obténgase la correlación parcial entre la integración moral y la heterogeneidad, controlando la movilidad. Calcúlese asimismo la parcial entre la integración moral y la movilidad, controlando la heterogeneidad. Respuesta, $-.51$; $-.63$.
- Obténgase la ecuación de mínimos cuadrados múltiple, tomando la integración moral como variable dependiente.
- ¿Qué son las ponderaciones beta? ¿Cómo se comparan con las parciales obtenidas en a)?
- Calcúlese la correlación múltiple, tomando la integración moral como variable dependiente. ¿Cómo pueden controlarse los cálculos? Respuesta, $R = .64$.
- Verifíquese la significación de las correlaciones múltiple y parcial calculadas en los apartados a) y d). Pónganse intervalos de confianza del 99 por ciento alrededor de cada una de estas correlaciones.

2. Escribanse fórmulas para $r_{37.12450}$, $R^2_{4.1285}$ y $r^2_{5(23).1407}$. Respuesta, b) $R^2_{4.1285} = R^2_{4.128} + r^2_{45.125}(1 - R^2_{4.128})$.

3. Escribanse las fórmulas de F que se emplearían para verificar el significado de cada una de las correlaciones del ejercicio 2 anterior.

$$\text{Respuesta, (c)} F = \frac{r^2_{5(23).1407}}{1 - r^2_{5(23).1407}} \frac{N-7}{2}$$

BIBLIOGRAFÍA

- Althausen, R. P.: "Multicollinearity and Non-Additive Regression Models", en H. M. Blalock (ed.), *Causal Models in the Social Sciences*, Aldine Publishing Company, Chicago, 1971, cap. 26.

2. Blalock, H. M.: *Causal Inferences in Nonexperimental Research*, University of North Carolina Press, Chapel Hill, 1964, cap. 3.
3. Blalock, H. M.: "Per Cent Non-white and Discrimination in the South", *American Sociological Review*, vol. 22, pp. 677-682, 1957.
4. Christ, Carl: *Econometric Models and Methods*, John Wiley & Sons, Inc., Nueva York, 1966, Parte III.
5. Cowden, D. J.: "A Procedure for Computing Regression Coefficients", *Journal of the American Statistical Association*, vol. 53, pp. 144-150, 1958.
6. Croxton, F. E., y D. J. Cowden: *Applied General Statistics*, 3ª ed. Prentice-Hall, Inc., Englewood Cliffs, N. J., 1967, cap. 21.
7. Davis, J. A.: "A Partial Coefficient for Goodman and Kruskal's Gamma", *Journal of the American Statistical Association*, vol. 62, pp. 189-193, 1967.
8. Draper, N. R., y H. Smith: *Applied Regression Analysis*, John Wiley & Sons, Inc., Nueva York, 1966, caps. 5-10.
9. Dwyer, P. S.: *Linear Computations*, John Wiley & Sons, Inc., Nueva York, 1951.
10. Gordon, Robert: "Issues in Multiple Regression", *American Journal of Sociology*, vol. 73, pp. 592-616, 1968.
11. Hagood, M. J., y D. O. Price: *Statistics for Sociologists*, Henry Holt and Company, Inc., Nueva York, 1952, cap. 25.
12. Johnston, J.: *Econometric Methods*, McGraw-Hill Book Company, Nueva York, 1963.
13. Kendall, M. G.: *Rank Correlation Methods*, Hafner Publishing Company, Inc., Nueva York, 1955, cap. 8.
14. Land, K. C.: "Principles of Path Analysis", en Edgar Borgatta (ed.), *Sociological Methodology 1969*, Jossey-Bass, Inc., Publishers, San Francisco, 1969, cap. 1.
15. Morris, R. N., "Multiple Correlation and Ordinally Scaled Data", *Social Forces*, vol. 48, pp. 299-311, 1970.
16. Quade, Dana: *Nonparametric Partial Correlation*, University of North Carolina, Institute of Statistics Mimeo Series, núm. 526, 1967.
17. Reynolds, H. T.: *Making Causal Inferences with Ordinal Data*, University of North Carolina, Institute for Research in Social Science, Chapel Hill, 1971.
18. Simon, H. A.: "Spurious Correlation: A Causal Interpretation", *Journal of the American Statistical Association*, vol. 49, pp. 467-479, 1954.
19. Somers, R. H.: "An Approach to the Multivariate Analysis of Ordinal Data", *American Sociological Review*, vol. 33, pp. 971-977, 1968.
20. Wilson, T. P.: "A Critique of Ordinal Variables", *Social Forces*, vol. 49, pp. 432-444, 1971.

XX. ANÁLISIS DE COVARIANCIA Y VARIABLES SIMULADAS

HEMOS estudiado el análisis de variancia en que una sola escala de intervalo puede relacionarse con una o más escalas nominales. En el capítulo anterior vimos cómo las técnicas de la correlación podían emplearse para relacionar cualquier número de escalas de intervalo. En el análisis de covariancia combinamos ahora las ideas básicas del análisis de variancia y del análisis de correlación, con objeto de tratar problemas que comportan más de una escala de intervalo en combinación con cualquier número de escalas nominales. Así, pues, el análisis de covariancia es una extensión teórica de estos dos procedimientos, que nos pone idealmente en condiciones de tratar problemas que comporten diversas combinaciones de escalas de intervalo y nominales.

Por desgracia, según veremos en seguida, los cálculos requeridos por el análisis de covariancia son muy fastidiosos si se realizan a mano o con una calculadora de escritorio, pero no se plantean problemas especiales si se dispone de programas de computación. En un terreno ideal cabe ampliar el procedimiento hasta incluir el manejo de un gran número de variables independientes nominales y de escalas de intervalos, a condición de que la variable dependiente sea una escala de intervalo. En la práctica, sin embargo, uno se encuentra limitado a tres o cuatro variables independientes por razón de que las interacciones de más elevado orden resultan muy numerosas pasado aquel límite. El análisis de la covariancia es, en su forma, equivalente a un procedimiento denominado de análisis por "variable simulada", que será explicado al final del capítulo. Este procedimiento equivale a una simple ampliación del modelo de regresión, y el estudio de ambos sistemas suministra una buena apreciación intuitiva de la relación existente entre el análisis de la variancia y la regresión.

En este capítulo limitaremos nuestra atención al caso de tres variables, en el que tenemos una escala nominal y dos escalas de intervalo. El problema básico del que nos ocuparemos es el de relacionar dos de dichas variables controlando en relación con la tercera. Si bien semejante control podría efectuarse tomando categorías de la variable de control y llevando a cabo análisis separados en el interior de esas clases, es posible, con todo, obtener una eficacia mucho mayor mediante el empleo de las técnicas del análisis de covariancia, a condición de que la interacción no sea significativa. En otros términos: el control puede efectuarse sin necesidad de tener que tomar un número sumamente grande de casos. Efectivamente, nos servimos de promedios ponderados y de procedimientos de ajuste, como lo hicimos en el caso de

la correlación parcial. Con todo, al servirnos del análisis de covariancia podemos obtener considerablemente más información de lo que fue el caso con la correlación parcial, ya que podemos desplegar correlaciones y estimaciones de declive separadas para cada categoría de la variable de control, pudiendo además buscar la interacción.

Hay dos tipos de situaciones de los que habremos de ocuparnos: 1) aquellas en las que relacionamos las dos escalas de intervalo, controlando en relación con la escala nominal, y 2) aquellas en que una de las escalas de intervalo es relacionada con la escala nominal, siendo la variable de control la otra escala de intervalo. Pese a que rara vez el interés se fijará en ambos tipos de problemas para una combinación dada de datos, será necesario, con todo, llevar a cabo la mayor parte del análisis requerido por el primer tipo de problema, incluso cuando el interés se centre principalmente en el segundo. Ésta es la razón de que procedamos primero con el tipo de problema en que se utiliza como control la variable de escala nominal.

XX.1. Relación de dos escalas de intervalo, control de la escala nominal

Los métodos básicos de correlación y regresión pueden emplearse para relacionar dos o más escalas de intervalo dentro de las categorías de la variable de control. Habiendo investigado cada una de las relaciones en el interior de las diversas categorías, puede resultar posible juntar los resultados, obteniendo coeficientes promedios de correlación intraclase y de mínimos cuadrados, a condición de que pueda suponerse que las relaciones son las mismas de una categoría a la siguiente. Si los resultados se pueden juntar, puede obtenerse una sola medida general que servirá como medida efectiva de resumen y será más segura como estimación que cualquiera de las medidas de las categorías separadas. El coeficiente promedio de correlación de intraclase puede interpretarse como directamente análogo al coeficiente de correlación parcial, ya que puede utilizarse para representar la relación entre las dos variables de escala de intervalo después de haber permitido que actúe la variable de control.

Hay dos pruebas de significación que hemos de practicar en este tipo de problema. La primera es una prueba para ver si el hecho de juntar los resultados de las diversas clases es o no legítimo. Aquí verificamos esencialmente la interacción, para ver si podemos o no suponer la misma naturaleza de relación (según la medida de las b) para todas las clases. Si no podemos, entonces el juntar no tendrá mucho objeto, y habremos de practicar análisis separados para cada una de las categorías de la variable de control. Pero si la reunión de los resultados parece justificada,

entonces seguimos adelante y obtenemos una correlación promedio de intraclase, y la segunda prueba que hagamos será para ver si dicho coeficiente es o no significativamente distinto de cero.

Como de costumbre, hemos de establecer algunos supuestos acerca de los métodos de muestreo y de las poblaciones de las que se han extraído los datos y, como podría esperarse, estos su-

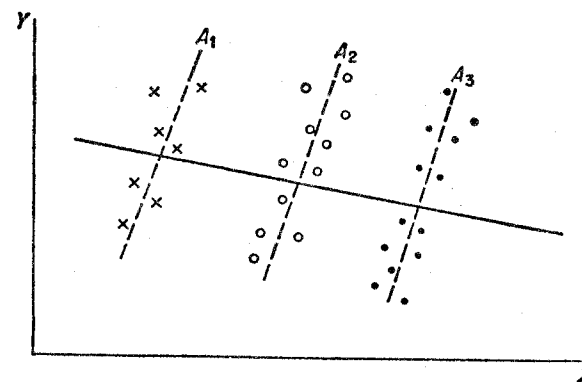


FIG. XX.1. Datos hipotéticos que indican una débil correlación total entre X y Y, pero correlaciones más fuertes dentro de las categorías de A.

puestos serán esencialmente los que requieren el análisis de variancia y el de correlación. En líneas generales, esto es lo que hacemos en el problema del primer tipo de análisis de covariancia. Veamos ahora más de cerca los detalles del procedimiento.

Con objeto de obtener una visión clara de lo que puede suceder cuando empleamos el análisis de covariancia para controlar en relación con una variable de escala nominal, consideremos dos tipos extremos de situaciones. En la figura XX.1 tenemos una situación en la que se da una correlación ligeramente negativa general o total (indicada por la recta continua) entre la variable dependiente Y y una variable independiente X. Si nos fijamos separadamente en cada una de las categorías (A_1 , A_2 y A_3) de la variable de control A, vemos que dentro de cada clase se da una relación positiva más bien fuerte entre X y Y. En este caso, las medias en X dentro de las diversas categorías son lo bastante diferentes como para oscurecer la relación básica entre X y Y.

Si fuéramos a superponer las medias de las tres categorías una sobre otra, moveríamos en realidad las ecuaciones de intraclase de tal modo que quedarán una encima de otra, con lo que obtendríamos entre X y Y una relación mucho más fuerte. En esencia, esto es lo que hacemos cuando obtenemos un coeficiente promedio de correlación intraclase. Una manera de representar-

se el proceso consiste en pensar en términos de haber ajustado nosotros las diferencias entre las categorías A, sacando la fuente de variación debida a la variable de control. Habiendo ajustado en relación con A mediante superposición de las medias de X y Y, podemos comparar ahora las relaciones entre X y Y dentro de las categorías, investigando las diferencias entre las pendientes in-

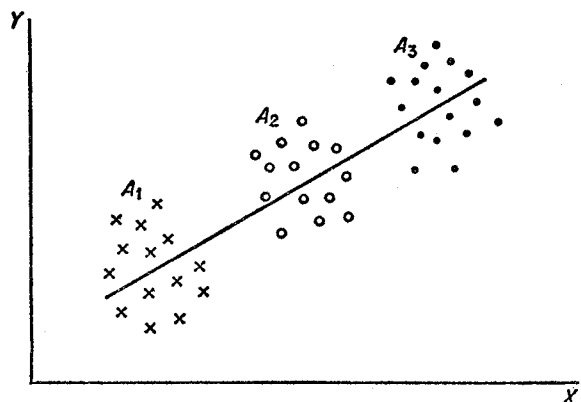


FIG. XX.2 Datos hipotéticos que indican una fuerte correlación total entre X y Y, pero correlaciones más débiles dentro de las categorías de A.

traclase (como lo indican las líneas de trazos). Sin duda, el hecho de superponer las medias afectará las a en cada una de las ecuaciones de mínimos cuadrados, pero dejará inalteradas estas pendientes y las r del interior de las clases.

La figura XX.2 representa una situación contrastante, en la que se dan relaciones extremadamente débiles al interior de las categorías de A, pero en donde la relación general entre X y Y es muy fuerte. La variable de control afecta nuevamente la relación entre X y Y, pero esta vez, si superpusiéramos las medias, no hallaríamos esencialmente relación alguna entre las dos escalas de intervalo. Tal vez no exista relación causal alguna entre X y Y, y la relación general se deba al hecho de que A produce cambios tanto en la una como en la otra de ellas. En tal caso, pues, consideraríamos que la relación entre X y Y era espuria.

En estos dos tipos generales de situación tendrá, por consiguiente, objeto controlar en relación con A. En el primer caso, la correlación parcial o intraclase será mayor en magnitud que la total; en el segundo, en cambio, será menor. Un diagrama de dispersión cuidadosamente construido, sirviéndonos de puntos de diversos colores para representar las distintas categorías de

la variable de control, indicará por lo regular si vale o no la pena molestarse en llevar a cabo un análisis de covariancia sobre la base de los datos disponibles. Si los resultados son semejantes a los de las figuras XX.1 o XX.2, valdrá probablemente la pena seguir adelante. Por otra parte, si los puntos de diversos colores se hallan distribuidos más o menos al azar en el diagrama de dispersión, de modo que las medias de las categorías no sean muy diferentes, no puede esperarse que el análisis de covariancia produzca resultados muy interesantes.

Al superponer las medias de una categoría sobre las de la otra, controlamos de hecho la magnitud de las medias en cuestión. En realidad, pues, medimos variaciones y covariaciones alrededor de las categorías individuales más que con respecto de las grandes medias. Se recordará que esto es exactamente lo que hicimos en el análisis de variancia al dividir la suma total de cuadrados en dos componentes. Una de estas componentes, la variación intraclase, comportaba desviaciones de las medias de clase en tanto que la segunda componente se refería a las desviaciones de las medias de clase en relación con la gran media o media total. Todo lo que ahora necesitamos hacer es extender los mismos procedimientos, descomponiendo la covariación total, o suma de productos, en porciones explicadas e inexplicadas. Nuestro razonamiento será exactamente paralelo al que empleamos en conexión con las sumas de cuadrados. Como quiera que:

$$X_{ij} - \bar{X}_{..} = (X_{ij} - \bar{X}_{.j}) + (\bar{X}_{.j} - \bar{X}_{..})$$

y

$$Y_{ij} - \bar{Y}_{..} = (Y_{ij} - \bar{Y}_{.j}) + (\bar{Y}_{.j} - \bar{Y}_{..})$$

podemos escribir:

$$(X_{ij} - \bar{X}_{..})(Y_{ij} - \bar{Y}_{..}) = [(X_{ij} - \bar{X}_{.j}) + (\bar{X}_{.j} - \bar{X}_{..})][(Y_{ij} - \bar{Y}_{.j}) + (\bar{Y}_{.j} - \bar{Y}_{..})]$$

Si sumamos todos los casos y efectuamos la multiplicación, obtenemos cuatro términos, de los cuales, sin embargo, los dos centrales se eliminan. Como resultado podemos escribir:

$$\sum_i \sum_j (X_{ij} - \bar{X}_{..})(Y_{ij} - \bar{Y}_{..}) = \sum_i \sum_j (X_{ij} - \bar{X}_{.j})(Y_{ij} - \bar{Y}_{.j})$$

Suma total de productos = suma de productos "dentro" (no explicada)

$$+ \sum_i \sum_j (\bar{X}_{.j} - \bar{X}_{..})(\bar{Y}_{.j} - \bar{Y}_{..})$$

+ suma de productos "entre" (explicada)

Aquí también, lo más práctico consiste en servirse de las fórmulas de cálculos de las sumas total y entre de productos, obteniendo la cifra interior por sustracción. Estas fórmulas de cálculo resultan ser exactamente análogas a las que se emplearon para obtener las sumas de cuadrados, excepto en que un valor de Y reemplaza a un valor de las X , de modo que obtenemos productos cruzados, y no cuadrados. Así tenemos, pues:

$$\text{Suma total de productos} = \sum_i \sum_j X_{ij} Y_{ij} - \frac{(\sum_i \sum_j X_{ij})(\sum_i \sum_j Y_{ij})}{N} \quad (\text{XX.1})$$

$$\text{Suma de productos entre} = \sum_j \frac{(\sum_i X_{ij})(\sum_i Y_{ij})}{N_j} - \frac{(\sum_i \sum_j X_{ij})(\sum_i \sum_j Y_{ij})}{N} \quad (\text{XX.2})$$

en donde N_j representa el número de casos en la clase j -ésima.

Lo mismo que en el caso de las sumas de cuadrados, el segundo término es la misma cantidad en ambas ecuaciones. Obsérvese asimismo que, en la fórmula de la suma entre de productos, la cantidad del numerador del primer término representa simplemente el producto de la suma de las X y de la suma de las Y para cada clase. La fórmula nos manda dividir dicho producto entre el número de casos y sumar luego todas las clases.

Hay una diferencia importante entre una suma de productos y una suma de cuadrados, en cuanto la primera puede tener un valor negativo. Así, pues, la covariación total puede ser negativa, en tanto que el valor entre podrá ser positivo. Esto significa, por supuesto, que cuando sustraigamos un número positivo de un número negativo, la suma dentro de productos resultante será un número negativo mayor.

Problema. Antes de seguir adelante será útil presentar un ejemplo numérico e indicar de qué modo los varios cálculos requeridos en el análisis de covariancia pueden llevarse a efecto en forma sistemática. El cuadro XX.1 muestra dichos cálculos para las siguientes variables:

Y (variable dependiente, escala de intervalo): medida de discriminación educativa contra los negros

X (variable independiente, escala de intervalo): porcentaje de negros¹

A (variable independiente, escala nominal): Estado.

Los datos fueron reunidos para una muestra aleatoria de 150 distritos del Sur, utilizando el censo de 1950. Supongamos, en esta parte del problema, que estamos interesados en estudiar la relación entre las marcas de discriminación y el porcentaje de negros, controlando en relación con el estado del distrito.

A primera vista, el cuadro XX.1 se presenta algo formidable, pero, si lo examinamos columna por columna, vemos que por lo menos las trece primeras nada contienen realmente de nuevo. En efecto, las columnas 2, 3, 5, 7, 9 y 11 contienen los datos básicos que se necesitan para todos los demás cálculos. Las columnas 2 a 6 y 7 a 10 sirven para obtener las sumas de cuadrados total, entre y dentro respectivamente de las variables dependiente e independiente. Sirviéndose de esta rutina de cálculos, se opera a través de la tabla simplemente, obteniendo los valores de cada hilera mediante el empleo de la fórmula indicada en la cabeza de cada columna. Así, por ejemplo, las cifras de la columna 6, que representan la suma de los cuadrados en Y , se obtienen sustrayendo la columna 4 de la 5. Por lo tanto, para Florida tenemos: $54\,989 = 3\,866\,409 - 3\,811\,420$. En esta forma obtenemos en la columna 6 la suma de cuadrados dentro de cada estado. Si estas cantidades se suman, obtenemos la suma de cuadrados dentro de clase, de modo que podemos inscribir esta misma cantidad en la hilera inferior de la columna 6. Obsérvese que esta rutina particular de cálculo difiere de la que utilizamos antes al tratar problemas de análisis de variancia, en que hemos obtenido la suma intracase de los cuadrados directamente, sustrayendo este valor del total, para obtener la suma de cuadrados entre. Así, por ejemplo, $1\,370\,555 = 2\,961\,762 - 1\,591\,207$.

Para obtener la suma total de cuadrados, utilizamos exactamente el mismo procedimiento en el caso de cada estado, o sea que sustraemos la columna 4 de la columna 5. Al proceder así nos servimos, por supuesto, de la fórmula:

$$\begin{aligned} \Sigma y^2 &= \Sigma Y^2 - \frac{(\Sigma Y)^2}{N} = 40\,399\,788 - \frac{(74\,938)^2}{150} \\ &= 40\,399\,788 - 37\,438\,026 = 2\,961\,762 \end{aligned}$$

Aquí, la N de la hilera de los totales es el número total de casos de la muestra (150).

Obsérvese que las filas de totales y sumas contienen exacta-

¹ En el cuadro XX.1, las cifras del porcentaje de negros se han multiplicado por 10 con objeto de evitar los decimales.

mente las mismas entradas en las columnas 3, 5, 7, 9 y 11, con puntuaciones burdas ΣY , ΣY^2 , ΣX , ΣX^2 y ΣXY . Pero las entradas difieren en las columnas 4, 8 y 12 referentes a los factores de corrección que han de ser restados para obtener Σy^2 , Σx^2 y Σxy . En realidad, las cifras de "sumas" no son del todo necesarias en las columnas 4, 8 y 12, excepto para comprobar los cálculos. Por ejemplo: la fórmula (6) = (5) - (4), es aplicable a la fila de sumas y así, como comprobación, observamos que

$$1\ 591\ 207 = 40\ 399\ 788 - 38\ 808\ 581$$

La cifra de sumas de la columna 4, a saber: 38 808 581, se obtuvo sumando los resultados correspondientes a cada estado, en tanto que la cifra de "totales", 37 438 026, se obtuvo utilizando el tamaño total de la muestra de 150. Así:

$$37\ 438\ 026 = (74\ 938)^2/150$$

Será útil en este lugar efectuar un número suficiente de cálculos en las columnas 2 a 6 y 7 a 10, de modo que el lector comprenda de qué se trata y se percate de que los resultados que obtiene con este nuevo método son exactamente los mismos (prescindiendo de los errores de redondeo) que los que hubiéramos obtenido con el método antiguo.

Las columnas 11 a 13 sirven para descomponer la variación en partes componentes, en forma análoga. Como se indica más arriba, las fórmulas son similares a las del análisis de variancia, excepto en que los cuadrados se sustituyen por productos, obteniendo por consiguiente la columna 13 sustrayendo la 12 de la 11, como lo indican las fórmulas de cálculo. Calculamos también la suma interior de productos directamente, y el valor entre por sustracción. De este modo, la covariación total es de 3 025 678, y la interior es de 1 744 189, lo que da 1 281 489 para la covariación entre-clases. Ocurre, en este caso, que las tres sumas de productos, lo mismo que los valores para todos los estados, son positivas, pero esto no será siempre necesariamente así. Hemos efectuado ahora el cálculo básico que vamos a necesitar para nuestra labor ulterior, habiendo obtenido las sumas total, explicada e inexplicada para y^2 , x^2 y xy . Nuestra atención puede fijarse ahora en las varias pruebas y medidas que se necesitan para llevar a cabo el análisis. Las columnas restantes del cuadro XX.1 se explicarán cuando lleguemos a ellas.

Prueba de la interacción. Se recordará que en el análisis en dos formas de variancia la primera prueba que efectuamos fue la del efecto de interacción. La razón de proceder a dicha prueba en primer lugar estaba en que si las dos variables independientes producen efectos distintos, al actuar en combinación, de los que

CUADRO XX.1. Cálculos para el análisis de covariancia *

Clase (1)	N_i (2)	Para calcular las sumas de cuadrados en Y (col. 6)				Para calcular las sumas de cuadrados en X (col. 10)			
		ΣY (3)	$(\Sigma Y)^2/N_i$ (4) = $(3)^2/(2)$	ΣY^2 (5)	$(\Sigma Y)^2/N_i$ (6) = $(3)^2/(2)$	ΣX (7)	$(\Sigma X)^2/N_i$ (8) = $(7)^2/(2)$	ΣX^2 (9)	$(\Sigma X)^2/N_i$ (10) = $(7)^2/(2)$
Florida	11	6 475	3 811 420	3 866 409	54 989	2 683	654 408	744 861	90 453
Alabama	8	4 030	2 030 112	2 168 898	138 786	3 367	1 417 086	1 964 231	547 145
Arkansas	10	4 608	2 123 366	2 223 740	100 374	3 211	1 031 052	1 236 701	205 649
Georgia	33	18 911	10 837 149	11 239 451	402 302	12 707	4 892 965	5 826 629	933 664
Kentucky	9	2 724	824 464	891 102	66 638	695	53 669	63 293	9 624
Louisiana	15	7 476	3 726 038	3 926 182	200 144	5 257	1 842 403	2 025 311	182 908
North Carolina	24	9 281	3 589 040	3 862 309	273 269	7 459	2 318 195	3 266 843	948 648
Mississippi	20	12 206	7 449 322	7 586 664	137 342	10 419	5 427 778	6 043 283	615 505
South Carolina	11	5 967	3 236 826	3 371 315	134 489	4 676	1 987 725	2 367 054	379 329
Tennessee	9	3 260	1 180 844	1 263 718	82 874	1 088	131 527	229 200	97 673
Sumas	150	74 938	38 808 581	40 399 788	1 591 207	51 562	19 756 808	23 767 406	4 010 598
Totales	150	74 938	37 438 026	40 399 788	2 961 762	51 562	17 724 266	23 767 406	6 043 140
Entre clase (explicada por A)					1 370 555				2 032 542
Dentro de clase (no explicada por A)					1 591 207				4 010 598

* Adaptada de [4], cuadro 74, pp. 486-487, con la amable autorización del editor.

CUADRO XX.1 [continuación]

Clase (1)	Para el cálculo de covariaciones (col. 13)			Pendientes	Explicada por X	No explicada por X	Para el cálculo de correlaciones	
	ΣXY (11)	$(\Sigma X)(\Sigma Y)/N_j$ (12) = (3)(7)/(2)	Σxy (13) = (11)-(12)	$b = \frac{\Sigma xy / \Sigma x^2}{(14) = (13)/(10)}$	$(\Sigma xy)^2 / \Sigma x^2$ (15) = (13)(14)	$\Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2}$ (16) = (6) - (15)	$r^2 = \frac{(\Sigma xy)^2}{\Sigma x^2 \Sigma y^2}$ (17) = (15)/(6)	$r = \frac{(18)}{\pm \sqrt{(17)}}$
Florida	1 601 644	1 579 311	22 333	24690	5 514	49 475	.10027	.317
Alabama	1 894 209	1 696 126	198 083	36203	71 712	67 074	.51671	.719
Arkansas	1 579 758	1 479 629	100 129	48689	48 752	51 622	.48570	.697
Georgia	7 765 621	7 281 881	483 740	51811	250 630	151 672	.62299	.789
Kentucky	217 349	210 353	6 996	72693	5 086	61 552	.07632	.276
Louisiana	2 700 374	2 620 089	80 285	43894	35 240	164 904	.17607	.420
North Carolina	3 203 824	2 884 457	319 367	33665	107 515	165 754	.39344	.627
Mississippi	6 620 545	6 358 716	261 829	42539	111 379	25 963	.81096	.900
South Carolina	2 737 694	2 536 517	201 177	53035	106 694	27 795	.79333	.891
Tennessee	464 348	394 098	70 250	71924	50 527	32 347	.60968	.781
Sumas	28 785 366	27 041 177	1 744 189		798 158			
Totales	28 785 366	25 759 688	3 025 678	.50068	1 514 896	1 446 866	.51148	.715
Entre clase (explicada por A)			1 281 489			614 189		
Dentro de clase (no explicada por A)				$b_w =$				
			1 744 189	.43489	758 530	832 677	.47670	.690

CUADRO XX.1 [conclusión]

Para calcular el ajuste de las \bar{Y} (col. 23)

Clase (1)	$\bar{X}_j = \Sigma X / N_j$ (19) = (7)/(2)	$x = \bar{X}_j - \bar{X}_{..}$ (20) = (19) - $\bar{X}_{..}$	$b_w x$ (21) = b_w (20)	$\bar{Y}_j = \Sigma Y / N_j$ (22) = (3)/(2)	$\bar{Y}'_j = \bar{Y}_j - b_w x$ (23) = (22) - (21)
Florida	243.909	- 99.838	- 43.42	588.64	632.06
Alabama	420.875	77.128	33.54	503.75	470.21
Arkansas	321.100	- 22.647	- 9.85	460.80	470.65
Georgia	385.060	41.313	17.97	573.06	555.09
Kentucky	77.222	-266.525	-115.91	302.67	418.58
Louisiana	350.467	6.720	2.92	498.40	495.48
North Carolina	310.792	- 32.955	- 14.33	386.71	401.04
Mississippi	520.950	177.203	77.06	610.30	533.24
South Carolina	425.091	81.344	35.38	542.45	507.07
Tennessee	120.889	-222.858	- 96.92	362.22	459.14
Sumas					
Totales	$\bar{X}_{..} = 343.747$			$\bar{Y}_{..} = 499.59$	

esperábamos sobre la base de sus efectos separados, tiene muy escaso objeto, teóricamente, estudiar los efectos de una de ellas controlando la otra. En otros términos: la relación entre una de las variables independientes y la variable dependiente difiere según el valor de la variable de control. Si tal es el caso, la relación deberá estudiarse separadamente en el interior de cada una de las categorías de la variable de control. En el análisis de covarianza nos enfrentamos a un problema similar, aunque, en lugar de pensar en términos del supuesto de adición, nos encontramos ahora comparando las pendientes de las ecuaciones de los mínimos cuadrados en el interior de cada una de las categorías. Observemos primero el paralelismo entre el supuesto de adición y el de pendientes iguales. Estaremos luego en mejores condiciones de comprender la naturaleza de la prueba de interacción en el análisis de covarianza.

En el capítulo XVI, que trata del análisis de variancia, nos serviremos del siguiente ejemplo numérico con objeto de ilustrar la adición:

	A_1	A_2	A_3
B_1	5	10	20
B_2	10	15	25
B_3	25	30	40

Se hizo observar que no era necesario suponer diferencias iguales entre las marcas de B_1 y B_2 , por una parte, y las de B_2 y B_3 , por la otra. Pero hubimos de suponer que las diferencias entre B_1 y B_2 eran las mismas para cada una de las categorías de A . Supongamos ahora que la variable B represente en realidad una variable X de escala de intervalo, que ha sido categorizada. Habremos de suponer que las relaciones entre X y la variable dependiente Y (representada por las marcas en el cuerpo de la tabla) son lineales dentro de cada una de las categorías de A . Una somera reflexión nos convencerá de que, situando adecuadamente las categorías de B a lo largo del eje de las X , podemos traducir la propiedad de adición en el enunciado de que las tres líneas de regresión presentan todas ellas la misma pendiente. La figura XX.3 indica esta relación. Vemos así que la prueba de adición es directamente análoga a la de la hipótesis de que las pendientes en el interior de las clases son iguales.

Al verificar la interacción en el análisis bimodal de variancia tomamos la cantidad de variación de la variable dependiente que no podía ser explicada por las dos escalas nominales al suponer adición. Esta cantidad se fragmentó luego en dos componentes, a saber: la cantidad que podía explicarse por la interacción, y la cantidad que permanecía inexplicada todavía por los efectos entre columnas, entre hileras y de la interacción. La ra-

zón de estas dos últimas componentes se utilizó para verificar la interacción. En el análisis de covarianza hacemos exactamente lo mismo, pero, según cabía esperar, nuestro procedimiento adopta una forma algo distinta. Acabamos de ver, en efecto, que el supuesto de adición es análogo al de que las pendientes de población dentro de cada una de las categorías son las mismas.

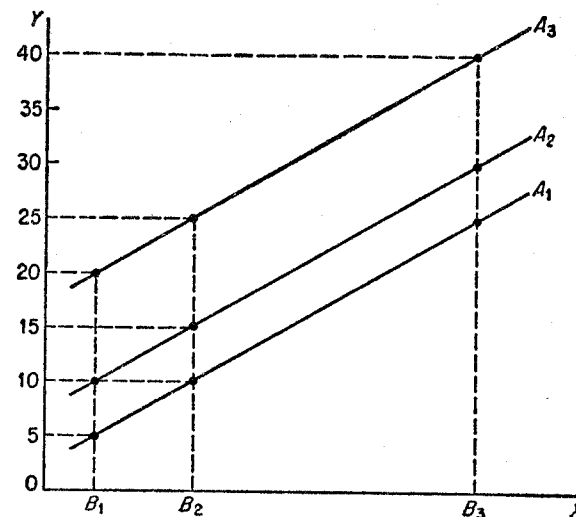


FIG. XX.3. Rectas de pendientes iguales, que indican no interacción.

Sin embargo, si se da un efecto significativo de interacción, esto supondrá una relación diferente para algunas por lo menos de las categorías. En otros términos: un determinado cambio de X producirá diferentes cambios de Y en las distintas clases de A . Si tomamos ahora la cantidad de variación de Y no explicada por X suponiendo pendientes iguales, podemos ver cuánta variación adicional podemos explicar por la interacción. Podemos luego verificar la interacción comparando la suma de cuadrados de ésta con el término de error.

¿Cómo determinamos la cantidad de variación que podemos atribuir a la interacción? Para contestar a esta pregunta, hemos de interrogarnos primero a nosotros mismos cuánta variación podríamos eventualmente esperar explicar sirviéndonos de modelos lineales dentro de cada una de las categorías de A . Manifiestamente, la ecuación individual de los mínimos cuadrados para cada categoría nos da el mejor ajuste que pueda esperarse de una recta, y el coeficiente de correlación calculado sobre la base de los datos de dicha categoría particular nos proporcionará

una medida de la bondad de ajuste. Podemos, pues, obtener para cada categoría cifras que representen la cantidad de variación de Y explicada por X , sirviéndonos de la recta que mejor se ajusta a los datos de dicha categoría particular. Al sumar las variaciones explicadas para cada una de las categorías, obtenemos la cantidad de variación efectivamente explicada por todas las ecuaciones distintas de los mínimos cuadrados. Y en forma análoga, al sumar las sumas inexplicadas de cuadrados, obtenemos la cantidad de variación de Y que permanece todavía sin explicar por esas líneas de mínimos cuadrados distintas.

En el cuadro XX.1, estos cálculos se han llevado a cabo en las columnas 15 y 16. En el caso de Florida, por ejemplo, la variación total de Y (columna 6) es de 54 989. De esta cantidad, 5 514 es explicada por la ecuación de mínimos cuadrados que mejor se adapta a los datos de Florida, permaneciendo inexplicada la de 49 475. De la variación total de Y (2 961 762), la cantidad de 798 158 representa la cantidad dejada inexplicada por esas ecuaciones de mínimos cuadrados separadas.

Hemos de preguntarnos a continuación cuánta variación queda inexplicada si se supone que no hay efecto de interacción. Si no lo hay, entonces todas las pendientes de las categorías de A serán iguales. Nuestra mejor apreciación de esta pendiente común consistirá en una apreciación conjunta, que es un promedio ponderado de las pendientes individuales en el interior de las clases. Estas pendientes se han calculado en la columna 14. La apreciación conjunta, o pendiente media dentro de las clases, se ha calculado asimismo en la columna 14, sirviéndonos de los datos interiores a las clases de las columnas 10 y 13. Así, el valor de 43489 se obtuvo dividiendo 1 744 189 entre 4 010 598.

Podemos comparar ahora las relativas capacidades de explicación de las distintas líneas de mínimos cuadrados interiores a las clases, cada una con una pendiente distinta, y un número de rectas trazadas a través de las medias de cada categoría, pero de igual pendiente todas ellas, esto es, la b promedio "dentro" de las clases (véase figura XX.4). De estas últimas líneas paralelas no puede esperarse que expliquen tanto de la variación total como las líneas realmente mejor ajustadas de cada una de las categorías; pero, si no se da efectivamente interacción alguna en los datos de la población, las distintas ecuaciones de regresión tendrán todas la misma pendiente, y podemos esperar que las líneas de mínimos cuadrados no difieran en cuanto a la pendiente en forma demasiado pronunciada. En otros términos: si no se da interacción, la serie de líneas paralelas de trazos se acercará con una aproximación relativamente buena a las ecuaciones de mínimos cuadrados reales de cada categoría. Toda vez que en tal caso el valor de la pendiente media dentro de las clases no será demasiado diferente de aquel de cada una de las pendientes par-

ticulares dentro de las clases, las líneas de trazos tendrán un poder explicativo casi tan grande como las continuas.

Debido a las fluctuaciones de la muestra, podemos esperar alguna interacción dentro de la muestra, aun si no la hay acaso entre la población. Las líneas continuas y de trazos nunca serán idénticas y, por consiguiente, estas últimas dejarán siempre algo

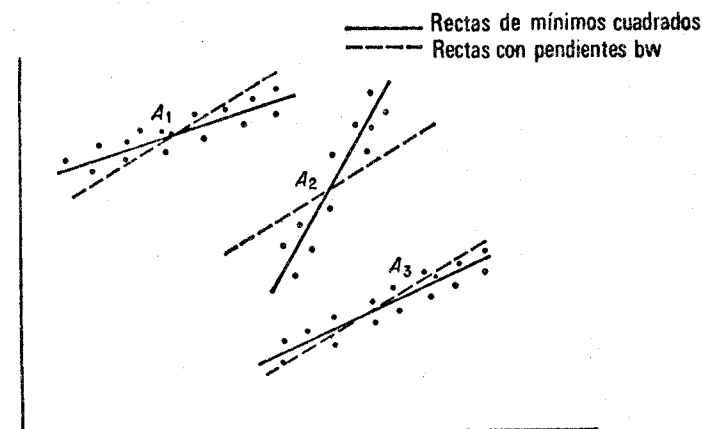


FIG. XX.4. Comparación entre rectas separadas de mínimos cuadrados y rectas a través de las medias de categorías, pero todas tienen la misma pendiente b_w .

más de variación sin explicar que las líneas individuales de mínimos cuadrados. La cuestión está ahora en saber si las líneas de mínimos cuadrados separadas difieren o no suficientemente entre sí, y por consiguiente de las líneas de trazos, para garantizar nuestra conclusión en el sentido de que la interacción es significativa desde el punto de vista estadístico.

En esta y en otras pruebas del análisis de covarianza hemos de establecer una serie de supuestos que son esencialmente los mismos requeridos por los análisis de variancia y regresión. Como de costumbre, hay que suponer una muestra aleatoria independiente. Hemos de suponer también normalidad bivariable entre X y Y dentro de cada una de las categorías de A . Además, hemos de suponer que las variancias de X y Y son las mismas dentro de todas las categorías de A .² En la prueba de interacción, nuestra hipótesis nula será, por supuesto, que cada una de las relaciones de categoría entre X y Y comporta la misma pendiente β .

² Una vez más resultará innecesario suponer la normalidad de las X en tanto las Y estén distribuidas normalmente (con variancias iguales) alrededor de las puntuaciones fijas X , dentro de cada categoría de la variable de escala nominal.

En el cuadro XX.2 se efectúa una prueba F en relación con la interacción. Tomamos la cantidad de variación de Y no explicada por X y A , suponiendo ausencia de interacción, o iguales pendientes de población. Esta cantidad puede encontrarse en el cuadro XX.1, recorriendo la hilera correspondiente al interior de las clases, hasta llegar a la columna inexplicada por X (columna 16).

CUADRO XX.2. Prueba de análisis de variancia para la interacción

	Suma de cuadrados	Grados de libertad	Estimación de la variancia	F
Inexplicada por X y A , suponiendo ausencia de in- teracción	832 677	$N - (k + 1) = 139$		
Explicada por interacción	34 519	$k - 1 = 9$	3 835.4	
Error	798 158	$N - 2k = 130$	6 139.7	< 1.0

Como quiera que la cifra de 832 677 se obtuvo sirviéndonos de la misma combinación de cifras que en el caso de la b promedio dentro de las clases, vemos que hemos supuesto esencialmente pendientes iguales al calcular esta suma inexplicada de cuadrados. Vimos también que la cantidad de 798 158 representa la cantidad de variación dejada sin explicar por las ecuaciones separadas de mínimos cuadrados. Por lo tanto, la diferencia entre estas dos cantidades representa la cantidad de variación que puede atribuirse a la interacción.

Para asociar grados de libertad con cada una de esas cantidades, contamos el número de coeficientes que se han estimado en las ecuaciones de mínimos cuadrados respectivas. Fijándonos primero en el término de error, o cantidad dejada sin explicar por las ecuaciones separadas de mínimos cuadrados, observamos que para cada una de estas ecuaciones separadas hubimos de calcular dos coeficientes (a y b). Por lo tanto, perderemos $2k$ grados de libertad, representando k el número de categorías de A . Así, pues, los grados de libertad asociados a dicho término serán $N - 2k$. Sin embargo, para servirnos de las líneas de trazos, sólo hubimos de calcular una sola pendiente, la b promedio dentro de las clases. Como quiera que, sin embargo, cada una de dichas líneas atraviesa un grupo diferente de medias de muestra, tenemos valores distintos de a para cada una de esas líneas. Hemos perdido, por consiguiente, $(k + 1)$ grados de libertad, y los grados de libertad asociados con este término serán $N - (k + 1)$, o sea $N - k - 1$. Podemos obtener luego los grados de libertad para el término de interacción sustrayendo, lo que nos da:

$$(N - k - 1) - (N - 2k) = k - 1$$

o uno menos que el número de categorías. Calculamos ahora F en la forma usual y concluimos que, toda vez que $F_{9,130} < 1.0$, no tenemos interacción significativa.

Como quiera que la interacción no resultó ser significativa, estamos justificados en reponer la pequeña cantidad de interacción de la muestra en el término de error, sirviéndonos en adelante de la cantidad de 832 677 como variación no explicada ni por X ni por A . Y toda vez que al proceder así nos hallamos en el extremo indebido de la prueba de interacción, hemos de comportarnos con cierta cautela. Sin embargo, con una N tan grande y un valor tan pequeño de F , no corremos ciertamente riesgo alguno al excluir en este problema particular la interacción.

Si ésta hubiera sido significativa, nuestro próximo paso habría consistido en averiguar el o los estados que discrepan de los demás. Esto se efectúa fácilmente consultando la columna de las b . Si resulta manifiesto que algunos estados producen el efecto de interacción y si pueden sugerirse buenas razones teóricas del porqué esto sea así, entonces será acaso posible excluir dichos estados y repetir la prueba con el resto. Pero si no destaca en esta forma estado alguno, tal vez será necesario seguir adelante analizando cada estado separadamente. En tal caso podrán eventualmente obtenerse valiosos datos teóricos preguntándose uno mismo por qué difiere la relación entre la discriminación y el porcentaje de negros de un estado a otro.

Una estrategia de posible uso cuando se dan diferencias apreciables entre los declives, consiste en ordenar por rangos las categorías (en nuestro caso, estados) en relación con las magnitudes de las pendientes, tratando a continuación de localizar alguna *variable* específica que esté sumamente correlacionada con dicha ordenación. Por ejemplo: tal vez cuando ordenamos los estados de abajo hacia arriba en relación con lo inclinado de los declives (aquí todos positivos), podremos observar que los estados con inclinaciones más pronunciadas tienden a ser los más urbanizados o los más industrializados. Si tal fuera el caso podríamos obtener una medida Z de urbanización (o industrialización), reemplazando la escala nominal "estado", con la Z , utilizando a continuación alguna alternativa específica a un modelo aditivo, tal como la función multiplicativa $Y = kX^aZ^b$. Tomando los logaritmos de ambos lados, esta función multiplicativa puede ser transformada en la ecuación aditiva $\log Y = \log k + b_1 \log X + b_2 \log Z$.

La correlación promedio dentro de clase. Habiendo establecido que no se da efecto de interacción significativo alguno estamos ahora justificados en agrupar las r individuales dentro de las clases para obtener un coeficiente de correlación promedio en el interior de las clases, que será análogo al coeficiente de correlación parcial. En otros términos: toda vez que estamos justifica-

dos en suponer una sola pendiente para todas las ecuaciones de regresión, podemos suponer asimismo que los coeficientes de correlación de la población serán también iguales, y que el valor común puede apreciarse juntando las r de la muestra para las varias clases. El coeficiente de correlación promedio intraclase, que podemos simbolizar como $r_{XY \cdot A}$, se calcula del mismo modo que la b promedio intraclase, sirviéndonos de los datos relativos dentro de las clases registrados en la hilera inferior del cuadro XX.1 (véanse las columnas 17 y 18). El cuadrado de dicho coeficiente puede interpretarse como la proporción de variación en Y que no es aplicada por A , pero sí por X . Así:

$$.47670 = (.690)^2 = \frac{758\,530}{1\,591\,207}$$

Si nos fijamos en las fórmulas empleadas para el cálculo de cada uno de estos números, veremos que la interpretación se deduce inmediatamente a partir de ellas. A título de control global de nuestros cálculos, la r promedio intraclase habría de resultar comparable en magnitud con las diversas r intraclase separadas. Como quiera que se trata esencialmente de un promedio ponderado, los estados con el mayor número de distritos ejercerán la mayor influencia en la determinación de su valor. Si alguna de las b en la columna (14) resulta ser negativa, las r comparables en la columna (18) deberán recibir desde luego signos negativos.

Si quisiéramos tener una medida análoga a la R múltiple, podríamos tomar la razón de la cantidad de variación explicada por X y A juntas a la suma total de cuadrados. En este problema, por ejemplo, hemos explicado $2\,961\,762 - 832\,677$ o $2\,129\,085$. Por lo tanto, hemos explicado $2\,129\,085/2\,961\,762$ o 71.9 por ciento de la variación. Hemos de recordar, sin embargo, que si queremos formar una R múltiple tomando la raíz cuadrada de dicho valor, el resultado será en parte una función del número promedio de casos dentro de las categorías de A (véase sección XVI.5).

Podemos efectuar la prueba de la significación de $r_{XY \cdot A}$ en la forma habitual. Primero dejamos que la variable de control A explique todo lo que puede. Dejamos luego actuar X sobre la variación no explicada, fragmentando esta última cantidad en dos componentes. La primera de éstas será la porción explicada por X , y la segunda será el término de error, que no es explicado ni por X ni por A (suponiendo que no se da interacción). Ya vimos que los grados de libertad del término de error serán $N - (k + 1)$. Los grados de libertad asociados a la variación inexplicada por A , que figura en la hilera al pie de la columna 6, serán, por supuesto, $N - k$ (véase sección XVI.1). Esto deja un grado de libertad asociado a la componente no explicada por A pero explicada

por X . Los resultados de esta prueba se hallan resumidos en el cuadro XX.3. Vemos, en esta forma, que la correlación promedio intraclase es significativa al nivel de .001.

CUADRO XX.3. Prueba de análisis de variancia para la significación de la correlación promedio intraclase ($\rho_{XY \cdot A}$)

	Suma de cuadrados	Grados de libertad	Estimación de la variancia	F
No explicada por A	1 591 207	$N - k = 140$		
No explicada por A , pero explicada por X	758 530	1	758 530	
Error (suponiendo ausencia de interacción)	832 677	$N - (k + 1) = 139$	5 990.5	126.6

Antes de terminar esta porción del capítulo en la que hemos estudiado la relación entre dos escalas de intervalo controlando en relación con la escala nominal, podemos establecer una comparación con el tipo de control efectuado por la correlación parcial. Sin duda, el control por medio del análisis de covariancia comporta considerablemente más trabajo que el empleo de la correlación parcial. Como se concibe fácilmente, las extensiones que comporten variables adicionales empezarán a requerir tantos cálculos, que por lo regular el análisis de covariancia no resultará practicable. Por otra parte, el análisis de covariancia nos proporciona más información que la correlación parcial. En efecto, podemos no sólo efectuar una prueba de interacción, sino que podemos investigar además las relaciones entre X y Y dentro de cada una de las categorías de las variables de control, comparando los diversos valores de r y b . Al servirnos de las correlaciones parciales, en cambio, sólo obtenemos la única medida comparable a la correlación promedio dentro de clase, y no podemos efectuar la prueba respecto de la interacción.

Vemos, pues, que el análisis de covariancia presenta cierto número de ventajas respecto de los análisis que emplean las correlaciones parciales, sobre todo en aquellos estudios en los que se pueda esperar que se da interacción. Así, pues, en algunos casos valdrá eventualmente la pena convertir una de las escalas de intervalo en escala nominal y de proceder con el análisis de covariancia, en lugar de la correlación parcial, aun a sabiendas de que perdemos así información con respecto al nivel de medición.

XX.2. *Relación de una escala de intervalo y una escala nominal, control de la escala de intervalo*

En el análisis de una forma de variancia relacionamos una escala de intervalo con una sola escala nominal, probando el significado de las diferencias entre las medias de las categorías de A . Con objeto de determinar la magnitud de la relación entre las dos variables, calculamos un coeficiente de correlación intraclase. Obtuvimos asimismo las medias de las diversas categorías que podían utilizarse con fines descriptivos para indicar las marcas relativas de una categoría con las otras. En el análisis cruzado de variancia pudimos controlar en relación con una escala nominal, averiguando la interacción. Sin embargo, nos vimos fuertemente limitados, ya que necesitábamos tener el mismo número de casos en cada subcasilla. En esta sección, en cambio, veremos situaciones en las que deseamos relacionar Y y A , pero en las que la variable de control es una escala de intervalo X .

Supóngase que nuestro interés se endereza ante todo en descubrir la relación entre las cuotas de discriminación y las subregiones del Sur, definidas por los diversos estados. Sin duda, los estados no constituyen las mejores clases de unidades para delinear subregiones, pero nos sirven con todo aquí con fines de ilustración. Es obvio que una variable como la del porcentaje de negros necesita ser controlada, ya que los diversos estados del Sur difieren considerablemente en cuanto a los porcentajes de sus minorías respectivas. Supóngase que dividiéramos el porcentaje de negros en categorías y procediéramos a efectuar análisis de variancia separados para cada una de ellas. Obsérvese que probablemente ni siquiera intentaríamos el análisis cruzado de variancia, debido a la necesidad de tener subclases iguales. Pero, ¿es que los análisis separados de variancia resuelven realmente nuestro problema? Al examinar los distritos de porcentajes bajos de minoría, encontraríamos inmediatamente que excluíamos prácticamente todos los distritos de Mississippi y Alabama, incluyendo en cambio prácticamente todos los de Kentucky y Tennessee. Por otra parte, habría a lo sumo uno o dos distritos de estos últimos estados con un alto porcentaje de negros. Así, pues, al tratar de controlar por este método, descartamos casi nuestro problema, por cuanto sólo unos pocos estados estarían representados en cada uno de los análisis separados. Los efectos de las subregiones o estados se confundirían así irremisiblemente con el porcentaje de negros. En efecto, no podemos mantener literalmente constante una de las variables, sin reducir al propio tiempo la variabilidad de la otra.

Si bien no podemos mantener la variable de control efectivamente constante, podemos con todo, sirviéndonos del análisis de covariancia, efectuar algunos ajustes en relación con sus efectos.

Concretamente: si estamos dispuestos a suponer que las regresiones de Y a X dentro de cada una de las categorías de A tienen una pendiente común que puede apreciarse por la b promedio intraclase, podemos apreciar el cambio producido en Y por un cambio dado de X . En otros términos: podemos formular algunas predicciones acerca de lo que ocurriría con las tasas de discriminación en cada estado si los porcentajes de la minoría fueran a cambiar. En particular, podemos preguntarnos a nosotros mismos, ¿qué ocurriría con estas tasas si los porcentajes distintos de negros fueran a igualarse? Esta clase de proceso sólo proporciona resultados hipotéticos, y esto ha de tenerse claramente presente. En efecto, no tratamos de obtener tasas de discriminación de los distintos estados manteniendo realmente constante el porcentaje de negros, sino que sólo podemos predecir lo que ocurriría si esto fuera efectivamente así y si las relaciones entre X y Y fueran efectivamente tales como se supone que son. Se concibe perfectamente que, si los negros fueran a redistribuirse a sí mismos en forma más uniforme entre los estados del Sur, las relaciones particulares halladas entre X y Y ya no se verificarían. No obstante, un procedimiento de ajustes de esta clase puede conducir a menudo a comprobaciones útiles.

Si puede presumirse que no se da efecto alguno de interacción, ya vimos que la mejor manera de apreciar las pendientes comunes de las ecuaciones de regresión dentro de las clases es por medio de la b promedio intraclase calculada en el cuadro XX.1. Podemos describir ahora el procedimiento que vamos a utilizar. Nos gustaría ajustar cada una de las medias $\bar{Y}_{.j}$ de las clases de tal manera que se tuviera en cuenta el hecho de que las medias en X difieren asimismo de un estado a otro. Con fines de comodidad supondremos que todas las $\bar{X}_{.j}$ están ajustadas respecto de la gran media de las X . Eso comporta el desplazar la media de X para cada clase en una distancia de $(\bar{X}_{..} - \bar{X}_{.j})$. La figura XX.5 indica esta diferencia como el largo de la base del triángulo. Pero sabemos que para obtener la cantidad de cambio en Y para un cambio dado de X hemos de multiplicar el cambio de X por la b promedio intraclase. Por consiguiente, $\bar{Y}_{.j}$ cambia en la cantidad de $b_w(\bar{X}_{..} - \bar{X}_{.j})$, en donde nos servimos del símbolo b_w para representar la pendiente promedio intraclase. El valor ajustado de las medias de Y puede encontrarse ahora añadiendo dicho incremento a la media original de Y .

Así, pues, dejando que $\bar{Y}'_{.j}$ represente el valor ajustado, tenemos:

$$\begin{aligned}\bar{Y}'_{.j} &= \bar{Y}_{.j} + b_w(\bar{X}_{..} - \bar{X}_{.j}) \\ &= \bar{Y}_{.j} - b_w(\bar{X}_{.j} - \bar{X}_{..})\end{aligned}\quad (XX.3)$$

La segunda de estas formas, que sólo comporta la inversión del orden de las $\bar{X}_{..}$ y $\bar{X}_{.j}$, y el cambio correspondiente de signos, es la forma que se ha empleado para el cálculo de la \bar{Y} ajustada en el cuadro XX.1. Obsérvese que, en este ejemplo concreto, la pendiente es positiva, siéndolo también el cambio de $\bar{X}_{.j}$ a $\bar{X}_{..}$, tal como lo muestra la figura XX.5. Los mismos resultados alge-

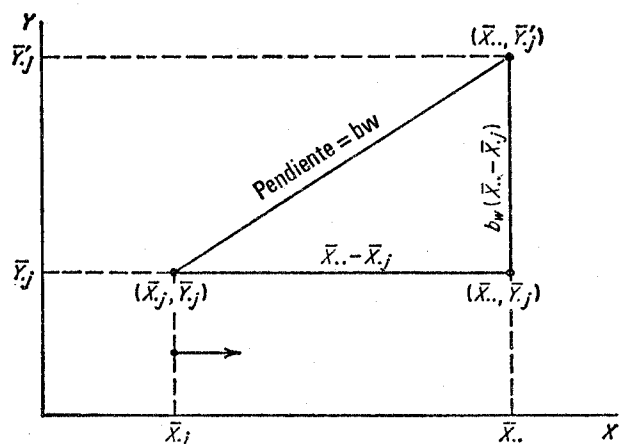


FIG. XX.5. Interpretación geométrica de los cálculos de las medias Y ajustadas.

braicos se verifican exactamente en el caso de ser la pendiente negativa, o cuando el valor de X decrece. A estas alturas deberíamos estar ya en condiciones de convencernos por cuenta propia de que esto es efectivamente así.

La figura XX.6 ayudará a entender lo que hemos hecho al ajustar los valores medios de Y . En efecto, hemos desplazado cada una de las medias de las clases, paralelamente a la pendiente de la b promedio intraclase, a una posición en la que todas las \bar{X} son iguales, a la gran media de las X . Las \bar{Y} ajustadas pueden hallarse a lo largo de la línea de trazos, correspondiente a la gran media de las X . Las magnitudes relativas de las medias en Y pueden resultar considerablemente alteradas. En la figura XX.6, los valores no ajustados de \bar{Y} son tales que la media de A_1 queda ligeramente debajo de A_2 , la cual, a su vez, es sustancialmente menor que A_3 . Obsérvese, con todo, que A_1 tiene un valor \bar{X} muy pequeño. Toda vez que la pendiente se ha representado como positiva, el ajuste respecto de X tiene por efecto aumentar el valor de Y en el caso de A_1 . Por otra parte, el proceso de ajuste reduce los valores de Y tanto para A_2 como para A_3 , ya que estas

dos categorías tienen valores de X relativamente grandes. Como resultado de ello, la \bar{Y} ajustada para A_1 es efectivamente mayor que la ajustada para A_2 , y el valor de A_3 es mucho más vecino del de A_1 .

Si volvemos al cuadro XX.1, columnas 22 y 23, observaremos el efecto del ajuste del porcentaje de negros sobre las tasas de

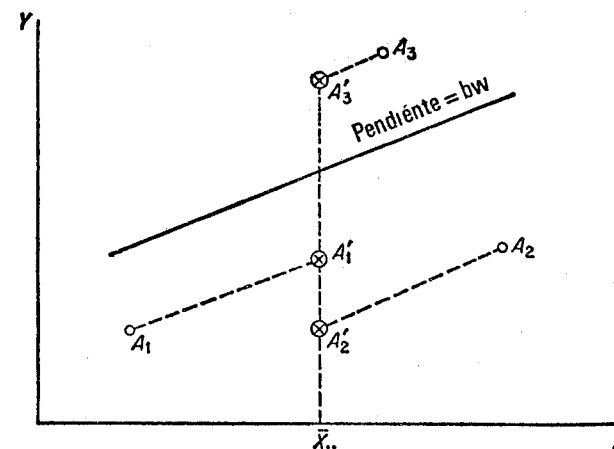


FIG. XX.6. Interpretación geométrica de las medias ajustadas de Y mediante deslizamiento de las medias de categorías paralelamente a la recta de pendientes b_w .

discriminación. Florida, en efecto, que tiene un porcentaje relativamente bajo de negros, destaca ahora con tasas ajustadas muy altas, en tanto que estados como los de Mississippi y Tennessee concuerdan ahora más con los estados restantes. Obsérvese también que las diferencias totales entre estados se han reducido considerablemente.

Ya se mencionó en este mismo capítulo que el análisis de covarianza resultará útil si los diagramas de dispersión revelan que las diversas medias de las clases en X son muy diferentes en valor. Esto puede apreciarse en la figura XX.6. Si las medias de las clases hubieran estado agrupadas densamente en X alrededor de la gran media, las bases, y por consiguiente, también los lados de los triángulos habrían sido muy cortos. En otros términos: el hecho de ajustar respecto de X no produciría un efecto muy pronunciado, ya que, en realidad, el ajuste efectivamente requerido era muy pequeño. Y si todas las medias de las clases hubieran sido exactamente iguales en X , habríamos tenido efectivamente un control en relación con X . Sólo cuando las medias de las clases en X son muy diferentes podemos espe-

rar que el ajuste produzca un efecto notable. Expresado en otra forma: ha de haber una relación relativamente fuerte entre X y A , las dos variables independientes.

Para que el ajuste valga la pena se requiere además otra cosa. En efecto, si la b promedio intraclase hubiera sido muy pequeña numéricamente, se habría requerido un cambio muy grande en X para producir un ligero cambio en Y . Así, pues, si entre X y Y dentro de las clases de A sólo se da una relación pequeña o nula, no tendrá objeto alguno ajustar en relación con X . Estas observaciones concuerdan, por supuesto, con el sentido común, el cual nos dice que no se obtiene gran ventaja controlando respecto de una variable que no esté relacionada con las dos variables que nos interesan. Sin duda, si X sólo se relaciona con la variable dependiente, se la puede controlar como influencia perturbadora. Sin embargo, podemos ver en la figura XX.6 que, a menos que se den algunas variaciones con respecto a X entre las categorías de A , el ajuste no tendrá gran objeto.

Para servirnos de la b promedio intraclase hubimos de suponer que no se daba efecto alguno de interacción. Por lo tanto, será necesario llevar a cabo la prueba de la interacción así como los cálculos de b antes de seguir adelante con el proceso de ajuste. Si la interacción resulta significativa, el problema es más complicado y queda fuera del objeto del presente texto. En determinadas circunstancias será acaso posible ajustar sirviéndose de las pendientes individuales dentro de las clases. Sin embargo, la interpretación ha de efectuarse con prudencia. Supóngase, por ejemplo, que la pendiente de Mississippi resultaba ser totalmente distinta de la de Tennessee. ¿Podríamos en tal caso servirnos legítimamente de sus líneas individuales de mínimos cuadrados para ajustar los valores de Y ? Esto requeriría suponer que Mississippi mantiene esencialmente los mismos tipos de discriminación a medida que va perdiendo negros. Sin embargo, el hecho de que otros estados muestren relaciones distintas con el porcentaje de negros sugiere que el supuesto puede no ser legítimo. El hecho de que se haya demostrado que existe interacción deberá hacernos muy cautos en cuanto a conjeturar lo que ocurriría realmente si cambiaran las X . Por otra parte, si encontramos esencialmente la misma relación en cada uno de los estados entre el porcentaje de negros y la discriminación, o sea ausencia de interacción, estamos más confiados en el sentido de que el ajuste no nos extraviará demasiado.

Hemos de plantear ahora la cuestión relativa a la significación de las diferencias entre las medias ajustadas de Y . Las diferencias entre las medias no ajustadas podrán o no haber sido significativas, pero esto no implica, con todo, que el mismo resultado se verifique en relación con los valores ajustados. Tal vez, en efecto, el hecho de ajustar respecto de X pueda haber tenido

como consecuencia el juntar más los valores de Y . O tal vez estén ahora más separados. Hemos efectuado una tarea descriptiva, la de obtener efectivamente las figuras ajustadas, de modo que puedan desplegarse con fines de comparación. Y hemos de verificar ahora la hipótesis nula de que, en la población, las medias ajustadas de Y son todas las mismas. Los supuestos en relación

CUADRO XX.4. Prueba de análisis de variancia para la significación de las diferencias entre medias ajustadas

	Suma de cuadrados	Grados de libertad	Estimación de la variancia	F
Inexplicada por X	1 446 866	$N - 2 = 148$		
Inexplicada por X , pero explicada por A	614 189	$k - 1 = 9$	68 243	
Error (suponiendo ausencia de interacción)	832 677	$N - (k + 1) = 139$	5 990.5	11.39

con esta prueba son los usuales. Hemos de suponer, en efecto, muestras aleatorias independientes, normalidad y variancias iguales de las \bar{Y} ajustadas, y hemos de establecer asimismo los supuestos requeridos por el análisis de regresión, es decir, una distribución normal bivariable de X y Y dentro de cada categoría de A .

Afortunadamente, no hemos de volver a calcular las sumas de cuadrados sirviéndonos de los propios valores ajustados. En efecto, podemos llevar a cabo una prueba de análisis de variancia empleando el procedimiento familiar de dejar que la variable de control explique primero todo lo que puede de la variación. Toda vez que nuestra variable de control es ahora X , tomamos como nueva suma total de cuadrados la cantidad de variación no explicada por aquella. Fragmentamos luego esta cantidad en la porción explicada por A y la porción que no ha sido explicada por las dos variables. Los grados de libertad asociados a cada una de esas cantidades ya se han determinado. Los resultados de la prueba de F están resumidos en el cuadro XX.4. Vemos, en esta forma, que las diferencias ajustadas, si bien menores que las originales, son significativas al nivel de .001. Concluimos, pues, que si bien el hecho de ajustar en relación con el porcentaje de negros reduce las diferencias de las tasas de discriminación entre los estados, estas diferencias no desaparecen con todo por completo en el proceso.

Por fin, podemos eventualmente querer calcular una correlación parcial de intraclase entre Y y A , controlando en relación con X . Esto puede ser recomendable, con objeto de obtener una

mejor indicación del grado de relación entre las dos variables de la que pueden indicar las diferencias entre las medias ajustadas. Fijándonos simplemente en estas diferencias ajustadas no podemos obtener una idea muy buena de sus magnitudes relativas a las diferencias dentro de las categorías, y por ello una correlación parcial de intraclass puede resultar útil. Generalizando la noción de la correlación intraclass podemos escribir:

$$r_{iYA \cdot X} = \frac{V_b - V_e}{V_b + (\bar{n} - 1)V_e}$$

en donde V_b = estimación entre clase (no explicada por X ; explicada por A)

V_e = estimación del error (inexplicada por X y A)

\bar{n} = número promedio de casos por clase, calculado conforme a la ecuación (XVI.2).

Aquí nos interesa la estimación entre clase de la variancia de las Y ajustadas. Nuestra estimación del error tiene en cuenta que X ha explicado ya todo lo que podía de la variación en Y .

Numéricamente obtenemos, pues:

$$\begin{aligned}\bar{n} &= \frac{1}{k-1} \left(\sum_{i=1}^k N_i - \frac{\sum_{i=1}^k N_i^2}{\sum_{i=1}^k N_i} \right) \\ &= \frac{1}{9} \left(150 - \frac{2858}{150} \right) \\ &= \frac{1}{9} (150 - 19.05) = 14.55\end{aligned}$$

y

$$\begin{aligned}r_{iYA \cdot X} &= \frac{68\,243 - 5\,990.5}{68\,243 + 13.55(5\,990.5)} \\ &= \frac{62\,252.5}{149\,414} = .417\end{aligned}$$

XX.3. Extensiones del análisis de covariancia

La adición de una segunda escala nominal complicará el análisis de covariancia, debido al requisito de subclases iguales. Desde el punto de vista práctico, esto significa de hecho que este tipo de extensión no resultará practicable, excepto en estudios que com-

porten esquemas de experimento en los que el control del número de casos sea posible. Sin embargo, si añadimos una o más escalas de intervalo, la extensión es sencilla en principio, pese a que introducirá un número considerable de cálculos adicionales. Habremos de añadir nuevas columnas a la tabla de cálculo. En particular, habrá una columna que indique la cantidad de variación dejada sin explicar por las dos escalas de intervalo (X y Z) actuando simultáneamente. Nos vemos envueltos, en esta forma, en ecuaciones de mínimos cuadrados múltiples para cada una de las categorías de A . Para obtener medias ajustadas de Y , por ejemplo, habremos de ajustar en relación con X y Z sirviéndonos de las dos b promedios parciales dentro de las clases. En lugar de desplazar ahora las medias de las categorías paralelamente a una línea de mínimos cuadrados, habremos de deslizarlas paralelamente a un plano medio interior a las clases. Para verificar la significación de las Y ajustadas, dejaremos primero que X y Z expliquen de Y todo lo que puedan, permitiendo a continuación a A actuar sobre el remanente.

Como quiera que no hemos agotado ni con mucho el tema relativo al análisis de covariancia, el lector podrá, si lo desea, consultar las referencias que se relacionan más abajo acerca de otras aplicaciones y extensiones del método general aquí expuesto. En particular, cuando el número de las categorías de A es muy grande, a veces resulta muy útil investigar la regresión de las medias de categorías de Y en las medias de X , considerando así de hecho cada categoría como un caso. Así, por ejemplo, en el problema que hemos considerado, podríamos desear estudiar la relación entre X y Y sirviéndonos como unidades de los estados, en lugar de los distritos, y tratando las \bar{X} y las \bar{Y} de cada estado como marcas particulares. En la mayoría de los problemas que interesan a los sociólogos, sin embargo, el número de categorías de A será demasiado reducido para justificar semejante análisis, siendo ésta la razón de que el tema no se examine en el presente capítulo.

XX.4. Análisis de la variable simulada

Tanto en el análisis de variancia como en el de la covariancia nuestra atención estaba centrada en el proceso de dividir en varios componentes las sumas de cuadrados y las sumas de productos, en tanto que en el análisis de regresión lo estaba más bien en calcular los coeficientes de una ecuación. Resultará útil reunir ahora ambas ideas, demostrando la forma en que una combinación de las escalas de intervalo y nominales puede ser también manejada bajo el formato de la regresión. Recuérdese que en análisis por dos métodos de la variancia se indicó que es posible utilizar un modelo aditivo de la forma

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

en tanto que en la regresión múltiple utilizamos ecuaciones de la forma siguiente

$$Y_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon_i$$

Aparte las diferencias en notación, que no deben preocuparnos, observamos dos diferencias obvias en estas ecuaciones: 1) En el modelo simplificado de regresión no tenemos en cuenta la interacción, y 2) El modelo aditivo, en el caso del análisis de variancia, no contiene ninguna X que represente escalas de intervalo. Observamos empero que el modelo de regresión no plantea restricción alguna a las X en cuanto a su distribución de frecuencia, aunque *cabe* suponer que aquéllas se encuentren distribuidas normalmente. En particular, algunas X , o todas ellas, podrían recibir marcas de 1 o 0, y vamos a ver cómo el hacerlo nos permitiría manejar las escalas nominales como casos especiales. Hemos observado, asimismo, que algunas de las X pueden ser producto de otras X (podemos, por ejemplo, hacer $X_3 = X_1 X_2$), y mediante este dispositivo podremos manejar factores de interacción en el contexto de la regresión. El análisis de variancia puede así ser considerado como un caso especial del análisis de regresión y viceversa, resumiendo los dos bajo un solo modelo matemático general.

Para que la explicación no se haga demasiado abstracta, supongamos que estamos tratando con una variable dependiente de escala de intervalo Y ; dos variables independientes de escala de intervalo, X_1 y X_2 , y una sola escala nominal compuesta por cuatro categorías. Supongamos que Y representa el ingreso, a la edad de 35 años; X_1 sus años de estudio; X_2 la puntuación relativa a su *status* ocupacional, y Z_i una variable (que más abajo describimos) que representa la región en la que radica el sujeto. Si hay cuatro regiones (Nordeste, Sur, Medio Oeste y Oeste), podemos utilizar tres Z_i , como sigue:

$$Z_1 = 1, \text{ si el sujeto reside en el Nordeste} \\ = 0 \text{ en otro caso}$$

$$Z_2 = 1, \text{ si el sujeto reside en el Sur} \\ = 0, \text{ en otro caso}$$

$$y \quad Z_3 = 1, \text{ si el sujeto reside en el Medio Oeste} \\ = 0, \text{ en otro caso}$$

La "variable" Z_i se denomina *variable simulada*, ya que las marcas de 1 y 0 son asignadas arbitrariamente. Podríamos en realidad haber utilizado un grupo distinto de marcas, pero el empleo de 1 y 0 mantendrá el análisis dentro de la mayor sencillez.

Obsérvese que no hay necesidad de usar una Z_4 que tome el valor de 1 en el caso de ser Oeste o el de 0 en otro caso, ya que, si conocemos los valores de Z_1 , Z_2 y Z_3 , sabremos con certeza el de Z_4 . En particular, todos los sujetos del Oeste recibirán marcas de 0 en las tres Z . En tanto nos ocupemos de una sola escala nominal, y en tanto, asimismo, no introduzcamos una constante α en la ecuación, será posible incluir en ésta la Z_4 . Si por el contrario, retenemos α , o si tenemos más de una escala nominal, e intentamos retener Z para todas las categorías, veremos que los procedimientos de mínimos cuadrados se vendrán abajo, debido al hecho de que, en este ejemplo, Z_4 es una función lineal perfecta de Z_1 , Z_2 y Z_3 . Podemos de hecho ver que $Z_4 = 1 - (Z_1 + Z_2 + Z_3)$. En la práctica, si tomamos la costumbre de "suprimir" siempre una categoría de cada escala nominal, estaremos listos para aplicar mínimos cuadrados bajo los supuestos habituales. Resultará que la categoría suprimida (en este caso el Oeste) formará una base de comparación con las categorías restantes.

Podemos ahora integrar una ecuación estimativa como sigue:

$$Y = a + b_1 X_1 + b_2 X_2 + c_1 Z_1 + c_2 Z_2 + c_3 Z_3$$

Interpretemos esta ecuación antes de introducir otro modelo más complejo, en el que se tenga en cuenta la interacción. Supongamos que estamos ocupándonos de un sujeto del Oeste, para el cual $Z_1 = Z_2 = Z_3 = 0$. En este caso la ecuación se reduce a

$$Y = a + b_1 X_1 + b_2 X_2$$

Si comparamos ahora este individuo con otro del Nordeste, para el cual $Z_1 = 1$, y $Z_2 = Z_3 = 0$, veremos que para este segundo sujeto la ecuación contendrá un término adicional $c_1 Z_1 = c_1(1) = c_1$, pudiendo considerar que ha sido agregado a a . Así, para el sujeto del Nordeste tenemos

$$Y = (a + c_1) + b_1 X_1 + b_2 X_2$$

y vemos que c_1 puede ser interpretado como la *diferencia* entre los puntos de corte entre las dos regiones. De forma análoga, c_2 puede ser interpretado como la diferencia entre los puntos de corte entre los individuos del Sur por comparación con los del Oeste. En este sentido, la categoría suprimida representa un grupo de comparación para las tres categorías restantes. En términos causales puede interpretarse la c_1 como los incrementos o decrementos en los ingresos en que se incurriría si todos los individuos hubiesen de emigrar del Oeste a las demás regiones.

Consideremos a continuación el caso en que deseamos tener en

cuenta las interacciones entre las regiones y X_1 o X_2 . Para mayor claridad limitaremos nuestra atención a X_1 , eliminando X_2 de la ecuación. En el caso del análisis de covarianza vimos que la interacción aparecía como una diferencia entre los declives de categoría dentro. Esto puede manejarse en función de la formulación de la variable simulada mediante la introducción de factores de la forma $d_{ij}X_iZ_j$. En el caso de una variable independiente X_1 , y tres Z_i , nuestra ecuación será:

$$Y = a + b_1X_1 + c_1Z_1 + c_2Z_2 + c_3Z_3 + d_{11}X_1Z_1 + d_{12}X_1Z_2 + d_{13}X_1Z_3$$

En el caso de un sujeto del Oeste, para el cual $Z_1 = Z_2 = Z_3 = 0$, la ecuación se reduce a $Y = a + b_1X_1$. Para el individuo del Nordeste, en cambio, la ecuación será:

$$Y = a + b_1X_1 + c_1Z_1 + d_{11}X_1Z_1 = (a + c_1) + (b_1 + d_{11})X_1$$

ya que $Z_1 = 1$ para todas las personas de aquella región. Si comparamos las ecuaciones de los individuos del Nordeste con nuestra ecuación "estándar", correspondiente a los del Oeste, no sólo tendremos una diferencia de corte c_1 , sino también una diferencia en declives. Esto permite interpretar d_{11} como el incremento (o decremento) que agregamos al declive de la relación entre X_1 y Y cuando los individuos se desplazan del Oeste al Nordeste. Pueden darse interpretaciones análogas a d_{12} y d_{13} , y si estos coeficientes se separan de cero en forma significativa, deduciremos que hay presente una interacción en la población. Por otra parte, un examen de las magnitudes de la d_{ij} puede resultar útil para comprender dicha interacción.

Acabamos de considerar el caso en que sólo hay un intervalo y una variable nominal independiente, y los resultados de este análisis serán idénticos a los obtenidos al aplicar el análisis de la covarianza. Los procedimientos de cálculo son muy sencillos, siempre que se disponga de programas de computación capaces de manejar problemas de regresión múltiple. Basta utilizar las marcas de las variables de escalas de intervalo, tal como aparecen, convirtiendo sus escalas nominales en variables simuladas Z_i , analizando éstas a continuación en forma idéntica a lo que se habría hecho en el caso de la regresión múltiple. Cada uno de los coeficientes b_i , c_j y d_{ij} puede ser investigado para determinar la significancia. Pueden obtenerse correlaciones múltiples y parciales, y así sucesivamente. Si, por ejemplo, se desea medir el poder explicativo de la región, con un control para todas las X_i , se comenzará por comprobar si es posible prescindir de las interacciones. Si ello es posible, puede calcularse una parcial múltiple que relacione Y con todas las Z_j tomadas en conjunto, con un control para todas las X_i .

Si se desean utilizar dos o más escalas nominales, pueden seguirse dos estrategias alternativas, ambas sencillas. Una posibilidad consiste en combinar las dos escalas en una simple escala nominal, procediendo a continuación como antes se indica. Si se desea, por ejemplo, estudiar las interacciones de raza y sexo con la educación X_1 , para medir en qué forma resulta afectado el ingreso Y , pueden utilizarse las cuatro combinaciones negro-varón (Z_1), negra-hembra (Z_2), blanca-hembra (Z_3) y blanco-varón (suprimida), comparando así las tres combinaciones raza-sexo restantes con el grupo blanco-varón como grupo estándar. La segunda alternativa consiste en utilizar dos variables simuladas distintas, una para el sexo y otra para la raza. Si hacemos $Z_1 = 1$ para todos los negros, y $W_1 = 1$ para todas las hembras, podremos introducir explícitamente interacciones de primer orden entre X_1 y raza, con sólo agregar un factor que abarque el producto X_1Z_1 , pudiendo, de manera análoga, utilizar el factor X_1W_1 para averiguar la interacción ingreso-sexo. Podríamos también manejar una interacción raza-sexo utilizando el producto W_1Z_1 , que sería igual a la unidad sólo en el caso de las hembras negras. Podemos además manejar las interacciones, de más elevado orden, raza-sexo-ingreso, mediante un factor igual al producto $X_1W_1Z_1$.

Si se cuenta con dos escalas nominales con categorías r y c , respectivamente, habrá $(r-1)$ y $(c-1)$ categorías no suprimidas, y necesitaremos $(r-1)(c-1)$ términos para mejorar todas las interacciones de los dos factores. Podemos, pues, expresar Y como una función de los efectos principales de la variable de fila, de los efectos principales de una variable de columna y de una serie de factores de interacción. Podemos así tratar el análisis por dos métodos de la variancia, como un caso especial del análisis mediante variable simulada, y no necesitaremos suponer igual número de casos en todas las subcasillas, ya que estamos aceptando las intercorrelaciones entre las variables independientes. Como ocurría en el caso del análisis de regresión, habremos de pagar el precio de una ambigüedad teórica, resultante de la superposición en la variación, la que será "explicada" por las dos variables independientes correlacionadas. Como ejercicio, puede resultar útil imaginar de nuevo los problemas discutidos en el capítulo del análisis de la variancia, dentro de este nuevo concepto de las variables simuladas.

XX.5. Observaciones finales

Hemos cubierto cierto número de aproximaciones estadísticas al análisis multivariado, aunque algunos temas más especializados se han quedado sin tratar. El problema que tal vez es el fundamental del análisis multivariado, consecuencia de la falta de teorías bien específicas que dicten de manera precisa los pormenores

seguir, es el de encontrar métodos relativamente sistemáticos para hacer frente a diversos tipos de complicaciones. La tarea básica consiste en eliminar tantas de dichas complicaciones como sea posible, pero sólo una vez que nosotros mismos hayamos descubierto su existencia y valorado su importancia. La estrategia general consiste en disponer un grupo de prioridades ordenadas en principio, tratando a continuación de eliminar en primer lugar aquellas complicaciones potenciales en las que estemos menos interesados, avanzando a continuación hacia un análisis más intensivo, que incluya aquellas que ocupan el centro del propio interés teórico y que en forma empírica resulten las más importantes.

Hay varios tipos de complejidades que han sido mencionados sólo de paso. Entre ellos la posibilidad realista de encontrar varios tipos de error tanto en las mediciones aleatorias como en las no aleatorias. Como hemos visto, los primeros han recibido cierto grado de atención en la bibliografía estadística, en tanto los últimos han permanecido virtualmente ignorados hasta hace muy poco tiempo. Se encuentra un segundo tipo de complejidad en la investigación no experimental realista, en la que es necesario tener presente una causación recíproca. Hemos supuesto que la elección de variable dependiente no es problemática, y que no hay efecto de retroalimentación de las variables dependientes a las independientes. Aunque hemos aceptado la posibilidad de variables independientes intercorrelacionadas, no hemos examinado modelos que traten de tener en cuenta estas intercorrelaciones, tomando algunas de las variables "independientes" como función de las otras. Estos temas serán tratados en un volumen posterior, habiendo sido estudiado en gran detalle por los econométristas en conexión con modelos de ecuaciones simultáneas. (Véanse especialmente, Christ [2], y Johnston [6]).

Un tipo de complicación, que ha sido estudiado, abarca la adición a una ecuación de variables explicativas, las que, como acabamos de hacer notar, pueden estar intercorrelacionadas. Se ha observado que siempre que dichas intercorrelaciones sean altas en relación con las correlaciones con la o las variables dependientes, resultará especialmente difícil separar sus efectos componentes. Por ello, una forma a que deben ajustarse siempre las simplificaciones, es la de reducir hasta un número razonable las variables explicativas. Esto se logra mediante cierto número de artificios. Uno de éstos consiste en separar las variables en "bloques", tratando solamente éstos como diferenciados. O bien, puede construirse una sola marca para la totalidad del bloque (por ejemplo, *status* socioeconómico), o pueden usarse medidas tales como el coeficiente de correlación múltiple parcial, para determinar los efectos del bloque en su conjunto. Junto a estas operaciones, puramente estadísticas, debe incluirse una cuidadosa concep-

ción teórica, relativa a la naturaleza del particular bloque de variables que hemos formado. A tal fin pueden ser usadas las técnicas del análisis de factor múltiple, análisis de grupo, análisis de estructura latente, análisis de clasificación múltiple y correlación canónica.

Suele darse el caso de que un investigador sea capaz de reunir sus variables independientes en varios grupos, de acuerdo con sus intereses teóricos. Figurarán en primer lugar aquellas variables en las que se centra su interés principal. A continuación un grupo de las variables independientes que se proponga usar como variables de control. Estas son las variables que espera han de tener mayor efecto sobre las variables de su interés primordial, pero que en términos de su propio esquema investigativo serán consideradas como "variables perjudiciales". No pueden ser ignoradas, pero en teoría tendrán poco interés. Habrá por fin un grupo de variables, grupo tal vez muy grande, que se considere que tienen relativamente menor importancia, o que han sido sugeridas como variables con las que hay que contar en caso de que se observe que las restantes tienen escaso valor explicativo. En los estudios exploratorios es razonable incluir estas variables, ya que las orientaciones teóricas son por lo general muy vagas. La estrategia básica del análisis, en el caso de este tercer grupo de variables, consiste en comenzar por ver cuántas de ellas pueden ser eliminadas desde luego. Las que no estén en este caso podrán ser transferidas al segundo grupo. Lo importante es que, al tratar de reducir la amplitud del análisis, deberá trabajarse de afuera hacia adentro, por así decir. Trátese primero de eliminar las complicaciones. En este caso, tal intento consiste en librarse de aquellas variables que sólo muestran un poder explicativo marginal. En general, y a menos que se disponga de amplios recursos económicos, muchas de tales variables serán eliminadas automáticamente si las correlaciones de orden cero con las variables dependientes son despreciables, o si las variables se encuentran altamente asociadas con otras variables independientes cuyo interés sea más fundamental.

Las posibles no linealidades constituyen otra forma de complejidad que deberá ser siempre investigada en el caso de las escalas de intervalo, pudiendo ser evaluadas aproximadamente en el caso de los datos ordinales. Es muy cierto que todas las relaciones bivariadas (incluso las que existen entre variables independientes) deberán ser rutinariamente investigadas en relación con la no linealidad, comparando para ello E^2 con r^2 . Si tal diferencia es estadísticamente significativa pero numéricamente pequeña (debido a que se trata de una muestra muy grande), será necesario resolver si el incremento explicado, al tener en cuenta la no linealidad, justifica el aumento en la complejidad. La solución dependerá de las peculiares prioridades de la investigación,

y el lugar central que ocupe esta relación particular con vistas al análisis consiguiente. Por ejemplo: si hay una relación no lineal embebida en un complejo grupo de relaciones, entre, tal vez, tanto como diez o quince variables, quizá no valga la pena aceptar el aumento en la complejidad. Si, por el contrario, no hay más de tres o cuatro variables mayores, y es la variable dependiente la que concentra nuestra atención, puede resultar justificado el refinamiento. En tal caso deberá tratarse de especificar una función matemática razonablemente sencilla (por ejemplo: logarítmica, parabólica o exponencial) que explique casi tanto de la variancia como la función no lineal completamente irrestricta (es decir: sin restricciones en las medias de las categorías), cuyo poder explicativo es medido por E^2 . En otras palabras, no basta con indicar que una relación importante *no es lineal*. Deberá indicarse su forma específica, haciendo una prueba para ver si tal forma (por ejemplo, una parábola) se ajusta mejor, en forma significativa, que una línea recta. La posibilidad adicional de que una forma particular de la relación varíe también con el nivel de otras variables (lo que supone una interacción), deberá ser investigada asimismo. Por ejemplo: una relación puede ser logarítmica para hombres, y lineal para mujeres. Cuando en un sistema se dan tantas como ocho o nueve variables, el número de posibles complejidades de este tipo aumenta en progresión geométrica, a medida que va agregándose una variable más. Habitualmente, sin embargo, la mayor parte de las complejidades potenciales no llegan a hacerse presentes.

Por último, debe investigarse siempre la posibilidad de interacciones o relaciones no aditivas entre las variables independientes. Con variables independientes múltiples se darán numerosas interacciones de orden elevado, las que prácticamente siempre son ignoradas en el análisis. Una estrategia razonable consiste en buscar todas las posibles interacciones de dos variables. La mayoría de éstas, según se verá, son sin duda despreciables. Podrán hacerse pruebas de significancia de grupos enteros de interacciones, utilizando para ello los coeficientes parciales múltiples. Supongamos, por ejemplo, que se cuenta con cuatro variables independientes X_1 , X_2 , X_3 y X_4 . Podrían sumarse a la ecuación de regresión todos los productos cruzados posibles $X_i X_j$, comprobando si este grupo de variables agrega en forma significativa al valor de la variancia explicada. Si no ocurre así, todas las interacciones podrán ser omitidas. Si, por el contrario, se produce efecto, al menos, algunas de ellas podrán tal vez ser eliminadas.

Cuando se encuentra un número razonablemente grande de interacciones significativas de dos variables, puede considerarse justificado buscar otras interacciones de orden elevado. El supuesto, en este caso, es el de que no aparecerán interacciones

de orden elevado si se observó la ausencia de interacciones de orden inferior. Las bases teóricas de tal supuesto pueden no estar lo suficientemente claras, pero, en términos puramente empíricos, el supuesto parece razonable. Es cierto que si uno hallase amplias interacciones de tercero y cuarto órdenes en ausencia de interacciones de primer orden, resultaría ciertamente difícil encontrar una explicación teórica del hecho. Tal vez podría defenderse, a medias, el ignorar las interacciones de dos factores en ausencia de efectos principales, pero, por lo menos en el caso de las variables de baja prioridad, los efectos principales casi cero justificarán de ordinario el que se descuide el estudio de las interacciones en las que intervienen dichas variables. Hay ocasiones, por supuesto, en que uno puede encontrarse desorientado, pero resulta inevitable en el análisis multivariado el verse obligado a tomar por algunos atajos.

La cuestión principal por recalcar es la de que la búsqueda de interacciones (y no linealidades) debe ser no sólo sistemática, sino rutinaria. No debe correrse un "albur" buscando sólo un subgrupo selecto de posibles interacciones, en tanto se desconoce el resto de éstas. En raras ocasiones las teorías de las ciencias sociológicas (y sus afines) son lo bastante explícitas y precisas como para especificar y predecir tales interacciones (especialmente las de orden superior) con anterioridad a la recopilación de datos. La falla principal de esta clase de "barrido" aplicado al análisis de los datos es por supuesto la de que sólo por casualidad habrá de hacerse visible cierto número de interacciones significativas. Por lo tanto, cuando se las encuentra deberá comprobarse si están o no *diseñadas* en forma sistemática. ¿Tendrán a abarcar, por ejemplo, sólo dos o tres de las variables?

Todo lo anterior lleva implícito que siempre está presente el riesgo de sobreanalizar los propios datos, particularmente cuando el número de parámetros por estimar comienza a aproximarse al tamaño total de la muestra, o cuando se observa un gran número de complejidades de más bien escasa importancia. Hay por supuesto una cierta tensión entre la necesidad de simplificar, por un lado, y el contar con una mayor fuerza explicativa, por el otro. No hay normas rígidas para escoger entre ellas, por razón sobre todo de que el número de tipos de complejidades es grande.

Los estudios varían considerablemente en cuanto al grado en que son principalmente explicativos, o definidamente teóricos. Varían también en cuanto a la calidad de las mediciones, como anteriormente pudo observarse. Cuando las medidas son burdas y la teoría débil, pero se cuenta con un buen número de variables explicativas potenciales, podrán llevarse a cabo los análisis exploratorios mediante el empleo de procedimientos rutinarios de cálculo (Sonquist y Morgan [8]). Cuando se cuenta con una teoría más explícita, resultan recomendables las técnicas de las ecua-

ciones simultáneas. Si el tamaño de la muestra es adecuado, es aconsejable la estrategia de dividir (al azar) la muestra en mitades, o incluso en tercios. Puede así llevarse a cabo un estudio puramente exploratorio con la primera submuestra, utilizando los datos, así obtenidos, para desarrollar las explicaciones teóricas, las que a continuación podrán ser comprobadas usando el resto de los datos. De esta forma pueden adaptarse con gran flexibilidad las técnicas estadísticas multivariadas a las necesidades del momento, utilizándolas para el desarrollo de las propias teorías y la comprobación de éstas.

EJERCICIOS

1. Compruébense tantos cálculos del cuadro XX.1 como sean necesarios para comprender cómo se obtuvieron las cifras en cuestión.
2. Tómense los datos del ejercicio 1, cap. XVII y descompóngase el índice de heterogeneidad en las siguientes categorías: 10.0 a 14.9, 15.0 a 19.9, 20.0 a 24.9, 25.0 a 29.9 y 30.0 a 49.9. Designando la integración moral con Y , la movilidad con X y la heterogeneidad con A :
 - a. Verifíquese la interacción. Respuesta, $F = 2.17$.
 - b. Obténgase $r_{XY \cdot A}$ y verifíquese la significación. Respuesta, $F = 13.6$.
 - c. Ajústense las medias de las categorías en Y en relación con diferencias respecto de X .
 - d. Verifíquese la significación de las diferencias entre las \bar{Y} ajustadas. Respuesta, $F = 2.71$.
 - e. Obténgase la correlación parcial intraclass $r_{iYA \cdot X}$.
3. Lévese a cabo un análisis con variable simulada sobre los datos del ejercicio 2, aceptando la interacción, y compárense los resultados con los del análisis de covariancia.

BIBLIOGRAFÍA

1. Boyle, R. P.: "Path Analysis and Ordinal Data", *American Journal of Sociology*, vol. 75, pp. 461-480, 1970.
2. Christ, Carl: *Econometric Models and Methods*, John Wiley & Sons, Inc., Nueva York, 1966, Parte III.
3. Dixon, W. J., y F. J. Massey: *Introduction to Statistical Analysis*, 3ª ed., McGraw-Hill Book Company, Nueva York, 1969, cap. 12.
4. Hagood, M. J., y D. O. Price: *Statistics for Sociologists*, Henry Holt and Company, Inc., Nueva York, 1952, cap. 24.
5. Johnson, P. O.: *Statistical Methods in Research*, Prentice-Hall, Inc., Englewood Cliffs, N. J., 1949, caps. 10 y 11.
6. Johnston, J.: *Econometric Methods*, McGraw-Hill Book Company, Nueva York, 1963.
7. Schuessler, Karl: "Covariance Analysis in Sociological Research", en Edgar Borgatta (ed.), *Sociological Methodology 1969*, Jossey-Bass, Inc., Publishers, San Francisco, 1969, cap. 7.

8. Sonquist, J. A., y J. N. Morgan: *The Detection of Interaction Effects*, Institute for Social Research, University of Michigan, Ann Arbor, 1964.
9. Suits, Daniel: "The Use of Dummy Variables in Regression Equations", *Journal of the American Statistical Association*, vol. 52, pp. 548-551, 1957.

Quinta Parte

MUESTREO

XXI. MUESTREO

TODAS las pruebas que hemos examinado, lo mismo que los procedimientos empleados para obtener intervalos de confianza, han requerido el supuesto de muestreo aleatorio, y de hecho el lector puede haberse formado acaso la impresión de que la muestra aleatoria era la única clase respetable de muestra utilizada por el estadígrafo, lo cual dista mucho de ser así. Existen, en efecto, cuatro tipos básicos de muestreo de probabilidad que se examinan en el presente capítulo, a saber: la muestra aleatoria, la muestra sistemática, la muestra estratificada y la muestra por conglomerados. Según veremos, es posible servirse de la inducción estadística con cada uno de estos cuatro tipos de muestreo de probabilidad, si bien es cierto, por desgracia, que al presente estamos muy limitados por lo que se refiere al número de tipos diferentes de pruebas que pueden efectuarse con muestras no fortuitas de probabilidad. Sobre todo en el caso de las muestras agrupadas, nuestros cálculos se hacen al propio tiempo mucho más complicados. Por lo tanto, en un texto general como el presente, será imposible hacer mucho más que indicar algunas consideraciones generales de estrategia para escoger el tipo de muestreo que resulte más apropiado en vista de una situación determinada.

Acabamos de indicar que hay cuatro tipos básicos de muestreo de probabilidad, uno de los cuales es el muestreo aleatorio. ¿Qué es, pues, la muestra de probabilidad? La característica distintiva de la muestra de probabilidad es que todo individuo ha de tener una probabilidad *conocida* de quedar incluido en la muestra. En la muestra al azar, ya vimos que todas las combinaciones de individuos tienen iguales posibilidades de figurar. Pero al formular inducciones estadísticas no es absolutamente necesario que todas las probabilidades sean iguales, ya que, si la probabilidad de selección es conocida, será posible ajustar en relación con probabilidades desiguales mediante algún procedimiento de ponderación de una clase u otra. Es esencial, sin embargo, que las probabilidades sean conocidas, con objeto de llegar a los pesos apropiados. Si las probabilidades se desconocen, será imposible servirse legítimamente de la inducción estadística. En efecto, con un muestreo carente de probabilidad, podemos acaso obtener una muestra efectivamente muy representativa, pero no estaremos en condiciones de apreciar los riesgos de error implicados. Después de describir y comparar cada uno de los cuatro tipos de muestreo de probabilidad, examinaremos brevemente algunos casos en los que es probable que se obtengan muestras sin probabilidad.

XXI.1. Muestreo aleatorio sencillo

Se ha recalcado que en el muestreo al azar no sólo ha de tener cada individuo una oportunidad igual de ser seleccionado, sino que todas las combinaciones han de ser además igualmente probables. Hemos indicado también que por lo regular resulta más conveniente seleccionar sin reposición. Los especialistas de la selección suelen designar la muestra de esta clase como "muestra sencilla aleatoria". Obsérvese que después de cada extracción sucesiva la probabilidad para un individuo de ser seleccionado aumenta ligeramente debido al hecho de que quedarán cada vez menos individuos no seleccionados en la población. Si en relación con una extracción determinada las probabilidades de todos los individuos restantes en cuanto a ser seleccionados son iguales, independientemente de los individuos seleccionados anteriormente, entonces tenemos una muestra sencilla aleatoria. En efecto, tenemos independencia de una extracción a la siguiente, excepto en cuanto al hecho de que ningún individuo puede ser seleccionado dos veces.

¿Por cuál procedimiento mecánico se seleccionan las muestras al azar? Se supone a veces erróneamente que casi todo método de selección de "cara o cruz" dará una muestra al azar. Esto dista mucho de ser así. En efecto, tales métodos conducen casi invariablemente a una muestra sesgada, debido al elemento humano implicado. Con objeto de asegurarnos que todos los individuos, incluidos los atípicos o los que son difíciles de localizar, tienen efectivamente la misma posibilidad de aparecer en la muestra, hemos de observar por lo regular muchas condiciones al proceder a la extracción. Primero, en efecto, hemos de asegurarnos de que cada individuo figura en la lista y de que aparece únicamente en ella una sola vez. Podemos luego asociar un número a cada puesto de la lista y servirnos de algún procedimiento mecánico, por el estilo del que se emplea en el juego de *bingo*, con objeto de asegurar probabilidades iguales de selección. Examinemos primero algunos problemas que pueden presentarse en relación con la misma confección de la lista, o lo que los especialistas en muestreo denominan "armazón de la muestra".

Podrá pensarse acaso que el hecho de obtener una lista es por lo regular asunto de poca monta. Sin embargo, en la mayoría de los casos prácticos esto no es así. A menudo ni siquiera existen listas. Por ejemplo: no existe lista alguna de los residentes de los Estados Unidos o del estado de Michigan. Es casi seguro que tampoco habrá lista alguna de los negros o los niponorte-americanos que viven en una localidad determinada. Y si no hay lista, puede resultar muy costoso confeccionar una. Si tal es el caso, hay otros métodos de muestreo de probabilidad que son preferibles a la muestra simple aleatoria. Por otra parte, puede acaso

haber listas, pero es posible que no estén al día. Algunos individuos podrán no estar incluidos, mientras que otros habrán dejado entretanto de pertenecer a la población considerada. Los directorios locales, que a primera vista parecen constituir la fuente ideal para aquel que desea extraer una muestra aleatoria de los residentes, pueden acaso resultar tan anticuados en el momento de su publicación, que ya prácticamente no sirvan. Los individuos acabados de llegar estarán excluidos de la lista y no tendrán, por consiguiente, probabilidad alguna de ser seleccionados. Y en la medida en que dichas personas difieran del resto de la población en cuanto a las características objeto del estudio, el investigador obtendrá una muestra sesgada y resultados engañosos. Otras listas, tales como los directorios telefónicos o los registros de los vehículos o automóviles, pueden estar sesgadas en el sentido de que los grupos de ingresos inferiores estarán probablemente subrepresentados. Cabe decir, por lo tanto, que por mucho que una lista parezca estar cuidadosamente confeccionada, deberemos investigar siempre hasta qué punto resulta apropiada. Una lista deficiente puede resultar peor que la falta total de la misma, si conduce a una muestra excepcionalmente sesgada.

¿Qué podemos hacer si la lista es inadecuada? Si la lista es completa pero contiene duplicaciones, el problema es relativamente sencillo, a condición, por supuesto, que las duplicaciones se puedan descubrir fácilmente. Por ejemplo: si la lista comprende a todos los niños de una escuela determinada y queremos seleccionar una muestra aleatoria de los *padres*, descubriremos, sin duda alguna, que ciertos padres tienen más de un niño que va a dicha escuela. Por consiguiente, si damos a la ficha de cada *niño* la misma probabilidad de selección, algunos padres tendrán mayores probabilidades de ser seleccionados que otros. Con objeto de remediar esta situación, podríamos descartar las fichas de todos los hermanos de padre y madre menos una, o podríamos seleccionar un padre solamente en el caso de que fuera seleccionada la ficha de su hijo mayor, descartándolo, en cambio, si salía la ficha de alguno de sus otros hijos.

Debe observarse que si se seleccionara el segundo o el tercer hijo de Jones y no incluyéramos, por consiguiente, a Jones en la muestra, no sería legítimo remplazar a Jones por el padre que figurara a continuación en la lista. En efecto, si se hiciera así, las personas de fichas vecinas a las de los padres de más de un niño tendrían mayores probabilidades de selección. El procedimiento correcto consistirá, en tal caso, en prescindir de Jones y pasar a la próxima ficha seleccionada por métodos de probabilidad. Otra alternativa, posible teóricamente pero susceptible de crear problemas adicionales para el análisis, consistiría en incluir a Jones si salía la ficha de cualquiera de sus hijos, pero

atribuyéndole menor peso en el análisis. Así, por ejemplo, si tiene tres hijos y, por lo tanto, tres veces la probabilidad general de ser seleccionado, daríase a sus marcas una tercera parte del peso atribuido al padre de un solo hijo.

En la mayoría de los problemas, sin embargo, lo más probable es que la lista sea incompleta o que incluya nombres de individuos que ya no son miembros de la población. Aquí será también posible depurar la lista hasta que sea correcta. Pero si esto no es practicable, podrá resultar deseable redefinir la población ligeramente, para adaptarla a la lista. Supóngase que se sabe que una lista de empleados es completa y exacta a la fecha del día primero del año. En lugar de obtener los nombres de todas las personas empleadas desde entonces, será tal vez posible limitar nuestra atención a las personas que trabajaban en la empresa con anterioridad a la fecha en cuestión y que siguen trabajando en ella. Luego, todas las personas incluidas en la muestra pero que resultan haber dejado entretanto la empresa podrán descartarse. Obsérvese, sin embargo, que la población estudiada no constará de *todos* los empleados presentes, y el lector ha de percartarse bien de ello.

Una vez obtenida una lista correcta, es relativamente sencillo extraer una muestra aleatoria. Teóricamente podría emplearse toda una serie de medios mecánicos para asegurar probabilidades iguales de selección. Podría utilizarse un juego de naipes bien barajado, o extraerse números de un sombrero. Tal vez una esfera con bolas numeradas daría resultados más seguros, debido a la tendencia de las cartas o los pedacitos de papel a pegarse cuando se los baraja o mezcla. En realidad, sin embargo, el investigador no necesita seguir un proceso tan complicado, ya que se han confeccionado con tal objeto tablas de números aleatorios. Estas tablas se han confeccionado sirviéndose de medios mecánicos como los que se acaban de indicar. Así, por ejemplo, podría ponerse un número igual de bolas con los dígitos 0, 1, 2, ..., 9 en una cesta y proceder a extraerlas, reponiéndolas y mezclándolas cada vez a fondo. Los dígitos resultantes podrían luego servir para confeccionar un cuadro de números al azar, como el del cuadro B del apéndice 2.

Al servirnos de un cuadro de números aleatorios, no importa que sigamos las columnas de arriba abajo o que procedamos a través de las hileras, ni que empecemos con una de las columnas o hileras con preferencia a otra, a condición, sin embargo, que nuestra decisión se adopte antes de examinar los datos. Para ilustrar el empleo del cuadro de números aleatorios, supongamos que se quiere extraer una muestra de tamaño 100 de una población que consta de 736 individuos. Toda vez que el número 736 consta de tres dígitos, resultará conveniente escoger tres columnas adyacentes (cualesquiera), eligiendo otras tres al lle-

gar al pie de la página. Supóngase, por ejemplo, que decidimos servirnos de las tres primeras columnas de la primera página del cuadro B. Como primer caso de la muestra escogemos el primer número que aparece entre 001 y 736. Este número es 100. En otros términos: el centésimo individuo figurará en la muestra. Seguimos ahora las columnas 1 a 3 abajo, y obtenemos los números 375 y 084. Llegamos luego al número 990. Esto correspondería al individuo noningentésimo nonagésimo de la población, pero, como quiera que este individuo no existe, pasamos al próximo número, que es 128.

Después de un rato empezamos a encontrarnos con números que ya han sido seleccionados. Toda vez que estamos seleccionando sin reposición, hemos de omitir dichos números, hasta haber seleccionado finalmente los 100 casos. Esto es todo lo que hay que hacer. La razón de que el proceso sea tan sencillo y que pueda decidirse arbitrariamente el empleo de las columnas o hileras está, por supuesto, en que los números que figuran en el cuadro son totalmente aleatorios. De hecho, es casi imposible servirse de una de estas tablas incorrectamente, a menos que se repitan las columnas (o las hileras) o que se haga trampa, decidiendo, por ejemplo, que se quiere obtener en la muestra el caso ducentésimo decimonono y buscando deliberadamente una columna que lo contenga.

Corrección de la muestra sin reposición. Ya se mencionó en el capítulo IX, relativo a la probabilidad, que cuando se saca una muestra sin reposición violamos el supuesto de independencia y que, estrictamente hablando, hemos de modificar, por consiguiente, nuestras fórmulas para tener en cuenta dicho hecho. Por lo regular, esto no constituye problema grave alguno, ya que la muestra seleccionada no es más que un pequeño fragmento de la población y, por lo tanto, la probabilidad de que un individuo determinado resulte seleccionado dos o más veces es más bien pequeña. Sin embargo, si la muestra llega a comprender hasta un quinto de los individuos de la población, será conveniente introducir factores de corrección, siempre que tales factores sean conocidos. Por desgracia, sólo se conocen factores exactos de corrección para los tipos de problemas más sencillos. Con todo, este hecho sólo raramente resulta perturbador, ya que, si fuéramos a seleccionar una muestra que comportara el tercio o la mitad de la población, estaríamos de todos modos en condiciones de seleccionar también la población entera. El empleo del factor de corrección para fórmulas que comportan el error estándar de la media se examina más abajo. En casos más complicados, habremos de referirnos a algún texto especial sobre muestreo, aunque probablemente no se encuentre en los mismos un examen de los factores de corrección aplicado a las diversas pruebas no paramétricas. Por otra parte, dichos textos tienen su

mayor aplicación a las muestras pequeñas, en los que el problema de la reposición reviste menor importancia.

La fórmula que habremos de aplicar efectivamente para la corrección del error estándar de la media, si seleccionamos sin reposición, es la siguiente:

$$\sigma_{\bar{x}} = \sqrt{1-f} \frac{\sigma}{\sqrt{N}} \quad (\text{XXI.1})$$

en donde f representa la *fracción de muestreo*, o sea la razón del número de casos de la muestra con respecto de la población. Si designamos el tamaño de la muestra como N y el de la población como M , podemos escribir el factor de corrección como:

$$\sqrt{1 - \frac{N}{M}}$$

Se echa de ver inmediatamente que si el tamaño de la muestra es relativamente pequeña en comparación con M , el valor del factor de corrección se hace aproximadamente igual a la unidad, y tiene, por consiguiente, escaso objeto o nulo el servirse de él. Así, por ejemplo, si se selecciona una muestra de 500 de una población de 10 000, la fracción de muestreo es $1/20$, y el valor del factor de corrección es de .975. Obsérvese que, toda vez que el factor de corrección ha de ser menor que la unidad tratándose de poblaciones finitas, el valor corregido del error estándar será siempre menor que el de la cifra sin corregir. Así, pues, si deseamos un error estándar pequeño, como suele ser el caso, nos encontraremos del lado conservador no sirviéndonos de la corrección. A menos que la fracción de muestreo sea del orden de un quinto o más, raramente la tenemos en cuenta.

Este mismo factor de corrección puede emplearse en otras fórmulas que comportan errores estándar de medias o proporciones. Así, si hubiéramos de servirnos de una estimación, recurriríamos a la fórmula:

$$\hat{\sigma}_{\bar{x}} = \sqrt{1-f} \left(\frac{\hat{\sigma}}{\sqrt{N}} \right) = \sqrt{1-f} \left(\frac{s}{\sqrt{N-1}} \right) \quad (\text{XXI.2})$$

En una prueba de diferencia de medias, habrá dos fracciones de selección, y la fórmula básica de la estimación del error estándar de la diferencia de las medias será:

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{(1-f_1) \frac{\hat{\sigma}_1^2}{N_1} + (1-f_2) \frac{\hat{\sigma}_2^2}{N_2}} \quad (\text{XXI.3})$$

XXI.2. Muestreo sistemático

Otro tipo de muestreo de uso muy frecuente se confunde fácilmente con el de muestreo aleatorio y, de hecho, se emplea a menudo como intercambiable con éste. En el muestreo sistemático, en lugar de servirnos de un cuadro de números aleatorio, recorreremos simplemente una lista y tomamos cada k -ésimo individuo, empezando con un caso escogido aleatoriamente entre los primeros k individuos. Así, por ejemplo, si deseáramos seleccionar una muestra de 90 personas de entre una lista de 1 800, tomaríamos cada vigésima persona de la lista. Sin embargo, nuestra primera elección ha de determinarse por algún procedimiento al azar, como el empleo, por ejemplo, de un cuadro de números aleatorios. Supóngase que se eligiera el individuo undécimo. Entonces la muestra constaría de los individuos 11, 31, 51, 71, 91, ...

El muestreo sistemático es manifiestamente mucho más sencillo que el muestreo aleatorio, siempre que la lista sea sumamente larga o que haya que extraer una muestra muy grande. Si nos pudiéramos servir legítimamente, por ejemplo, de un directorio telefónico o del directorio de una ciudad, se concibe fácilmente la dificultad de buscar los individuos 512, 1 078 y 15 324. Si la ordenación empleada en la compilación de la lista puede considerarse esencialmente como al azar con respecto a la variable que se está midiendo, el muestreo sistemático será equivalente al muestreo sencillo aleatorio. Por ejemplo: la mayoría de las listas están confeccionadas por orden alfabético. Los apellidos, por supuesto, no son casuales. Así, pues, un marido y mujer registrados separadamente no tendrían prácticamente probabilidad alguna de figurar juntos en la muestra, a menos que su apellido fuera sumamente común. Algunos grupos étnicos tienen una proporción elevada de nombres que empiezan con la misma letra (O'Brien, O'Neil, etcétera). En realidad, en el caso de las listas alfabéticas tenemos algo que se aproxima al muestreo estratificado (véase más adelante), en el que los grupos étnicos presentan cierta tendencia a unirse. El hecho de tomar cada k -ésimo individuo tiene, por lo tanto, probabilidades de proporcionar una representación apropiada de cada grupo. En la práctica, sin embargo, como quiera que la ordenación alfabética es esencialmente irrelevante desde el punto de vista de la mayoría de las variables estudiadas, no solemos por lo regular correr riesgo alguno al considerar la muestra sistemática como equivalente al muestreo sencillo aleatorio. Sin embargo, se han desarrollado para el primero algunas fórmulas algo distintas, que parten de supuestos diferentes. En la mayoría de los casos, con todo, apenas valdrá la pena tomarse ese trabajo adicional.

Hay dos tipos de situaciones en las que la selección sistemática produce sesgos considerables. Afortunadamente, éstas no se pre-

sentan con frecuencia en los problemas sociológicos. *Primera:* los individuos pueden haberse ordenado de manera que se produzca una tendencia. En efecto, si las personas se han registrado por profesiones, prestigio, o edad, la posición de la salida al azar puede afectar los resultados. Supóngase, por ejemplo, que la fracción de muestreo sea de $1/30$. Dos personas pueden extraer muestras sistemáticas con partidas aleatorias muy diferentes. Una partida al azar de dos, por ejemplo, dará una marca promedio considerablemente más elevada (si los individuos están ordenados de mayor a menor) que la de 27, ya que cada individuo de la primera muestra estará situado 25 lugares adelante de la persona correspondiente de la segunda muestra. Si se observa un sesgo de esta clase, habrá que mezclar algo la lista, o servirse de una "partida media" (o sea, empezando con los individuos 15 o 16).

El *segundo* tipo de situación que hay que evitar es aquel en que la lista presenta cierta característica periódica o cíclica correspondiente a la fracción de muestreo. Así, por ejemplo, en un edificio o una casa de departamentos cada octavo de éstos forma esquina. Si éste es algo mayor que los restantes, cabe esperar que sus ocupantes difieran asimismo. Por lo tanto, si se da el caso de que la fracción de muestreo sea también de $1/8$, podría obtenerse un muestreo con todo de esquinas o, inversamente, sin ninguna, según el punto de partida aleatorio. Con objeto de evitar esta trampa, podría cambiarse ligeramente la fracción de muestreo, tomándola como $1/7$ o $1/9$, o cabría servirse de varios puntos de partida tomados al azar. Así, por ejemplo, una vez seleccionados diez departamentos, podría escogerse otro número aleatorio y seleccionar otras diez residencias, extraer otro número, y así sucesivamente.

La selección sistemática se emplea a menudo en combinación con otros procedimientos en los estudios sociales, debido a su sencillez. Al practicante inexperto de una encuesta, por ejemplo, es más fácil indicarle que visite cada tercera casa de una manzana que decirle que emplee un cuadro de números aleatorios. Sin embargo, lo mismo que en el caso del muestreo sencillo aleatorio, la lista ha de ser completa y precisa. Si el que realiza la encuesta omitiera los departamentos más pequeños o algunas residencias de las avenidas lejanas, podrían resultar de ello graves errores. Es importante percatarse de que en toda clase de muestreo de probabilidad han de darse tanto un elemento fortuito como algún tipo de registro completo. Sin embargo, según veremos en seguida, la naturaleza de las listas requeridas puede diferir de un propósito a otro, siendo algunas de ellas mucho más fáciles de obtener que otras. El investigador ha de examinar siempre su lista cuidadosamente y ha de saber cómo se ha confeccionado y cuáles son sus defectos.

XXI.3. Muestreo estratificado

En tanto que en términos de ahorro de costos o de problemas de análisis las diferencias entre las muestras aleatorias sencillas y las sistemáticas son por lo regular relativamente secundarias, los otros dos tipos básicos de muestreo, en cambio, difieren de los primeros que acabamos de ver en algunos aspectos fundamentales. Según veremos, tanto el muestreo estratificado como el conglomerado pueden emplearse en determinadas circunstancias para aumentar la eficacia del diseño de muestreo. En otros términos: pueden concebirse para conseguir mayor precisión con los mismos costos o bien, si se prefiere, cuestan menos y comportan la misma precisión. Se verá también que ambos procedimientos requieren fórmulas distintas de aquellas de las que nos hemos servido anteriormente.

En el muestreo estratificado dividimos primero todos los individuos en grupos o categorías y seleccionamos luego muestras independientes dentro de cada estrato. Es importante que los estratos se definan de tal modo que cada individuo figure en uno y sólo en uno de ellos. En los tipos de muestreo estratificado más sencillos y de empleo más frecuente, tomamos una muestra aleatoria sencilla o sistemática de cada uno de los estratos. Las fracciones de muestreo de los distintos estratos pueden ser iguales, en cuyo caso hablamos de un muestreo estratificado *proporcional*, o puede tratarse, por el contrario, de un muestreo estratificado *no proporcional*.

Una de las razones en cuya virtud estratificamos a menudo una muestra es la de que pueden haberse empleado métodos o listas de muestreo para cada estrato. Así, por ejemplo, los estratos pueden consistir en fábricas, escuelas o dormitorios distintos, cada uno de los cuales se ha estudiado en momentos distintos por distintas personas. Es posible que hubiera sido totalmente impracticable combinar las listas de los distintos estratos seleccionando luego una muestra sencilla aleatoria de todos ellos. Otra razón importante de la estratificación, frente a la muestra aleatoria sencilla, consiste en la reducción de los casos requeridos para la obtención de un determinado grado de precisión. En la medida en que los estratos son homogéneos con respecto a las variables estudiadas, podemos mejorar la eficacia del diseño. Al examinar las muestras estratificadas proporcionales y no proporcionales, apreciaremos mejor algunas de las ventajas particulares de este muestreo frente al muestreo sencillo aleatorio.

Muestreo estratificado proporcional. El muestreo estratificado proporcional se emplea a menudo para asegurarse una muestra más representativa de la que cabría esperar de las muestras aleatorias sencillas o sistemáticas. Supóngase, por ejemplo, que hay 600 protestantes, 300 católicos y 100 judíos en una población

determinada. Si se fuera a extraer una muestra aleatoria de tamaño 100, no esperaríamos ciertamente obtener exactamente 60 protestantes, 30 católicos y 10 judíos. La proporción de los judíos, en particular, podría resultar fácilmente o demasiado grande o demasiado pequeña. Supóngase ahora que nos interesaba estudiar alguna variable, tal como la asistencia a la iglesia, íntimamente ligada a la confesión. Supóngase, además, que nos interesaba estimar el número promedio de veces que las personas de la población asistían a la iglesia. Resulta fácil ver intuitivamente que una muestra estratificada con proporción al tamaño en el que las fracciones de muestreo fueran de 1/10 para los tres estratos (o sea que constaran de 60 protestantes, 30 católicos y 10 judíos) nos proporcionaría, por lo regular, resultados más seguros que la muestra sencilla aleatoria.

Tenemos aquí, en efecto, un problema análogo al del análisis de variancia. En la muestra aleatoria hay dos fuentes de variación. Puede haber errores de selección *dentro* de cada estrato, y puede haberlos *entre* los estratos en relación con los respectivos números seleccionados. No sólo podríamos seleccionar judíos o católicos muy atípicos, sino que podríamos seleccionar, además, demasiados o muy pocos de cada tipo. En el muestreo estratificado, en cambio, hemos eliminado la variación entre estratos y nos queda sólo la variación dentro. Si los estratos fueran totalmente homogéneos, el muestreo proporcional nos daría siempre resultados correctos, en tanto que no sería así con el muestreo sencillo aleatorio. Por otra parte, si los estratos fueran tan homogéneos como podría esperarse del azar, nada ganaríamos estratificando. En otros términos: si las diferencias entre los grupos son pequeñas en comparación con las diferencias dentro, la estratificación de nada sirve. Así, pues, la ventaja resultante de estratificar es proporcional en líneas generales a la correlación de intraclass entre las dos variables. Por consiguiente, si el criterio en favor de la estratificación se relaciona muy íntimamente con la variable estudiada, la ventaja puede ser acaso considerable. Al ganar el control sobre el número de casos de cada estrato, cosa que no era posible en el muestreo aleatorio, podemos asegurarnos mayor precisión en relación con un tamaño determinado de la muestra.

No debe por ello esperarse demasiado del muestreo estratificado proporcional. Si el tamaño de la muestra es relativamente grande, esperamos, por supuesto, que el solo factor azar nos asegure aproximadamente proporciones correctas de cada estrato. Y como quiera que los problemas de análisis no se complican demasiado a consecuencia de la muestra estratificada, poco perdemos en realidad al estratificar. Por lo regular, no es ni necesario ni practicable esforzarse en obtener un solo criterio "mejor" para estratificar. Para obtener una muestra estratificada pro-

porcional, hay que conocer los tamaños de los estratos de población, y sólo será posible, por supuesto, estratificar conforme a variables a cuyo propósito la información se desprende de las listas en el momento de la extracción de la muestra. Esto significa a menudo que nos vemos limitados a variables tan sencillas como el sexo, la edad, la ocupación o el área de residencia. Algunas de estas variables pueden incluso utilizarse combinadas, si se desea, si bien rara vez resultará ventajoso estratificar con más de dos o tres variables a un tiempo. Sin embargo, como quiera que la estratificación constituye un procedimiento tan sencillo, sus posibilidades deberían examinarse siempre.

Muestreo estratificado no proporcional. En el muestreo estratificado no proporcional nos servimos de distintas fracciones de muestreo para manipular el número de casos seleccionado, con objeto de aumentar todavía más la eficacia del diseño. Hay diversos tipos de situaciones en los que esta forma de muestreo resulta indicada. A menudo, en efecto, nuestro interés puede centrarse más en las diversas subpoblaciones representadas por los estratos que en la población conjunta misma. Supóngase, por ejemplo, que deseáramos *comparar* los tres grupos religiosos principales en relación con la asistencia a la iglesia. Es obvio que tanto el muestreo sencillo aleatorio como el estratificado proporcional nos darían demasiados pocos judíos en la muestra para poder establecer comparaciones significativas. Por consiguiente, podríamos acaso decidir seleccionar números iguales de cada grupo, dando así a cada judío una probabilidad de selección igual a tres veces la de los católicos y seis veces la de los protestantes. Si seleccionáramos 50 de cada grupo, las fracciones de selección respectivas serían así de 1/12, 1/6 y 1/2. Y si luego quisiéramos generalizar a la población entera con objeto de apreciar la cifra media de la asistencia, habríamos de ponderar las medias de los tres estratos, a fin de compensar el hecho de que los judíos han sido sobreseleccionados. Este procedimiento de ponderación se describe más adelante.

Pero incluso si nuestro objetivo está en generalizar a la población entera y no en comparar diversas subpoblaciones, aun así puede resultar indicado servirnos de la muestra estratificada no proporcional, siempre que: 1) las desviaciones estándar dentro de los distintos estratos difieran considerablemente entre sí, o 2) que los costos de reunir los datos varíen sustancialmente de un estrato a otro. Habrá siempre una *distribución óptima* en relación con la cual el propósito del muestreo presentará una eficacia máxima. En otros términos: habrá una determinada combinación de fracciones de muestreo que proporcionará el menor error de muestreo al menor costo posible. Y podemos obtener esta distribución óptima si *hacemos la fracción de muestreo de cada estrato directamente proporcional a la desviación*

estándar dentro del estrato e inversamente proporcional a la raíz cuadrada del costo de cada caso dentro de su estrato. Veamos intuitivamente por qué esto es así, examinando primero la cuestión de las desviaciones estándar.

Si un determinado estrato particular es excepcionalmente homogéneo con respecto a la variable estudiada, no será necesario extraer de la misma una muestra muy grande para conseguir un grado determinado de precisión. Por otra parte, será indicado tomar una muestra mucho mayor de un estrato muy heterogéneo. Como quiera que nuestra precisión conjunta vendrá determinada ante todo por el grado de precisión del eslabón más débil de la cadena, por así decir, importa que no tengamos uno o dos estratos con errores grandes de selección. Esto es particularmente así si los estratos suelen ser grandes. No tendría objeto, en efecto, sostener una precisión perfecta en algunos de los estratos más pequeños, con un error de muestreo muy grande, en cambio, en otro estrato. Por consiguiente, si tomamos relativamente más casos de los estratos heterogéneos y menos, en cambio, de los homogéneos, podemos salir del paso con menos casos. Según se demuestra matemáticamente, las fracciones de selección deseadas son proporcionales a las desviaciones estándar respectivas, y no a las variancias.

Conviene hacer aquí una advertencia. En efecto, un determinado estrato podrá ser acaso muy homogéneo en relación con una de las variables estudiadas y muy heterogéneo, en cambio, en relación con otra. Toda vez que los proyectos de investigación comportan por lo regular más de una variable, puede acaso resultar muy difícil encontrar distribuciones que sean óptimas, o aproximadamente tales, para más de una variable a la vez. Y de hecho, un diseño muy eficaz en relación con una variable puede acaso ser sumamente ineficaz en relación con otra. Por consiguiente, lo mejor será consultar un especialista en materia de muestreo, percatándose bien de cuáles son las variables importantes, antes de servirse de la distribución no proporcional. En caso de duda, la estimación proporcional será mucho más segura.

Hasta aquí las consideraciones relativas a los costos no se han tenido en cuenta, debido al hecho de que hemos venido suponiendo implícitamente que los costos de la reunión de datos eran iguales para todos los individuos. Supóngase, sin embargo, que esto no sea así, y que algunos estratos comporten costos más elevados que otros. Diferentes administradores, por ejemplo, pueden permitir acaso el empleo de diversas técnicas de recopilación de datos, o tal vez las condiciones materiales de los diversos estratos sean tales que la encuesta tome más tiempo en uno de ellos que en los demás. En igualdad de los demás factores, será obviamente menos costoso seleccionar un número relativamente mayor de casos de los estratos más baratos. Puede de-

mostrarse matemáticamente que la distribución óptima se obtendrá si las fracciones de muestreo se toman inversamente proporcionales a la raíz cuadrada de los factores de los costos.

Obsérvese que en el caso especial en que todos los costos sean iguales y en que todas las desviaciones estándar dentro de los estratos sean asimismo iguales, las fracciones de muestreo serán asimismo iguales, y tenemos así la situación en que la estratificación proporcional nos dé la distribución óptima. En general, suele ser indicado seguir la regla de servirse de la estratificación proporcional, a menos que las diferencias de costos sean muy grandes, o a menos que las desviaciones estándar de los estratos sean sustancialmente diferentes. Según veremos más adelante, el empleo del muestreo no proporcional tiende a complicar los problemas del análisis y debería, por consiguiente, descartarse, a menos que presentara ventajas realmente muy claras.

Hasta aquí todavía no nos hemos enfrentado a una cuestión importante. En efecto, ¿cómo podemos servirnos de cálculos de costos y de las desviaciones estándar relativas, siendo así que estos elementos no se conocen todavía en el momento de extraer la muestra? La respuesta obvia es que han de apreciarse, lo mismo que hemos de efectuar anticipaciones lógicas en relación con los valores de determinados parámetros antes de apreciar el tamaño de la muestra que necesitaremos. Hemos de tener presente, sin embargo, que la clase de estimaciones que necesitamos no es del tipo de las que formulamos a partir de las estadísticas de las muestras. Sin duda, sería posible efectuar un estudio de ensayo con objeto de obtener dichas estimaciones, pero, a menos que el estudio haya de ser sumamente vasto y costoso, semejante gasto de dinero no será probablemente conveniente. Por lo tanto, nuestras estimaciones han de basarse en la experiencia de los peritos o en estudios anteriores. Con todo, la situación no es tan difícil como parece, ya que resulta por lo regular posible obtener aproximaciones muy satisfactorias de la distribución óptima mediante anticipaciones muy generales en cuanto a los costos y las desviaciones estándar. En otros términos: si existe alguna razón para sospechar que se dan diferencias sensibles entre los estratos en relación con el uno o el otro de los factores en cuestión, una anticipación inteligente nos dará probablemente un diseño casi tan eficaz como el que se obtendría con valores exactos.

Cálculos relativos a las muestras estratificadas. Cuando calculamos estimaciones de medias y estimamos errores estándar a partir de muestras estratificadas, hemos de calcular valores separados para cada uno de los estratos y ponderarlos luego de acuerdo con el tamaño relativo del estrato en la población. Si indicamos con W_i (*weight* = peso) el peso del i -ésimo estrato de la población y ponemos $\sum W_i = 1$, reduciendo así los pesos a pro-

porciones, podemos escribir la fórmula para la estimación de la media de la población como sigue:

$$\bar{X} = \sum_{i=1}^k W_i \bar{X}_i$$

en donde las \bar{X}_i representan las medias de cada uno de los k estratos. Esta fórmula es tal como la esperaríamos. Dice simplemente que si un estrato es tres veces mayor que otro, su media habrá de recibir un peso tres veces mayor.

Si se ha empleado el muestreo estratificado *proporcional* y dejamos que N_i y M_i indiquen respectivamente los tamaños de la muestra y de la población en relación con el estrato i -ésimo, entonces, por definición, todos los N_i/M_i serán iguales a N/M . Pero, como quiera, que, para el estrato i -ésimo.

$$\bar{X}_i = \frac{\sum_{j=1}^{N_i} X_{ij}}{N_i}$$

y también

$$W_i = \frac{M_i}{M} = \frac{N_i}{N}$$

tenemos:

$$\bar{X} = \sum_{i=1}^k \frac{N_i}{N} \frac{\sum_{j=1}^{N_i} X_{ij}}{N_i} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{N_i} X_{ij}$$

Esta doble suma total significa simplemente que hemos sumado todas las X . Y toda vez que luego dividimos esta suma entre el número total de casos para obtener \bar{X} , vemos así que, en el caso de muestra estratificada proporcional, podríamos haber obtenido la estimación de μ exactamente en la misma forma que en el caso de la muestra aleatoria simple. Por esta razón designamos la estratificación proporcional como *autoponderada*. En otros términos: cada estrato ha recibido su propio peso. Y si la estratificación no ha sido proporcional, entonces hemos de multiplicar cada \bar{X}_i por el peso de dicho estrato en la población.

Al apreciar el error estándar de la media, nuestros cálculos no resultan tan sencillos. En efecto, hemos de apreciar primero el error estándar para cada estrato y juntar luego los resultados, como lo hicimos en la prueba de la diferencia de las medias y en el análisis de variancia. Se recordará que, en lugar de sumar desviaciones estándar, operamos con las variancias y las sumas

de cuadrados. Hemos también de llevar al cuadrado los pesos W_i . La fórmula de la *variancia* anticipada de la media puede, pues, escribirse, en el caso de la muestra estratificada, como:

$$\hat{\sigma}_{\bar{X}}^2 = \sum W_i^2 \hat{\sigma}_{\bar{X}_i}^2$$

en donde $\hat{\sigma}_{\bar{X}_i}^2$ indica una estimación de la variancia de la media dentro del estrato i -ésimo. Podemos obtener el error estándar

CUADRO XXI.1. Datos para calcular estimaciones de parámetros de muestras estratificadas

	Distrito			Total
	1	2	3	
Tamaño del distrito (M_i)	10 000	15 000	25 000	50 000 (= M)
Peso (W_i)	.20	.30	.50	1.00
Tamaño de la muestra (N_i)	50	50	50	150 (= N)
Media de la muestra (\bar{X}_i)	3 100	4 300	3 800	
Desviación estándar de la muestra (s_i)	500	400	300	

anticipado de la media extrayendo la raíz cuadrada de la expresión anterior y calculando luego la estadística t como se hizo antes.

Supóngase, por ejemplo, que hay tres distritos y que los datos de los mismos pueden resumirse como en el cuadro XXI.1. Obsérvese que hemos obtenido una muestra no proporcionada, ya que nos hemos servido de fracciones de muestreo desiguales. Supongamos que nos servimos del muestreo simple aleatorio dentro de cada estrato y que las muestras se extrajeron independientemente. Los errores estándar anticipados, prescindiendo del factor $1 - f$, son

Distrito I: $\frac{s_1}{\sqrt{N_1 - 1}} = \frac{500}{\sqrt{49}} = 71.4$

Distrito II: $\frac{s_2}{\sqrt{N_2 - 1}} = \frac{400}{\sqrt{49}} = 57.1$

Distrito III: $\frac{s_3}{\sqrt{N_3 - 1}} = \frac{300}{\sqrt{49}} = 42.9$

La media y la variancia anticipadas serán por consiguiente:

$$\bar{X} = .20(3\ 100) + .30(4\ 300) + .50(3\ 800) = 3\ 810$$

$$y \quad \hat{\sigma}_{\bar{X}}^2 = (.20)^2(71.4)^2 + (.30)^2(57.1)^2 + (.50)^2(42.9)^2 \\ = 957.5$$

Si bien los cálculos de las medias y las proporciones son sencillos en el caso de las muestras estratificadas, ha de reconocerse, con todo, que no se pueden emplear legítimamente las diversas pruebas no paramétricas, las pruebas para la significación de la correlación, el análisis de covariancia, etcétera, sin una modificación sustancial. Por desgracia, no suelen encontrarse estudios de estos problemas en los textos sobre muestreo. Sabemos cómo tratar problemas estadísticos complicados si podemos suponer la forma más sencilla de muestra: el muestreo aleatorio. En el caso de diseños más complicados, podemos tratar los más sencillos de los problemas estadísticos, tales como estimación de medias o proporciones, cálculo de intervalos de confianza para medias y proporciones, pruebas de diferencias de medias, etcétera. Pero existe un vacío, con todo, cuando se llega a técnicas estadísticas más complejas, de diseños de muestreo más complicados.

XXI.4. Muestreo por conglomerados

En el muestreo estratificado dividimos la población en grupos que llamamos estratos y seleccionamos de entre *cada* estrato. En ocasiones resulta ventajoso dividir la población en un gran número de porciones o conjuntos llamados conglomerados y seleccionar *entre* éstos. Así, por ejemplo, podríamos acaso dividir una ciudad en varios centenares de distritos electorales y seleccionar luego 40 distritos como muestra. Semejante diseño de muestreo se designa como muestreo conglomerado y se emplea frecuentemente en los estudios sociales, con objeto de reducir los costos inherentes a la recopilación de datos. Según veremos en seguida, el objeto del muestreo conglomerado consiste en seleccionar grupos lo más *heterogéneos* posible y lo suficientemente pequeños para reducir los costos, tales como gastos de viaje, etcétera, inherentes a la encuesta.

En la muestra conglomerada no seleccionamos nuestros elementos directamente. En lugar de ellos seleccionamos grupos o conjuntos de elementos. En el diseño de conglomerados más sencillos podríamos servirnos del muestreo aleatorio entre grupos, seleccionando luego cada individuo de los grupos incluidos en la muestra de éstos. Semejante diseño se designa a menudo como diseño de conglomerados de fase única, ya que en el proceso la selección tiene lugar una sola vez. En la selección de fa-

ses múltiples, por otra parte, el diseño puede ser mucho más complicado. Podríamos tomar primero una muestra simple aleatoria de distritos electorales de la ciudad. Y luego, podríamos tomar una muestra sencilla aleatoria de manzanas censales (aglomeraciones más pequeñas) dentro de cada distrito. Finalmente, podrían darse instrucciones al que realizara la encuesta, en el sentido de que visite cada tercera vivienda dentro de las manzanas incluidas e interrogue a cada segundo adulto dentro de aquéllas. De este modo, los procedimientos de muestreo pueden introducir el proceso de selección en ciertos números de puntos distintos. Por supuesto, es esencial en el muestreo probabilista que se dé en el procedimiento algún elemento al azar. Pueden calcularse fracciones de muestreo que produzcan muestras no sesgadas, de modo que cada individuo de la población tenga una probabilidad igual de figurar en la muestra. Sin embargo, con este procedimiento no será posible asegurar una selección independiente. En efecto, las personas del mismo conglomerado tendrán por lo regular más probabilidades de aparecer juntas en la misma muestra que los miembros de conglomerados diferentes. Y de hecho, el objetivo global del muestreo por conglomerados consiste precisamente en asegurar que esto ocurra.

Será instructivo comparar el muestreo conglomerado tanto con el sencillo aleatorio como con el estratificado. Para simplificar la cuestión, supongamos que nos servimos de un diseño de muestra por conglomerado de fase única en el que los conglomerados se seleccionan aleatoriamente, incluyendo luego cada individuo de los conglomerados seleccionados en la muestra total. ¿En qué difiere el muestreo por conglomerado del estratificado? Obsérvese que si bien ambos comportan la división de la población en grupos, implican con todo, en cierto sentido, operaciones de selección opuestas. En efecto, en la muestra estratificada *seleccionamos individuos* dentro de cada estrato. Estamos, por lo tanto, seguros de que cada estrato está representado por cierto número de casos. Nuestros errores de selección comportan en este caso variabilidad *dentro* de los estratos. Queremos, por consiguiente, que éstos sean en sí mismos lo más homogéneos posible y lo más diferentes posible unos de otros.

En el muestreo por conglomerado (de fase única), en cambio, no tenemos fuente alguna de error dentro del conglomerado, porque nos servimos en cada caso del mismo. Toda vez que sólo tomamos una muestra de conglomerados, nuestro error comporta variabilidad *entre* los conglomerados. Si las medias de los conglomerados difieren considerablemente en comparación con la variabilidad dentro de los mismos, corremos el riesgo de obtener un conglomerado muy poco usual en nuestra muestra de conglomerados. Si esto ocurriera efectivamente y si los conglomerados fueran homogéneos, nuestro error de muestra podría

ser considerable. Pero si los conglomerados son heterogéneos en sí mismos en comparación con las diferencias entre ellos, podemos salir adelante con pocos conglomerados relativamente grandes. Supóngase, en el caso extremo, que cada conglomerado fuera heterogéneo y que, en contraste con ello, las diferencias entre las medias de los conglomerados fueran insignificantes. En tal caso podríamos seleccionar simplemente un conglomerado muy grande y obtener una excelente muestra. En cambio, si los conglomerados fueran totalmente homogéneos, sólo necesitaríamos un caso en cada uno de ellos. Tratamos, pues, en esta forma, de obtener estratos homogéneos, pero en cambio, conglomerados heterogéneos, siendo que la razón de la variancia en la estrategia es la diferencia en cuanto al punto de extracción de la muestra.

Comparemos ahora el muestreo por conglomerados con el muestreo simple aleatorio. En casi todos los ejemplos que encontraremos, las muestras de conglomerados serán *menos eficaces* (o sea que producirán mayores errores de selección) que las muestras aleatorias sencillas *de igual tamaño*. Sin embargo, según veremos dentro de poco, es posible que *cueste* bastante menos obtener muestras de conglomerados. Nuestro problema será esencialmente el de equilibrar los costos y la eficiencia. ¿Cómo comparamos, pues, la eficiencia relativa de dos diseños? Ésta se mide de la manera más adecuada en términos del tamaño del error estándar de la estimación; un error pequeño indicando una eficiencia alta. Según vimos, es deseable obtener conglomerados que sean lo más heterogéneos posible. Esta noción intuitiva puede traducirse en una fórmula que comporte el coeficiente de correlación intraclase. Puede demostrarse que la razón de las variancias de las estimaciones de μ para las muestras por conglomerados y los muestreos aleatorios es aproximadamente:

$$\frac{\sigma_{\bar{X}_o}^2}{\sigma_{\bar{X}_R}^2} = 1 + \rho_i(\bar{N} - 1)$$

en donde $\sigma_{\bar{X}_o}^2$ y $\sigma_{\bar{X}_R}^2$ representan respectivamente las variancias de las medias de las muestras de conglomerado y sencilla aleatoria, ρ_i representa la correlación intraclase de la población, y \bar{N} es el número medio de casos en cada uno de los conglomerados.

Obsérvese que la razón de las variancias será por lo regular mayor que la unidad, lo que indica variancias mayores (y de aquí también mayores errores estándar) para el muestreo por conglomerados. La expresión será mayor que la unidad, a menos que $\bar{N} = 1$, o $\rho_i \leq 0$. Es obvio que, si $\bar{N} = 1$, el muestreo de conglomerados se reduce al caso especial de la muestra aleatoria, ya que cada conglomerado consta de un solo caso. La co-

relación de intraclase es, por supuesto, una medida de homogeneidad. Si el conglomerado es más homogéneo de lo que podría esperarse al azar, ρ_i será mayor que cero y, cuanto más homogéneo sea el conglomerado, tanto mayor será el valor de ρ_i . Se concibe que ρ_i sea negativo. Pero esto requeriría que el conglomerado fuera más homogéneo de lo que se esperaría por azar. En conjunto, las clases de conglomerados que solemos por lo regular escoger con fines prácticos serán casi siempre tan homogéneos por lo menos como se esperaría por azar.

Vemos que si $\rho_i > 0$, cuanto mayor es el número de casos \bar{N} del conglomerado, tanto mayor es la razón de las variancias y por consiguiente, tanto menor la eficiencia relativa del diseño de conglomerados. Esto puede verse intuitivamente. En efecto, si un grupo es perfectamente homogéneo, no necesitamos muchos casos para obtener una estimación precisa de su media. Podríamos en tal caso tomar una muestra muy pequeña del conglomerado, destinando el dinero ahorrado al estudio de conglomerados adicionales. Son dos factores, pues, los que determinan la eficiencia relativa del diseño de conglomerados, a saber: el grado de homogeneidad dentro del conglomerado y el tamaño del conglomerado mismo. Deseamos seleccionar de los conglomerados homogéneos sólo unos cuantos casos; si son heterogéneos, podemos tomar más casos de cada conglomerado, sin perjuicio grave de la eficiencia.

Como ya lo hemos indicado, la selección de conglomerados resulta por lo regular más económica que el muestreo aleatorio. Supóngase, por ejemplo, que se trataba de obtener una muestra a escala nacional para estudiar las preferencias electorales o las tasas de fecundidad. En primer lugar, no se dispondrá de lista alguna de adultos, y el costo de confeccionar una resultaría prohibitivo. En cambio, se dispone de listas de distritos. Será ciertamente mucho menos costoso extraer una muestra aleatoria (o sistemática o estratificada) de los distritos y operar únicamente con los distritos efectivamente seleccionados. Es probable que incluso dentro de cada distrito una muestra aleatoria no sería conveniente. Existe todavía otro factor de ahorro manifiesto. Será sin duda mucho menos costoso mandar entrevistadores a 50 distritos, por ejemplo, que esparcirlos por todo el país. En una muestra sencilla aleatoria, tal vez sólo resultarían seleccionadas 10 personas en el estado de Montana. Con el muestreo por conglomerados, en cambio, es posible preparar eficazmente a entrevistadores locales, y cada uno de ellos puede conseguir un número relativamente grande de encuestas sin incurrir en gastos exorbitantes de viaje. Las muestras de conglomerados efectuadas a los niveles de los estados, los distritos o las ciudades reunirán todos ellos las mismas ventajas, aunque, sin duda, en menor grado.

Toda muestra general comporta cierto número de costos. Y son éstos, y no el número de casos, los que ponen límites al estudio. Hay ciertos costos fijos que son independientes del diseño de muestreo y del número de los casos seleccionados. En relación con nuestros fines, éstos pueden ignorarse, ya que pueden sustraerse simplemente del total de los fondos disponibles. Y hay otros costos, luego, que resultan de la confección efectiva de las listas de unidades a seleccionar. Como acabamos de ver, el muestreo por conglomerados reduce a menudo estos últimos costos considerablemente. Otros costos, todavía, son directamente proporcionales al número de casos definitivamente seleccionados. El salario pagado al entrevistador mientras habla al interrogado, los costos de clasificación de los datos y ciertos costos de cálculos, todos ellos corresponden a esta categoría.

Otros costos, en cambio, serán proporcionales al número de conglomerados seleccionados. La mayoría de los gastos de viaje, incluidas las llamadas telefónicas, son de este tipo. Resultará más económico mandar a un individuo a un determinado distrito por varios días, y luego a otro, que hacerlo viajar por todo el estado, con el único resultado, acaso, de hallar que las personas por interrogar no están en casa la primera vez que las visita. En términos generales, si los costos de viaje y demás que dependen del número de los conglomerados seleccionados son muy elevados en comparación con los que varían directamente con el número de los casos, el muestreo por conglomerados resultará más económico que el sencillo aleatorio. Así, por ejemplo, en la prospección de un área grande que comporte encuestas muy breves, el muestreo por conglomerados puede resultar indicado. En cambio, si las encuestas duran cada una varias horas, la muestra sencilla aleatoria podrá ser más apropiada, a condición que los costos de la confección de listas no sean prohibitivos.

Así, pues, al decidir el diseño a utilizar, hay que sopesar las consideraciones relativas a los costos con las relativas a la eficiencia del diseño. Y habría que servirse del método que dé un error estándar menor a un costo determinado. Toda vez que no es necesario tomar a cada individuo dentro de los conglomerados de la muestra, el muestreo de fase múltiple puede constituir un compromiso aceptable. Tenemos entonces el problema complicado de escoger un diseño óptimo, en el que hemos de decidir el número de fases en las que el muestreo vaya a ser usado, el número de conglomerados por usar y el número de casos por seleccionar dentro de cada conglomerado. El problema se complica además por el hecho de que la mayoría de los estudios comportarán indudablemente no una sola, sino cierto número de variables, no siendo además todos los conglomerados del mismo tamaño. Con objeto de descartar las dudas al respecto, siempre será prudente consultar a un especialista en

materia de muestreo antes de tomar una decisión en cuanto al diseño. En efecto, cuando se llega al análisis de los datos, un planeamiento cuidadoso puede traducirse no sólo en costos menores, sino que puede redundar además en un número menor de problemas.

Antes de terminar esta sección de muestreo por conglomerados, conviene una vez más hacer una advertencia. Las fórmulas expuestas en este texto no pueden utilizarse en el muestreo por conglomerados. Como ya se indicó, los errores introducidos por el hecho de servirse de fórmulas de muestreo sencillo aleatorio en relación con los datos reunidos de muestras de conglomerados pueden resultar muy graves. Estos errores, en efecto, no son del orden de magnitud de aquellos que se introducen sirviéndose de la tabla normal, por ejemplo, en lugar de la tabla t , sino que pueden ser mucho mayores. En lugar de tener significancia al nivel de .05, el verdadero nivel (obtenido por las fórmulas correctas de la selección de conglomerados) puede llegar a .50 (véase [3]). Si deseamos descartar la hipótesis nula, rara vez nos encontraremos del lado conservador, si es que llegamos a alguna sirviéndonos de las fórmulas de muestra aleatoria con datos agrupados. Se recordará, además, que los muestreos por conglomerados son menos eficientes que el muestreo sencillo aleatorio del mismo tamaño. Por consiguiente, las fórmulas del muestreo sencillo aleatorio *subestimarán* los verdaderos errores estándar. O dicho en otra forma: una muestra de conglomerados de un tamaño determinado puede ser el equivalente, en términos de eficiencia, de una muestra sencilla aleatoria mucho menor. Así, por ejemplo, una muestra de conglomerados de tamaño 800 puede equivaler en términos de eficiencia a un muestreo sencillo aleatorio de 500. Por consiguiente, si se emplean las fórmulas del muestreo sencillo aleatorio con una N de 800, tenemos más probabilidades de obtener significación que sirviéndonos de los procedimientos correctos.

Así, pues, hemos de proceder con la mayor cautela al analizar datos provenientes de muestras por conglomerados. No deberemos servirnos de estadísticas tales como la χ^2 -cuadrada, a menos que el especialista en materia de muestreo pueda ayudarnos a introducir los factores de corrección apropiados. El problema no es tan grave con las muestras estratificadas debido, si más no, a que las muestras estratificadas son más eficientes que las sencillas al azar. En efecto, un muestreo estratificado de un tamaño dado puede igualar en eficiencia una muestra aleatoria mayor, de modo que el investigador se encontrará siempre, con aquél, del lado conservador en cuanto a descartar la hipótesis nula. Con todo, esto no es siempre así, de modo que la cautela se impone en todos los casos.

XXI.5. *Muestreo sin probabilidad*

Veamos ahora brevemente algunas situaciones en las que se ha empleado el muestreo sin probabilidad. El mayor inconveniente de ésta está en que no obtenemos con ella una estimación válida de nuestros riesgos de error. Por lo tanto, la inducción estadística no está legitimada, y no debería utilizarse. Esto no significa, con todo, que el muestreo sin probabilidad no resulte apropiado alguna vez. En efecto, en los estudios de exploración, cuyo principal objetivo está en obtener nociones valiosas que puedan llevarnos en última instancia a hipótesis verificables, el muestreo de probabilidad puede o resultar demasiado caro o conducir a conocimientos más limitados. Así, por ejemplo, podemos tal vez querer interrogar a personas que estén en una posición particularmente favorable para proporcionar información. O podemos querer acaso interrogar casos extremos, susceptibles de procurarnos las diferencias más notables. Si hacemos esto, no tenemos derecho legítimo alguno, por supuesto, a verificar la significancia de las diferencias entre extremos, a menos que tratemos de generalizar a una población compuesta exclusivamente de tales personas. El hecho de que conozcamos indudablemente estudios en los que se han efectuado pruebas estadísticas de casos extremos de esta clase, no significa, con todo, que el procedimiento sea legítimo. Pero no puede negarse, sin embargo, que pueden obtenerse conocimientos útiles a partir de comparaciones de esta clase.

Se emplean en ocasiones los métodos sin probabilidad cuando el propósito está en formular generalizaciones acerca de una población muestreada. Tales métodos se sirven invariablemente ya sea del criterio del entrevistador en cuanto a los individuos a incluir, o permiten que un individuo de la muestra se seleccione aparte del estudio sobre alguna base no fortuita.

Los muestreos de cuota empleados a menudo en las encuestas de la opinión pública parecen ser similares, a primera vista, a los muestreos estratificados. Se dan al investigador determinadas "cuotas" que ha de llenar. Ha de tener tantas o cuantas mujeres de más de 40 años, tantas o cuantas personas con un ingreso menor de \$ 3 000, o cierto porcentaje de católicos. Pero se deja a su discreción *cuáles* mujeres de más de cuarenta años o cuáles católicos quiera interrogar. Y como quiera que es humano, es probable que seleccione aquellas personas que le resulte más cómodo visitar. Si va a sus casas, probablemente sólo seleccionará aquellas personas que se encuentran en ellas en aquel momento. Incluso si se da cuenta de semejante tendencia selectiva, le resultará difícil corregirla adecuadamente. Un entrevistador sumamente responsable podrá incluso sobreesleccionar acaso a personas que rara vez se encuentran en la casa,

o a individuos de las clases inferiores, a los que los demás entrevistadores pasarán a menudo por alto. Tal vez una persona bien entrenada llegue a hacerse muy experta en el empleo de su discreción. Pero será difícil, por no decir imposible, saberlo. Y si cualquier grupo sobreesleccionado o, respectivamente, subseleccionado suele presentar diferencias pronunciadas con respecto a otros en relación con la variable objeto del estudio, el muestreo podrá resultar gravemente sesgado. Y lo que es peor, no hay forma de apreciar exactamente cuán sesgado pueda estar.

Siempre que las listas sean incompletas o que deba considerarse un gran porcentaje de personas como no respondientes, tenemos de hecho otro ejemplo de muestreo carente de probabilidad. Si en el caso de un cuestionario remitido por correo recibimos un 50 por ciento de respuestas, podemos acaso introducir sesgos graves, debido al hecho de que las personas que no contestan pueden tal vez ser significativamente diferentes de las que devuelven el cuestionario. Así, pues, aunque inicialmente nos hayamos tomado la pena de obtener un muestreo de probabilidad, algunos individuos no tendrán en realidad oportunidad alguna de verse incluidos en la muestra definitiva, porque se han descartado ellos mismos negándose a contestar. De ahí que sea sumamente importante hacer seguir un cuestionario remitido por correo de una o varias tarjetas postales, con objeto de obtener un porcentaje mayor de respuestas. Y en forma análoga, el entrevistador ha de aprender a insistir y ha de esperar y hacer varias llamadas para conseguir un número de respuestas suficiente. Es obvio, por lo demás, que un sesgo sustancial no se deja compensar por medio de una muestra mayor.

XXI.6. *Errores no de muestreo y tamaño de la muestra*

Incluso si se ha puesto el mayor cuidado en concebir un estudio que reúna todos los requisitos de un buen muestreo, siempre se tendrán, con todo, algunos errores ajenos a ésta. La teoría de las probabilidades nos permite apreciar los riesgos de errores de selección, o sea de aquellos errores introducidos en virtud del hecho de que las muestras varían de una a otra. Los errores no de muestreo, en cambio, son errores de medición. En efecto, en un estudio que comporte una entrevista o un cuestionario, habrá siempre errores de respuesta. En algunos casos, tal como en la edad de las personas, por ejemplo, puede darse un conjunto de errores que conduzcan a un sesgo manifiesto. En otros ejemplos, en cambio, los errores de respuesta podrán deberse más o menos al azar. Y los propios sesgos del entrevistador pueden afectar sus resultados.

En este texto no podemos entrar a estudiar detalladamente las clases de posibles errores no de muestreo. Vale la pena, sin em-

bargo, mencionar un punto sumamente importante. No se gana nada en reducir los errores de muestreo por debajo de cierto nivel, en comparación con los errores no de muestreo. Si estos dos tipos de errores pueden suponerse independientes uno de otro, la situación se puede representar por medio de un diagrama, como el de la figura XXI.1. El error total es así una función de dos

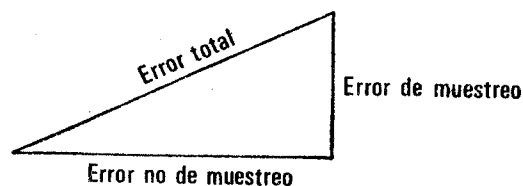


Fig. XXI.1. Relación entre el error total y los errores de muestreo y no de muestreo.

fuentes independientes de error, y no se puede reducir sustancialmente, a menos que se controlen simultáneamente los dos tipos. Si las equivocaciones ajenas al muestreo, tales como los errores de respuesta o de entrevista, son grandes, no tiene objeto tomar una muestra grande con el propósito de reducir el error estándar de la estimación, ya que el error total estará determinado en primer término por el largo de la base del triángulo. Y en forma análoga, si se desea hacer todo lo posible para reducir los errores no de muestreo a un mínimo, será contraproducente servirse de una muestra pequeña, con lo que se tendrá un error mayor de muestreo. Deberá, por consiguiente, mantenerse un equilibrio apropiado entre los errores de muestreo y los no de muestreo. El cuidado en la investigación limita el tamaño efectivo de la muestra y viceversa. Por desgracia, los errores no de muestreo son por lo regular difíciles de apreciar. Sin embargo, si los errores pueden ser apreciados, el diseño total más eficaz será aquel con respecto al cual los dos lados del triángulo sean iguales. Conviene tener presente este hecho.

GLOSARIO

Muestreo por conglomerados
Fracción de muestreo
Muestreo sencillo al azar
Muestreo estratificado
Muestreo sistemático

BIBLIOGRAFÍA

1. Cochran, W. G., *Sampling Techniques*, John Wiley & Sons, Inc., Nueva York, 1953.

2. Hansen, M. H., W. N. Hurwitz y W. G. Madow, *Sample Survey Methods and Theory*, vol. 1, John Wiley & Sons, Inc., Nueva York, 1953.
3. Kish, L., "Confidence Intervals for Clustered Samples", *American Sociological Review*, vol. XXII, pp. 154-165, abril de 1957.
4. Kish, L., "Selection of the Sample", en L. Festinger y D. Katz (eds.), *Research Methods in the Behavioral Sciences*, The Dryden Press, Inc., Nueva York, 1953, cap. 5.
5. Kish, L.: *Survey Sampling*, John Wiley & Sons, Inc., Nueva York, 1965.
6. Lazerwitz, Bernard: "Sampling Theory and Procedures", en H. M. Blalock y Ann B. Blalock (ed.), *Methodology in Social Research*, McGraw-Hill Book Company, Nueva York, 1968, cap. 8.

APÉNDICES

APÉNDICE 1

RESUMEN DE OPERACIONES ALGEBRAICAS

TODA vez que la mayoría de los estudiantes habrán olvidado mucho de lo que aprendieron en el álgebra elemental, un breve resumen de algunas de las operaciones algebraicas básicas les resultará sin duda útil. Algunas de estas reglas serán expuestas más adelante en forma muy concisa. Si se necesita un repaso más extenso, deberá consultarse un texto sobre álgebra.

Una de las cosas básicas que hay que recordar a propósito de las operaciones aritméticas y algebraicas es que el *orden* en que dichas operaciones se efectúan reviste suma importancia. En términos generales, en presencia de una expresión relativamente complicada se opera del interior al exterior. Convendrá retener más o menos las siguientes reglas.

1. *Desarrollo de una suma o una diferencia al cuadrado.*

$$(a + b)^2 = a^2 + 2ab + b^2 \neq a^2 + b^2$$

$$(a - b)^2 = a^2 - 2ab + b^2 \neq a^2 - b^2$$

Lo inverso se verifica al tratar con raíces cuadradas

$$\sqrt{a^2 + 2ab + b^2} = \sqrt{(a + b)^2} = a + b$$

Definitivamente *no es cierto* que

$$\sqrt{a^2 + b^2} = a + b$$

2. *División entre una suma o una diferencia.* Aun siendo cierto que

$$\frac{a + b}{c} = \frac{a}{c} + \frac{b}{c}$$

no podemos con todo simplificar tan fácilmente las expresiones

$$\frac{a}{b + c} \quad \text{o} \quad \frac{a}{b - c}$$

Así, por ejemplo:

$$\frac{a}{b + c} \neq \frac{a}{b} + \frac{a}{c}$$

3. *División entre una fracción.* Si el denominador es él mismo una fracción, podemos poner el denominador del denominador en el numerador como sigue:

$$\frac{a}{b/c} = a \frac{c}{b} = \frac{ac}{b}$$

Y en forma análoga: $\frac{a/b}{c/d} = \frac{a}{b} \frac{d}{c} = \frac{ad}{bc}$

$$y \quad \frac{a}{b/(c+d)} = a \frac{c+d}{b} = \frac{a(c+d)}{b}$$

4. *Multiplicación de potencias.* Si tenemos el producto de un número elevado a la potencia a y el mismo número a la potencia b , podemos *sumar* los exponentes. Así, por ejemplo:

$$X^a X^b = X^{a+b} \quad y \quad X^3 X^2 = X^5$$

Pero: $X^a + X^b \neq X^{a+b}$ y $X^3 + X^2 = X^2(X+1) \neq X^5$

Y en forma análoga, al dividir sustraemos los exponentes:

$$\frac{X^a}{X^b} = X^{a-b} \quad y \quad \frac{X^3}{X^2} = X^1 = X$$

En particular:

$$\frac{X^a}{X^a} = X^{a-a} = X^0 = 1$$

Así, pues, cualquier número real (excepto cero) elevado a la potencia 0 es igual a 1.

5. *Exponentes negativos.* Un número elevado a una potencia negativa puede escribirse como su número recíproco elevado a la potencia positiva. Por ejemplo:

$$X^{-a} = \frac{1}{X^a} \quad y \quad X^{-2} = \frac{1}{X^2}$$

6. *Supresión o adición de paréntesis.* Aquí seguimos la regla de proceder de dentro para afuera. Un signo negativo antepuesto a un paréntesis significa que cada término dentro del paréntesis ha de cambiar de signo al suprimirse el paréntesis. O sea:

$$a(b-c) = ab - ac$$

$$y \quad -[a - (b-c)] = -[a - b + c] = -a + b - c$$

$$y \quad a - [b - (c-d)^2] = a - [b - (c^2 - 2cd + d^2)] \\ = a - [b - c^2 + 2cd - d^2] \\ = a - b + c^2 - 2cd + d^2$$

Y en forma análoga, hemos de cambiar los signos de todas las cantidades que introduzcamos en un paréntesis si éste va precedido del signo negativo. Así, pues:

$$a - b - c = a - (b + c)$$

$$y \quad a - b + c - d = (a - b) + (c - d) = -(b - a) - (d - c)$$

Empleo de los signos de suma total. En estadística es necesario con frecuencia servirse de fórmulas que comportan sumas de numerosas cantidades. A título de sustituto taquigráfico de la plena escritura de cada una de dichas sumas nos servimos de la letra griega Σ (sigma mayúscula), que indica la suma total. A manera de regla general, siempre que dicho signo aparece significa que todas las cantidades que figuran a su derecha han de sumarse. En lugar de servimos de letras totalmente distintas para cada una de las cantidades a sumar (*v.gr.*, a, b, c, d, e, f, \dots), por lo regular hacemos uso de una sola letra (generalmente X, Y o Z), junto con un subíndice i, j o k , que puede tomar cualquier valor numérico que deseemos. Por lo regular, aunque no siempre, la primera marca se representará por medio del símbolo X_1 , la segunda por X_2 , y así sucesivamente. Nos servimos, pues, de Σ como sigue:

$$\sum_{i=1}^N X_i = X_1 + X_2 + X_3 + \dots + X_N$$

Las notaciones arriba y abajo de la Σ se emplean para indicar que i toma todos los valores sucesivos 1, 2, 3, etcétera hasta N . En forma análoga, podríamos escribir:

$$\sum_{i=3}^8 X_i = X_3 + X_4 + X_5 + X_6 + X_7 + X_8$$

En este último caso, los símbolos nos indican que hemos de adicionar las marcas de las observaciones tres a ocho.

Si seguimos las reglas generales del álgebra, podemos derivar ciertas reglas que han de aplicarse a las sumas totales. La ma-

yoría de estas reglas se enunciarán con poca o ninguna explicación, ya que resultan obviamente de la definición de Σ y de reglas muy sencillas del álgebra.

1.
$$\sum_{i=1}^N X_i^2 = X_1^2 + X_2^2 + X_3^2 + \cdots + X_N^2$$
2.
$$\sum_{i=1}^N X_i Y_i = X_1 Y_1 + X_2 Y_2 + X_3 Y_3 + \cdots + X_N Y_N$$
3.
$$\begin{aligned} \sum_{i=1}^N (X_i + Y_i) &= (X_1 + Y_1) + (X_2 + Y_2) + \cdots + (X_N + Y_N) \\ &= (X_1 + X_2 + \cdots + X_N) + (Y_1 + Y_2 + \cdots + Y_N) \\ &= \sum_{i=1}^N X_i + \sum_{i=1}^N Y_i \end{aligned}$$
4.
$$\sum_{i=1}^N (X_i - Y_i) = \sum_{i=1}^N X_i - \sum_{i=1}^N Y_i \quad (\text{véase 3})$$
5.
$$\begin{aligned} \sum_{i=1}^N (X_i + Y_i)^2 &= \sum_{i=1}^N (X_i^2 + 2X_i Y_i + Y_i^2) \\ &= \sum_{i=1}^N X_i^2 + \sum_{i=1}^N 2X_i Y_i + \sum_{i=1}^N Y_i^2 \\ &\neq \sum_{i=1}^N X_i^2 + \sum_{i=1}^N Y_i^2 \end{aligned}$$

Nota: El factor 2 puede ponerse delante del segundo término, lo que da: $2 \sum_{i=1}^N X_i Y_i$ (véase 6).

6. Si k es una constante:

$$\begin{aligned} \sum_{i=1}^N kX_i &= kX_1 + kX_2 + \cdots + kX_N \\ &= k(X_1 + X_2 + \cdots + X_N) = k \sum_{i=1}^N X_i \end{aligned}$$

$$\begin{aligned} 7. \left(\sum_{i=1}^N X_i \right)^2 &= (X_1 + X_2 + \cdots + X_N)^2 \\ &= X_1^2 + X_2^2 + \cdots + X_N^2 + 2X_1 X_2 \\ &\quad + 2X_1 X_3 + \cdots + 2X_{N-1} X_N \\ &\neq X_1^2 + X_2^2 + \cdots + X_N^2 \end{aligned}$$

En otros términos: hemos de distinguir entre

$$\sum_{i=1}^N X_i^2 \quad \text{y} \quad \left(\sum_{i=1}^N X_i \right)^2$$

En ocasiones podrá resultar asimismo conveniente expresar una suma en términos de una doble suma total sobre dos índices i y j . Cada cantidad a sumar puede escribirse con un doble subíndice (ij). La cantidad $\sum_{i=1}^N \sum_{j=1}^M X_{ij}$ significa que primero sumamos el segundo subíndice j de 1 a M , y luego, de dentro a fuera, sumamos i de 1 a N . Así, pues:

$$\begin{aligned} \sum_{i=1}^N \sum_{j=1}^M X_{ij} &= \sum_{i=1}^N (X_{i1} + X_{i2} + X_{i3} + \cdots + X_{iM}) \\ &= (X_{11} + X_{12} + \cdots + X_{1M}) + (X_{21} + X_{22} + \cdots \\ &\quad + X_{2M}) + \cdots + (X_{N1} + X_{N2} + \cdots + X_{NM}). \end{aligned}$$

Y en forma análoga:

$$\sum_{i=1}^N \left(\sum_{j=1}^M X_{ij} \right)^2 = \sum_{i=1}^N (X_{i1} + X_{i2} + \cdots + X_{iM})^2$$

Operaciones con números muy grandes o muy pequeños. Al operar con números muy grandes o muy pequeños, sobre todo al elevar al cuadrado o al extraer la raíz cuadrada, resulta a menudo conveniente servirse de las potencias de 10. Toda vez que $10^1 = 10$, $10^2 = 100$, $10^3 = 1\,000$, etcétera, contando el número de lugares ya sea a la derecha o a la izquierda del decimal podemos escribir cualquier cifra como número entre 0 y 10 multiplicado por una determinada potencia de 10. Así, por ejemplo:

$$13 = 1.3(10) = 1.3 \times 10^1$$

$$138 = 1.38(100) = 1.38 \times 10^2$$

$$1\,382 = 1.382(1\,000) = 1.382 \times 10^3$$

$$1\,382\,461 = 1.382461 \times 10^6$$

$$.13 = \frac{1.3}{10} = 1.3 \times 10^{-1}$$

$$.013 = \frac{1.3}{100} = \frac{1.3}{10^2} = 1.3 \times 10^{-2}$$

$$.00013 = \frac{1.3}{10\,000} = 1.3 \times 10^{-4}$$

Si queremos elevar al cuadrado la cantidad 1 382, tendremos:

$$\begin{aligned} 1\,382^2 &= (1.382 \times 10^3)^2 = 1.382^2 \times 10^6 \\ &= 1.9099 \times 1\,000\,000 = 1\,909\,900 \end{aligned}$$

Resulta así mucho más fácil colocar el punto decimal.

Al extraer la raíz cuadrada, lo más sencillo consiste en servirse de potencias pares de 10. Toda vez que $\sqrt{100} = \sqrt{10^2} = 10$ y $\sqrt{10\,000} = \sqrt{10^4} = 10^2 = 100$ y, en general,

$$\sqrt{10^{2k}} = 10^k$$

en tanto que

$$\sqrt{1\,000} = \sqrt{10^3} = 10\sqrt{10} \quad \text{y} \quad \sqrt{100\,000} = \sqrt{10^5} = 100\sqrt{10}$$

vemos que resulta siempre posible sacar potencias pares de diez fuera del radical, en tanto que esto no es así con las potencias impares. Por consiguiente, al extraer una raíz cuadrada, podemos contar el número de *pares* de dígitos a derecha o izquierda del decimal y expresar la cantidad original como número entre 1 y 100 multiplicado por una potencia par de 10.

$$13 = 1.3(10) = 1.3 \times 10^1$$

$$138 = 1.38(100) = 1.38 \times 10^2$$

$$1\,382 = 1.382(1\,000) = 1.382 \times 10^3$$

$$1\,382\,461 = 1.382461 \times 10^6$$

$$.13 = \frac{1.3}{10} = 1.3 \times 10^{-1}$$

$$.013 = \frac{1.3}{100} = \frac{1.3}{10^2} = 1.3 \times 10^{-2}$$

$$.00013 = \frac{1.3}{10\,000} = 1.3 \times 10^{-4}$$

CUADROS

CUADRO A. Cuadro de cuadrados y raíces cuadradas

Número	Cuadrado	Raíz Cuadrada	Número	Cuadrado	Raíz Cuadrada
1	1	1.0000	31	9 61	5.5678
2	4	1.4142	32	10 24	5.6569
3	9	1.7321	33	10 89	5.7446
4	16	2.0000	34	11 56	5.8310
5	25	2.2361	35	12 25	5.9161
6	36	2.4495	36	12 96	6.0000
7	49	2.6458	37	13 69	6.0828
8	64	2.8284	38	14 44	6.1644
9	81	3.0000	39	15 21	6.2450
10	1 00	3.1623	40	16 00	6.3246
11	1 21	3.3166	41	16 81	6.4031
12	1 44	3.4641	42	17 64	6.4807
13	1 69	3.6056	43	18 49	6.5574
14	1 96	3.7417	44	19 36	6.6332
15	2 25	3.8730	45	20 25	6.7082
16	2 56	4.0000	46	21 16	6.7823
17	2 89	4.1231	47	22 09	6.8557
18	3 24	4.2426	48	23 04	6.9282
19	3 61	4.3589	49	24 01	7.0000
20	4 00	4.4721	50	25 00	7.0711
21	4 41	4.5826	51	26 01	7.1414
22	4 84	4.6904	52	27 04	7.2111
23	5 29	4.7958	53	28 09	7.2801
24	5 76	4.8990	54	29 16	7.3485
25	6 25	5.0000	55	30 25	7.4162
26	6 76	5.0990	56	31 36	7.4833
27	7 29	5.1962	57	32 49	7.5498
28	7 84	5.2915	58	33 64	7.6158
29	8 41	5.3852	59	34 81	7.6811
30	9 00	5.4772	60	36 00	7.7460

FUENTE: H. Sorenson, *Statistics for Students Psychology and Education*, McGraw-Hill Book Company, Nueva York, 1936, cuadro 72, pp. 347-359, con la amable autorización del autor.

CUADROS

CUADRO A [continuación]

Número	Cuadrado	Raíz Cuadrada	Número	Cuadrado	Raíz Cuadrada
61	37 21	7.8102	101	1 02 01	10.0499
62	38 44	7.8740	102	1 04 04	10.0995
63	39 69	7.9373	103	1 06 09	10.1489
64	40 96	8.0000	104	1 08 16	10.1980
65	42 25	8.0623	105	1 10 25	10.2470
66	43 56	8.1240	106	1 12 36	10.2956
67	44 89	8.1854	107	1 14 49	10.3441
68	46 24	8.2462	108	1 16 64	10.3923
69	47 61	8.3066	109	1 18 81	10.4403
70	49 00	8.3666	110	1 21 00	10.4881
71	50 41	8.4261	111	1 23 21	10.5357
72	51 84	8.4853	112	1 25 44	10.5830
73	53 29	8.5440	113	1 27 69	10.6301
74	54 76	8.6023	114	1 29 96	10.6771
75	56 25	8.6603	115	1 32 25	10.7238
76	57 76	8.7178	116	1 34 56	10.7703
77	59 29	8.7750	117	1 36 89	10.8167
78	60 84	8.8318	118	1 39 24	10.8628
79	62 41	8.8882	119	1 41 61	10.9087
80	64 00	8.9443	120	1 44 00	10.9545
81	65 61	9.0000	121	1 46 41	11.0000
82	67 24	9.0554	122	1 48 84	11.0454
83	68 89	9.1104	123	1 51 29	11.0905
84	70 56	9.1652	124	1 53 76	11.1355
85	72 25	9.2195	125	1 56 25	11.1803
86	73 96	9.2736	126	1 58 76	11.2250
87	75 69	9.3274	127	1 61 29	11.2694
88	77 44	9.3808	128	1 63 84	11.3137
89	79 21	9.4340	129	1 66 41	11.3578
90	81 00	9.4868	130	1 69 00	11.4018
91	82 81	9.5394	131	1 71 61	11.4455
92	84 64	9.5917	132	1 74 24	11.4891
93	86 49	9.6437	133	1 76 89	11.5326
94	88 36	9.6954	134	1 79 56	11.5758
95	90 25	9.7468	135	1 82 25	11.6190
96	92 16	9.7980	136	1 84 96	11.6619
97	94 09	9.8489	137	1 87 69	11.7047
98	96 04	9.8995	138	1 90 44	11.7473
99	98 01	9.9499	139	1 93 21	11.7898
100	1 00 00	10.0000	140	1 96 00	11.8322

CUADROS

CUADRO A [continuación]

Número	Cuadrado	Raíz Cuadrada	Número	Cuadrado	Raíz Cuadrada
141	1 98 81	11.8743	181	3 27 61	13.4536
142	2 01 64	11.9164	182	3 31 24	13.4907
143	2 04 49	11.9583	183	3 34 89	13.5277
144	2 07 36	12.0000	184	3 38 56	13.5647
145	2 10 25	12.0416	185	3 42 25	13.6015
146	2 13 16	12.0830	186	3 45 96	13.6382
147	2 16 09	12.1244	187	3 49 69	13.6748
148	2 19 04	12.1655	188	3 53 44	13.7113
149	2 22 01	12.2066	189	3 57 21	13.7477
150	2 25 00	12.2474	190	3 61 00	13.7840
151	2 28 01	12.2882	191	3 64 81	13.8203
152	2 31 04	12.3288	192	3 68 64	13.8564
153	2 34 09	12.3693	193	3 72 49	13.8924
154	2 37 16	12.4097	194	3 76 36	13.9281
155	2 40 25	12.4499	195	3 80 25	13.9642
156	2 43 36	12.4900	196	3 84 16	14.0000
157	2 46 49	12.5300	197	3 88 09	14.0357
158	2 49 64	12.5693	198	3 92 04	14.0712
159	2 52 81	12.6095	199	3 96 01	14.1067
160	2 56 00	12.6491	200	4 00 00	14.1421
161	2 59 21	12.6886	201	4 04 01	14.1774
162	2 62 44	12.7279	202	4 08 04	14.2127
163	2 65 69	12.7671	203	4 12 09	14.2478
164	2 68 96	12.8062	204	4 16 16	14.2829
165	2 72 25	12.8452	205	4 20 25	14.3178
166	2 75 56	12.8841	206	4 24 36	14.3527
167	2 78 89	12.9228	207	4 28 49	14.3875
168	2 82 24	12.9615	208	4 32 64	14.4222
169	2 85 61	13.0000	209	4 36 81	14.4568
170	2 89 00	13.0384	210	4 41 00	14.4914
171	2 92 41	13.0767	211	4 45 21	14.5258
172	2 95 84	13.1149	212	4 49 44	14.5602
173	2 99 29	13.1529	213	4 53 69	14.5945
174	3 02 76	13.1909	214	4 57 96	14.6287
175	3 06 25	13.2288	215	4 62 25	14.6629
176	3 09 76	13.2665	216	4 66 56	14.6969
177	3 13 29	13.3041	217	4 70 89	14.7309
178	3 16 84	13.3417	218	4 75 24	14.7648
179	3 20 41	13.3791	219	4 79 61	14.7986
180	3 24 00	13.4164	220	4 84 00	14.8324

CUADRO A [continuación]

Número	Cuadrado	Raíz Cuadrada	Número	Cuadrado	Raíz Cuadrada
221	4 88 41	14.8661	261	6 81 21	16.1555
222	4 92 84	14.8997	262	6 86 44	16.1864
223	4 97 29	14.9332	263	6 91 69	16.2173
224	5 01 76	14.9666	264	6 96 96	16.2481
225	5 06 25	15.0000	265	7 02 25	16.2788
226	5 10 76	15.0333	266	7 07 56	16.3095
227	5 15 29	15.0665	267	7 12 89	16.3401
228	5 19 84	15.0997	268	7 18 24	16.3707
229	5 24 41	15.1327	269	7 23 61	16.4012
230	5 29 00	15.1658	270	7 29 00	16.4317
231	5 33 61	15.1987	271	7 34 41	16.4621
232	5 38 24	15.2315	272	7 39 84	16.4924
233	5 42 89	15.2643	273	7 45 29	16.5227
234	5 47 56	15.2971	274	7 50 76	16.5529
235	5 52 25	15.3297	275	7 56 25	16.5831
236	5 56 96	15.3623	276	7 61 76	16.6132
237	5 61 69	15.3948	277	7 67 29	16.6433
238	5 66 44	15.4272	278	7 72 84	16.6733
239	5 71 21	15.4596	279	7 78 41	16.7033
240	5 76 00	15.4919	280	7 84 00	16.7332
241	5 80 81	15.5242	281	7 89 61	16.7631
242	5 85 64	15.5563	282	7 95 24	16.7929
243	5 90 49	15.5885	283	8 00 89	16.8226
244	5 95 36	15.6205	284	8 06 56	16.8523
245	6 00 25	15.6525	285	8 12 25	16.8819
246	6 05 16	15.6844	286	8 17 96	16.9115
247	6 10 09	15.7162	287	8 23 69	16.9411
248	6 15 04	15.7480	288	8 29 44	16.9706
249	6 20 01	15.7797	289	8 35 21	17.0000
250	6 25 00	15.8114	290	8 41 00	17.0294
251	6 30 01	15.8430	291	8 46 81	17.0587
252	6 35 04	15.8745	292	8 52 64	17.0880
253	6 40 09	15.9060	293	8 58 49	17.1172
254	6 45 16	15.9374	294	8 64 36	17.1464
255	6 50 25	15.9687	295	8 70 25	17.1756
256	6 55 36	16.0000	296	8 76 16	17.2047
257	6 60 49	16.0312	297	8 82 09	17.2337
258	6 65 64	16.0624	298	8 88 04	17.2627
259	6 70 81	16.0935	299	8 94 01	17.2916
260	6 76 00	16.1245	300	9 00 00	17.3205

CUADRO A [continuación]

Número	Cuadrado	Raíz Cuadrada	Número	Cuadrado	Raíz Cuadrada
301	9 06 01	17.3494	341	11 62 81	18.4662
302	9 12 04	17.3781	342	11 69 64	18.4932
303	9 18 09	17.4069	343	11 76 49	18.5203
304	9 24 16	17.4356	344	11 83 36	18.5472
305	9 30 25	17.4642	345	11 90 25	18.5742
306	9 36 36	17.4929	346	11 97 16	18.6011
307	9 42 49	17.5214	347	12 04 09	18.6279
308	9 48 64	17.5499	348	12 11 04	18.6548
309	9 54 81	17.5784	349	12 18 01	18.6815
310	9 61 00	17.6068	350	12 25 00	18.7083
311	9 67 21	17.6352	351	12 32 01	18.7350
312	9 73 44	17.6635	352	12 39 04	18.7617
313	9 79 69	17.6918	353	12 46 09	18.7883
314	9 85 96	17.7200	354	12 53 16	18.8149
315	9 92 25	17.7482	355	12 60 25	18.8414
316	9 98 56	17.7764	356	12 67 36	18.8680
317	10 04 89	17.8045	357	12 74 49	18.8944
318	10 11 24	17.8326	358	12 81 64	18.9209
319	10 17 61	17.8606	359	12 88 81	18.9472
320	10 24 00	17.8885	360	12 96 00	18.9737
321	10 30 41	17.9165	361	13 03 21	19.0000
322	10 36 84	17.9444	362	13 10 44	19.0263
323	10 43 29	17.9722	363	13 17 69	19.0526
324	10 49 76	18.0000	364	13 24 96	19.0788
325	10 56 25	18.0278	365	13 32 25	19.1050
326	10 62 76	18.0555	366	13 39 56	19.1311
327	10 69 29	18.0831	367	13 46 89	19.1572
328	10 75 84	18.1108	368	13 54 24	19.1833
329	10 82 41	18.1384	369	13 61 61	19.2094
330	10 89 00	18.1659	370	13 69 00	19.2354
331	10 95 61	18.1934	371	13 76 41	19.2614
332	11 02 24	18.2209	372	13 83 84	19.2873
333	11 08 89	18.2483	373	13 91 29	19.3132
334	11 15 56	18.2757	374	13 98 76	19.3391
335	11 22 25	18.3030	375	14 06 25	19.3649
336	11 28 96	18.3303	376	14 13 76	19.3907
337	11 35 69	18.3576	377	14 21 29	19.4165
338	11 42 44	18.3848	378	14 28 84	19.4422
339	11 49 21	18.4120	379	14 36 41	19.4679
340	11 56 00	18.4391	380	14 44 00	19.4936

CUADRO A [continuación]

Número	Cuadrado	Raíz Cuadrada	Número	Cuadrado	Raíz Cuadrada
381	14 51 61	19.5192	421	17 72 41	20.5183
382	14 59 24	19.5448	422	17 80 84	20.5426
383	14 66 89	19.5704	423	17 89 29	20.5670
384	14 74 56	19.5959	424	17 97 76	20.5913
385	14 82 25	19.6214	425	18 06 25	20.6155
386	14 89 96	19.6469	426	18 14 76	20.6398
387	14 97 69	19.6723	427	18 23 29	20.6640
388	15 05 44	19.6977	428	18 31 84	20.6882
389	15 13 21	19.7231	429	18 40 41	20.7123
390	15 21 00	19.7484	430	18 49 00	20.7364
391	15 28 81	19.7737	431	18 57 61	20.7605
392	15 36 64	19.7990	432	18 66 24	20.7846
393	15 44 49	19.8242	433	18 74 89	20.8087
394	15 52 36	19.8494	434	18 83 56	20.8327
395	15 60 25	19.8746	435	18 92 25	20.8567
396	15 68 16	19.8997	436	19 00 96	20.8806
397	15 76 09	19.9249	437	19 09 69	20.9045
398	15 84 04	19.9499	438	19 18 44	20.9284
399	15 92 01	19.9750	439	19 27 21	20.9523
400	16 00 00	20.0000	440	19 36 00	20.9762
401	16 08 01	20.0250	441	19 44 81	21.0000
402	16 16 04	20.0499	442	19 53 64	21.0238
403	16 24 09	20.0749	443	19 62 49	21.0476
404	16 32 16	20.0998	444	19 71 36	21.0713
405	16 40 25	20.1246	445	19 80 25	21.0950
406	16 48 36	20.1494	446	19 89 16	21.1187
407	16 56 49	20.1742	447	19 98 09	21.1424
408	16 64 64	20.1990	448	20 07 04	21.1660
409	16 72 81	20.2237	449	20 16 01	21.1896
410	16 81 00	20.2485	450	20 25 00	21.2132
411	16 89 21	20.2731	451	20 34 01	21.2368
412	16 97 44	20.2978	452	20 43 04	21.2603
413	17 05 69	20.3224	453	20 52 09	21.2838
414	17 13 96	20.3470	454	20 61 16	21.3073
415	17 22 25	20.3715	455	20 70 25	21.3307
416	17 30 56	20.3961	456	20 79 36	21.3542
417	17 38 89	20.4206	457	20 88 49	21.3776
418	17 47 24	20.4450	458	20 97 64	21.4009
419	17 55 61	20.4695	459	21 06 81	21.4243
420	17 64 00	20.4939	460	21 16 00	21.4476

CUADRO A [continuación]

Número	Cuadrado	Raíz Cuadrada	Número	Cuadrado	Raíz Cuadrada
461	21 25 21	21.4709	501	25 10 01	22.3830
462	21 34 44	21.4942	502	25 20 04	22.4054
463	21 43 69	21.5174	503	25 30 09	22.4277
464	21 52 96	21.5407	504	25 40 16	22.4499
465	21 62 25	21.5639	505	25 50 25	22.4722
466	21 71 56	21.5870	506	25 60 36	22.4944
467	21 80 89	21.6102	507	25 70 49	22.5167
468	21 90 24	21.6333	508	25 80 64	22.5389
469	21 99 61	21.6564	509	25 90 81	22.5610
470	22 09 00	21.6795	510	26 01 00	22.5832
471	22 18 41	21.7025	511	26 11 21	22.6053
472	22 27 84	21.7256	512	26 21 44	22.6274
473	22 37 29	21.7486	513	26 31 69	22.6495
474	22 46 76	21.7715	514	26 41 96	22.6716
475	22 56 25	21.7945	515	26 52 25	22.6936
476	22 65 76	21.8174	516	26 62 56	22.7156
477	22 75 29	21.8403	517	26 72 89	22.7376
478	22 84 84	21.8632	518	26 83 24	22.7596
479	22 94 41	21.8861	519	26 93 61	22.7816
480	23 04 00	21.9089	520	27 04 00	22.8035
481	23 13 61	21.9317	521	27 14 41	22.8254
482	23 23 24	21.9545	522	27 24 84	22.8473
483	23 32 89	21.9773	523	27 35 29	22.8692
484	23 42 56	22.0000	524	27 45 76	22.8910
485	23 52 25	22.0227	525	27 56 25	22.9129
486	23 61 96	22.0454	526	27 66 76	22.9347
487	23 71 69	22.0681	527	27 77 29	22.9565
488	23 81 44	22.0907	528	27 87 84	22.9783
489	23 91 21	22.1133	529	27 98 41	23.0000
490	24 01 00	22.1359	530	28 09 00	23.0217
491	24 10 81	22.1585	531	28 19 61	23.0434
492	24 20 64	22.1811	532	28 30 24	23.0651
493	24 30 49	22.2036	533	28 40 89	23.0868
494	24 40 36	22.2261	534	28 51 56	23.1084
495	24 50 25	22.2486	535	28 62 25	23.1301
496	24 60 16	22.2711	536	28 72 96	23.1517
497	24 70 09	22.2935	537	28 83 69	23.1733
498	24 80 04	22.3159	538	28 94 44	23.1948
499	24 90 01	22.3383	539	29 05 21	23.2164
500	25 00 00	22.3607	540	29 16 00	23.2379

CUADRO A [continuación]

Número	Cuadrado	Raíz Cuadrada	Número	Cuadrado	Raíz Cuadrada
541	29 26 81	23.2594	581	33 75 61	24.1039
542	29 37 64	23.2809	582	33 87 24	24.1247
543	29 48 49	23.3024	583	33 98 89	24.1454
544	29 59 36	23.3238	584	34 10 56	24.1661
545	29 70 25	23.3452	585	34 22 25	24.1868
546	29 81 16	23.3666	586	34 33 96	24.2074
547	29 92 09	23.3880	587	34 45 69	24.2281
548	30 03 04	23.4094	588	34 57 44	24.2487
549	30 14 01	23.4307	589	34 69 21	24.2693
550	30 25 00	23.4521	590	34 81 00	24.2899
551	30 36 01	23.4734	591	34 92 81	24.3105
552	30 47 04	23.4947	592	35 04 64	24.3311
553	30 58 09	23.5160	593	35 16 49	24.3516
554	30 69 16	23.5372	594	35 28 36	24.3721
555	30 80 25	23.5584	595	35 40 25	24.3926
556	30 91 36	23.5797	596	35 52 16	24.4131
557	31 02 49	23.6008	597	35 64 09	24.4336
558	31 13 64	23.6220	598	35 76 04	24.4540
559	31 24 81	23.6432	599	35 88 01	24.4745
560	31 36 00	23.6643	600	36 00 00	24.4949
561	31 47 21	23.6854	601	36 12 01	24.5153
562	31 58 44	23.7065	602	36 24 04	24.5357
563	31 69 69	23.7276	603	36 36 09	24.5561
564	31 80 96	23.7487	604	36 48 16	24.5764
565	31 92 25	23.7697	605	36 60 25	24.5967
566	32 03 56	23.7908	606	36 72 36	24.6171
567	32 14 89	23.8118	607	36 84 49	24.6374
568	32 26 24	23.8328	608	36 96 64	24.6577
569	32 37 61	23.8537	609	37 08 81	24.6779
570	32 49 00	23.8747	610	37 21 00	24.6982
571	32 60 41	23.8956	611	37 33 21	24.7184
572	32 71 84	23.9165	612	37 45 44	24.7385
573	32 83 29	23.9374	613	37 57 69	24.7588
574	32 94 76	23.9583	614	37 69 96	24.7790
575	33 06 25	23.9792	615	37 82 25	24.7992
576	33 17 76	24.0000	616	37 94 56	24.8193
577	33 29 29	24.0208	617	38 06 89	24.8395
578	33 40 84	24.0416	618	38 19 24	24.8596
579	33 52 41	24.0624	619	38 31 61	24.8797
580	33 64 00	24.0832	620	38 44 00	24.8998

CUADRO A [continuación]

Número	Cuadrado	Raíz Cuadrada	Número	Cuadrado	Raíz Cuadrada
621	38 56 41	24.9199	661	43 69 21	25.7099
622	38 68 84	24.9399	662	43 82 44	25.7294
623	38 81 29	24.9600	663	43 95 69	25.7488
624	38 93 76	24.9800	664	44 08 96	25.7682
625	39 06 25	25.0000	665	44 22 25	25.7876
626	39 18 76	25.0200	666	44 35 56	25.8070
627	39 31 29	25.0400	667	44 48 89	25.8263
628	39 43 84	25.0599	668	44 62 24	25.8457
629	39 56 41	25.0799	669	44 75 61	25.8650
630	39 69 00	25.0998	670	44 89 00	25.8844
631	39 81 61	25.1197	671	45 02 41	25.9037
632	39 94 24	25.1396	672	45 15 84	25.9230
633	40 06 89	25.1595	673	45 29 29	25.9422
634	40 19 56	25.1794	674	45 42 76	25.9615
635	40 32 25	25.1992	675	45 56 25	25.9808
636	40 44 96	25.2190	676	45 69 76	26.0000
637	40 57 69	25.2389	677	45 83 29	26.0192
638	40 70 44	25.2587	678	45 96 84	26.0384
639	40 83 21	25.2784	679	46 10 41	26.0576
640	40 96 00	25.2982	680	46 24 00	26.0768
641	41 08 81	25.3180	681	46 37 61	26.0960
642	41 21 64	25.3377	682	46 51 24	26.1151
643	41 34 49	25.3574	683	46 64 89	26.1343
644	41 47 36	25.3772	684	46 78 56	26.1534
645	41 60 25	25.3969	685	46 92 25	26.1725
646	41 73 16	25.4165	686	47 05 96	26.1916
647	41 86 09	25.4362	687	47 19 69	26.2107
648	41 99 04	25.4558	688	47 33 44	26.2298
649	42 12 01	25.4755	689	47 47 21	26.2488
650	42 25 00	25.4951	690	47 61 00	26.2679
651	42 38 01	25.5147	691	47 74 81	26.2869
652	42 51 04	25.5343	692	47 88 64	26.3059
653	42 64 09	25.5539	693	48 02 49	26.3249
654	42 77 16	25.5734	694	48 16 36	26.3439
655	42 90 25	25.5930	695	48 30 25	26.3629
656	43 03 36	25.6125	696	48 44 16	26.3818
657	43 16 49	25.6320	697	48 58 09	26.4008
658	43 29 64	25.6515	698	48 72 04	26.4197
659	43 42 81	25.6710	699	48 86 01	26.4386
660	43 56 00	25.6905	700	49 00 00	26.4575

CUADROS

CUADRO A [continuación]

Número	Cuadrado	Raíz Cuadrada	Número	Cuadrado	Raíz Cuadrada
701	49 14 01	26.4764	741	54 90 81	27.2213
702	49 28 04	26.4953	742	55 05 64	27.2397
703	49 42 09	26.5141	743	55 20 49	27.2580
704	49 56 16	26.5330	744	55 35 36	27.2764
705	49 70 25	26.5518	745	55 50 25	27.2947
706	49 84 36	26.5707	746	55 65 16	27.3130
707	49 98 49	26.5895	747	55 80 09	27.3313
708	50 12 64	26.6083	748	55 95 04	27.3496
709	50 26 81	26.6271	749	56 10 01	27.3679
710	50 41 00	26.6458	750	56 25 00	27.3861
711	50 55 21	26.6646	751	56 40 01	27.4044
712	50 69 44	26.6833	752	56 55 04	27.4226
713	50 83 69	26.7021	753	56 70 09	27.4408
714	50 97 96	26.7208	754	56 85 16	27.4591
715	51 12 25	26.7395	755	57 00 25	27.4773
716	51 26 56	26.7582	756	57 15 36	27.4955
717	51 40 89	26.7769	757	57 30 49	27.5136
718	51 55 24	26.7955	758	57 45 64	27.5318
719	51 69 61	26.8142	759	57 60 81	27.5500
720	51 84 00	26.8328	760	57 76 00	27.5681
721	51 98 41	26.8514	761	57 91 21	27.5862
722	52 12 84	26.8701	762	58 06 44	27.6043
723	52 27 29	26.8887	763	58 21 69	27.6225
724	52 41 76	26.9072	764	58 36 96	27.6405
725	52 56 25	26.9258	765	58 52 25	27.6586
726	52 70 76	26.9444	766	58 67 56	27.6767
727	52 85 29	26.9629	767	58 82 89	27.6948
728	52 99 84	26.9815	768	58 98 24	27.7128
729	53 14 41	27.0000	769	59 13 61	27.7308
730	53 29 00	27.0185	770	59 29 00	27.7489
731	53 43 61	27.0370	771	59 44 41	27.7669
732	53 58 24	27.0555	772	59 59 84	27.7849
733	53 72 89	27.0740	773	59 75 29	27.8029
734	53 87 56	27.0924	774	59 90 76	27.8209
735	54 02 25	27.1109	775	60 06 25	27.8388
736	54 16 96	27.1293	776	60 21 76	27.8568
737	54 31 69	27.1477	777	60 37 29	27.8747
738	54 46 44	27.1662	778	60 52 84	27.8927
739	54 61 27	27.1846	779	60 68 41	27.9106
740	54 76 00	27.2029	780	60 84 00	27.9285

CUADROS

CUADRO A [continuación]

Número	Cuadrado	Raíz Cuadrada	Número	Cuadrado	Raíz Cuadrada
781	60 99 61	27.9464	821	67 40 41	28.6531
782	61 15 24	27.9643	822	67 56 84	28.6705
783	61 30 89	27.9821	823	67 73 29	28.6880
784	61 46 56	28.0000	824	67 89 76	28.7054
785	61 62 25	28.0179	825	68 06 25	28.7228
786	61 77 96	28.0357	826	68 22 76	28.7402
787	61 93 69	28.0535	827	68 39 29	28.7576
788	62 09 44	28.0713	828	68 55 84	28.7750
789	62 25 21	28.0891	829	68 72 41	28.7924
790	62 41 00	28.1069	830	68 89 00	28.8097
791	62 56 81	28.1247	831	69 05 61	28.8271
792	62 72 64	28.1425	832	69 22 24	28.8444
793	62 88 49	28.1603	833	69 38 89	28.8617
794	63 04 36	28.1780	834	69 55 56	28.8791
795	63 20 25	28.1957	835	69 72 25	28.8964
796	63 36 16	28.2135	836	69 88 96	28.9137
797	63 52 09	28.2312	837	70 05 69	28.9310
798	63 68 04	28.2489	838	70 22 44	28.9482
799	63 84 01	28.2666	839	70 39 21	28.9655
800	64 00 00	28.2843	840	70 56 00	28.9828
801	64 16 01	28.3019	841	70 72 81	29.0000
802	64 32 04	28.3196	842	70 89 64	29.0172
803	64 48 09	28.3373	843	71 06 49	29.0345
804	64 64 16	28.3549	844	71 23 36	29.0517
805	64 80 25	28.3725	845	71 40 25	29.0689
806	64 96 36	28.3901	846	71 57 16	29.0861
807	65 12 49	28.4077	847	71 74 09	29.1033
808	65 28 64	28.4253	848	71 91 04	29.1204
809	65 44 81	28.4429	849	72 08 01	29.1376
810	65 61 00	28.4605	850	72 25 00	29.1548
811	65 77 21	28.4781	851	72 42 01	29.1719
812	65 93 44	28.4956	852	72 59 04	29.1890
813	66 09 69	28.5132	853	72 76 09	29.2062
814	66 25 96	28.5307	854	72 93 16	29.2233
815	66 42 25	28.5482	855	73 10 25	29.2404
816	66 58 56	28.5657	856	73 27 36	29.2575
817	66 74 89	28.5832	857	73 44 49	29.2746
818	66 91 24	28.6007	858	73 61 64	29.2916
819	67 07 61	28.6082	859	73 78 81	29.3087
820	67 24 00	28.6356	860	73 96 00	29.3258

CUADRO A [continuación]

Número	Cuadrado	Raíz Cuadrada	Número	Cuadrado	Raíz Cuadrada
861	74 13 21	29.3428	901	81 18 01	30.0167
862	74 30 44	29.3598	902	81 36 04	30.0333
863	74 47 69	29.3769	903	81 54 09	30.0500
864	74 64 96	29.3939	904	81 72 16	30.0666
865	74 82 25	29.4109	905	81 90 25	30.0832
866	74 99 56	29.4279	906	82 08 36	30.0998
867	75 16 89	29.4449	907	82 26 49	30.1164
868	75 34 24	29.4618	908	82 44 64	30.1330
869	75 51 61	29.4788	909	82 62 81	30.1496
870	75 69 00	29.4958	910	82 81 00	30.1662
871	75 86 41	29.5127	911	82 99 21	30.1828
872	76 03 84	29.5296	912	83 17 44	30.1993
873	76 21 29	29.5466	913	83 35 69	30.2159
874	76 38 76	29.5635	914	83 53 96	30.2324
875	76 56 25	29.5804	915	83 72 25	30.2490
876	76 73 76	29.5973	916	83 90 56	30.2655
877	76 91 29	29.6142	917	84 08 89	30.2820
878	77 08 84	29.6311	918	84 27 24	30.2985
879	77 26 41	29.6479	919	84 45 61	30.3150
880	77 44 00	29.6648	920	84 64 00	30.3315
881	77 61 61	29.6816	921	84 82 41	30.3480
882	77 79 24	29.6985	922	85 00 84	30.3645
883	77 96 89	29.7153	923	85 19 29	30.3809
884	78 14 56	29.7321	924	85 37 76	30.3974
885	78 32 25	29.7489	925	85 56 25	30.4138
886	78 49 96	29.7658	926	85 74 76	30.4302
887	78 67 69	29.7825	927	85 93 29	30.4467
888	78 85 44	29.7993	928	86 11 84	30.4631
889	79 03 21	29.8161	929	86 30 41	30.4795
890	79 21 00	29.8329	930	86 49 00	30.4959
891	79 38 81	29.8496	931	86 67 61	30.5123
892	79 56 64	29.8664	932	86 86 24	30.5287
893	79 74 49	29.8831	933	87 04 89	30.5450
894	79 92 36	29.8998	934	87 23 56	30.5614
895	80 10 25	29.9166	935	87 42 25	30.5778
896	80 28 16	29.9333	936	87 60 96	30.5941
897	80 46 09	29.9500	937	87 79 69	30.6105
898	80 64 04	29.9666	938	87 98 44	30.6268
899	80 82 01	29.9833	939	88 17 21	30.6431
900	81 00 00	30.0000	940	88 36 00	30.6594

CUADRO A [conclusión]

Número	Cuadrado	Raíz Cuadrada	Número	Cuadrado	Raíz Cuadrada
941	88 54 81	30.6757	971	94 28 41	31.1609
942	88 73 64	30.6920	972	94 47 84	31.1769
943	88 92 49	30.7083	973	94 67 29	31.1929
944	89 11 36	30.7246	974	94 86 76	31.2090
945	89 30 25	30.7409	975	95 06 25	31.2250
946	89 49 16	30.7571	976	95 25 76	31.2410
947	89 68 09	30.7734	977	95 45 29	31.2570
948	89 87 04	30.7896	978	95 64 84	31.2730
949	90 06 01	30.8058	979	95 84 41	31.2890
950	90 25 00	30.8221	980	96 04 00	31.3050
951	90 44 01	30.8383	981	96 23 61	31.3209
952	90 63 04	30.8545	982	96 43 24	31.3369
953	90 82 09	30.8707	983	96 62 89	31.3528
954	91 01 16	30.8869	984	96 82 56	31.3688
955	91 20 25	30.9031	985	97 02 25	31.3847
956	91 39 36	30.9192	986	97 21 96	31.4006
957	91 58 49	30.9354	987	97 41 69	31.4166
958	91 77 64	30.9516	988	97 61 44	31.4325
959	91 96 81	30.9677	989	97 81 21	31.4484
960	92 16 00	30.9839	990	98 01 00	31.4643
961	92 35 21	31.0000	991	98 20 81	31.4802
962	92 54 44	31.0161	992	98 40 64	31.4960
963	92 73 69	31.0322	993	98 60 49	31.5119
964	92 92 96	31.0483	994	98 80 36	31.5278
965	93 12 25	31.0644	995	99 00 25	31.5436
966	93 31 56	31.0805	996	99 20 16	31.5595
967	93 50 89	31.0966	997	99 40 09	31.5753
968	93 70 24	31.1127	998	99 60 04	31.5911
969	93 89 61	31.1288	999	99 80 01	31.6070
970	94 09 00	31.1448	1000	100 00 00	31.6228

CUADRO B. *Números aleatorios*

10 09 73 25 33	76 52 01 35 86	34 67 35 48 76	80 95 00 91 17	39 29 27 49 45
37 54 20 48 05	64 89 47 42 96	24 80 52 40 37	20 63 61 04 02	00 82 29 16 65
08 42 26 89 53	19 64 50 93 03	23 20 90 25 60	15 95 33 47 64	35 08 03 36 06
99 01 90 25 29	09 37 67 07 15	38 31 13 11 65	88 67 67 43 97	04 43 62 76 59
12 80 79 99 70	80 15 73 61 47	64 03 23 66 53	98 95 11 68 77	12 17 17 68 33
66 06 57 47 17	34 07 27 68 50	36 69 73 61 70	65 81 33 98 85	11 19 92 91 70
31 06 01 08 05	45 57 18 24 06	35 30 34 26 14	86 79 90 74 39	23 40 30 97 32
85 26 97 76 02	02 05 16 56 92	68 66 57 48 18	73 05 38 52 47	18 62 38 85 79
63 57 33 21 35	05 32 54 70 48	90 55 35 75 48	28 46 82 87 09	83 49 12 56 24
73 79 64 57 53	03 52 96 47 78	35 80 83 42 82	60 93 52 03 44	35 27 38 84 35
98 52 01 77 67	14 90 56 86 07	22 10 94 05 58	60 97 09 34 33	50 50 07 39 98
11 80 50 54 31	39 80 82 77 32	50 72 56 82 48	29 40 52 42 01	52 77 56 78 51
83 45 29 96 34	06 28 89 80 83	13 74 67 00 78	18 47 54 06 10	68 71 17 78 17
88 68 54 02 00	86 50 75 84 01	36 76 66 79 51	90 36 47 64 93	29 60 91 10 62
99 59 46 73 48	87 51 76 49 69	91 82 60 89 28	93 78 56 13 68	23 47 83 41 13
65 48 11 76 74	17 46 85 09 50	58 04 77 69 74	73 03 95 71 86	40 21 81 65 44
80 12 43 56 35	17 72 70 80 15	45 31 82 23 74	21 11 57 82 53	14 38 55 37 63
74 35 09 98 17	77 40 27 72 14	43 23 60 02 10	45 52 16 42 37	96 28 60 26 55
69 91 62 68 03	66 25 22 91 48	36 93 68 72 03	76 62 11 39 90	94 40 05 64 18
09 89 32 05 05	14 22 56 85 14	46 42 75 67 88	96 29 77 88 22	54 38 21 45 98
91 49 91 45 23	68 47 92 76 86	46 16 28 35 54	94 75 08 99 23	37 08 92 00 48
80 33 69 45 98	26 94 03 68 58	70 29 73 41 35	53 14 03 33 40	42 05 08 23 41
44 10 48 19 49	85 15 74 79 54	32 97 92 65 75	57 60 04 08 81	22 22 20 64 13
12 55 07 37 42	11 10 00 20 40	12 86 07 46 97	96 64 48 94 39	28 70 72 58 15
63 60 64 93 29	16 50 53 44 84	40 21 95 25 63	43 65 17 70 82	07 20 73 17 90
61 19 69 04 46	26 45 74 77 74	51 92 43 37 29	65 39 45 95 93	42 58 26 05 27
15 47 44 52 66	95 27 07 99 53	59 36 78 38 48	82 39 61 01 18	33 21 15 94 66
94 55 72 85 73	67 89 75 43 87	54 62 24 44 31	91 19 04 25 92	92 92 74 59 73
42 48 11 62 13	97 34 40 87 21	16 86 84 87 67	03 07 11 20 59	25 70 14 66 70
23 52 37 83 17	73 20 88 98 37	68 93 59 14 16	26 25 22 96 63	05 52 28 25 62
04 49 35 24 94	75 24 63 38 24	45 86 25 10 25	61 96 27 93 35	65 33 71 24 72
00 54 99 76 54	64 05 18 81 59	96 11 96 38 96	54 69 28 23 91	23 28 72 95 29
35 96 31 53 07	26 89 80 93 54	33 35 13 54 62	77 97 45 00 24	90 10 33 93 33
59 80 80 83 91	45 42 72 68 42	83 60 94 97 00	13 02 12 48 92	78 56 52 01 06
46 05 88 52 36	01 39 09 22 86	77 28 14 40 77	93 91 08 36 47	70 61 74 29 41
32 17 90 05 97	87 37 92 52 41	05 56 70 70 07	86 74 31 71 57	85 39 41 18 38
69 23 46 14 06	20 11 74 52 04	15 95 66 00 00	18 74 39 24 23	97 11 89 63 38
19 56 54 14 30	01 75 87 53 79	40 41 92 15 85	66 67 43 68 06	84 96 28 52 07
45 15 51 49 38	19 47 60 72 46	43 66 79 45 43	59 04 79 00 33	20 82 66 95 41
94 86 43 19 94	36 16 81 08 51	34 88 88 15 53	01 54 03 54 56	05 01 45 11 76

FUENTE: The Rand Corporation, *A Million Random Digits*, Free Press, Glencoe, Ill., 1955, pp. 1-3, con la amable autorización del editor.

CUADRO B [continuación]

98 08 62 48 26	45 24 02 84 04	44 99 90 88 96	39 09 47 34 07	35 44 13 18 80
33 18 51 62 32	41 94 15 09 49	89 43 54 85 81	88 69 54 19 94	37 54 87 30 43
80 95 10 04 06	96 38 27 07 74	20 15 12 33 87	25 01 62 52 98	94 62 46 11 71
79 75 24 91 40	71 96 12 82 96	69 86 10 25 91	74 85 22 05 39	00 38 75 95 79
18 63 33 25 37	98 14 50 65 71	31 01 02 46 74	05 45 56 14 27	77 93 89 19 36
74 02 94 39 02	77 55 73 22 70	97 79 01 71 19	52 52 75 80 21	80 81 45 17 48
54 17 84 56 11	80 99 33 71 43	05 33 51 29 69	56 12 71 92 55	36 04 09 03 24
11 66 44 98 83	52 07 98 48 27	59 38 17 15 39	09 97 33 34 40	88 46 12 33 56
48 32 47 79 28	31 24 96 47 10	02 29 53 68 70	32 30 75 75 46	15 02 00 99 94
69 07 49 41 38	87 63 79 19 76	35 58 40 44 01	10 51 82 16 15	01 84 87 69 38
09 18 82 00 97	32 82 53 95 27	04 22 08 63 04	83 38 98 73 74	64 27 85 80 44
90 04 58 54 97	51 98 15 06 54	94 93 88 19 97	91 87 07 61 50	68 47 66 46 59
73 18 95 02 07	47 67 72 52 69	62 29 06 44 64	27 12 46 70 18	41 36 18 27 60
75 76 87 64 90	20 97 18 17 49	90 42 91 22 72	95 37 50 58 71	93 82 34 31 78
54 01 64 40 56	66 28 13 10 03	00 68 22 73 98	20 71 45 32 95	07 70 61 78 13
08 35 86 99 10	78 54 24 27 85	13 66 15 88 73	04 61 89 75 53	31 22 30 84 20
28 30 60 32 64	81 33 31 05 91	40 51 00 78 93	32 60 46 04 75	94 11 90 18 40
53 84 08 62 33	81 59 41 36 28	51 21 59 02 90	28 46 66 87 95	77 76 22 07 91
91 75 75 37 41	61 61 36 22 69	50 26 39 02 12	55 78 17 65 14	83 48 34 70 55
89 41 59 28 94	00 39 75 83 91	12 60 71 76 46	48 94 97 23 06	94 54 13 74 08
77 51 30 38 20	86 83 42 99 01	68 41 48 27 74	51 90 81 39 80	72 89 35 55 07
19 50 23 71 74	69 97 92 02 88	55 21 02 97 73	74 28 77 52 51	65 34 46 74 15
21 81 85 93 13	93 27 88 17 57	05 68 67 31 56	07 08 28 50 46	31 85 33 84 52
51 47 46 64 99	68 10 72 36 21	94 04 99 13 45	42 83 60 91 91	08 00 74 54 49
99 55 96 83 31	62 53 52 41 70	69 77 71 28 30	74 81 97 81 42	43 86 07 28 34
33 71 34 80 07	93 58 47 28 69	51 92 66 47 21	58 30 32 98 22	93 17 49 39 72
85 27 48 68 93	11 30 32 92 70	28 83 43 41 37	73 51 59 04 00	71 14 84 36 43
84 13 38 96 40	44 03 55 21 66	73 85 27 00 91	61 22 26 05 61	62 32 71 84 23
56 73 21 62 34	17 39 59 61 31	10 12 39 16 22	85 49 65 75 60	81 60 41 88 80
65 13 85 68 06	87 64 88 52 61	34 31 36 58 61	45 87 52 10 69	85 64 44 72 77
38 00 10 21 76	81 71 91 17 11	71 60 29 29 87	74 21 96 40 49	65 58 44 96 98
37 40 29 63 97	01 30 47 75 86	56 27 11 00 86	47 32 46 26 05	40 03 03 74 38
97 12 54 03 48	87 08 33 14 17	21 81 53 92 50	75 23 76 20 47	15 50 12 95 78
21 82 64 11 34	47 14 33 40 72	64 63 88 59 02	49 13 90 64 41	03 85 65 45 52
73 13 54 27 42	95 71 90 90 35	85 79 47 42 85	08 78 98 81 56	64 69 11 92 02
07 63 87 79 29	03 06 11 80 72	96 20 74 41 58	23 82 19 95 38	04 71 36 69 94
60 52 88 34 41	07 95 41 98 14	59 17 52 06 95	05 53 35 21 39	61 21 20 64 55
83 59 63 56 55	06 95 89 29 83	05 12 80 97 19	77 43 35 37 83	92 30 15 04 98
10 85 06 27 46	99 59 91 05 07	13 49 90 63 19	53 07 57 18 39	06 41 01 93 62
39 82 09 89 52	43 62 26 31 47	64 42 18 08 14	43 80 00 93 51	31 02 47 31 67

CUADRO B [continuación]

59 58 00 64 73	75 56 97 88 00	88 83 55 44 86	23 76 20 61 56	04 11 10 84 08
38 50 80 73 41	23 79 34 87 63	00 82 29 70 22	17 71 90 42 07	95 95 44 99 53
30 69 27 06 68	94 68 81 61 27	56 19 68 00 91	82 06 76 34 00	05 46 26 92 00
65 44 39 56 59	18 28 82 74 37	49 83 22 40 41	08 33 76 56 76	96 29 99 08 36
27 26 75 02 64	13 19 27 22 94	07 47 74 46 06	17 98 54 89 11	97 34 13 03 58
91 30 70 69 91	19 07 22 42 10	36 69 95 37 28	28 82 53 57 93	28 97 66 62 52
68 43 49 46 88	84 47 31 36 22	62 12 69 84 08	12 84 38 25 90	09 81 59 81 46
48 90 81 58 77	54 74 52 45 91	35 70 00 47 54	83 82 45 26 92	54 13 05 51 60
06 91 34 51 97	42 67 27 86 01	11 83 30 95 28	63 01 19 89 01	14 97 44 03 44
10 45 51 60 19	14 21 03 37 12	91 34 23 78 21	83 32 58 08 51	43 66 77 08 83
12 88 39 73 43	65 02 76 11 84	04 28 50 13 92	17 97 41 50 77	90 71 22 67 09
21 77 83 09 76	38 80 73 69 61	31 64 94 20 96	63 28 10 20 23	08 81 64 74 49
19 52 35 95 15	65 12 25 96 59	88 28 36 82 58	69 57 21 37 98	16 43 59 15 29
67 24 55 26 70	35 58 31 65 63	79 24 68 68 86	76 46 33 42 22	26 65 59 08 02
60 58 44 73 77	07 50 03 79 92	45 13 42 65 29	26 76 08 36 37	41 32 64 43 44
53 85 34 13 77	36 06 69 48 50	58 83 87 38 59	49 36 47 33 31	93 24 04 36 42
24 63 73 87 36	74 38 48 93 42	62 62 30 79 92	12 36 91 86 01	03 74 28 38 73
83 08 01 24 51	38 99 22 28 15	07 75 95 17 77	97 37 72 76 85	51 97 23 78 67
16 44 42 43 34	38 15 19 90 73	27 49 37 09 39	85 13 03 25 52	54 84 65 47 59
60 79 01 81 57	57 17 86 57 62	11 16 17 85 76	45 81 95 29 79	65 13 00 48 60
03 99 11 04 61	93 71 61 68 94	66 08 32 46 53	84 60 95 82 32	88 61 81 91 61
38 55 59 55 54	32 88 65 97 80	08 35 66 08 60	29 73 54 77 62	71 29 92 38 53
17 54 67 37 04	92 05 24 62 15	55 12 12 92 81	59 07 60 79 36	27 95 45 89 09
32 64 35 28 61	95 81 90 68 31	00 91 19 89 36	76 35 59 37 79	80 86 30 05 14
69 57 26 87 77	39 51 03 59 05	14 06 04 06 19	29 54 96 96 16	33 56 46 07 80
24 12 26 66 91	27 69 90 64 94	14 84 54 66 72	61 95 87 71 00	90 89 97 57 54
61 19 63 02 31	92 96 26 17 73	41 83 95 53 82	17 26 77 09 43	78 03 87 02 67
30 53 22 17 04	10 27 41 22 02	39 68 52 33 09	10 06 16 88 29	55 98 66 64 85
03 78 89 75 99	76 86 72 07 17	74 41 66 31 66	35 20 83 33 74	87 53 90 88 23
48 22 86 33 79	85 78 34 76 19	63 15 26 74 33	35 66 35 29 72	16 81 86 03 11
60 36 59 46 53	35 07 53 39 43	42 61 42 92 97	01 91 82 83 16	98 95 37 32 31
83 79 94 24 02	56 62 33 44 42	34 99 44 13 74	70 07 11 47 36	09 95 81 80 66
32 96 00 74 05	36 40 98 32 32	99 38 54 16 00	11 13 30 75 86	15 91 70 62 53
19 32 25 38 45	57 32 05 26 06	66 49 76 86 46	78 13 56 65 59	19 64 09 94 13
11 22 09 47 47	07 39 93 74 08	48 50 92 39 29	27 48 24 54 76	85 24 43 51 59
31 75 15 72 60	68 98 00 53 39	15 47 04 83 55	88 65 12 25 96	03 15 21 92 21
88 49 29 93 82	14 46 40 45 04	20 09 49 89 77	74 84 39 34 13	22 10 97 85 06
30 93 44 77 44	07 48 18 38 28	73 78 80 65 33	28 59 73 04 05	94 20 52 03 80
22 88 84 88 93	27 49 99 87 48	60 53 04 51 28	74 02 28 46 17	82 03 71 02 68
78 21 21 69 93	35 90 29 13 86	44 37 21 54 89	65 74 11 40 14	87 48 13 72 20

CUADRO B [conclusión]

41 84 98 45 47	46 85 05 23 26	34 67 75 83 00	74 91 06 43 45	19 32 58 15 49
46 35 23 30 49	69 24 89 34 60	45 30 50 75 21	61 31 83 18 55	14 41 37 09 51
11 08 79 62 94	14 01 33 17 92	59 74 78 72 77	76 50 33 45 13	39 66 37 75 44
52 70 10 83 37	56 39 38 73 15	16 52 06 96 76	11 65 49 98 93	02 18 16 81 01
57 27 53 68 98	81 30 44 85 85	68 65 22 73 76	92 85 25 58 66	88 44 80 35 84
20 85 77 31 56	70 28 42 43 26	79 37 59 52 20	01 15 96 32 67	10 62 24 83 91
15 63 38 49 24	90 41 59 36 14	33 52 12 66 65	55 82 34 76 41	86 22 53 17 04
92 69 44 82 97	39 90 40 21 15	59 58 94 90 67	66 82 14 15 75	49 76 70 40 37
77 61 31 90 19	88 15 20 00 80	20 55 49 14 09	96 27 74 82 57	50 81 69 78 16
33 68 83 24 86	45 13 46 35 45	59 40 47 20 59	43 94 75 16 80	43 85 25 96 93
25 16 30 18 89	70 01 41 50 21	41 29 06 73 12	71 85 71 59 57	68 97 11 14 03
65 25 10 76 29	37 23 93 32 95	05 87 00 11 19	92 78 42 63 40	18 47 76 56 22
36 81 54 36 25	18 63 73 75 09	82 44 49 90 05	04 92 17 37 01	14 70 79 39 97
64 39 71 16 92	05 32 78 21 62	20 24 78 17 59	45 19 72 53 32	83 74 52 25 67
04 51 52 56 24	95 09 66 79 46	48 46 08 55 58	15 19 11 87 82	16 93 03 33 61
83 76 16 08 73	43 25 38 41 45	60 83 32 59 83	01 29 14 13 49	20 36 80 71 26
14 38 70 63 45	80 85 40 92 79	43 52 90 63 18	38 38 47 47 61	41 19 63 74 80
51 32 19 22 46	80 08 87 70 74	88 72 25 67 36	66 16 44 94 31	66 91 93 16 78
72 47 20 00 08	80 89 01 80 02	94 81 33 19 00	54 15 58 34 36	35 35 25 41 31
05 46 65 53 06	93 12 81 84 64	74 45 79 05 61	72 84 81 18 34	79 98 26 84 16
39 52 87 24 84	82 47 42 55 93	48 54 53 52 47	18 61 91 36 74	18 61 11 92 41
81 61 61 87 11	53 34 24 42 76	75 12 21 17 24	74 62 77 37 07	58 31 91 59 97
07 58 61 61 20	82 64 12 28 20	92 90 41 31 41	32 39 21 97 63	61 19 96 79 40
90 76 70 42 35	13 57 41 72 00	69 90 26 37 42	78 46 42 25 01	18 62 79 08 72
40 18 82 81 93	29 59 38 86 27	94 97 21 15 98	62 09 53 67 87	00 44 15 89 97
34 41 48 21 57	84 88 75 50 87	19 15 20 00 23	12 30 28 07 83	32 62 46 86 91
63 43 97 53 63	44 98 91 68 22	36 02 40 09 67	76 37 84 16 05	65 96 17 34 88
67 04 80 90 70	93 39 94 55 47	94 45 87 42 84	05 04 14 98 07	20 28 83 40 60
79 49 50 41 46	52 16 29 02 86	54 15 83 42 43	46 97 83 54 82	59 36 29 59 38
91 70 43 05 52	04 73 72 10 31	75 05 19 30 29	47 66 56 43 82	99 78 29 34 78

CUADRO C. Areas bajo la curva normal

Fracciones del área total (10 000) bajo la curva normal, correspondientes a distancias entre la media y las ordenadas situadas a Z unidades de desviación estándar de la media.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0000	0040	0080	0120	0159	0199	0239	0279	0319	0359
0.1	0398	0438	0478	0517	0557	0596	0636	0675	0714	0753
0.2	0793	0832	0871	0910	0948	0987	1026	1064	1103	1141
0.3	1179	1217	1255	1293	1331	1368	1406	1443	1480	1517
0.4	1554	1591	1628	1664	1700	1736	1772	1808	1844	1879
0.5	1915	1950	1985	2019	2054	2088	2123	2157	2190	2224
0.6	2257	2291	2324	2357	2389	2422	2454	2486	2518	2549
0.7	2580	2612	2642	2673	2704	2734	2764	2794	2823	2852
0.8	2881	2910	2939	2967	2995	3023	3051	3078	3106	3133
0.9	3159	3186	3212	3238	3264	3289	3315	3340	3365	3389
1.0	3413	3438	3461	3485	3508	3531	3554	3577	3599	3621
1.1	3643	3665	3686	3713	3729	3749	3770	3790	3810	3830
1.2	3849	3869	3888	3907	3925	3944	3962	3980	3997	4015
1.3	4032	4049	4066	4083	4099	4115	4131	4147	4162	4177
1.4	4192	4207	4222	4236	4251	4265	4279	4292	4306	4319
1.5	4332	4345	4357	4370	4382	4394	4406	4418	4430	4441
1.6	4452	4463	4474	4485	4495	4505	4515	4525	4535	4545
1.7	4554	4564	4573	4582	4591	4599	4608	4616	4625	4633
1.8	4641	4649	4656	4664	4671	4678	4686	4693	4699	4706
1.9	4713	4719	4726	4732	4738	4744	4750	4758	4762	4767
2.0	4773	4778	4783	4788	4793	4798	4803	4808	4812	4817
2.1	4821	4826	4830	4834	4838	4842	4846	4850	4854	4857
2.2	4861	4865	4868	4871	4875	4878	4881	4884	4887	4890
2.3	4893	4896	4898	4901	4904	4906	4909	4911	4913	4916
2.4	4918	4920	4922	4925	4927	4929	4931	4932	4934	4936
2.5	4938	4940	4941	4943	4945	4946	4948	4949	4951	4952
2.6	4953	4955	4956	4957	4959	4960	4961	4962	4963	4964
2.7	4965	4966	4967	4968	4969	4970	4971	4972	4973	4974
2.8	4974	4975	4976	4977	4977	4978	4979	4980	4980	4981
2.9	4981	4982	4983	4984	4984	4984	4985	4985	4986	4986
3.0	4986.5	4987	4987	4988	4988	4988	4989	4989	4989	4990
3.1	4990.0	4991	4991	4991	4992	4992	4992	4992	4993	4993
3.2	4993.129									
3.3	4995.166									
3.4	4996.631									
3.5	4997.674									
3.6	4998.409									
3.7	4998.922									
3.8	4999.277									
3.9	4999.519									
4.0	4999.683									
4.5	4999.966									
5.0	4999.997133									

FUENTE: Harold O. Rugg, *Statistical Methods Applied to Education*, Houghton Mifflin Company, Boston, 1917, apéndice al cuadro III, pp. 389-390, con la amable autorización del editor.

CUADRO D. Distribución de t

df	Nivel de significación para la prueba de una sola cola					
	.10	.05	.025	.01	.005	.0005
	Nivel de significación para la prueba de dos colas					
	.20	.10	.05	.02	.01	.001
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.935
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291

FUENTE: El cuadro D es una abreviación del cuadro III de *Statistical Tables for Biological, Agricultural and Medical Research* (ed. 1948), de R. A. Fisher y F. Yates, publicada por Oliver & Boyd, Ltd., Edimburgo y Londres, con la autorización de los autores y editores.

CUADRO E. Valores críticos de r en la prueba de las secuencias $P = .05$

En la prueba de las secuencias de las muestras, cualquier valor de r igual o menor que el que figura en el cuerpo del cuadro es significativo al nivel de .05 con dirección no anticipada, o al nivel .025 con dirección anticipada.

$N_2 \backslash N_1$	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
4			2																
5		2	2	3															
6		2	3	3	3														
7		2	3	3	4	4													
8	2	2	3	3	4	4	5												
9	2	2	3	4	4	5	5	6											
10	2	3	3	4	5	5	6	6	6										
11	2	3	3	4	5	5	6	6	7	7									
12	2	3	4	4	5	6	6	7	7	8	8								
13	2	3	4	4	5	6	6	7	8	8	9	9							
14	2	3	4	5	5	6	7	7	8	8	9	9	10						
15	2	3	4	5	6	6	7	8	8	9	9	10	10	11					
16	2	3	4	5	6	6	7	8	8	9	10	10	11	11	11				
17	2	3	4	5	6	7	7	8	9	9	10	10	11	11	12	12			
18	2	3	4	5	6	7	8	8	9	10	10	11	11	12	12	13	13		
19	2	3	4	5	6	7	8	8	9	10	10	11	12	12	13	13	14	14	
20	2	3	4	5	6	7	8	9	9	10	11	11	12	12	13	13	14	14	15

FUENTE: F. S. Swed y C. Eisenhart, "Tables for Testing Randomness of Grouping in a Sequence of Alternatives", *Annals of Mathematical Statistics*, vol. 14, pp. 83-86, 1943, con la amable autorización de los autores y el editor.

CUADRO F. Cuadro de probabilidades asociadas a valores tan pequeños como los valores observados de U en la prueba de Mann-Whitney (con dirección anticipada) *

$N_2 = 3$					$N_2 = 4$				
$N_1 \backslash U$	1	2	3		$N_1 \backslash U$	1	2	3	4
0	.250	.100	.050		0	.200	.067	.028	.014
1	.500	.200	.100		1	.400	.133	.057	.029
2	.750	.400	.200		2	.600	.267	.114	.057
3		.600	.350		3		.400	.200	.100
4			.500		4		.600	.314	.171
5			.650		5			.429	.243
					6			.571	.343
					7				.443
					8				.557

$N_2 = 5$						$N_2 = 6$					
$N_1 \backslash U$	1	2	3	4	5	$N_1 \backslash U$	1	2	3	4	5
0	.167	.047	.018	.008	.004	0	.143	.036	.012	.005	.002
1	.333	.095	.036	.016	.008	1	.286	.071	.024	.010	.004
2	.500	.190	.071	.032	.016	2	.428	.143	.048	.019	.009
3	.667	.286	.125	.056	.028	3	.571	.214	.083	.033	.015
4		.429	.196	.095	.048	4		.321	.131	.057	.026
5		.571	.286	.143	.075	5		.429	.190	.086	.041
6			.393	.206	.111	6		.571	.274	.129	.063
7			.500	.278	.155	7			.357	.176	.089
8			.607	.365	.210	8			.452	.238	.123
9				.452	.274	9			.548	.305	.165
10				.548	.345	10				.381	.214
11					.421	11				.457	.268
12					.500	12				.545	.331
13					.579	13					.396
						14					.465
						15					.535
						16					.409
						17					.469
						18					.531

FUENTE: H. B. Mann y D. R. Whitney, "On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other", *Annals of Mathematical Statistics*, vol. 18, pp. 52-54, 1947, con la amable autorización de los autores y el editor.

* Si la dirección no ha sido anticipada se duplicarán las probabilidades.

CUADROS

CUADRO F [continuación]

 $N_2 = 7$

$\begin{smallmatrix} N_1 \\ U \end{smallmatrix}$	1	2	3	4	5	6	7
0	.125	.028	.008	.003	.001	.001	.000
1	.250	.056	.017	.006	.003	.001	.001
2	.375	.111	.033	.012	.005	.002	.001
3	.500	.167	.058	.021	.009	.004	.002
4	.625	.250	.092	.036	.015	.007	.003
5		.333	.133	.055	.024	.011	.006
6		.444	.192	.082	.037	.017	.009
7		.556	.258	.115	.053	.026	.013
8			.333	.158	.074	.037	.019
9			.417	.206	.101	.051	.027
10			.500	.264	.134	.069	.036
11			.583	.324	.172	.090	.049
12				.394	.216	.117	.064
13				.464	.265	.147	.082
14				.538	.319	.183	.104
15					.378	.223	.130
16					.438	.267	.159
17					.500	.314	.191
18					.562	.365	.228
19						.418	.267
20						.473	.310
21						.527	.355
22							.402
23							.451
24							.500
25							.549

CUADROS

CUADRO F [conclusión]

 $N_2 = 8$

$\begin{smallmatrix} N_1 \\ U \end{smallmatrix}$	1	2	3	4	5	6	7	8
0	.111	.022	.006	.002	.001	.000	.000	.000
1	.222	.044	.012	.004	.002	.001	.000	.000
2	.333	.089	.024	.008	.003	.001	.001	.000
3	.444	.133	.042	.014	.005	.002	.001	.001
4	.556	.200	.067	.024	.009	.004	.002	.001
5		.267	.097	.036	.015	.006	.003	.001
6		.356	.139	.055	.023	.010	.005	.002
7		.444	.188	.077	.033	.015	.007	.003
8		.556	.248	.107	.047	.021	.010	.005
9			.315	.141	.064	.030	.014	.007
10			.387	.184	.085	.041	.020	.010
11			.461	.230	.111	.054	.027	.014
12			.539	.285	.142	.071	.036	.019
13				.341	.177	.091	.047	.025
14				.404	.217	.114	.060	.032
15				.467	.262	.141	.076	.041
16				.533	.311	.172	.095	.052
17					.362	.207	.116	.065
18					.416	.245	.140	.080
19					.472	.286	.168	.097
20					.528	.331	.198	.117
21						.377	.232	.139
22						.426	.268	.164
23						.475	.306	.191
24						.525	.347	.221
25							.389	.253
26							.433	.287
27							.478	.323
28							.522	.360
29								.399
30								.439
31								.480
32								.520

CUADRO G. Cuadro de valores críticos de U en la prueba de Mann-WhitneyValores críticos de U a $\alpha = .001$ con dirección anticipada, o a $\alpha = .002$ con dirección sin anticipar.

$N_1 \backslash N_2$	9	10	11	12	13	14	15	16	17	18	19	20
1												
2												
3									0	0	0	0
4		0	0	0	1	1	1	2	2	3	3	3
5	1	1	2	2	3	3	4	5	5	6	7	7
6	2	3	4	4	5	6	7	8	9	10	11	12
7	3	5	6	7	8	9	10	11	13	14	15	16
8	5	6	8	9	11	12	14	15	17	18	20	21
9	7	8	10	12	14	15	17	19	21	23	25	26
10	8	10	12	14	17	19	21	23	25	27	29	32
11	10	12	15	17	20	22	24	27	29	32	34	37
12	12	14	17	20	23	25	28	31	34	37	40	42
13	14	17	20	23	26	29	32	35	38	42	45	48
14	15	19	22	25	29	32	36	39	43	46	50	54
15	17	21	24	28	32	36	40	43	47	51	55	59
16	19	23	27	31	35	39	43	48	52	56	60	65
17	21	25	29	34	38	43	47	52	57	61	66	70
18	23	27	32	37	42	46	51	56	61	66	71	76
19	25	29	34	40	45	50	55	60	66	71	77	82
20	26	32	37	42	48	54	59	65	70	76	82	88

FUENTE: D. Aule, "Extended Tables for the Mann-Whitney Statistics", *Bulletin of the Institute of Educational Research at Indiana University*, vol. 1, núm. 2, cuadros 1, 3, 5 y 7, 1953, con la amable autorización del editor; tal como ha sido adaptada por S. Siegel, en *Nonparametric Statistics*, McGraw-Hill Book Company, Nueva York, 1956, cuadro K.

CUADRO G [continuación]

Valores críticos de U a $\alpha = .01$ con dirección anticipada, o a $\alpha = .02$ con dirección sin anticipar.

$N_1 \backslash N_2$	9	10	11	12	13	14	15	16	17	18	19	20
1												
2					0	0	0	0	0	0	1	1
3	1	1	1	2	2	2	3	3	4	4	4	5
4	3	3	4	5	5	6	7	7	8	9	9	10
5	5	6	7	8	9	10	11	12	13	14	15	16
6	7	8	9	11	12	13	15	16	18	19	20	22
7	9	11	12	14	16	17	19	21	23	24	26	28
8	11	13	15	17	20	22	24	26	28	30	32	34
9	14	16	18	21	23	26	28	31	33	36	38	40
10	16	19	22	24	27	30	33	36	38	41	44	47
11	18	22	25	28	31	34	37	41	44	47	50	53
12	21	24	28	31	35	38	42	46	49	53	56	60
13	23	27	31	35	39	43	47	51	55	59	63	67
14	26	30	34	38	43	47	51	56	60	65	69	73
15	28	33	37	42	47	51	56	61	66	70	75	80
16	31	36	41	46	51	56	61	66	71	76	82	87
17	33	38	44	49	55	60	66	71	77	82	88	93
18	36	41	47	53	59	65	70	76	82	88	94	100
19	38	44	50	56	63	69	75	82	88	94	101	107
20	40	47	53	60	67	73	80	87	93	100	107	114

CUADRO G [continuación]

Valores críticos de U a $\alpha = .025$ con dirección anticipada, o a $\alpha = .05$ con dirección sin anticipar.

$N_1 \backslash N_2$	9	10	11	12	13	14	15	16	17	18	19	20
1												
2	0	0	0	1	1	1	1	1	2	2	2	2
3	2	3	3	4	4	5	5	6	6	7	7	8
4	4	5	6	7	8	9	10	11	11	12	13	13
5	7	8	9	11	12	13	14	15	17	18	19	20
6	10	11	13	14	16	17	19	21	22	24	25	27
7	12	14	16	18	20	22	24	26	28	30	32	34
8	15	17	19	22	24	26	29	31	34	36	38	41
9	17	20	23	26	28	31	34	37	39	42	45	48
10	20	23	26	29	33	36	39	42	45	48	52	55
11	23	26	30	33	37	40	44	47	51	55	58	62
12	26	29	33	37	41	45	49	53	57	61	65	69
13	28	33	37	41	45	50	54	59	63	67	72	76
14	31	36	40	45	50	55	59	64	67	74	78	83
15	34	39	44	49	54	59	64	70	75	80	85	90
16	37	42	47	53	59	64	70	75	81	86	92	98
17	39	45	51	57	63	67	75	81	87	93	99	105
18	42	48	55	61	67	74	80	86	93	99	106	112
19	45	52	58	65	72	78	85	92	99	106	113	119
20	48	55	62	69	76	83	90	98	105	112	119	127

CUADRO G [conclusión]

Valores críticos de U a $\alpha = .05$ con dirección anticipada, o a $\alpha = .10$ con dirección sin anticipar.

$N_1 \backslash N_2$	9	10	11	12	13	14	15	16	17	18	19	20
1											0	0
2	1	1	1	2	2	2	3	3	3	4	4	4
3	3	4	5	5	6	7	7	8	9	9	10	11
4	6	7	8	9	10	11	12	14	15	16	17	18
5	9	11	12	13	15	16	18	19	20	22	23	25
6	12	14	16	17	19	21	23	25	26	28	30	32
7	15	17	19	21	24	26	28	30	33	35	37	39
8	18	20	23	26	28	31	33	36	39	41	44	47
9	21	24	27	30	33	36	39	42	45	48	51	54
10	24	27	31	34	37	41	44	48	51	55	58	62
11	27	31	34	38	42	46	50	54	57	61	65	69
12	30	34	38	42	47	51	55	60	64	68	72	77
13	33	37	42	47	51	56	61	65	70	75	80	84
14	36	41	46	51	56	61	66	71	77	82	87	92
15	39	44	50	55	61	66	72	77	83	88	94	100
16	42	48	54	60	65	71	77	83	89	95	101	107
17	45	51	57	64	70	77	83	89	96	102	109	115
18	48	55	61	68	75	82	88	95	102	109	116	123
19	51	58	65	72	80	87	94	101	109	116	123	130
20	54	62	69	77	84	92	100	107	115	123	130	138

CUADRO H. Cuadro de valores críticos de T en la prueba de pares asociados y órdenes provistos de signo, de Wilcoxon

N	Nivel de significación, dirección anticipada		
	.025	.01	.005
	Nivel de significación, dirección sin anticipar		
	.05	.02	.01
6	0	—	—
7	2	0	—
8	4	2	0
9	6	3	2
10	8	5	3
11	11	7	5
12	14	10	7
13	17	13	10
14	21	16	13
15	25	20	16
16	30	24	20
17	35	28	23
18	40	33	28
19	46	38	32
20	52	43	38
21	59	49	43
22	66	56	49
23	73	62	55
24	81	69	61
25	89	77	68

FUENTE: F. Wilcoxon, *Some Rapid Approximate Statistical Procedures*, American Cyanamid Company, Nueva York, 1949, cuadro I, p. 13, con la amable autorización del autor y el editor; tal como ha sido adaptada por S. Siegel en *Nonparametric Statistics*, McGraw-Hill Book Company, Nueva York, 1956, cuadro G.

CUADRO I. Distribución de la χ^2

Probabilidad

df	.99	.98	.95	.90	.80	.70	.50	.30	.20	.10	.05	.02	.01	.001
1	.0157	.0203	.0238	.0288	.0344	.0413	.0509	.0638	.0798	.0993	.1213	.1486	.1846	.2615
2	.0203	.0240	.0270	.0309	.0349	.0398	.0464	.0549	.0658	.0798	.0978	.1193	.1456	.1846
3	.0240	.0270	.0295	.0329	.0363	.0401	.0454	.0524	.0607	.0714	.0854	.1020	.1213	.1548
4	.0270	.0295	.0315	.0344	.0373	.0406	.0449	.0504	.0574	.0668	.0798	.0943	.1113	.1344
5	.0295	.0315	.0330	.0354	.0378	.0406	.0439	.0489	.0554	.0638	.0750	.0879	.1020	.1213
6	.0315	.0330	.0344	.0363	.0383	.0406	.0434	.0479	.0539	.0607	.0700	.0814	.0943	.1093
7	.0330	.0344	.0354	.0373	.0393	.0413	.0441	.0484	.0539	.0607	.0688	.0798	.0913	.1048
8	.0344	.0354	.0363	.0383	.0401	.0424	.0454	.0499	.0549	.0607	.0688	.0798	.0913	.1038
9	.0354	.0363	.0373	.0393	.0413	.0434	.0464	.0509	.0554	.0607	.0688	.0798	.0913	.1020
10	.0363	.0373	.0383	.0401	.0424	.0441	.0474	.0514	.0554	.0607	.0688	.0798	.0913	.1013
11	.0373	.0383	.0393	.0413	.0434	.0454	.0484	.0524	.0564	.0607	.0688	.0798	.0913	.1000
12	.0383	.0393	.0401	.0424	.0441	.0464	.0494	.0534	.0564	.0607	.0688	.0798	.0913	.1000
13	.0393	.0401	.0413	.0434	.0454	.0474	.0504	.0544	.0574	.0607	.0688	.0798	.0913	.1000
14	.0401	.0413	.0424	.0441	.0464	.0484	.0514	.0554	.0584	.0607	.0688	.0798	.0913	.1000
15	.0413	.0424	.0434	.0454	.0474	.0494	.0524	.0564	.0594	.0607	.0688	.0798	.0913	.1000
16	.0424	.0434	.0441	.0464	.0484	.0504	.0534	.0574	.0607	.0607	.0688	.0798	.0913	.1000
17	.0434	.0441	.0454	.0474	.0494	.0514	.0544	.0584	.0607	.0607	.0688	.0798	.0913	.1000
18	.0441	.0454	.0464	.0484	.0504	.0524	.0554	.0594	.0607	.0607	.0688	.0798	.0913	.1000
19	.0454	.0464	.0474	.0494	.0514	.0534	.0564	.0607	.0607	.0607	.0688	.0798	.0913	.1000
20	.0464	.0474	.0484	.0504	.0524	.0544	.0574	.0607	.0607	.0607	.0688	.0798	.0913	.1000
21	.0474	.0484	.0494	.0514	.0534	.0554	.0584	.0607	.0607	.0607	.0688	.0798	.0913	.1000
22	.0484	.0494	.0504	.0524	.0544	.0564	.0594	.0607	.0607	.0607	.0688	.0798	.0913	.1000
23	.0494	.0504	.0514	.0534	.0554	.0574	.0607	.0607	.0607	.0607	.0688	.0798	.0913	.1000
24	.0504	.0514	.0524	.0544	.0564	.0584	.0607	.0607	.0607	.0607	.0688	.0798	.0913	.1000
25	.0514	.0524	.0534	.0554	.0574	.0594	.0607	.0607	.0607	.0607	.0688	.0798	.0913	.1000
26	.0524	.0534	.0544	.0564	.0584	.0607	.0607	.0607	.0607	.0607	.0688	.0798	.0913	.1000
27	.0534	.0544	.0554	.0574	.0594	.0607	.0607	.0607	.0607	.0607	.0688	.0798	.0913	.1000
28	.0544	.0554	.0564	.0584	.0607	.0607	.0607	.0607	.0607	.0607	.0688	.0798	.0913	.1000
29	.0554	.0564	.0574	.0594	.0607	.0607	.0607	.0607	.0607	.0607	.0688	.0798	.0913	.1000
30	.0564	.0574	.0584	.0607	.0607	.0607	.0607	.0607	.0607	.0607	.0688	.0798	.0913	.1000

Para valores mayores de df, la expresión $\sqrt{2\chi^2} - \sqrt{2df} - 1$ puede utilizarse como una desviación normal, con variancia de unidad, recordando que la probabilidad para χ^2 corresponde a la de una sola cola de la curva normal.

FUENTE: El cuadro I es una reimpresión del cuadro IV de *Statistical Tables for Biological, Agricultural and Medical Research* (ed. 1948), de R. A. Fisher y F. Yates, publicada por Oliver & Boyd, Ltd., Edimburgo y Londres, con autorización de los autores y los editores.

CUADRO J. Distribución de F
p = .05

n_1 n_2	1	2	3	4	5	6	8	12	24	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.0	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.00	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83	1.61	1.25
∞	3.84	2.99	2.60	2.37	2.21	2.09	1.94	1.75	1.52	1.00

Los valores de n_1 y n_2 representan los grados de libertad asociados a las estimaciones mayores y menores respectivamente de la variancia.

FUENTE: El cuadro J es una abreviación del cuadro V de *Statistical Tables for Biological, Agricultural and Medical Research* (ed. 1948), de R. A. Fisher y Yates, publicada por Oliver & Boyd, Ltd., Edimburgo y Londres, con autorización de los autores y los editores.

CUADRO J [continuación]
p = .01

n_1 n_2	1	2	3	4	5	6	8	12	24	∞
1	4052	4999	5403	5625	5764	5859	5981	6108	6234	6366
2	98.49	99.01	99.17	99.25	99.30	99.33	99.36	99.42	99.46	99.50
3	34.12	30.81	29.46	28.71	28.24	27.91	27.49	27.05	26.60	26.12
4	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.37	13.93	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.27	9.89	9.47	9.02
6	13.74	10.92	9.78	9.15	8.75	8.47	8.10	7.72	7.31	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.47	6.07	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67	5.28	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.11	4.73	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.71	4.33	3.91
11	9.65	7.20	6.22	5.67	5.32	5.07	4.74	4.40	4.02	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.16	3.78	3.36
13	9.07	6.70	5.74	5.20	4.86	4.62	4.30	3.96	3.59	3.16
14	8.86	6.51	5.56	5.03	4.69	4.46	4.14	3.80	3.43	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.67	3.29	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.55	3.18	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.79	3.45	3.08	2.65
18	8.28	6.01	5.09	4.58	4.25	4.01	3.71	3.37	3.00	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.63	3.30	2.92	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.23	2.86	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.51	3.17	2.80	2.36
22	7.94	5.72	4.82	4.31	3.99	3.76	3.45	3.12	2.75	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.41	3.07	2.70	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.03	2.66	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.32	2.99	2.62	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.29	2.96	2.58	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.26	2.93	2.55	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.23	2.90	2.52	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.20	2.87	2.49	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84	2.47	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.66	2.29	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.50	2.12	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.66	2.34	1.95	1.38
∞	6.64	4.60	3.78	3.32	3.02	2.80	2.51	2.18	1.79	1.00

Los valores de n_1 y n_2 representan los grados de libertad asociados a las estimaciones mayores y menores respectivamente de la variancia.

CUADRO J [conclusión]
p = .001

n_1 n_2	1	2	3	4	5	6	8	12	24	∞
1	405284	500000	540379	562500	576405	585937	598144	610607	623497	636819
2	998.5	999.0	999.2	999.2	999.3	999.3	999.4	999.4	999.5	999.5
3	167.5	148.5	141.1	137.1	134.6	132.8	130.6	128.3	125.9	123.5
4	74.14	61.25	56.18	53.44	51.71	50.53	49.00	47.41	45.77	44.05
5	47.04	36.61	33.20	31.09	29.75	28.84	27.64	26.42	25.14	23.78
6	35.51	27.00	23.70	21.90	20.81	20.03	19.03	17.99	16.89	15.75
7	29.22	21.69	18.77	17.19	16.21	15.52	14.63	13.71	12.73	11.59
8	25.42	18.49	15.83	14.39	13.49	12.86	12.04	11.19	10.30	9.34
9	22.86	16.39	13.90	12.56	11.71	11.13	10.37	9.57	8.72	7.81
10	21.04	14.91	12.55	11.28	10.48	9.92	9.20	8.45	7.64	6.76
11	19.69	13.81	11.56	10.35	9.58	9.05	8.35	7.63	6.85	6.00
12	18.64	12.97	10.80	9.63	8.89	8.33	7.71	7.00	6.25	5.42
13	17.81	12.31	10.21	9.07	8.35	7.86	7.21	6.52	5.78	4.97
14	17.14	11.78	9.73	8.62	7.92	7.43	6.80	6.13	5.41	4.60
15	16.59	11.34	9.34	8.25	7.57	7.09	6.47	5.81	5.10	4.31
16	16.12	10.97	9.00	7.94	7.27	6.81	6.19	5.55	4.85	4.06
17	15.72	10.66	8.73	7.68	7.02	6.56	5.96	5.32	4.63	3.85
18	15.38	10.39	8.49	7.46	6.81	6.35	5.76	5.13	4.45	3.67
19	15.08	10.13	8.28	7.26	6.61	6.18	5.59	4.97	4.29	3.52
20	14.82	9.95	8.10	7.10	6.46	6.02	5.44	4.82	4.15	3.38
21	14.59	9.77	7.94	6.95	6.32	5.88	5.31	4.70	4.03	3.26
22	14.38	9.61	7.80	6.81	6.19	5.76	5.19	4.58	3.92	3.15
23	14.19	9.47	7.67	6.69	6.08	5.65	5.09	4.48	3.82	3.05
24	14.03	9.34	7.55	6.59	5.98	5.55	4.99	4.39	3.74	2.97
25	13.88	9.22	7.45	6.49	5.88	5.46	4.91	4.31	3.66	2.89
26	13.74	9.12	7.33	6.41	5.80	5.38	4.83	4.24	3.59	2.82
27	13.61	9.02	7.27	6.33	5.73	5.31	4.76	4.17	3.52	2.75
28	13.50	8.93	7.19	6.25	5.66	5.24	4.69	4.11	3.46	2.70
29	13.39	8.85	7.12	6.19	5.59	5.18	4.64	4.05	3.41	2.64
30	13.29	8.77	7.05	6.12	5.53	5.12	4.58	4.00	3.36	2.59
40	12.61	8.25	6.60	5.70	5.13	4.73	4.21	3.64	3.01	2.23
60	11.97	7.76	6.17	5.31	4.76	4.37	3.87	3.31	2.69	1.90
120	11.38	7.31	5.79	4.95	4.42	4.04	3.55	3.02	2.40	1.56
∞	10.83	6.91	5.42	4.62	4.10	3.74	3.27	2.74	2.13	1.00

Los valores de n_1 y n_2 representan los grados de libertad asociados a las estimaciones mayores y menores respectivamente de la variancia.

CUADRO K. Valores de z para valores dados de r

r	.000	.001	.002	.003	.004	.005	.006	.007	.008	.009
.000	.0000	.0010	.0020	.0030	.0040	.0050	.0060	.0070	.0080	.0090
.010	.0100	.0110	.0120	.0130	.0140	.0150	.0160	.0170	.0180	.0190
.020	.0200	.0210	.0220	.0230	.0240	.0250	.0260	.0270	.0280	.0290
.030	.0300	.0310	.0320	.0330	.0340	.0350	.0360	.0370	.0380	.0390
.040	.0400	.0410	.0420	.0430	.0440	.0450	.0460	.0470	.0480	.0490
.050	.0501	.0511	.0521	.0531	.0541	.0551	.0561	.0571	.0581	.0591
.060	.0601	.0611	.0621	.0631	.0641	.0651	.0661	.0671	.0681	.0691
.070	.0701	.0711	.0721	.0731	.0741	.0751	.0761	.0771	.0782	.0792
.080	.0802	.0812	.0822	.0832	.0842	.0852	.0862	.0872	.0882	.0892
.090	.0902	.0912	.0922	.0933	.0943	.0953	.0963	.0973	.0983	.0993
.100	.1003	.1013	.1024	.1034	.1044	.1054	.1064	.1074	.1084	.1094
.110	.1105	.1115	.1125	.1135	.1145	.1155	.1165	.1175	.1185	.1195
.120	.1206	.1216	.1226	.1236	.1246	.1257	.1267	.1277	.1287	.1297
.130	.1308	.1318	.1328	.1338	.1348	.1358	.1368	.1379	.1389	.1399
.140	.1409	.1419	.1430	.1440	.1450	.1460	.1470	.1481	.1491	.1501
.150	.1511	.1522	.1532	.1542	.1552	.1563	.1573	.1583	.1593	.1604
.160	.1614	.1624	.1634	.1644	.1655	.1665	.1676	.1686	.1696	.1706
.170	.1717	.1727	.1737	.1748	.1758	.1768	.1779	.1789	.1799	.1810
.180	.1820	.1830	.1841	.1851	.1861	.1872	.1882	.1892	.1903	.1913
.190	.1923	.1934	.1944	.1954	.1965	.1975	.1986	.1996	.2007	.2017
.200	.2027	.2038	.2048	.2059	.2069	.2079	.2090	.2100	.2111	.2121
.210	.2132	.2142	.2153	.2163	.2174	.2184	.2194	.2205	.2215	.2226
.220	.2237	.2247	.2258	.2268	.2279	.2289	.2300	.2310	.2321	.2331
.230	.2342	.2353	.2363	.2374	.2384	.2395	.2405	.2416	.2427	.2437
.240	.2448	.2458	.2469	.2480	.2490	.2501	.2511	.2522	.2533	.2543
.250	.2554	.2565	.2575	.2586	.2597	.2608	.2618	.2629	.2640	.2650
.260	.2661	.2672	.2682	.2693	.2704	.2715	.2726	.2736	.2747	.2758
.270	.2769	.2779	.2790	.2801	.2812	.2823	.2833	.2844	.2855	.2866
.280	.2877	.2888	.2898	.2909	.2920	.2931	.2942	.2953	.2964	.2975
.290	.2986	.2997	.3008	.3019	.3029	.3040	.3051	.3062	.3073	.3084
.300	.3095	.3106	.3117	.3128	.3139	.3150	.3161	.3172	.3183	.3195
.310	.3206	.3217	.3228	.3239	.3250	.3261	.3272	.3283	.3294	.3305
.320	.3317	.3328	.3339	.3350	.3361	.3372	.3384	.3395	.3406	.3417
.330	.3428	.3439	.3451	.3462	.3473	.3484	.3495	.3507	.3518	.3530
.340	.3541	.3552	.3564	.3575	.3586	.3597	.3609	.3620	.3632	.3643
.350	.3654	.3666	.3677	.3689	.3700	.3712	.3723	.3734	.3746	.3757
.360	.3769	.3780	.3792	.3803	.3815	.3826	.3838	.3850	.3861	.3873
.370	.3884	.3896	.3907	.3919	.3931	.3942	.3954	.3966	.3977	.3989
.380	.4001	.4012	.4024	.4036	.4047	.4059	.4071	.4083	.4094	.4106
.390	.4118	.4130	.4142	.4153	.4165	.4177	.4189	.4201	.4213	.4225
.400	.4236	.4248	.4260	.4272	.4284	.4296	.4308	.4320	.4332	.4344
.410	.4356	.4368	.4380	.4392	.4404	.4416	.4429	.4441	.4453	.4465
.420	.4477	.4489	.4501	.4513	.4526	.4538	.4550	.4562	.4574	.4587
.430	.4599	.4611	.4623	.4636	.4648	.4660	.4672	.4685	.4697	.4710
.440	.4722	.4735	.4747	.4760	.4772	.4784	.4797	.4809	.4822	.4835
.450	.4847	.4860	.4872	.4885	.4897	.4910	.4923	.4935	.4948	.4961
.460	.4973	.4986	.4999	.5011	.5024	.5037	.5049	.5062	.5075	.5088
.470	.5101	.5114	.5126	.5139	.5152	.5165	.5178	.5191	.5204	.5217
.480	.5230	.5243	.5256	.5279	.5292	.5305	.5318	.5331	.5344	.5357
.490	.5361	.5374	.5387	.5400	.5413	.5427	.5440	.5453	.5466	.5480

FUENTE: Albert E. Waugh, *Statistical Tables and Problems*, McGraw-Hill Book Company, Nueva York, 1952, cuadro A11, pp. 40-41, con la amable autorización del autor y el editor.

CUADRO K [conclusión]

r	.000	.001	.002	.003	.004	.005	.006	.007	.008	.009
500	5493	5506	5520	5533	5547	5560	5573	5587	5600	5614
510	5627	5641	5654	5668	5681	5695	5709	5722	5736	5750
520	5763	5777	5791	5805	5818	5832	5846	5860	5874	5888
530	5901	5915	5929	5943	5957	5971	5985	5999	6013	6027
540	6042	6056	6070	6084	6098	6112	6127	6141	6155	6170
550	6184	6198	6213	6227	6241	6256	6270	6285	6299	6314
560	6328	6343	6358	6372	6387	6401	6416	6431	6446	6460
570	6475	6490	6505	6520	6535	6550	6565	6579	6594	6610
580	6625	6640	6655	6670	6685	6700	6715	6731	6746	6761
590	6777	6792	6807	6823	6838	6854	6869	6885	6900	6916
600	6931	6947	6963	6978	6994	7010	7026	7042	7057	7073
610	7089	7105	7121	7137	7153	7169	7185	7201	7218	7234
620	7250	7266	7282	7299	7315	7332	7348	7364	7381	7398
630	7414	7431	7447	7464	7481	7497	7514	7531	7548	7565
640	7582	7599	7616	7633	7650	7667	7684	7701	7718	7736
650	7753	7770	7788	7805	7823	7840	7858	7875	7893	7910
660	7928	7946	7964	7981	7999	8017	8035	8053	8071	8089
670	8107	8126	8144	8162	8180	8199	8217	8236	8254	8273
680	8291	8310	8328	8347	8366	8385	8404	8423	8442	8461
690	8480	8499	8518	8537	8556	8576	8595	8614	8634	8653
700	8673	8693	8712	8732	8752	8772	8792	8812	8832	8852
710	8872	8892	8912	8933	8953	8973	8994	9014	9035	9056
720	9076	9097	9118	9139	9160	9181	9202	9223	9245	9266
730	9287	9309	9330	9352	9373	9395	9417	9439	9461	9483
740	9505	9527	9549	9571	9594	9616	9639	9661	9684	9707
750	9730	9752	9775	9799	9822	9845	9868	9892	9915	9939
760	9962	9986	1.0010	1.0034	1.0058	1.0082	1.0106	1.0130	1.0154	1.0179
770	1.0203	1.0228	1.0253	1.0277	1.0302	1.0327	1.0352	1.0378	1.0403	1.0428
780	1.0454	1.0479	1.0505	1.0531	1.0557	1.0583	1.0609	1.0635	1.0661	1.0688
790	1.0714	1.0741	1.0768	1.0795	1.0822	1.0849	1.0876	1.0903	1.0931	1.0958
800	1.0986	1.1014	1.1041	1.1070	1.1098	1.1127	1.1155	1.1184	1.1212	1.1241
810	1.1270	1.1299	1.1329	1.1358	1.1388	1.1417	1.1447	1.1477	1.1507	1.1538
820	1.1568	1.1599	1.1630	1.1660	1.1692	1.1723	1.1754	1.1786	1.1817	1.1849
830	1.1870	1.1913	1.1946	1.1979	1.2011	1.2044	1.2077	1.2111	1.2144	1.2178
840	1.2212	1.2246	1.2280	1.2315	1.2349	1.2384	1.2419	1.2454	1.2490	1.2526
850	1.2561	1.2598	1.2634	1.2670	1.2708	1.2744	1.2782	1.2819	1.2857	1.2895
860	1.2934	1.2972	1.3011	1.3050	1.3089	1.3129	1.3168	1.3209	1.3249	1.3290
870	1.3331	1.3372	1.3414	1.3456	1.3498	1.3540	1.3583	1.3626	1.3670	1.3714
880	1.3758	1.3802	1.3847	1.3892	1.3938	1.3984	1.4030	1.4077	1.4124	1.4171
890	1.4219	1.4268	1.4316	1.4366	1.4415	1.4465	1.4516	1.4566	1.4618	1.4670
900	1.4722	1.4775	1.4828	1.4883	1.4937	1.4992	1.5047	1.5103	1.5160	1.5217
910	1.5275	1.5334	1.5393	1.5453	1.5513	1.5574	1.5636	1.5699	1.5762	1.5825
920	1.5890	1.5956	1.6022	1.6089	1.6157	1.6226	1.6296	1.6366	1.6438	1.6510
930	1.6584	1.6659	1.6734	1.6811	1.6888	1.6967	1.7047	1.7129	1.7211	1.7295
940	1.7380	1.7467	1.7555	1.7645	1.7736	1.7828	1.7923	1.8019	1.8117	1.8216
950	1.8318	1.8421	1.8527	1.8635	1.8745	1.8857	1.8972	1.9090	1.9210	1.9333
960	1.9459	1.9588	1.9721	1.9857	1.9996	2.0140	2.0287	2.0439	2.0595	2.0756
970	2.0923	2.1095	2.1273	2.1457	2.1649	2.1847	2.2054	2.2269	2.2494	2.2729
980	2.2976	2.3223	2.3507	2.3796	2.4101	2.4426	2.4774	2.5147	2.5550	2.5988
990	2.6497	2.6996	2.7587	2.8257	2.9031	2.9945	3.1063	3.2504	3.4534	3.8002

r
.9999 4.95172
.99999 6.10303

INDICE DE FIGURAS

IV.1. Histograma de intervalos iguales	61
IV.2. Histograma de intervalos desiguales y alturas proporcionales a las frecuencias	62
IV.3. Histograma de intervalos desiguales y áreas proporcionales a las frecuencias	62
IV.4. Polígono de frecuencia	63
IV.5. Ojiva que representa una distribución de frecuencia acumulativa	64
V.1. Relación entre la asimetría y las posiciones relativas de la media y la mediana	83
V.2. Una distribución bimodal	86
VII.1. Comparaciones de curvas lisas con histogramas de amplitudes diferentes de intervalo	105
VII.2. Comparación de las áreas debajo de la curva y debajo del rectángulo	106
VII.3. Forma general de la curva normal	107
VII.4. Comparación de curvas normales de igual desviación estándar pero de medias diferentes	108
VII.5. Comparación de dos curvas normales de medias iguales pero con desviaciones estándar diferentes	109
VII.6. Comparación de una curva normal con curvas de su misma desviación estándar pero distintas en cuanto a las cimas	109
VII.7. Áreas debajo de la curva normal	110
VII.8. Comparación de las formas estándar y general de la curva normal	112
VII.9. Curva normal, con porción achurada representando el área en una sola cola	113
VII.10. Curva normal, con porciones achuradas presentando áreas en ambas colas	114
VII.11. Curva normal con porción achurada, representando el área entre dos ordenadas	114
IX.1. Oscilación de la proporción de éxitos aproximándose al límite de .50	130
IX.2. Representación geométrica de probabilidades, con áreas proporcionales a $P(A)$, $P(B)$ y $P(A \& B)$	136

X.1. Comparación de las regiones críticas de pruebas de una sola cola y de dos colas, empleando el nivel de significación de .05	174
XI.1. Comparación de las distribuciones normales de muestreo para muestras de tamaño diferente	188
XI.2. Comparación entre las distribuciones de la población y de la muestra	189
XI.3. Distribución de la población de las probabilidades de obtener caras de 1, 2, 3, 4, 5 o 6 con un dado perfecto	192
XI.4. Distribución de muestreo de las medias de las caras, con dados perfectos y muestras de tamaño 2	192
XI.5. Distribución de muestreo de las medias de las caras, con dados perfectos y muestras de tamaño 3	193
XI.6. Distribución normal de muestras, con área achurada representando una región crítica de una sola cola al nivel de significación de .05	197
XII.1. Comparación de las distribuciones de muestreo de una estimación sesgada, con alta eficiencia, y una estimación no sesgada, de eficacia menor	214
XII.2. Comparación de intervalos de confianza con la distribución de muestreo de la media, mostrando por qué los intervalos de confianza del 95 por ciento comprenden μ el 95 por ciento de las veces	217
XII.3. Distribución de intervalos variables de confianza con respecto a un valor fijo del parámetro μ	218
XII.4. Comparación de un intervalo de confianza del 95 por ciento con pruebas de hipótesis al nivel de .05 mostrando el no rechazo de la media hipotética μ_1 , que queda dentro del intervalo, y el descarte de la μ_2 hipotética, que queda fuera del intervalo	220
XIV.1. Funciones de potencia para pruebas de dos colas, con $\alpha = .05$, para muestras de tamaño variable	258
XIV.2. Derivación de la fuerza como función de $(\mu - \mu_0)$	260
XIV.3. Comparación de funciones de potencia para pruebas de una y dos colas, con $\alpha = .05$. a) Rechace si $Z > 1.645$. b) Rechace si $Z < -1.645$. c) Rechace si $Z > 1.96$ o si $Z < -1.96$	261
XVII.1. Forma general de la regresión de Y sobre X , o curso de las medias de los valores de Y para valores fijos de X	380

XVII.2. La ecuación lineal de regresión, mostrando interpretaciones geométricas de α y β	382
XVII.3. La distribución normal bivariable	385
XVII.4. Diagrama de dispersión y recta de mínimos cuadrados	386
XVII.5. Ecuación de mínimos cuadrados, que minimiza las sumas de los cuadrados de las distancias verticales y estima la regresión de Y sobre X	388
XVII.6. Diagrama de dispersión y recta de mínimos cuadrados para los datos del cuadro XVII.1	392
XVII.7. Diagrama de dispersión que muestra las diferentes fuerzas y direcciones de las relaciones entre X y Y	393
XVII.8. Diagrama de dispersión de una relación no lineal perfecta, en que $r = 0$	395
XVII.9. Diagramas de dispersión que muestran los efectos posibles de valores extremos de X	398
XVII.10. Diagrama de dispersión que no muestra relación alguna dentro de un recorrido limitado de variación de X , pero con relación positiva sobre el recorrido total.	399
XVII.11. Representación geométrica que muestra las desviaciones respecto de la media \bar{Y} como una suma de desviaciones respecto de la recta de mínimos cuadrados y desviaciones de la recta de mínimos cuadrados respecto de la \bar{Y}	408
XVIII.1. Representación geométrica del hecho de que la hipótesis de $\beta = 0$ es equivalente a la hipótesis $\mu_1 = \mu_2 = \dots = \mu_k$	415
XVIII.2. Banda de confianza con respecto de la recta de mínimos cuadrados	422
XVIII.3. Ecuación logarítmica de mínimos cuadrados de la forma $Y = a + b \log X$	427
XVIII.4. Comparación de las desviaciones respecto de la recta de mínimos cuadrados con las desviaciones respecto de las medias de las categorías.	428
XIX.1. Interpretación geométrica de la regresión múltiple de Y sobre X_1 y X_2	449
XIX.2. Plano de mínimos cuadrados, que reduce al mínimo las sumas de las desviaciones al cuadrado en la dimensión vertical Y	450
XIX.3. Rectas de mínimos cuadrados indicando los residuos entre: a) Y y X_2 , y b) entre X_1 y X_2	451

XIX.4. Las seis flechas causales posibles entre X , Y y Z	462
XIX.5. Relaciones causales posibles entre X , Y y Z , tomando a Y como variable dependiente y excluyendo la causalidad en dos direcciones	463
XIX.6. Formas de polinomios de segundo, tercero y cuarto grados	480
XIX.7. Datos hipotéticos con una parábola de mejor ajuste	481
XX.1. Datos hipotéticos que indican una débil correlación total entre X y Y , pero correlaciones más fuertes dentro de las categorías de A	493
XX.2. Datos hipotéticos que indican una fuerte correlación total entre X y Y , pero correlaciones más débiles dentro de las categorías de A	494
XX.3. Rectas de pendientes iguales, que indican no interacción	503
XX.4. Comparación entre rectas separadas de mínimos cuadrados y rectas a través de las medias de categorías, pero todas tienen la misma pendiente b_w	505
XX.5. Interpretación geométrica de los cálculos de las medias Y ajustadas	512
XX.6. Interpretación geométrica de las medias ajustadas de Y mediante deslizamiento de las medias de categorías paralelamente a la recta de pendientes b_w	513
XXI.1. Relación entre el error total y los errores de muestreo y no de muestreo	554

INDICE DE CUADROS

III.1. Número de delincuentes y de no delincuentes en dos localidades hipotéticas	44
III.2. Proporciones de delincuentes y de no delincuentes en dos localidades hipotéticas	44
III.3. Distribución de los números y porcentajes de casos tratados por tres agencias hipotéticas de servicios domésticos	46
III.4. Distribución en porcentajes de los casos tratados por tres agencias hipotéticas de servicios domésticos, con los porcentajes dispuestos verticalmente	47
III.5. Distribución en porcentajes de los casos tratados por tres agencias hipotéticas de servicios domésticos, con los porcentajes calculados horizontalmente	48
IV.1. Distribución de la frecuencia, con datos agrupados en intervalos de 5 por ciento	55
IV.2. Distribución de la frecuencia, con datos agrupados en intervalos de 10 por ciento	55
IV.3. Distribución de las frecuencias con datos agrupados en intervalos de 20 por ciento	56
IV.4. Distribución de frecuencia acumulativa	60
V.1. Cálculo de la media de datos agrupados por el método largo	75
V.2. Cálculo de la media de datos agrupados por el método corto	77
V.3. Cálculo de la media de datos agrupados por el método corto y de las desviaciones graduales	78
V.4. Cálculo de la mediana de datos agrupados	79
VI.1. Cálculo de la desviación estándar utilizando datos agrupados	100
XIV.1. Cálculos para la prueba de dos muestras de Smirnov	280
XIV.2. Cálculos de la prueba de Wilcoxon de pares asociados	282
XV.1. Cálculos de la χ -cuadrada	294
XV.2. Cálculo de la χ -cuadrada sirviéndose de la fórmula	295

XV.3. Cálculo de la χ -cuadrada para una tabla de contingencia de 3×3	299
XV.4. Cuadro maestro para correlacionar cuatro variables	322
XV.5. Serie de tablas de contingencia que relacionan dos variables con dos controles simultáneos	323
XV.6.	324
XVI.1. Datos para el análisis de variancia	333
XVI.2. Representación simbólica de los datos para el análisis de la variancia	335
XVI.3. Cálculos para el análisis de la variancia	342
XVI.4. Datos para el análisis de variancia en dos formas	350
XVI.5. Cálculos para el análisis de variancia, de dos formas con prueba de interacción	357
XVI.6. Cálculos para el análisis de variancia en dos formas, con la interacción añadida dentro del término de error	359
XVI.7. Datos y cálculos para el análisis de variancia con rangos de Kruskal-Wallis	366
XVI.8. Datos y cálculos para la prueba de Friedman	369
XVII.1. Datos para un problema de correlación	390
XVII.2. Datos clasificados cruzados para obtener correlaciones de datos agrupados	404
XVII.3. Cálculos de la correlación de datos agrupados	405
XVII.4. Relaciones numéricas entre r , r^2 , $1 - r^2$ y $\sqrt{1 - r^2}$	411
XVIII.1. Prueba de análisis de variancia de la hipótesis $\rho=0$	416
XVIII.2. Prueba de análisis de variancia para el caso de no linealidad	429
XVIII.3. Cálculo del coeficiente de Spearman de la correlación de rango	435
XVIII.4. Datos comparados para el cálculo de la tau de Kendall a partir de datos agrupados	440
XIX.1. Prueba de análisis de variancia para la significación de la correlación múltiple	485
XIX.2. Prueba de análisis de variancia para la significación de la correlación parcial $r_{13.2}$	486
XX.1. Cálculos para el análisis de covariancia	499-501
XX.2. Prueba de análisis de variancia para la interacción	506

XX.3. Prueba de análisis de variancia para la significación de la correlación promedio intraclase ($\rho_{XY.A}$)	509
XX.4. Prueba de análisis de variancia para la significación de las diferencias entre medias ajustadas	515
XXI.1. Datos para calcular estimaciones de parámetros de muestras estratificadas	545
A. Cuadro de cuadrados y raíces cuadradas	565-577
B. Números aleatorios	578-581
C. Áreas bajo la curva normal	582
D. Distribución de t	583
E. Valores críticos de r en la prueba de las secuencias $P = .05$	584
F. Cuadro de probabilidades asociadas a valores tan pequeños como los valores observados de U en la prueba de Mann-Whitney (con dirección anticipada)	585-587
G. Cuadro de valores críticos de U en la prueba de Mann-Whitney	588-591
H. Cuadro de valores críticos de T en la prueba de pares asociados y órdenes provistos de signo, de Wilcoxon	592
I. Distribución de la χ^2	593
J. Distribución de F	594-596
K. Valores de la z para valores dados de r	597-598

ÍNDICE GENERAL

Prefacio	9
--------------------	---

Primera Parte

INTRODUCCIÓN

I. Introducción: objetivos y límites de la estadística	15
I.1. Funciones de la estadística	16
I.2. El lugar de la estadística en el proceso de la investigación	19
I.3. Advertencia	20
Bibliografía	21
II. Teoría, medición y matemáticas	22
II.1. Teoría e hipótesis: definiciones operativas	22
II.2. El nivel de medición: escalas nominales, ordinales y de intervalo	26
II.3. Medición y estadística	32
II.4. Organización del libro	37
Bibliografía	40

Segunda Parte

ESTADÍSTICA DESCRIPTIVA UNIVARIADA

III. Escalas nominales: proporciones, porcentajes y razones	43
III.1. Proporciones	43
III.2. Porcentajes	45
III.3. Razones	49
Bibliografía	52
IV. Escalas de intervalo: distribuciones de frecuencia y representación gráfica	53
IV.1. Distribuciones de frecuencia: agrupamiento de los datos	53
IV.2. Distribuciones de frecuencia acumulativa	60
IV.3. Presentación gráfica: histogramas, polígonos de frecuencia y ojivas	61
Bibliografía	66

ÍNDICE GENERAL

V. Escalas de intervalo: medidas de tendencia central	67
V.1. La media aritmética	67
V.2. La mediana	71
V.3. Cálculo de la media y la mediana de datos agrupados	73
V.4. Comparación de la media y la mediana	81
V.5. Otras medidas de tendencia central	85
V.6. Deciles, cuartiles y percentiles	86
Bibliografía	88
VI. Escalas de intervalo: medidas de dispersión	90
VI.1. El recorrido	90
VI.2. La desviación cuartil	92
VI.3. La desviación media	92
VI.4. La desviación estándar	93
VI.5. El coeficiente de variabilidad	101
VI.6. Otras medidas resumidas	102
Bibliografía	103
VII. La distribución normal	104
VII.1. Distribuciones de frecuencias finitas <i>versus</i> infinitas	104
VII.2. Forma general de la curva normal	107
VII.3. Áreas bajo la curva normal	109
VII.4. Ilustraciones suplementarias del empleo de la tabla normal	113
Bibliografía	116

Tercera Parte

ESTADÍSTICA INDUCTIVA

VIII. Introducción a la estadística inductiva	119
VIII.1. Estadística y parámetros	119
VIII.2. Pasos en la verificación de una hipótesis	120
VIII.3. La falacia de afirmar el consecuente	123
VIII.4. La forma de las hipótesis estadísticas	124
Bibliografía	127
IX. Probabilidad	128
IX.1. Probabilidad <i>a priori</i>	129
IX.2. Propiedades matemáticas de las probabilidades	133

IX.3. Permutas	145
IX.4. Valores esperados	151
IX.5. Independencia y muestreo aleatorio	153
<i>Bibliografía</i>	159
X. Pruebas de hipótesis: la distribución binomial	160
X.1. La distribución de muestreo binomial	160
X.2. Pasos en las pruebas estadísticas	164
X.3. Aplicaciones de la binomial	177
X.4. Extensiones del binomio	181
X.5. Sumario	183
<i>Bibliografía</i>	186
XI. Pruebas de muestras simples que implican medias y proporciones	187
XI.1. Distribución en muestreo de las medias	187
XI.2. Prueba para la media de la población, conociendo σ	194
XI.3. La distribución t de Student	199
XI.4. Pruebas que comportan proporciones	204
<i>Bibliografía</i>	210
XII. Estimación de punto e intervalo	211
XII.1. Estimación del punto	212
XII.2. Estimación del intervalo	215
XII.3. Intervalos de confianza para otros tipos de problemas	221
XII.4. Determinación del tamaño de la muestra	224
<i>Bibliografía</i>	227

Cuarta Parte

ESTADÍSTICAS BIVARIADAS Y MULTIVARIADAS

XIII. Pruebas de dos muestras: diferencia de las medias y las proporciones.	231
XIII.1. Prueba de la diferencia de las medias	231
XIII.2. Diferencia de proporciones	240
XIII.3. Intervalos de confianza	245
XIII.4. Muestras dependientes: pares asociados	246
XIII.5. Comentarios a propósito de los esquemas experimentales y pruebas de significación	248
<i>Bibliografía</i>	255

XIV. Escalas ordinales: pruebas no paramétricas de dos muestras.	256
XIV.1. Fuerza y eficiencia de la fuerza	257
XIV.2. La prueba de las secuencias (<i>runs</i>) de Wald-Wolfowitz	263
XIV.3. La prueba de Mann-Whitney o de Wilcoxon	269
XIV.4. La prueba de Kolmogorov-Smirnov	277
XIV.5. La prueba de Wilcoxon de pares asociados y órdenes provistos de signo	280
XIV.6. Resumen	284
<i>Bibliografía</i>	288
XV. Escalas nominales: problemas de contingencia	289
XV.1. La prueba de la χ -cuadrada	289
XV.2. La prueba exacta de Fisher	301
XV.3. Medidas de la fuerza de la relación	306
XV.4. Control de otras variables	319
<i>Bibliografía</i>	330
XVI. Análisis de la variancia.	332
XVI.1. Análisis simple de la variancia	332
XVI.2. Comparación de medias específicas	343
XVI.3. Análisis bimodal de la variancia	349
XVI.4. Alternativas no paramétricas del análisis de variancia	365
XVI.5. Medidas de asociación: correlación intraclass	370
<i>Bibliografía</i>	376
XVII. Correlación y regresión	377
XVII.1. Regresión lineal y mínimos cuadrados	377
XVII.2. Correlación	393
<i>Bibliografía</i>	413
XVIII. Correlación y regresión [<i>conclusión</i>]	414
XVIII.1. Prueba de significación e intervalos de confianza	414
XVIII.2. Correlación no lineal y regresión	415
XVIII.3. Efectos de los errores de medición	416
XVIII.4. Escalas ordinales: correlación de rangos	417
<i>Bibliografía</i>	418
XIX. Correlación múltiple y parcial	418

INDICE GENERAL

XIX.1. Regresión múltiple y mínimos cuadrados . . .	447
XIX.2. Correlación parcial . . .	451
XIX.3. Correlación parcial e interpretaciones causales . . .	461
XIX.4. Mínimos cuadrados múltiples y los coeficientes beta . . .	469
XIX.5. Correlación múltiple . . .	473
XIX.6. Regresión múltiple y no linealidad . . .	479
XIX.7. Pruebas de significación e intervalos de confianza . . .	484
<i>Bibliografía</i> . . .	489
XX. Análisis de covariancia y variables simuladas . . .	491
XX.1. Relación de dos escalas de intervalo, control de la escala nominal . . .	492
XX.2. Relación de una escala de intervalo y una escala nominal, control de la escala de intervalo . . .	510
XX.3. Extensiones del análisis de covariancia . . .	516
XX.4. Análisis de la variable simulada . . .	517
XX.5. Observaciones finales . . .	521
<i>Bibliografía</i> . . .	526

Quinta Parte

MUESTREO

XXI. Muestreo . . .	531
XXI.1. Muestreo aleatorio sencillo . . .	532
XXI.2. Muestreo sistemático . . .	537
XXI.3. Muestreo estratificado . . .	539
XXI.4. Muestreo por conglomerados . . .	546
XXI.5. Muestreo sin probabilidad . . .	552
XXI.6. Errores no de muestreo y tamaño de la muestra . . .	553
<i>Bibliografía</i> . . .	554

APÉNDICES

I. Resumen de operaciones algebraicas . . .	559
Cuadros . . .	565
<i>Indice de figuras</i> . . .	599
<i>Indice de cuadros</i> . . .	603

Este libro se terminó de imprimir y encuadernar en el mes de junio de 1994 en los talleres de Encuadernación Progreso, S. A. de C. V. (IEPSA), Calz. de San Lorenzo, 244; 09830 México, D. F.
Se tiraron 2 000 ejemplares.

FONDO DE CULTURA ECONÓMICA

ECONOMÍA

- Adelman, Irma Glicman. *Teorías del desarrollo económico.*
Adizes, Ichak. *Autogestión: la práctica yugoslava.*
Aldrighetti, Angelo. *Técnica bancaria.*
Anderson, Charles W. *Política y cambio económico en América Latina.*
Apter, David Edward. *Una teoría política del desarrollo.*
Arrow, Kenneth; Hahn J. y F. H. *Análisis general competitivo.*
Ayza, Juan; Fichet, Gerard y González, Norberto. *Integración económica y sustitución de importaciones en América Latina.*
Bagehot, Walter. *Lombard Street. El mercado monetario de Londres.*
Balassa, Bela A. *Futuro comercial de los países en desarrollo.*
Banco de México, S. A. *La distribución del ingreso en México.*
Banco Interamericano de Desarrollo. *Factores para la integración latinoamericana.*
Bangs, Robert. *Financiamiento del desarrollo económico.*
Baran, Paul Alexander. *La economía política del crecimiento.*
Becker, Gary S. *Teoría económica.*
Benetti, Carlo. *La acumulación en los países capitalistas subdesarrollados.*
Benham, Frederick Charles. *Curso superior de economía.*
Bennett, Hugh Hammond. *Elementos de conservación del suelo.*
Bethel; Atwater; Smith y Stackman. *Organización y dirección industrial.*
Bettelheim, Charles. *Planificación y crecimiento acelerado.*
Bierman, Harold. *Temas de contabilidad de costos y toma de decisiones.*
Bierman, H. y Smidt, Seymour. *El presupuesto de bienes de capital.*
Capstick, Margaret. *La economía de la agricultura.*
Carrillo Arronte, Ricardo. *Ensayo analítico metodológico de planificación interregional en México.*
Ceceña Cervantes, José Luis. *Introducción a la economía política de la planificación económica nacional.*
Coontz, Sidney H. *Teorías de la población y su interpretación económica.*
Cramer, Jan Salomon. *Econometría empírica.*
Crum, William L. y Schumpeter, Joseph A. *Elementos de matemáticas para economistas y estadígrafos.*
Currie, Lauchlin. *Desarrollo económico acelerado.*

- Chandler, Lester Vernon. *Introducción a la teoría monetaria.*
Chenery, Hollis B. y Clark, Paul G. *Economía interindustrial.*
Chevalier, François. *La formación de los latifundios en México.*
Darin-Drabkin, Haim. *La otra sociedad.*
David, Jacques Henry. *La política monetaria.*
Dell, Sidney Samuel. *Bloques de comercio y mercados comunes.*
Dobb, Maurice Herbert. *Economía política y capitalismo.*
Dopfer, Kurt. *La economía del futuro.*
Dunning, John H. *La empresa multinacional.*
Eells, Richard y Clarence, Walton. *Fundamentos conceptuales de los negocios.*
Estey, James Arthur. *Tratado sobre los ciclos económicos.*
Fajnzylber, Fernando y Martínez Tarragó, Trinidad. *Las empresas transnacionales.*
Farrell, Barry R. *América Latina y Canadá frente a la política exterior de los Estados Unidos.*
Ferguson, John Maxwell. *Historia de la economía.*
Fernández y Fernández, Ramón y Acosta, Ricardo. *Política agrícola; ensayos sobre normas para México.*
Ferrer, Aldo. *Economía internacional contemporánea.*
Finley, M. I. *La economía de la antigüedad.*
Firth, Raymond. *Temas de antropología económica.*
Flores, Edmundo. *Dentro y fuera del desarrollo.*
Flores, Edmundo. *Tratado de economía agrícola.*
Flores de la Peña, Horacio. *Los obstáculos al desarrollo económico.*
Flores de la Peña, Horacio. *Teoría y práctica del desarrollo.*
Fox, Willard M. *Investigación de mercados. Interpretación y aplicación.*
Furtado, Celso. *Dialéctica del desarrollo.*
Griffin, Keith B. y Enos, John L. *La planificación en el desarrollo.*
Gutiérrez, Alfredo F. *Los estados financieros y su análisis.*
Hansen, Alvin Harvey. *Política fiscal y ciclo económico.*
Harrod, Roy Forbes. *Vida de John Maynard Keynes.*
Heilbroner, Robert Louis. *La formación de la sociedad económica.*
Hempel, Edward Henry. *Dirección de plantas industriales.*
Hicks, John Richard. *Valor y capital.*
Hinterhuber, Giovanni. *Política de inversiones en la industria.*
Hirschman, Albert O. *Salida, voz y lealtad.*
Jaguaribe, Helio. *Desarrollo económico y político.*
James, Émile. *Historia del pensamiento económico en el siglo xx.*
Jiménez Castro, Wilburg. *Administración pública para el desarrollo integral.*
Jöhr, Walter Adolf y Singer, W. H. *El papel del economista como asesor oficial.*
Kaldor, Nicholas. *Impuesto al gasto.*
Kalecki, Michal. *Economía socialista y mixta.*

Kaplan, Marcos. *Corporaciones públicas multinacionales para el desarrollo y la integración de la América Latina.*

Katz, Jorge M. *Importación de tecnología, aprendizaje e industrialización dependiente.*

Keynes, John Maynard. *Teoría general de la ocupación, el interés y el dinero.*

Kuklinski, Antoni R. *Polos y centros de crecimiento en la planificación regional.*

Kurihara, Kenneth Kenkichi. *Teoría monetaria y política pública.*

Lambert, Denis-Clair y Martin, Jean-Marie. *América Latina. Economías y sociedades.*

Lang, M. F. *El monopolio estatal del mercurio en el México colonial. 1550-1710.*

Lange, Oskar Richard. *Introducción a la econometría.*

Le Breton, Preston P. *Administración general.*

Lee, Robert D. y Johnson, Ronald W. *El gobierno y la economía*

Leemans, A. F. *Cómo reformar la administración pública.*

Lekachman, Robert. *Teoría general de Keynes.*

Levinson, Jerome y Onís, Juan de. *La alianza extraviada.*

Lewis, William Arthur. *Teoría del desarrollo económico.*

Little, Ian Malcolm David; Scitovsky, Tibor y Scott, Maurice. *Industria y comercio en algunos países en desarrollo.*

López Zamora, Emilio. *El agua, la tierra, los hombres de México.*

McCarty, Harold Hull y Lindberg, James B. *Introducción a la geografía económica.*

McFarland, Dalton Edwards. *Administración de personal.*

Macón, Jorge y Merino Mañón, José. *Contribución de mejoras en América Latina.*

Maddison, Angus. *Progreso y política económica en los países en vías de desarrollo.*

Malthus, Thomas Robert. *Principios de economía política.*

Marczewski, Jean. *¿Crisis de la planificación socialista?*

Martínez del Campo, Manuel. *Factores en el proceso de industrialización.*

Martínez de Navarrete, Ifigenia. *Alimentación básica y desarrollo agroindustrial.*

Marx, Carlos. *El capital* (3 vols.)

Meade, James Edward. *Una teoría neoclásica del crecimiento económico.*

Mellor, John Williams. *Economía del desarrollo agrícola.*

Miller, Eric A. J. *Desarrollo integral del medio rural.*

Muñoz Amato, Pedro. *Introducción a la administración pública.*

Myrdal, Gunnar. *Teoría económica y regiones subdesarrolladas.*

Nevin, Edward. *Fondos de capital en los países subdesarrollados.*

Newlyn, Walter T. *Teoría monetaria.*

North, Douglas C. y Miller, Roger LeRoy. *El análisis económico de la usura, el crimen, la pobreza, etcétera.*

Nurkse, Ragnar. *Problemas de formación de capital en los países insuficientemente desarrollados.*

Papanek, Gustav Fritz. *Teoría y práctica de la política del desarrollo.*

Peacock, Alan T. y Shaw, G. K. *La teoría económica de la política fiscal.*

Powelson, John Palen. *Contabilidad económica.*

Prebisch, Raúl. *Introducción a Keynes.*

Rautenstrauch, W. y Williers, R. *El presupuesto en el control de las empresas.*

Reynolds, Clark Winton. *La economía mexicana.*

Ricardo, David. *Cartas, 1810-1815. Obras y correspondencia.* VI.

Ricardo, David. *Cartas, 1821-1823. Obras y correspondencia.* IX.

Ricardo, David. *Principios de economía política y tributación.*

Robinson, Joan. *Introducción a la economía moderna.*

Robinson, Joan. *Ensayos sobre análisis económico.*

Robinson, Joan. *Ensayos sobre la teoría del crecimiento económico.*

Roll, E. *Historia de las doctrinas económicas.*

Rostow, W. W. *Las etapas del crecimiento económico.*

Sáenz Quiroga, Eladio. *Matemáticas para economistas.*

Salinas, Alberto. *La reforma administrativa.*

Sanz de Santamaría, Carlos. *Revolución silenciosa.*

Sayers, Richard Sidney. *La banca moderna.*

Scott, Honor Minturn. *Curso elemental de economía.*

Schell, Ervin Haskell. *Técnica del control ejecutivo.*

Schiavo-Campo, S. y Singer, H. W. *Perspectivas de desarrollo económico.*

Schickele, Rainer. *Tratado de política agrícola.*

Schultz, Theodore William. *La organización económica de la agricultura.*

Schumpeter, Joseph Alois. *Historia del análisis económico.* II.

Schumpeter, Joseph Alois. *Teoría del desenvolvimiento económico.*

Sée, Henri. *Orígenes del capitalismo moderno.*

Seers, Dudley y Joy, Leonard. *El desarrollo de un mundo dividido.*

Sen, Amartya Kumar. *La selección de técnicas.*

Sepúlveda, Bernardo, y Chumacero, Antonio. *La inversión extranjera en México.*

Shackle, George Lennox Sharman. *La naturaleza del pensamiento económico.*

Shackle, George Lennox Sharman. *Epistémica y economía.*

Siegel, Barry N. *Agregados económicos y política pública.*

Sik, Ota. *La tercera vía.*

Silva Herzog, Jesús. *Antología del pensamiento económico.*

Silva Herzog, Jesús. *Historia del pensamiento económico-social. De la Antigüedad al siglo xvi.*

Silva Herzog, Jesús. *El pensamiento económico, social y político de México, 1810-1964.*

Sirkin, Gerald. *Introducción a la teoría macroeconómica.*

Sloan, Pat. *Marx y la economía ortodoxa.*

- Snarely, W. P. *Teoría de los sistemas económicos.*
 Solís Manjarrez, Leopoldo. *Controversias sobre el crecimiento y la distribución.*
 Solow, Robert Merton. *La teoría del crecimiento.*
 Somers, Harold Milton. *Finanzas públicas e ingreso nacional.*
 Spiegelman, Mortimer. *Introducción a la demografía.*
 Stark, Werner. *Historia de la economía en su relación con el desarrollo social.*
 Strachey, John. *El capitalismo contemporáneo.*
 Strachey, John. *El fin del imperio.*
 Sweezy, Paul M. *Teoría del desarrollo capitalista.*
 Tinbergen, Jan. *La planeación del desarrollo.*
 Tobin, James. *Política económica nacional.*
 Tozzi, Glauco. *Economistas griegos y romanos.*
 Trejo Reyes, Saúl. *Industrialización y empleo en México.*
 Urquidí, Víctor L. y Troeller, Ruth R. *El petróleo, la OPEP y la perspectiva internacional.*
 Urquidí, Víctor L. y Thorp, Rosemary. *América Latina en la economía internacional.*
 Vaitsos, Constantine V. *Distribución del ingreso y empresas transnacionales.*
 Vernon, Raymond. *Soberanía en peligro.*
 Villarreal, René. *El desequilibrio externo en la industrialización de México.*
 Weber, Max. *Historia económica general.*
 Weitz, Raanan. *Planeación rural de los países en desarrollo.*
 Weitz, Raanan y Applebaum, Levia. *De campesino a agricultor, una nueva estrategia del desarrollo rural.*
 Wheeler, Joseph Louis y Goldor, Herbert. *Administración práctica de bibliotecas públicas.*
 White, Eduardo J. *Empresas multinacionales latinoamericanas.*
 Wionczek, Miguel S.; Bueno, Gerardo M. y Navarrete, Jorge Eduardo. *La transferencia internacional de tecnología. El caso de México.*
 Wolfe, Marshall. *El desarrollo esquivo.*
 Yudelman, Montague. *El desarrollo agrícola y la integración económica de la América Latina.*
 Zamora, Francisco. *Introducción a la micro y macro dinámica económica.*
 Zamora, Francisco. *La sociedad económica moderna.*

Inventario: 300823
 Fecha: 07/09/97
 Clasificación: 311
 Sig. Top.: B.637e
 Donación: ☐
 Compra: ☒ Pal
 Canje: ☐

BIBLIOTECA CENTRAL - CAMPUS

Devolver puntualmente



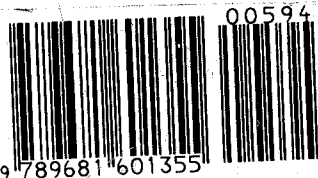
HUBERT M. BLALOCK

ESTADÍSTICA SOCIAL



El estudiante de ciencias sociales que se enfrente, en la etapa inmediatamente anterior a la obtención de su grado, o en la de estudios avanzados, al vasto campo de la realidad y sus problemas innumerables; el maestro que lo acompañe en ese enfrentamiento, y el especialista que eventualmente lo asesore no podrán soslayar los conocimientos de la ciencia de la estadística. Ésta se ha convertido en el instrumento más acabado para conocer de modo directo las coordenadas de un fenómeno, las magnitudes reales o posibles de una situación dada y los rasgos fundamentales de una configuración social, descrita sumariamente en el tiempo y en el espacio en la expresiva cuadrícula estadística. *Estadística social* es un manual completo sobre su tema para el estudiante descrito: le proporciona una introducción que señala dificultades y relaciones entre medida, teoría y estadística; lo capacita para examinar las medidas descriptivas con los métodos adecuados de cómputo que habrán de aplicarse al campo social estudiado, y lo familiariza con algunos de los conceptos básicos de la estadística inductiva.

Diseño: Carlos Haces/Fotografía de Rosa Lilia Martínez G.



FONDO DE CULTURA ECONÓMICA