FUNDAMENTOS DE Estadística EN LA INVESTIGACIÓN SOCIAL

Segunda edición

Jack Levin • William C. Levin

Alfaomega

OXFORD

Este es un libro de estadística concebido para estudiantes de carreras de ciencias sociales como sociología, psicología, ciencias políticas, trabajo social y educación, entre otras, que emplean la estadística en sus estudios diarios. Paso a paso explica los fundamentos de la estadística con un tratamiento claro y comprensible, considerando que la preparación matemática no ha sido tan intensa como en otras; áreas, además, se apoya en numerosos ejemplos desarrollados y ejercicios propuestos.

Dividido en tres partes:

- En la primera explica los métodos más empleados para la descripción, comparación y organización de los datos sin procesar: gráficas, medidas de tendencia central y de dispersión y variabilidad.
- La segunda aborda la curva normal y la generalización de muestras a poblaciones: desviación estándar, el modelo y la realidad, la probabilidad, métodos aleatorios y no aleatorios de muestreo, error estándar de la media, intervalos de confianza.
- La última parte estudia la toma de decisiones, pruebas de significancia, obtención de coeficientes de correlación y una introducción al análisis de regresión.

OXFORD
UNIVERSITY PRESS

Alfaomega Grupo Editor

ISBN 970-15-1054-2

Fundamentos de estadística en la investigación social

Traducción:

VIVIAN DEL VALLE
Facultad de Sociología
Universidad Nacional de Colombia
Bogotá, Colombia

Revisión Técnica:

HAROLDO ELORZA
Facultad de Psicología
Universidad Nacional Autónoma de México

Fundamentos de estadística en la investigación social

Segunda edición

Jack Levin y William C. Levin
Universidad de Northeastern

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Antonio Caso 142, San Rafael,
Delegación Cuauhtémoc, C.P. 06470, México, D.F.
Tel.: 5592 4277, Fax: 5705 3738, e-mail: oxford@oupmex.com.mx

Oxford University Press es un departamento de la Universidad de Oxford.
Promueve el objetivo de la Universidad relativo a la excelencia en la investigación, erudición
y educación mediante publicaciones en todo el mundo en

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong
Karachi Kuala Lumpur Madrid Melbourne Mexico City
Nairobi New Delhi Taipei Toronto Shanghai

Con oficinas en
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan South Korea Poland Portugal Singapore
Switzerland Thailand Turkey Ukraine Vietnam

Oxford es una marca registrada de Oxford University Press en el Reino Unido y otros países.
Publicado en México por Oxford University Press México, S.A. de C.V.

División: Universitaria
Área: Matemáticas

Producción: Antonio Figueredo Hurtado
Portada: Javier Perdomo

FUNDAMENTOS DE ESTADÍSTICA EN LA INVESTIGACIÓN SOCIAL

ISBN 968-6199-36-5

Traducido de la segunda edición en inglés de
ELEMENTARY STATISTICS IN SOCIAL RESEARCH. Workbook
Copyright © 1977, by Harper & Row Publishers, Inc.
ISBN 0-06-3150-12-3

*Alfaomega Grupo Editor es distribuidor exclusivo para todos los países de habla hispana
de esta coedición realizada entre Oxford University Press México, S.A. de C.V.
y Alfaomega Grupo Editor, S.A. de C.V.*

ISBN 970-15-1054-2

Alfaomega Grupo Editor, S.A. de C.V.
Pitágoras 1139, Col. Del Valle, 03100, México, D.F.

Impreso en México 8 9 0 1 2 3 4 5 6 7
Printed in Mexico 0 8 0 7 0 6 0 5 0 4

Esta obra se terminó de imprimir en octubre de 2004 en
Ediciones Culturales, S. A. de C. V.,
Av. 5 de Mayo Núm. 495, Col. Merced Gómez, 01600, México, D.F.,
sobre papel Bond Editor Alta Opacidad de 75 g.

El tiraje fue de 2 000 ejemplares.

Contenido

Págs.

Prefacio XI

Prólogo a la edición en español XIII

1. Razones por las que el investigador social emplea la Estadística 1

La naturaleza de la investigación social 1
¿Por qué probar hipótesis? 2
Las etapas de la investigación social 3
El uso de series de números en la investigación social 3
Funciones de la Estadística 7
Resumen 12

Parte I DESCRIPCION

2. Organización de datos 15

Distribuciones de frecuencia de datos nominales 15
Comparación de las distribuciones 16
Distribuciones de frecuencia simples de datos ordinales y por intervalos 20
Distribuciones de frecuencia agrupadas de datos por intervalos 21
Distribuciones acumuladas 24
Rango percentil 26
Resumen 29
Problemas 30

3. Gráficas 33

Gráficas de sectores 33
Gráficas de barras 34
Polígonos de frecuencia 35
Construcción de gráficas de barra y polígonos de frecuencia 36
La forma de una distribución de frecuencia 37
Resumen 38

4. Medidas de tendencia central 39

La moda 39
La mediana 40
La media 42
Comparación entre la moda, la mediana y la media 44
Obtención de la moda, la mediana y la media de una distribución de frecuencia agrupada 49
Resumen 51
Problemas 52

5. Medidas de dispersión o variabilidad 55

El rango 56
La desviación media 56
La desviación estándar 59
Comparación entre el rango, la desviación media y la desviación estándar 66
Cálculo del rango, de desviación media y la desviación estándar de los datos agrupados 67
Resumen 70
Problemas 70

Parte II DE LA DESCRIPCION A LA TOMA DE DECISIONES

6. La curva normal 75

Características de la curva normal 76
Curvas normales: el modelo y la realidad 76
El área bajo la curva normal 78
Aclarando la desviación estándar: un ejemplo 79
El uso de la Tabla B 81
Puntajes estándar y la curva normal 83

Probabilidad, curva normal 85
Resumen 91
Problemas 91

7. Muestras y poblaciones 93

Métodos de muestreo 94
Error de muestreo 99
Distribución muestral de medias 100
Error estándar de la media 106
Intervalos de confianza 107
Estimación de proporciones 113
Resumen 115
Problemas 116

Parte III LA TOMA DE DECISIONES

8. Comprobación de diferencias entre medias 121

La hipótesis nula: Ninguna diferencia entre las medias 121
La hipótesis de investigación: una diferencia entre medias 122
Distribución muestral de diferencias de medias 123
Contrastación de las hipótesis con la distribución de diferencias 126
Niveles de confianza 130
Error estándar de la diferencia 132
Comparaciones entre muestras pequeñas 136
Comparaciones entre muestras de diferente tamaño 140
Comparación de la misma muestra medida dos veces 143
Requisitos para el uso de los puntajes z y la razón t 145
Resumen 146
Problemas 146

9. Análisis de varianza 150

La lógica del análisis de varianza 151
Las sumas de cuadrados 152
La media cuadrática 158
La razón F 159
Una comparación múltiple de medias 164
Requisitos para el uso de la razón F 166
Resumen 167
Problemas 167

10. Chi cuadrada y otras pruebas no paramétricas 169

Chi cuadrada como prueba de significancia 170
Cálculo de la chi cuadrada 171
Cómo buscar las frecuencias esperadas 173
Una fórmula 2 × 2 para calcular la chi cuadrada 178
Correcciones para pequeñas frecuencias esperadas 179
Comparando varios grupos 181
Requisitos para el uso de la chi cuadrada 185
La prueba de la mediana 186
Análisis de varianza de dos direcciones por rangos de Friedman 189
Análisis de varianza en una dirección por rangos de Kruskal-Wallis 192
Resumen 194
Problemas 195

11. Correlación 200

La fuerza de la correlación 200
Dirección de la correlación 201
Correlación curvilínea 202
El coeficiente de correlación 203
Un coeficiente de correlación para datos por intervalos 204
Una fórmula para calcular el *r* de Pearson 207
Análisis de regresión 212
Coeficiente de correlación para los datos ordenados 217
La gamma de Goodman y Kruskal 223
Coeficiente de correlación para datos nominales organizados en una tabla de 2 × 2 231
Coeficiente de correlación para datos nominales mayores que una tabla de 2 × 2 . . . 233
Resumen 236
Problemas 237

12. Aplicación de métodos estadísticos a problemas de investigación 241

Situaciones de investigación 242
Solución a las investigaciones 250

APENDICES 254

Apéndice A Revisión de algunos aspectos fundamentales de matemáticas 256

Trabajando con decimales 256

Empleando los números negativos 258
Cómo buscar raíces cuadradas con la tabla A 259

Apéndice B Tablas 261

Apéndice C Lista de fórmulas 291

Respuestas a los problemas seleccionados 296

Referencias 301

Indice 303

Prefacio

El objetivo de esta segunda edición de *Fundamentos de Estadística en la Investigación Social* es introducir a los alumnos de Sociología y campos afines en la Estadística. El texto está especialmente diseñado para aquellos estudiantes de Sociología, Ciencias Políticas, Trabajo Social, Psicología, Administración Pública y Educación, quienes no han tenido una preparación intensiva en Matemáticas y deben tomar su primer curso de Estadística.

El libro *no* pretende ser una obra de referencia exhaustiva, ni debe considerarse como el texto más adecuado para cursos avanzados en métodos estadísticos. Por el contrario, fue escrito y adaptado para satisfacer la manifiesta necesidad de un tratamiento comprensible y significativo de la Estadística básica. Con este fin, para cada tema importante del texto se presentan ejemplos detallados y explicados paso a paso de los procedimientos estadísticos.

El volumen se ha dividido en tres partes: La primera parte (Capítulos 2-5) enseña al estudiante algunos de los métodos más utilizados para la descripción y comparación de los datos sin procesar. La segunda parte (Capítulos 6-7) es una etapa de tránsito, debido a que conduce al estudiante del tema de la curva normal, como importante recurso descriptivo, al próximo capítulo en que la curva normal se emplea como base para la generalización de las muestras a las poblaciones. La tercera parte, que también sigue la línea de preparación para la toma de decisiones, contiene varias pruebas de significancia bien conocidas, procedimientos para la obtención de coeficientes de correlación y una introducción al análisis de regresión. En esta edición se han realizado algunos cambios importantes en relación con la primera edición. Se ha dado mayor énfasis a la estadística no paramétrica (Capítulo 10), al análisis del rango percentil, probabilidad, comparación múltiple de medias siguiendo un análisis de varianza, gamma y *r* de Pearson. Para establecer las aplicaciones de la estadística a la investigación, se ha agregado un nuevo capítulo (12), en el cual se pide a los estudiantes que seleccionen los pro-

cedimientos estadísticos apropiados a las distintas situaciones que se presentan en la investigación. Se ha incrementado el número de ejercicios al final de los capítulos. Finalmente, los apéndices se han aumentado para incluir un repaso de los fundamentos de las matemáticas y una lista de fórmulas.

Varias personas han contribuido de una manera significativa al desarrollo de esta segunda edición. El profundo análisis de Kenneth Pollinger en *Contemporary Sociology* suministró las bases para varias mejoras y adiciones. Estoy agradecido con Richard Sprunthall y con sus estudiantes del American International College (especialmente con Lynn Armold, Cheryl Janes, Jim Lynch, Claire Nolen y Gary Zera), quienes me hicieron notar la presencia, en la edición anterior, de varias inexactitudes y errores de apreciación. Debo especial agradecimiento a las siguientes personas por sus análisis críticos a mis revisiones: George Bowlby, James Elliot, Roy Hansen, C. Lincoln Johnson, Carol Owen, Lawrence Rosen, Norman Roth, Ellen Bouchard Ryan y Larry Siegel. También estoy agradecido con Suzanne Johnson y Michael Wesbuch por los comentarios y sugerencias que nos han hecho en forma espontánea.

Finalmente, agradezco al Ejecutivo Literario del difunto Sir Roland A. Fisher, F.R.S., a Frank Yates, F.R.S., y a Oliver y Boyd Edinburgh por el permiso concedido para reproducir las Tablas III, IV, V y VI de su libro *Statistical Tables for Biological, Agricultural and Medical Research.*

Jack Levin

Prólogo a la edición en español

Nuestro objetivo, al traducir este libro de texto, es introducir en la metodología estadística al estudiante de Ciencias Sociales. La precisión, claridad y sencillez reflejadas en esta obra, son tres de las características más importantes del profesor Jack Levin. Estas cualidades pedagógicas son esenciales para una primera experiencia con la Estadística. Particularmente, pensamos en el caso de los estudiantes de cualquier área social que no poseen una base matemática sólida, pero que necesariamente deberán aplicar la Estadística en el curso de sus estudios y durante toda su actividad profesional.

No es aconsejable considerar a éste como un libro de texto para cursos avanzados de Estadística, pues fue diseñado para los dos primeros cursos elementales (*Estadística descriptiva* y *Estadística inferencial*) que sirven de fundamento en todas las áreas de las Ciencias Sociales.

En nuestra opinión se trata de un libro de gran valor didáctico para Latinoamérica que todo estudiante de Ciencias Sociales debe utilizar en su aprendizaje de los métodos estadísticos. Los ejemplos son muy actuales, amenos e interesantes; además se desarrollan en forma detallada, lo cual le imprime un valor pedagógico inapreciable.

Es importante mencionar que esta segunda edición revisada, del libro del profesor Levin, se realizó en 1977, después de treinta y seis años de experiencia pedagógica en el campo de la Estadística.

Sólo nos queda agradecer a los editores de HARLA su dedicación y esfuerzo para la publicación de esta obra, con lo cual se satisfacen las necesidades actuales de los estudiantes latinoamericanos.

Vivian del Valle y
Haroldo Elorza

1

Razones por las que el investigador social emplea la estadística

Todos nosotros tenemos algo de investigadores sociales. Casi diariamente hacemos "sabios pronósticos" relativos a los acontecimientos futuros de nuestra vida con el fin de predecir lo que sucederá ante nuevas situaciones o experiencias. A medida que aparecen estas situaciones, con frecuencia apoyamos o confirmamos nuestras ideas; otras veces, sin embargo, no somos tan afortunados y debemos experimentar desagradables consecuencias.

Tomemos en consideración algunos ejemplos familiares: podríamos invertir en el mercado de valores, votar por un candidato político que promete resolver problemas internos, apostar a los caballos, tomar medicinas para reducir las molestias de una gripe, jugar a los dados en un casino, tratar de conocer psicológicamente un poco a nuestros maestros en relación con un examen o aceptar una cita con un desconocido, confiando en la palabra de un amigo.

Algunas veces ganamos; algunas veces perdemos. Así, podríamos hacer una buena inversión en el mercado de valores, pero arrepentirnos de nuestra decisión electoral; ganar dinero en los juegos de azar, pero descubrir que nos hemos equivocado al tomar el remedio para nuestra enfermedad; resolver bien el examen, pero tener una desagradable sorpresa al asistir a la cita con el desconocido, y así sucesivamente. Desafortunadamente, es cierto que no todas nuestras predicciones diarias estarán apoyadas por la experiencia.

LA NATURALEZA DE LA INVESTIGACION SOCIAL

De una manera un tanto semejante, el científico social tiene ideas acerca de la naturaleza de la realidad social (a las cuales llama *hipótesis*), y, frecuentemente, comprueba sus ideas por medio de la investigación sistemática. Por ejemplo, podría presentar la hipótesis de que los niños socialmente aislados ven más televisión que

los niños que están bien integrados con sus grupos afines; podría hacer una encuesta en la cual se pregunte a ambos grupos de niños, los socialmente aislados y los bien integrados, acerca del tiempo que dedican a ver televisión. También podría plantear la hipótesis de que las familias, en donde sólo existe el padre y falta la madre o existe la madre y falta el padre, generan más delincuencia que las familias que cuentan con la presencia del padre y de la madre; podría, por último proceder a entrevistar muestras de delincuentes y no delincuentes para determinar si uno o ambos padres estuvieron presentes en su formación familiar.

Así, de un modo similar a su contraparte en las ciencias físicas, el investigador social con frecuencia investiga para comprender mejor los problemas y acontecimientos que se presentan en su especialidad. La investigación social toma muchas formas y puede ser empleada para investigar una amplia variedad de problemas. El investigador puede participar en la observación de una pandilla de delincuentes, en una encuesta de muestras de simpatías y de antipatías políticas, en un análisis de valores de la prensa clandestina o en un experimento para determinar los efectos que se producen al obligar a las familias a abandonar sus hogares y establecerlos en otros sitios con el fin de ceder este su espacio a las autopistas recientemente construidas.

¿POR QUE PROBAR HIPOTESIS?

Generalmente es conveniente, cuando no necesario, comprobar sistemáticamente nuestras hipótesis acerca de la naturaleza de la realidad social, aun aquéllas que parezcan lógicas, verdaderas o evidentes por sí mismas. Nuestras diarias "pruebas" de sentido común se basan generalmente en preconcepciones muy estrechas, cuando no parcializadas, y en experiencias personales que pueden conducirnos a aceptar conclusiones sin valor respecto a la naturaleza de los fenómenos sociales. Para demostrar este punto examinemos las siguientes hipótesis que fueron comprobadas en un gran número de soldados durante la Segunda Guerra Mundial. ¿Podría usted "predecir" estos resultados con base en sus experiencias cotidianas? ¿Cree que era necesario comprobarlos o parecen demasiado obvios y evidentes por sí mismos para una investigación sistemática?

1. Los hombres mejor educados mostraron más síntomas neuróticos que aquéllos con menos educación.
2. Los hombres procedentes de un medio rural generalmente se mostraron con mejor espíritu durante su vida militar que los soldados procedentes de la ciudad.
3. Los soldados del sur se aclimataron más fácilmente, en las calientes islas del Mar del Sur, que los soldados del Norte.
4. Mientras continuaba la guerra, los soldados estaban más ansiosos de regresar a los Estados Unidos de lo que lo estaban después de la rendición alemana.

Si usted cree que estas afirmaciones tienen suficiente sentido común como para

someterlas a una prueba sistemática, entonces tal vez le interesaría saber que cada afirmación es directamente opuesta a lo que se encontró en realidad. Los soldados deficientemente educados se mostraron más neuróticos que aquéllos con educación superior; a los del sur no se les notó mayor habilidad que a los del Norte en adaptarse a un clima tropical, y así sucesivamente.[1] Depender sólo del sentido común o de las experiencias cotidianas, *obviamente* tiene sus limitaciones.

LAS ETAPAS DE LA INVESTIGACION SOCIAL

El contrastar sistemáticamente nuestras ideas acerca de la naturaleza de la realidad social exige con frecuencia una investigación cuidadosamente planeada y ejecutada, en la cual:

1. Se reduce a una hipótesis contrastable, el problema que se va a estudiar, (por ejemplo las "familias con uno sólo de los padres, generan más delincuencia que las familias con los dos padres");
2. Se desarrolla un conjunto de instrumentos apropiados (por ejemplo, elaborar un cuestionario o un programa de entrevistas);
3. Se recogen los datos (esto es, el investigador puede ir al lugar del problema y hacer un censo o encuesta);
4. Se analizan los datos para apoyar su hipótesis inicial; y
5. Los resultados del análisis son interpretados y comunicados a un auditorio, por ejemplo, por medio de una conferencia o de un artículo en una revista.

Como veremos en los capítulos subsiguientes, el material presentado en este libro está más estrechamente relacionado con la etapa del análisis de los datos de la investigación (ver 4), en el cual los datos recogidos o reunidos por el investigador se analizan para apoyar su hipótesis inicial. Es en esta etapa de la investigación cuando los datos no procesados se tabulan, calculan, cuentan, resumen, reordenan, comparan o, en una palabra, *se organizan* para que podamos comprobar la exactitud o validez de nuestra hipótesis.

EL USO DE SERIES DE NUMEROS EN LA INVESTIGACION SOCIAL

Cualquiera que haya participado en la investigación social sabe que los problemas que se presentan en el análisis de los datos deben ser confrontados en las etapas de planeación de un proyecto de investigación, puesto que éstos (los datos) sustentan la naturaleza de las decisiones que se tomen en todas las demás etapas. Tales problemas afectan con frecuencia aspectos de diseño de la investigación y aun el

[1] Paul Lazarsfeld, "The American Soldier-An Expository Review", *Public Opinion Quarterly*, otoño, 1949, p. 380.

tipo de instrumentos que se emplearán al recoger los datos. Por esta razón, buscamos constantemente técnicas o métodos para mejorar la calidad del análisis de los mismos.

Muchos investigadores creen que es esencial emplear *mediciones*, o una serie de números en el análisis de los datos. Por consiguiente, los investigadores sociales han desarrollado mediciones para aplicarlas a una gama muy amplia de fenómenos, incluyendo prestigio ocupacional, actitudes políticas, autoritarismo, alienación, anomía, delincuencia, clase social, prejuicio, dogmatismo, conformidad, realización, etnocentrismo, buena vecindad, religiosidad, armonía matrimonial, movilidad ocupacional, urbanización, estatus socioeconómico* y fertilidad.

Los números tienen por lo menos tres funciones importantes para el investigador social, dependiendo del *nivel de medida* que emplee. Específicamente, las series de números se pueden usar:

1. para categorizar el nivel *nominal* de la medición
2. para determinar el *rango* o el *orden* al nivel *ordinal* de la medición
3. para *obtener montajes* al nivel de *intervalo* de la medición.

Antes de proceder a una discusión del papel de las estadísticas en la investigación social, detengámonos a examinar algunas de las principales características de estos niveles de medición, características que asumirán más tarde un considerable significado cuando tratemos de aplicar las técnicas estadísticas a situaciones particulares de investigación.

El nivel nominal

El nivel *nominal* de medición simplemente involucra el proceso de denominar o etiquetar; esto es, colocar los casos dentro de categorías y contar su frecuencia de ocurrencia. Para dar un ejemplo, podríamos usar una medida de nivel nominal para indicar cuántas de las personas entrevistadas tienen prejuicios hacia los portorriqueños y cuántas no. Como se muestran en la Tabla 1.1, podríamos interrogar a diez estudiantes de una clase dada y determinar que 5 pueden ser considerados como (1) con prejuicios y 5 pueden ser tomados como (2) sin prejuicios.

Otras medidas de nivel nominal en la investigación social son el sexo (femenino contra masculino), el estatus de bienestar social (los que lo reciben contra los que no lo reciben), los partidos políticos (conservador, liberal, independiente y socialista), el carácter social (de dirección interna, de otra dirección y tradicional), el modo de adaptación (conformidad, innovación, ritualismo, retiro, rebelión), la orientación en el tiempo (presente, pasado y futuro), y la urbanización (urbana, rural, suburbana), para mencionar sólo unas cuantas.

Al trabajar con los datos nominales debemos tener en cuenta que *cada caso debe colocarse en una sola categoría*. Esta exigencia indica que las categorías no

* N. del R. También conocido como estrato socioeconómico.

deben traslaparse ~~ni~~ *excluirse mutuamente.* Así, la raza de un entrevistado clasificada como "blanca" no puede clasificarse también como "negra"; al clasificarlo como "hombre" no se lo puede clasificar también como "mujer". La exigencia también indica que las categorías deben ser *exhaustivas* —debe haber un lugar para cada caso que se presente. Como una ilustración, imaginemos un estudio en el cual todas las personas entrevistadas se categorizaron por raza y se consideró solamente la blanca y la negra. ¿Dentro de qué grupo se categorizaría a un chino si apareciera entre los entrevistados? En este caso sería necesario aumentar el sistema original de categorías para incluir "orientales" o, suponiendo que la mayoría de los entrevistados fueran blancos o negros, incluir una categoría mixta en la cual se pudieran colocar tales excepciones.

El lector deberá notar que los datos nominales no se clasifican en un rango o escala por cualidades tales como mejor o peor, más alto o más bajo, más o menos. Queda claro entonces, que una medida nominal de sexo no explica si los hombres son "superiores" o "inferiores" a las mujeres. Los datos nominales únicamente se rotulan, algunas veces por nombre (hombres contra mujeres o personas con prejuicios contra las que no los tienen); otras veces por número (1 contra 2), pero siempre con el fin de agrupar los casos en categorías separadas para indicar semejanza o diferencia respecto a una cualidad o característica dada.

El nivel ordinal

Cuando el investigador va más allá de este nivel de medición y busca *ordenar* sus casos en términos del grado en que poseen una determinada característica, entonces está trabajando al nivel *ordinal* de medición. La naturaleza de la relación que existe entre categorías ordinales depende de la característica que el investigador trata de medir. Para dar un ejemplo conocido, el investigador podría clasificar a las personas con respecto al estatus socioeconómico como "clase baja", "clase media" y "clase alta". O, en lugar de clasificar a los estudiantes de una clase dada como con prejuicios *o* sin prejuicios, los podría clasificar de acuerdo con su grado de prejuicio hacia los portorriqueños, como se indica en la Tabla 1.2.

El nivel ordinal de medición nos da información acerca de la organización de las categorías, pero no indica *la magnitud de las diferencias* entre los números. Por ejemplo, el investigador social que emplea una medida de nivel ordinal, para estudiar el prejuicio contra los portorriqueños, *no sabe qué tanto más de prejuicios tiene una persona que otra.* En el ejemplo dado anteriormente, no es posible determinar hasta

TABLA 1.1 Actitudes hacia los portorriqueños (de diez estudiantes universitarios): datos nominales

Actitud hacia los portorriqueños	*Frecuencia*
1 = con prejuicios	5
2 = sin prejuicios	5
Total	10

TABLA 1.2 Actitudes hacia los portorriqueños (de diez estudiantes universitarios): datos ordinales

Estudiante	*Rango*
Julia	1. la que tiene más prejuicio
María	2. segunda
Jaime	3. tercero
José	4. cuarta
Laura	5. quinta
Juan	6. sexto
Fernando	7. séptimo
Aldo	8. octavo
Patricia	9. novena
Roberta	10. la que tiene menos prejuicio

qué punto Julia tiene más prejuicios que María o hasta qué grado Roberta muestra menos prejuicios que Patricia o Aldo. Esto se debe a que, en una escala ordinal, los intervalos entre los puntos o rangos no son conocidos o significativos. Por consiguiente, no es posible asignarle puntajes a casos localizados en puntos de la escala.

Nivel por intervalos

En contraste, el nivel de medición *por intervalos* nos indica tanto el orden de las categorías como la *distancia* exacta entre ellas. Las medidas por intervalos emplean *unidades constantes de medición* (por ejemplo, pesos o centavos, grados centígrados o Fahrenheit, metros o centímetros, minutos o segundos), las cuales proporcionan *intervalos iguales* entre los puntos de la escala.

De esta manera, una medición, por intervalos, del prejuicio hacia los portorriqueños –tal como respuestas a una serie de preguntas sobre los portorriqueños, clasificadas de 0 a 100 (donde 100 representa el más alto grado de prejuicio)– podría dar los datos que se observan en la Tabla 1.3 sobre los diez estudiantes de un determinado salón de clase.

TABLA 1.3 Actitudes hacia los portorriqueños (de diez estudiantes universitarios): datos por intervalos

Estudiante	*Puntuación*[a]
Julia	98
María	96
Jaime	95
José	94
Laura	22
Juan	21
Fernando	20
Aldo	15
Patricia	11
Roberto	6

[a] La puntuación más alta indica más prejuicio contra los portorriqueños

Como indica la Tabla 1.3, podemos ordenar a los estudiantes en términos de sus prejuicios y además indicar las distancias que los separan a unos de otros. Por ejemplo, es posible afirmar que Roberto es el menos prejuicioso de la clase ya que obtuvo el puntaje más bajo. También podemos decir que Roberto es ligeramente menos prejuicioso que Patricia o Aldo, y aun menos que Julia, María, Jaime o José, todos los cuales obtuvieron puntajes sumamente altos. Dependiendo del objetivo para el cual el estudio esté diseñado, podría ser importante determinar tal información, que no se encuentra disponible al nivel ordinal de medición.

FUNCIONES DE LA ESTADISTICA

El momento en el que el investigador social emplea números *cuantifica sus datos* a los niveles de medición nominal, ordinal o por intervalos – cuando es probable que emplee la estadística como un instrumento para (1) *la descripción* y (2) la *toma de decisiones.* Echemos ahora una mirada más de cerca a estas importantes funciones de la estadística.

Descripción

Para llegar a conclusiones o a obtener resultados, un investigador social con frecuencia estudia centenares, miles o aun cifras más altas de personas o grupos. Como caso extremo, la "Oficina de Censos" de los Estados Unidos lleva una lista completa de la población de los Estados Unidos en la cual se pone en contacto con más de 200 millones de personas. A pesar de la ayuda de numerosos procedimientos complejos

TABLA 1.4 Calificaciones de un examen de 80 estudiantes

72	83	91	29
38	89	49	36
43	60	67	49
81	52	76	62
79	62	72	31
71	32	60	73
65	28	40	40
59	39	58	38
90	49	52	59
83	48	68	60
39	65	54	75
42	72	52	93
58	81	58	53
56	58	77	57
72	45	88	61
63	52	70	65
49	63	61	70
81	73	39	79
56	69	74	37
60	75	68	46

diseñados para tal fin, constituye siempre una tarea descomunal describir y resumir las enormes cantidades de datos que se generan de los proyectos de investigación social.

Para dar un ejemplo cotidiano, las calificaciones de un examen de un grupo de sólo 80 estudiantes han sido enlistadas en la Tabla 1.4. ¿Ve algún sistema de referencia en estas calificaciones? ¿Puede describir estas calificaciones en pocas palabras? ¿En pocas frases? ¿Son, en conjunto, particularmente altas o bajas? Incluso usando los principios más elementales de la estadística descriptiva, como en los capítulos subsiguientes de este texto, es posible caracterizar la distribución de las calificaciones de exámenes de la Tabla 1.4 con bastante claridad y precisión, de modo que las tendencias o características generales del grupo se puedan descubrir más rápidamente y comunicar con mayor facilidad a cualquier persona. Primero, podríamos arreglar nuevamente las calificaciones en orden consecutivo (del más alto al más bajo) para reunirlas dentro de un número más pequeño de categorías. Como se muestra en la Tabla 1.5, esta *distribución de frecuencia agrupada* (la cual se estudiará en detalle en el Capítulo 2) presentaría las calificaciones dentro de categorías más amplias junto con el número o *frecuencia (f)* de estudiantes cuyas calificaciones cayeron dentro de estas categorías. Se puede ver fácilmente, por ejemplo, que 17 estudiantes recibieron calificaciones entre 60 y 69; solamente dos recibieron calificaciones entre 20 y 29.

Otro procedimiento útil (explicado en el Capítulo 3) sería el reorganizar las calificaciones gráficamente. Como se muestra en la Figura 1.1, podríamos colocar las categorías de calificaciones (desde 20-29 hasta 90-99) en un eje de la gráfica (esto es, *la línea base horizontal*) y sus números o frecuencias a lo largo de otra línea (esto es, *el eje vertical*).

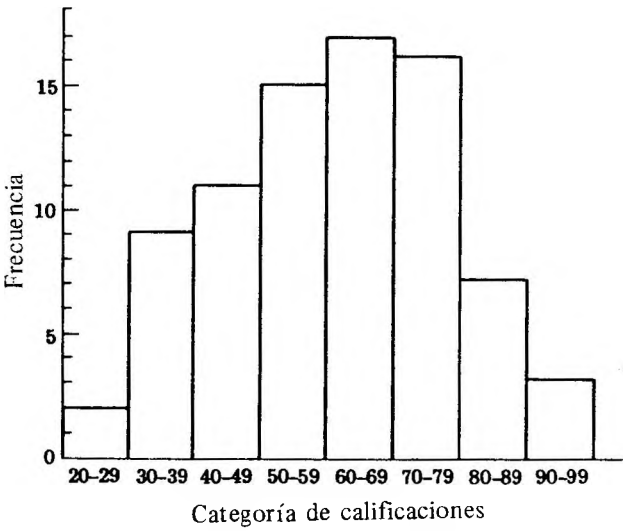
Este arreglo nos da una representación gráfica bastante fácil de visualizar (por ejemplo en la gráfica de barras), en la cual podemos ver que la mayoría de las calificaciones caen entre 50 y 80 y que relativamente pocas notas son: o mucho más altas o mucho más bajas.

Como lo explicaremos en el Capítulo 4, un método estadístico particularmente conveniente y útil —con el cual ya estamos más o menos familiarizados— es preguntar cuál es la calificación de la persona *promedio* en este grupo de 80 estudiantes. El promedio aritmético (o *media*) que se obtiene sumando la lista

TABLA 1.5 Calificaciones de examen de 80 estudiantes: una distribución de frecuencia agrupada

Calificaciones	*f*
90–99	3
80–89	7
70–79	16
60–69	17
50–59	15
40–49	11
30–39	9
20–29	2

FIGURA 1.1 Calificaciones de examen de 80 estudiantes, organizadas en una gráfica de barras



completa de las calificaciones y dividiendo esta suma entre el número de estudiantes, nos da una idea más clara de la tendencia del grupo en conjunto. El promedio aritmético en la presente ilustración es de 60,5 una calificación bastante baja si se compara con el promedio de clase con el que la mayoría de los estudiantes ya pueden estar familiarizados. Este grupo de 80 estudiantes dio en conjunto, un rendimiento aparentemente muy bajo:

Así, con la ayuda de recursos estadísticos, tales como las distribuciones de frecuencia agrupada, las gráficas y el promedio aritmético, es posible detectar y describir patrones o tendencias en las distribuciones de puntajes (por ejemplo en las calificaciones de la Tabla 1.4), las cuales, de otra manera, no hubieran sido advertidas por el observador casual. En el presente contexto, entonces, podemos definir la estadística como *un conjunto de técnicas para la reducción de datos cuantitativos (esto es, una serie de números) a un número pequeño de términos descriptivos más adecuados y de lectura más simple.*

La toma de decisiones

Con el fin de probar una hipótesis, es necesario, a menudo, ir más allá de la simple descripción; también es frecuentemente necesario hacer inferencias, esto es, tomar decisiones basándose en los datos recogidos solamente de una pequeña porción o *muestra* del grupo más grande que pensamos estudiar. Factores tales como costo, tiempo, y la necesidad de una supervisión adecuada, muchas veces impiden hacer una completa enumeración o lista del grupo completo (los investigadores sociales llaman *población o universo* a este grupo más grande, del cual se ha sacado una muestra).

TABLA 1.6 Uso de la mariguana, el sexo de los entrevistados: caso I

Uso de la mariguana	*Sexo de los entrevistados*	
	Masculino	*Femenino*
Número de los que la han probado	60	40
Número de los que no la han probado	40	60
Total	100	100

Como lo veremos en el Capítulo 7, cada vez que el investigador social prueba su hipótesis en una muestra, debe decidir si en verdad resulta correcto generalizar los resultados obtenidos con respecto a la población entera, de la cual se obtuvo la muestra. Del muestreo resulta inevitablemente el error, aun del muestreo que ha sido correctamente concebido y ejecutado. Este es el problema que se presenta al generalizar o *sacar inferencias* de la muestra a la población.[2]

La Estadística puede utilizarse con el fin de generalizar los resultados obtenidos en la investigación, con un alto grado de seguridad, de pequeñas muestras a poblaciones mayores. Para comprender mejor este objetivo de tomar decisiones en estadística y el concepto de generalizar de las muestras a las poblaciones, examinemos los resultados de un estudio hipotético que se llevó a cabo para probar la siguiente hipótesis:

Hipótesis: Es más probable que los universitarios hayan probado la mariguana, que las universitarias.

Los investigadores de este estudio decidieron probar su hipótesis en una universidad urbana en la cual había unos 20 000 estudiantes matriculados (10 000 hombres y 10 000 mujeres). Debido a los factores de costo y de tiempo no pudieron entrevistar a cada uno de los estudiantes de dicha universidad, pero obtuvieron, de la oficina de matriculación, una lista completa de los estudiantes. De esta lista escogieron uno de cada cien (mitad hombres y mitad mujeres) para la muestra y luego los entrevistaron miembros del grupo de investigación entrenados para este fin. Las personas encargadas de las entrevistas preguntaron a cada uno de los 200 participantes en la muestra si él o ella habían probado la mariguana y luego procedieron a registrar el sexo del estudiante como masculino o femenino. Los resultados de dicho estudio fueron tabulados por sexo y presentados en la Tabla 1.6.

[2] *Al estudiante:* El concepto de "error de muestreo" se estudiará con más detalle en el Capítulo 7. Sin embargo, para comprender mejor la inevitabilidad del error, cuando se muestrea de un grupo muy grande es posible que el estudiante desee hacer ahora la siguiente demostración. Refiriéndose a la Tabla 1.4, que contiene las calificaciones de una población de 80 estudiantes, seleccione, al "azar" (por ejemplo, cerrando los ojos y señalando), una muestra de una pocas calificaciones (por ejemplo 5) de la lista completa. Encuentre la calificación promedio sumando las cinco puntuaciones y dividiendo entre cinco el número total de calificaciones. Ya se ha indicado que la nota promedio del grupo completo de los 80 estudiantes fue de 60,5 ¿Hasta dónde difiere la muestra promedio del promedio de la clase 60,5 ? Pruebe esto en varias muestras más de algunas otras calificaciones escogidas al azar del grupo más grande. Con frecuencia se hallará que la muestra media diferirá casi siempre, al menos ligeramente, de la obtenida de la clase completa de 80 estudiantes. Esto es lo que para nosotros significa "*error de muestreo*".

Nótese que los resultados obtenidos de esta muestra de 200 estudiantes, como se presentan en la Tabla 1.6, están de acuerdo con la dirección de hipótesis formulada: 60 de cada 100 hombres informaron que habían probado la mariguana, mientras solamente 40 de cada 100 mujeres afirmaron que lo habían hecho. Claramente, en esta pequeña muestra, los hombres tuvieron más tendencia que las mujeres a fumar mariguana. Para nuestros propósitos, sin embargo, la pregunta más importante es si estas diferencias de sexo en el uso de la mariguana son lo suficientemente grandes como para generalizarlas confiadamente a una población de más de 20 000 estudiantes. ¿Representan, estos resultados, diferencias verdaderas en la población? ¿O hemos obtenido diferencias casuales entre hombres y mujeres debido estrictamente al error de muestreo –el error que ocurre cada vez que escogemos un grupo pequeño entre un grupo más grande?

Para ilustrar el problema de generalizar los resultados obtenidos, de muestras a poblaciones más grandes, imaginemos que los investigadores obtuvieron más bien los resultados que se muestran en la Tabla 1.7. Nótese que estos resultados están todavía en la dirección predicha por la hipótesis: 55 hombres en oposición a sólo 45 mujeres habían probado la mariguana. Pero aún estamos deseando generalizar estos resultados a una población universitaria más grande. ¿No es probable que una diferencia de esta magnitud (más hombres que mujeres) ocurriera simplemente por casualidad? ¿O podemos confiadamente decir que tales diferencias, relativamente pequeñas, reflejan una diferencia real entre hombres y mujeres sólo en el caso particular de esta universidad?

Ilustremos un poco más. Supongamos que los investigadores sociales hubiesen obtenido los datos que se muestran en la Tabla 1.8. Las diferencias entre hombres y mujeres mostradas en la tabla no podían haber sido más pequeñas y aún estar ceñidas a la dirección de la hipótesis: 51 hombres en contraste con 49 mujeres han fumado mariguana, sólo dos hombres más que mujeres. ¿Cuántos de nosotros estaríamos dispuestos a considerar *estos* resultados como una verdadera diferencia de población entre hombres y mujeres, más que como un producto de la casualidad o del error de muestreo? ¿Dónde trazaremos la línea? ¿En qué punto es lo suficientemente grande una diferencia de muestreo para que estemos dispuestos a tratarla como significativa o real? Con la ayuda de la estadística podemos tomar tales decisiones acerca de la relación entre muestras y poblaciones, con facilidad y un alto grado de confiabilidad.

A manera de ilustración, si hubiéramos empleado una de las pruebas estadísticas

TABLA 1.7 Uso de la mariguana según el sexo de los entrevistados: caso II

	Sexo de los entrevistados	
Uso de la mariguana	*Masculino*	*Femenino*
Personas que la han probado	55	45
Personas que no la han probado	45	55
Total	100	100

TABLA 1.8 Uso de la mariguana según el sexo de los entrevistados: caso III

Uso de la mariguana	*Sexo de los entrevistados*	
	Masculino	*Femenino*
Personas que la han probado	51	49
Personas que no la han probado	49	51
Total	100	100

de significado que se estudiarán más adelante en este texto (por ejemplo la Chi cuadrada; ver Capítulo 10), ya sabríamos que *solamente los resultados* de la Tabla 1.6 podrían generalizarse a la población de 20 000 universitarios – que 60 de cada 100 hombres, y solamente 40 de cada 100 mujeres, han probado la mariguana; este hecho es un hallazgo lo suficientemente sustancial como para aplicarlo a la población entera con un alto grado de confiabilidad. Nuestra prueba estadística nos dice que hay sólo un 5% de probabilidad de que estemos equivocados. Por contraste, los resultados presentados en las tablas 1.7 y 1.8 *son estadísticamente no significativos,* siendo el producto de un error de muestreo más que de las diferencias reales del sexo en el uso de la mariguana. De nuevo, empleando un criterio estadístico, concluimos que estos resultados no reflejan verdaderas diferencias de población, sino un mero error de muestreo.

Entonces, en el presente contexto, la Estadística es *un conjunto de técnicas para tomar decisiones que ayuden a los investigadores a hacer inferencias de las muestras a las poblaciones y, en consecuencia, a comprobar hipótesis relativas a la naturaleza de la realidad social.*

RESUMEN

Este capítulo relaciona nuestras predicciones diarias acerca de eventos futuros, con las experiencias del investigador social que emplea la Estadística como una ayuda para comprobar sus hipótesis acerca de la realidad social. La medición fue analizada en términos de datos nominales, ordinales y por intervalos. Se identificaron dos funciones principales de la Estadística con la etapa del análisis de los datos de la investigación social, posteriormente se discutieron e ilustraron brevemente:

1. La descripción (esto es, la reducción de datos cuantitativos a un número menor de términos descriptivos más convenientes), y
2. La toma de decisiones (esto es, hacer inferencias de muestras a poblaciones).

PARTE I

Descripción

2

Organización de datos

La recolección de datos implica un gran esfuerzo por parte del investigador social que busca aumentar sus conocimientos sobre el comportamiento humano. Para entrevistar o bien para sacar información a beneficiarios de la asistencia pública, estudiantes universitarios, drogadictos, residentes de viviendas públicas, homosexuales, personas de clase media, u otros, se requiere un grado de previsión, planificación cuidadosa y control o bien pasar algún tiempo en dicha situación.

Sin embargo, completar la recolección de datos es sólo el principio, en lo que concierne al análisis estadístico. La recolección de datos constituye la materia prima con que debe trabajar el investigador social si ha de analizar sus datos, obtener resultados y probar sus hipótesis sobre la naturaleza de la realidad social.

DISTRIBUCIONES DE FRECUENCIA DE DATOS NOMINALES

El carpintero transforma la madera en muebles; el cocinero convierte los alimentos crudos en los platos más apetitosos que se sirven a la mesa. Mediante un proceso similar, el investigador social, auxiliado por "recetas" –llamadas fórmulas y técnicas– intenta transformar sus datos crudos* en un conjunto de medidas significativas y organizadas que puedan utilizarse para probar su hipótesis inicial.

¿Qué puede hacer el investigador social para organizar los números desordenados que recoge de sus entrevistados? ¿Cómo se las arregla para transformar esta masa de datos en un resumen fácil de entender? El primer paso sería construir una *distribución de frecuencia* en forma de tabla.

TABLA 2.1 Estudiantes de ambos sexos concurrentes a una manifestación política de izquierda

Sexo del estudiante	*Frecuencia (f)*
Masculino	80
Femenino	20
Total	100

* N. del E. crudo significa "no procesados".

Examinemos la distribución de frecuencia en la Tabla 2.1. Nótese primero que la Tabla está encabezada por un *número* (2.1) y un *título* que da al lector una idea sobre la naturaleza de los datos presentados —"Estudiantes de ambos sexos concurrentes a una manifestación política de izquierda." Este es el arreglo estándar; toda tabla debe estar claramente titulada y, cuando se presente dentro de una serie, también debe estar marcada con un número.

Las distribuciones de frecuencia de los datos nominales consisten de dos columnas. Así, en la Tabla 2.1, la columna de la izquierda indica qué característica está siendo presentada (sexo del estudiante) y contiene las categorías de análisis (masculino y femenino). Una columna adyacente con el encabezado de "frecuencia" o "*f*", indica el número de casos en cada categoría (80 y 20 respectivamente), así como el número total de casos (N=100).

Una rápida mirada a la distribución de frecuencia, en dicha Tabla, revela claramente que a la manifestación de izquierda concurrieron muchos más hombres que mujeres —80 de los 100 estudiantes que asistieron eran hombres.

COMPARACION DE LAS DISTRIBUCIONES

Supongamos, sin embargo, que deseamos comparar los asistentes a la manifestación izquierdista con estudiantes similares en una manifestación derechista. La comparación entre distribuciones de frecuencia es un procedimiento que se utiliza a menudo para aclarar resultados y agregar información. La comparación particular que haga el investigador está determinada por la pregunta que busca contestar.

Volviendo a nuestra hipotética manifestación política, podríamos preguntar: ¿es probable que participen más estudiantes del sexo masculino, que del sexo femenino en manifestaciones tanto izquierdistas como derechistas? Para encontrar una respuesta podríamos comparar los 100 estudiantes asistentes a la manifestación izquierdista con otros 100 estudiantes de la misma universidad asistentes a una manifestación derechista. Imaginemos que obtenemos los datos mostrados en la Tabla 2.2.

Como se muestra en la tabla, 30 de 100 estudiantes en la manifestación derechista, pero sólo 20 de 100 estudiantes en la manifestación izquierdista, eran mujeres. Esto nos da considerablemente más información que la sola distribución de frecuencia con que empezamos (ver Tabla 2.1). Así, podemos afirmar ahora que los

TABLA 2.2 Estudiantes de ambos sexos asistentes a manifestaciones políticas de derecha e izquierda

	Asistencia a las manifestaciones	
	De izquierda	*De derecha*
Sexo del estudiante	*f*	*f*
Masculino	80	70
Femenino	20	30
Total	100	100

hombres, en esta universidad, participaron más que su contraparte femenina tanto en las manifestaciones izquierdistas como derechistas. Podemos afirmar también que, cuando las mujeres asistieron, tendieron a participar algo más en las manifestaciones derechistas que en las izquierdistas.

Proporciones y porcentajes

Cuando el investigador estudia distribuciones de igual tamaño total, los datos de frecuencia pueden utilizarse para hacer comparaciones entre los grupos. Así, el número de hombres asistentes a manifestaciones, de derecha y de izquierda, puede ser comparado directamente, ya que sabemos que había exactamente 100 estudiantes en cada manifestación. Sin embargo, generalmente no es posible estudiar distribuciones que tengan exactamente el mismo número de casos. Por ejemplo, ¿cómo podemos asegurarnos de que precisamente 100 estudiantes asistirán a ambas clases de manifestaciones políticas? Para aclarar tales resultados, necesitamos un método para *estandarizar distribuciones de frecuencia por tamaño* —una forma de comparar grupos a pesar de las diferencias en las frecuencias totales. Dos de los métodos más populares y útiles para estandarizar por tamaño y comparar distribuciones son la *proporción* y el *porcentaje*. La proporción compara el número de casos en una categoría dada con el tamaño total de la distribución. Podemos convertir cualquier frecuencia en una proporción *P*, dividiendo el número de casos en cualquier categoría dada *f* por el número total de casos en la distribución *N*.

O sea,
$$P = \frac{f}{N}$$

Por consiguiente, 10 hombres entre 40 estudiantes asistentes a una manifestación pueden expresarse en la proporción $P = \frac{10}{40} = 0{,}25$

A pesar de la utilidad de la proporción, mucha gente prefiere indicar el tamaño relativo de una serie de número en términos del *porcentaje*, la frecuencia de ocurrencia de una categoría *por cada 100 casos*. Para calcular un porcentaje, simplemente multiplicamos cualquier proporción dada por 100. Por fórmula,

$$\% = (100)\frac{f}{N}$$

Por consiguiente, 10 hombres de entre los 40 asistentes a una manifestación pueden expresarse en la proporción $P = \frac{10}{40} = 0{,}25$ o como un porcentaje

$$\% = (100)\frac{10}{40} = 25 \text{ por ciento.}$$

Así, el 25 por ciento de este grupo de 40 estudiantes son del sexo masculino. Para ilustrar la utilidad de los porcentajes al hacer comparaciones entre distribucio-

nes, examinemos la participación en manifestaciones políticas en una universidad predominantemente izquierdista.

Supongamos, por ejemplo, que la manifestación izquierdista atrajo a un gran número de estudiantes, digamos 1 352 mientras que la manifestación derechista atrajo a un número mucho más pequeño, digamos 183.

La Tabla 2.3 nos indica tanto las frecuencias como los porcentajes de asistencia a estas manifestaciones. Nótese la dificultad que existe para determinar rápidamente las diferencias de sexo en la asistencia sólo con los datos de frecuencia. En contraste, los porcentajes revelan claramente que las mujeres estuvieron igualmente representadas en las manifestaciones tanto de derecha como de izquierda. Específicamente, el 20% de los estudiantes asistentes a la manifestación izquierdista eran mujeres; el 20% de los estudiantes asistentes a la manifestación derechista eran mujeres.

TABLA 2.3 Estudiantes de ambos sexos asistentes a manifestaciones políticas de derecha e izquierda

	Asistencia a las manifestaciones			
	De izquierda		*De derecha*	
Sexo del estudiante	*f*	%	*f*	%
Masculino	1082	(80)	146	(8
Femenino	270	(20)	37	(2
Total	1352	(100)	183	(10

Razones *

Un método menos común, utilizado para estandarizar por tamaño, es la *razón,* que compara directamente el número de casos que caen dentro de una categoría (por ejemplo, hombres) con el número de casos que caen dentro de otra categoría (por ejemplo, mujeres). Así, puede obtenerse una razón de la siguiente manera, donde f_1 es igual a la frecuencia en cualquier categoría y f_2 es igual a la frecuencia en cualquier otra categoría:

$$\text{razón} = \frac{f_1}{f_2}$$

Si estuviéramos interesados en determinar la razón que haya de negros a blancos, podríamos comparar el número de negros entrevistados (f = 150) con el número de blancos entrevistados (f = 100) como $\frac{150}{100}$ Cancelando los factores comunes en el numerador y el denominador, es posible reducir la razón a su forma más simple, por ejemplo $\frac{150}{100} = \frac{3}{2}$ (había 3 entrevistados negros por cada 2 blancos).

* N. del E. Este término también se conoce como "cociente". El estudiante encontrará que en la práctica de campo se utilizan indistintamente.

El investigador podría aumentar la claridad de su razón dando la base (el denominador) de alguna forma comprensible. Por ejemplo, la *razón de sexo* a menudo empleada por los demógrafos, que buscan comparar el número de hombres y mujeres en cualquier población dada, se da generalmente como el número de hombres por cada 100 mujeres.

Para ilustrar, si la razón de hombres a mujeres es $\frac{150}{50}$ debería haber 150 hombres por cada 50 mujeres (o reduciendo, 3 hombres por cada mujer). Para obtener la terminología convencional de la razón de sexo, multiplicaríamos la razón por 100. Entonces.

$$\text{Razón de sexo} = (100)\frac{f \text{ hombres}}{f \text{ mujeres}} = \frac{(100)\ 150}{50} = 300$$

Resulta entonces que había 300 hombres en la población dada, por cada 100 mujeres.

Las razones ya no se usan extensamente en la investigación social, quizás por los siguientes motivos:

1. Se necesita un gran número de razones para describir distribuciones que tienen muchas categorías de análisis.
2. Puede ser difícil comparar razones basadas en números muy grandes.
3. Algunos investigadores sociales prefieren evitar las fracciones o decimales que generan las razones.

Tasas

Otra clase de razón, que tiende a ser utilizada más ampliamente por los investigadores sociales, se conoce como *tasa*. Los sociólogos analizan a menudo a las poblaciones en cuanto a las tasas de reproducción, muerte, crimen, divorcio, matrimonio, y otros. Sin embargo, mientras que la mayoría de las demás razones comparan el número de casos en cualquier subgrupo (categoría) con el número de casos en cualquier otro subgrupo (categoría), las tasas indican comparaciones entre el número de casos *reales* y el número de casos *potenciales*. Por ejemplo, para determinar la tasa de nacimientos para una determinada población, podríamos mostrar el número de nacimientos vivos reales, entre las mujeres en edad de concebir (aquellos miembros de la población que están expuestos al riesgo de concebir y que por lo tanto representan casos potenciales). De modo similar, para encontrar la tasa de divorcios, podríamos comparar el número real de divorcios con el número de matrimonios que ocurren durante algún periodo de tiempo (por ejemplo 1 año). Las tasas suelen darse en términos de una base de 1 000 casos potenciales. Así, las tasas de nacimiento se dan como el número de nacimientos por cada 1 000 mujeres; las tasas de divorcio podrían expresarse en términos del número de divorcios por cada 1 000 matrimonios. De este modo, si ocurren 500 nacimientos entre 4 000 mujeres en edad de concebir, resulta que hubo 125 nacimientos por cada 1 000 mujeres en edad de concebir.

$$\text{Tasa de nacimiento} = (1\ 000)\frac{f \text{ casos reales}}{f \text{ casos potenciales}} = \frac{(1\ 000)500}{4\ 000} = 125$$

Hasta ahora hemos discutido tasas que podrían ser útiles para hacer comparaciones entre diferentes poblaciones. Por ejemplo, podríamos buscar comparar tasas de nacimiento entre blancos y negros, entre mujeres de clase media y de clase baja, entre grupos religiosos o sociedades enteras, etc. Otra clase de tasa, la *tasa de cambio*, puede utilizarse para comparar la misma población en dos puntos a un tiempo. Al computar la tasa de cambio comparamos el cambio real entre el tiempo 1 y el tiempo 2, sirviendo como base el tamaño del periodo del tiempo 1. Así, una población que aumenta de 20 000 a 30 000 entre 1960 y 1970 experimentaría una tasa de cambio:

$$\frac{(100) \text{ tiempo } 2f - \text{tiempo } 1f}{\text{tiempo } 1f} = \frac{(100)\ 30\ 000 - 20\ 000}{20\ 000} = 50\%$$

En otras palabras, hubo un aumento de población del 50 por ciento en el periodo de 1960 a 1970.

Nótese que una tasa de cambio puede ser *negativa* si indica un crecimiento en tamaño en cualquier periodo dado. Por ejemplo, si una población cambia de 15 000 a 5 000 en un periodo de tiempo, la tasa de cambio sería:

$$\frac{(100)5\ 000 - 15\ 000}{15\ 000} = -67\%$$

DISTRIBUCIONES DE FRECUENCIA SIMPLES DE DATOS ORDINALES Y POR INTERVALOS

Dado que los datos nominales son colocados más bien dentro de una clasificación que dentro de una escala, las categorías de las distribuciones de nivel nominal no tienen que enlistarse en ningún orden en particular. Así, los datos sobre preferencias religiosas mostrados en la Tabla 2.4 se presentan de 3 formas diferentes, aunque igualmente aceptables.

TABLA 2.4 Distribución de preferencias religiosas mostrada de 3 maneras

Religión	*f*	*Religión*	*f*	*Religión*	*f*
Protestante	30	Católica	20	Judía	10
Católica	20	Judía	10	Protestante	30
Judía	10	Protestante	30	Católica	20
Total	60	Total	60	Total	60

En contraste, las categorías o puntajes en las distribuciones ordinales representan el grado en que está presente una característica en particular. El enlistado de tales categorías o puntajes en las distribuciones de frecuencia simples debe hacerse de modo que refleje ese orden.

Por este motivo, las categorías ordinales y por intervalos simpre se colocan en orden desde sus valores más altos hasta los más bajos. Por ejemplo, podríamos hacer una lista de las categorías de las clases sociales desde la más alta hasta la más baja (alta, media, baja) o podríamos situar los resultados de un examen semestral de biología, en orden consecutivo, de la nota más alta a la más baja.

La perturbación del orden de las categorías ordinales y por intervalos reduce la legibilidad de los hallazgos del investigador. Este efecto puede observarse en la Tabla 2.5, donde se han presentado las versiones tanto "correcta" como "incorrecta" de una distribución de "Actitudes Hacia la Guerra". ¿Qué versión encuentra el lector más fácil de leer?

TABLA 2.5 Una distribución de frecuencia de actitudes hacia la guerra: Presentación correcta e incorrecta

Actitud hacia la guerra	*f*
Ligeramente favorable	2
Algo desfavorable	10
Fuertemente favorable	0
Ligeramente desfavorable	4
Fuertemente desfavorable	21
Algo favorable	1
Total	38

Incorrecta

Actitud hacia la guerra	*f*
Fuertemente favorable	0
Algo favorable	1
Ligeramente favorable	2
Ligeramente desfavorable	4
Algo desfavorable	10
Fuertemente desfavorable	21
Total	38

Correcta

DISTRIBUCIONES DE FRECUENCIA AGRUPADAS DE DATOS POR INTERVALOS

Los puntajes a nivel de intervalos se extienden a veces sobre un amplio rango (puntajes más altos menos los más bajos), haciendo que la distribución de frecuencia simple que resulta, sea más larga y difícil de leer. Cuando ocurren tales instancias, pocos casos pueden caer en cada categoría y el patrón del grupo se vuelve borroso. Para ilustrar, la distribución colocada en la Tabla 2.6 contiene valores que varían de 50 a 99 y tiene casi cuatro columnas de longitud.

Para aclarar nuestra presentación, podríamos construir una *distribución de frecuencia agrupada*, condensando los puntajes separados en un número de categorías o grupos más pequeños, donde cada uno contenga más de un puntaje. Cada categoría o grupo, en una distribución agrupada, es conocido como un *intervalo de clase*, cuyo *tamaño* está determinado por el número de puntaje que contenga.

Las calificaciones de exámenes de 71 estudiantes, presentadas originalmente en la Tabla 2.6, se vuelven a ordenar en una distribución de frecuencia agrupada, mostrada en la Tabla 2.7. Aquí encontramos 10 intervalos de clase, cada uno de tamaño 5. Así, el intervalo de clase más alta (95-99) contiene los 5 puntajes 95, 96, 97, 98 y 99. De manera similar, el intervalo 70-74 es de tamaño 5 y contiene los puntajes 70, 71, 72, 73 y 74.

Límites de clase

De acuerdo con su tamaño, cada intervalo de clase tiene un *límite superior* y *un límite inferior.* A primera vista, los puntajes más alto y más bajo, en cualquier categoría, *parecen* ser tales límites. Así, podríamos razonablemente esperar que los límites superior e inferior del intervalo 60-64 sean 64 y 60 respectivamente. En este caso, sin embargo, *nos equivocaríamos,* ya que 60 y 64 no son en realidad los límites del intervalo 60-64.

Muchos lectores se estarán preguntando, "¿por qué no? ". Para encontrar una respuesta examinemos un problema que podría surgir si fuéramos a definir límites de clase en términos de los puntajes más altos y más bajos en cualquier intervalo. Supongamos que tratáramos de colocar números que contienen valores fraccionarios (fracciones decimales) en la distribución de frecuencia mostrada en la Tabla 2.7. ¿Dónde podríamos categorizar el puntaje 62,3? Muchos estaríamos de acuerdo en que pertenece al intervalo 60-64. Pero, ¿qué hay con el puntaje 69,4? ¿Y con el número 54,2 o 94,6? El lector podría darse cuenta que los puntajes más altos y más bajos en un intervalo dejarán separaciones entre grupos adyacentes, en tal forma que algunos valores fraccionarios no pueden asignarse a ningún intervalo de clase en la distribución y deben excluirse del todo.

A diferencia de los puntajes más altos y más bajos en un intervalo, los *límites de clase* se localizan en el punto medio situado entre los intervalos de clase adyacentes, y por tanto, sirven para cerrar las separaciones entre ellos (ver Fig. 2.1). Así, el límite superior del intervalo 90-94 es 94,5 y el límite inferior del intervalo 95-99 es también 94,5 Asimismo, 59,5 sirve como límite superior del intervalo 55-59 y como límite inferior del intervalo 60-64. El lector podría preguntar; ¿qué pasa con el valor 59,5 valor que cae *exactamente* a la mitad de las separaciones entre intervalos de clase vecinos? Deberíamos incluir este puntaje en el intervalo 55-59 o en el intervalo 60-64? Este problema se resuelve generalmente *redondeando al número par más cercano.* Por ejemplo, 59,5 estaría situado en el intervalo 60-64; 84,5 estaría incluido en el intervalo 80-84. Como veremos, debe determinarse la posición de los límites de clase para trabajar con ciertos procedimientos estadísticos.

El punto medio

Otra característica de cualquier intervalo de clase es su *punto medio,* que definimos como el puntaje medio en el intervalo de clase. Un método simple y rápido

para encontrar el punto medio es buscar el punto donde cualquier intervalo dado puede dividirse en dos partes iguales. Tomando algunos ejemplos, 50 es el punto medio del intervalo 48-52; 3,5 es el punto medio del intervalo 2,5. El punto medio puede ser calculado a partir de los puntajes más altos a los más bajos en cualquier intervalo.

$$\frac{\text{puntaje más bajo} + \text{puntaje más alto}}{2} = \frac{48 + 52}{2} = 50$$

TABLA 2.6 Distribución de frecuencia de calificaciones de exámenes finales para 71 estudiantes

Calificación	*f*	*Calificación*	*f*	*Calificación*	*f*	*Calificación*	*f*
99	0	85	2	71	4	57	0
98	1	84	1	70	9	56	1
97	0	83	0	69	3	55	0
96	1	82	3	68	5	54	1
95	1	81	1	67	1	53	0
94	0	80	2	66	3	52	1
93	0	79	8	65	0	51	1
92	1	78	1	64	1	50	1
91	1	77	0	63	2	Total	71
90	0	76	2	62	0		
89	1	75	1	61	0		
88	0	74	1	60	2		
87	1	73	1	59	3		
86	0	72	2	58	1		

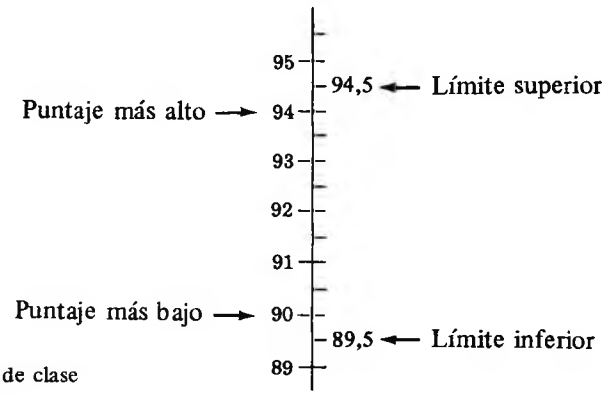
TABLA 2.7 Distribución de frecuencia agrupada de calificaciones de exámenes finales para 71 estudiantes

Intervalo de clase	*f*
95–99	3
90–94	2
85–89	4
80–84	7
75–79	12
70–74	17
65–69	12
60–64	5
55–59	5
50–54	4
Total	71

Determinación del número de intervalos

Para presentar datos por intervalos en una distribución de frecuencia agrupada, el investigador social debe considerar el número de categorías que desea emplear. Los

FIGURA 2.1 Puntajes más alto y más bajo contra los límites inferior y superior del intervalo de clase 90-94



:extos generalmente aconsejan usar de 5 a 20 intervalos. A este respecto, sería :onveniente recortar que las distribuciones de frecuencia agrupadas se emplean para 'evelar o enfatizar el patrón de un grupo. Muchos o muy pocos intervalos de clase)odrían confundir ese patrón y por tanto trabajar en contra del investigador que)usca darle claridad a su análisis. Además, reducir los valores de los puntajes ndividuales a un número innecesariamente pequeño de intervalos puede sacrificar nucha de la precisión —precisión que se había logrado originalmente conociendo la dentidad de puntajes individuales en la distribución. En suma, entonces, el investiga- lor decide generalmente sobre el número de intervalos, basándose en su propio :onjunto de datos y en sus objetivos personales, factores que pueden variar conside- ablemente de una investigación a otra.

)ISTRIBUCIONES ACUMULADAS

A veces, es deseable presentar frecuencias de una manera acumulada, especialmente :uando buscamos localizar la posición de un caso en relación con la actuación otal de un grupo. Las *frecuencias acumuladas* se definen como el número total de :asos que tengan cualquier puntaje dado *o uno que sea más bajo.* Así, la frecuencia icumulada (*fa*) para cualquier categoría (o intervalo de clase) se obtiene sumando la 'recuencia en esa categoría a la frecuencia total para todas las categorías abajo de :lla. En el caso de los puntajes del consejo universitario en la Tabla 2.8, vemos que a frecuencia (*f*) asociada con el intervalo de clase 301-350 es 12. Esta es también la 'recuencia acumulada para este intervalo, ya que ningún miembro del grupo obtuvo nenos de 301. La frecuencia en el próximo intervalo de clase 351-400 es 33, nientras que la frecuencia acumulada para este intervalo es 45 (33 + 12). Por lo anto, encontramos que 33 estudiantes ganaron puntajes del consejo universitario :ntre 351 y 400, pero que 45 recibieron puntajes de 400 *o menos.* Podríamos :ontinuar con este procedimiento, obteniendo frecuencias acumuladas para todos los ntervalos de clase hasta llegar a la parte más alta, 751-800, cuya frecuencia

acumulada (336) es igual al número total de casos, ya que ningún miembro del grupo logró puntajes sobre 800.

Además de la frecuencia acumulada, también podemos construir una distribución que indique *porcentajes acumulados (c%)*, o sea el tanto por ciento de casos que tengan cualquier puntaje o uno más bajo. Para calcular el porcentaje acumulado, modificamos la fórmula para porcentaje (%) introducida anteriormente en este capítulo, como sigue:

$$c\% = (100)\frac{fa}{N}$$

donde

fa = la frecuencia acumulada en cualquier categoría
N = el número total de casos en la distribución

Aplicando la fórmula anterior, a los datos de la Tabla 2.8, encontramos que el porcentaje de estudiantes que lograron puntajes de 350 o menos fue

$$c\% = (100)\frac{12}{336}$$
$$= (100)0{,}0357$$
$$= 3{,}57$$

El porcentaje que recibió puntajes de 400 o menos fue $c\% = (100)\frac{45}{336}$
$= (100)0{,}1339$
$= 13{,}39$

El porcentaje que alcanzó puntajes de 450 o menos fue $c\% = (100)\frac{93}{336}$
$= (100)0{,}2768$
$= 27{,}68$

En la Tabla 2.9 se muestra una distribución de porcentajes acumulados basada en los datos de la Tabla 2.8.

TABLA 2.8 Distribución de frecuencia acumulada de puntajes del Consejo Universitario para 336 estudiantes

Intervalo de clase	*f*	*fa*
751–800	6	336
701–750	25	330
651–700	31	305
601–650	30	274
551–600	35	244
501–550	55	209
451–500	61	154
401–450	48	93
351–400	33	45
301–350	12	12
Total	336	

TABLA 2.9 Distribución de porcentajes acumulados de puntajes del Consejo Universitario para 336 estudiantes (basado en los datos de la Tabla 2.8)

Intervalo de clase	*fa*	*c%*
751–800	336	100%
701–750	330	98.21
651–700	305	90.77
601–650	274	81.55
551–600	244	72.62
501–550	209	62.20
451–500	154	45.83
401–450	93	27.68
351–400	45	13.39
301–350	12	3.57

RANGO PERCENTIL

Supongamos que usted logró un puntaje de 80 en un examen de estadística. Para determinar exactamente qué tan bien lo ha hecho, podría ser de ayuda saber cómo se compara con los puntajes de otros en la clase que hayan tomado el mismo examen. ¿Lograron, la mayoría de los demás estudiantes, puntajes del orden de 80 y 90? Si fue así, su propia calificación puede no ser muy alta. O, ¿la mayoría de los demás recibió puntajes del orden de 60 y 70? Si fue así, un puntaje de 80 puede muy bien estar entre los más altos de su clase.

Con la ayuda de la distribución de porcentajes acumulados, podemos hacer comparaciones precisas entre cualquier caso individual y el grupo donde éste ocurre. Específicamente, podemos encontrar el *rango percentil* de un puntaje, un solo número que indique el porcentaje de casos en una distribución que cae por debajo de un puntaje dado. Por ejemplo, si un puntaje de 80 tiene un rango percentil de 95, entonces el 95% de los estudiantes en este curso de estadística recibieron puntajes de examen más bajo que 80 (sólo un 5% sacó puntajes arriba de 80). Sin embargo, si un puntaje de 80 tiene un rango percentil de 45, entonces sólo un 45% recibió puntajes de examen abajo de 80 (55% logró puntajes arriba de 80). Por fórmula,

$$\text{Rango Percentil} = \begin{array}{c}c\% \text{ abajo del} \\ \text{límite inferior} \\ \text{del intervalo} \\ \text{crítico}\end{array} + \left[\frac{\text{puntaje} - \begin{array}{c}\text{límite inferior del} \\ \text{intervalo crítico}\end{array}}{\text{tamaño del intervalo crítico}} \left(\begin{array}{c}\% \text{ en el} \\ \text{intervalo} \\ \text{crítico}\end{array}\right)\right]$$

A fin de ilustrar el procedimiento para obtener el rango percentil, busquemos el rango percentil para un puntaje de 620 en la distribución en la Tabla 2.8. Antes de aplicar la fórmula debemos localizar primero el *intervalo crítico,* el intervalo de clase en que aparece un puntaje de 620. Como se muestra más abajo, el intervalo crítico para el presente problema es 601-650:

Intervalo de clase	
751–800	
701–750	
651–700	
601–650	← Intervalo de clase en que ocurre el puntaje 620
551–600	
501–550	
451–500	
401–450	
351–400	
301–350	

Hay varias características del intervalo crítico que debemos determinar antes de aplicar la fórmula para rango percentil:

1. El límite inferior del intervalo crítico. Este es el punto que está a la mitad, entre el intervalo crítico, 601-650, y el intervalo de clase inmediatamente abajo de él, 551-600. El límite inferior de 601-650 es es 600,5.
2. El tamaño del intervalo crítico. Este está determinado por el número de puntajes dentro del intervalo de clase 601–650. El tamaño del intervalo crítico es 50, ya que contiene valores desde 601 hasta 650.
3. El porcentaje dentro del intervalo crítico. Para determinar el porcentaje dentro de cualquier intervalo de clase, dividimos el número de casos en ese intervalo de clase (f) entre el número total de casos en la distribución N y multiplicamos por 100 nuestra respuesta. Por fórmula.

$$\begin{aligned} \% &= (100)\frac{f}{N} \\ &= (100)\frac{30}{336} \\ &= (100)0{,}089 \\ &= 8{,}93 \end{aligned}$$

Por lo tanto, vemos que el 8,93 por ciento de estos puntajes del consejo universitario cayeron dentro del intervalo de clase 601–650.

4. El porcentaje acumulado abajo del límite inferior del intervalo crítico. Podemos leer $c\%$ directamente de la distribución de porcentaje acumulado en la Tabla 2.9. Subiendo por la columna $c\%$ de la tabla, vemos que el 72,62 por ciento de los puntajes caen abajo del intervalo crítico. Este es el porcentaje acumulado asociado con el intervalo de clase que cae inmediatamente abajo del intervalo crítico.

Ahora estamos preparados para aplicar la fórmula para rango percentil:

$$\text{Rango percentil} = 72{,}62 + \left[\frac{620 - 600{,}5}{50}(8{,}93)\right]$$

$$= 72{,}62 + \left[\frac{19{,}50}{50}(8{,}93)\right]$$

$$= 72{,}62 + (0{,}39)\ (8{,}93)$$
$$= 72{,}62 + 3{,}48$$
$$= 76{,}10$$

Resulta que ligeramente más del 76% recibió un puntaje más bajo de 620. Sólo el 23,90% logró puntajes por encima de esta cifra. ~~Como una ilustración más busquemos el rango percentil para un puntaje de 92 en la siguiente distribución de puntajes:~~

Intervalo de clase	*f*	*fa*	*c%*
90–99	6	49	100%
80–89	8	43	87,76
70–79	12	35	71,43
60–69	10	23	46,94
50–59	7	13	26,53
40–49	6	6	12,24
	N = 49		

Como se muestra más adelante, el intervalo crítico para un puntaje de 92 es 90-99:

Intervalo de clase	
90–99	← Intervalo de clase en que ocurre un puntaje de 92
80–89	
70–79	
60–69	
50–59	
40–49	

Las siguientes son las características del intervalo crítico que debemos determinar:

1. El límite inferior del intervalo crítico es 89,5.
2. El tamaño del intervalo crítico es 10, ya que hay 10 valores de puntajes dentro de él desde el 90 hasta el 99 (90, 91, 92, 93, 94, 95, 96, 97, 98, 99)
3. El porcentaje dentro del intervalo crítico es 12,24. Por fórmula:

$$\% = (100)\frac{f}{N}$$

$$= (100)\frac{6}{49}$$

$$= (100)0{,}1224$$
$$= 12{,}24$$

4. El porcentaje acumulado bajo el límite inferior puede encontrarse desde la columna *c%*, refiriéndose al intervalo de clase inmediatamente bajo el intervalo crítico. El porcentaje acumulado asociado al intervalo de clase 80-89 es 87,76.

Ahora estamos listos para sustituir en la fórmula para rango percentil:

$$\begin{aligned}\text{Rango percentil} &= 87{,}76 + \left[\frac{92 - 89{,}5}{10}(12{,}24)\right] \\ &= 87{,}76 + \left[\frac{2{,}50}{10}(12{,}24)\right] \\ &= 87{,}76 + (0{,}25)\ (12{,}24) \\ &= 87{,}76 + 3{,}06 \\ &= 90{,}82\end{aligned}$$

Casi el 91% recibió un puntaje más bajo de 92. Sólo el 9,18% obtuvo un puntaje más alto.

La escala de rangos percentiles consta de 100 unidades. Hay ciertos rangos a lo largo de la escala que tienen nombres específicos. Los *deciles* dividen la escala de rangos percentiles entre diez. Así, si un puntaje está localizado en el primer decil (rango percentil = 10), sabemos que el 10% de los casos caen abajo de él; si un puntaje está en el segundo decil (rango percentil = 20), entonces el 20% de los casos caen abajo de él, etc. Los rangos percentiles que dividen la escala en 4 partes se conocen como *cuartiles*. Si un puntaje está localizado en el primer cuartil (rango percentil = 25), sabemos que el 25% de los casos caen abajo de él; si un puntaje está en el segundo cuartil (rango porcentil = 50), el 50% de los casos caen abajo de él; y si un puntaje está en el tercer cuartil (rango percentil = 75), el 75% de los casos caen abajo de él (ver Figura 2.2)

FIGURA 2.2 Escala de rangos percentiles dividida por deciles y cuartiles

Rango Percentil	*Decil*	*Cuartil*
90 =	9o.	
85		
80 =	8o.	
75 =		3o.
70 =	7o.	
65		
60 =	6o.	
55		
50 =	5o.	2o.
45		
40 =	4o.	
35		
30 =	3o.	
25 =		1o.
20 =	2o.	
15		
10 =	1o.	

RESUMEN

En este capítulo se nos presentaron algunas de las técnicas básicas utilizadas por el investigador social para organizar el conjunto de números crudos que recoge de sus

entrevistados. Las distribuciones de frecuencia y los métodos para comparar tales distribuciones de datos nominales (proporciones, porcentajes, razones y tasas) fueron discutidos y ejemplificados. Con respecto a los datos ordinales y por intervalos, se examinaron las características de las distribuciones de frecuencia simples, agrupadas y acumuladas. Finalmente, se presentó el procedimiento para obtener el rango porcentil de un porcentaje no procesado.

PROBLEMAS

1. De la siguiente tabla, que representa la agudeza visual de los televidentes y no televidentes, encontrar (a) el porcentaje de no televidentes con alta agudeza visual, (b) el porcentaje de televidentes con alta agudeza visual; la proporción de no televidentes con alta agudeza visual y (d) la proporción de televidentes con alta agudeza visual.

Agudeza visual en televidentes y no televidentes

	Estatus visual	
	No televidentes	*Televidentes*
Agudeza visual	*f*	*f*
Alta	93	46
Baja	90	127
Total	183	173

2. De la siguiente tabla, que representa estructuras familiares para niños negros y blancos, encontrar (a) el porcentaje de niños negros con familias de padre y madre, (b) el porcentaje de niños blancos con familias de padre y madre, (c) la proporción de niños negros con familias de padre y madre y (d) la proporción de niños blancos con familias de padre y madre

Estructura familiar para niños negros y blancos

	Raza del niño	
	Negra	*Blanca*
Estructura familiar	*f*	*f*
(Padre o Madre)	53	59
(Padre y Madre)	130	167
Total	183	226

3. En un grupo de 4 televidentes con alta gudeza visual y 24 con baja agudeza visual, ¿cuál es la razón de televidentes con agudeza visual alta y baja?
4. En un grupo de 125 hombres y 80 mujeres, ¿cuál es la razón de hombres a mujeres?
5. En un grupo de 15 niños negros y 20 niños blancos, ¿cuál es la razón de negros a blancos?
6. Si ocurren 300 nacimientos, entre 3 500 mujeres en edad de concebir, ¿cuál es la tasa de nacimiento?
7. ¿Cuál es la tasa de cambio para un aumento de población de 15 000 en 1950 a 25 000 en 1970?
8. Convertir la siguiente distribución de porcentajes a una distribución de frecuencia que contenga cuatro intervalos de clase, y (a) determinar el tamaño de los intervalos de clase, (b) indicar los límites superior e inferior de cada intervalo de clase, (c) identificar el punto medio de cada intervalo de clase, (d) encontrar la frecuencia acumulada por cada intervalo de clase, y (e) encontrar el porcentaje acumulado para cada intervalo de clase.

Puntajes	*f*
12	3
11	4
10	4
9	5
8	6
7	5
6	4
5	3
4	2
3	1
2	1
1	2
	$N = 40$

9. En la siguiente distribución de puntajes, encontrar el rango percentil para (a) un puntaje de 75 y (b) un puntaje de 52.

Intervalo de clase	*f*	*fa*
90–99	6	48
80–89	9	42
70–79	10	33
60–69	10	23
50–59	8	13
40–49	5	5
	$N = 48$	

10. En la siguiente distribución de puntajes, encontrar el rango percentil para (a) un puntaje de 36 y (b) un puntaje de 18.

Intervalo de clase	*f*
40–44	5
35–39	5
30–34	8
25–29	9
20–24	10
15–19	8
10–14	6
5–9	5
	$N = 56$

3
Gráficas

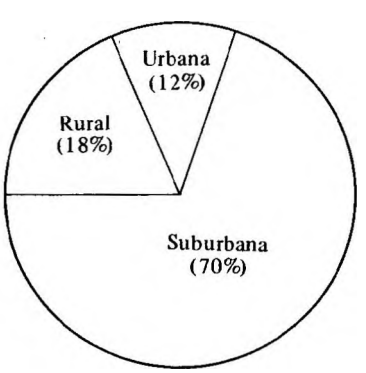
Sabemos muy bien que las columnas de números evocan temor, aburrimiento, apatía e incomprensión. Algunas personas parecen no tener interés en la información estadística presentada en forma tabular, pero podrían prestarle mucha atención a los mismos puntajes si les fueran presentados en forma de gráfica o cuadro. Como resultado, muchos investigadores comerciales y autores populares prefieren usar gráficas en contraposición a las tablas. Por motivos semejantes, los investigadores sociales usan frecuentemente gráficas tales como las gráficas de sectores, gráficas de barra y polígonos de frecuencia en un esfuerzo por aumentar el interés de sus hallazgos.

GRAFICAS DE SECTORES

Uno de los métodos gráficos más simples es el de la *gráfica de sectores,* una gráfica circular cuyos segmentos suman 100 por ciento. Las gráficas de sectores son particularmente útiles para visualizar las diferencias en frecuencia entre algunas categorías de nivel nominal. Para ilustrar. La Figura 3.1 presenta una población de 2 000 estudiantes universitarios de extracción urbana, suburbana o rural. Nótese que

FIGURA 3.1 Población de 2 000 estudiantes universitarios de extracción urbana, suburbana y rural

Extracción del estudiante	*f*	%
Urbana	240	(12)
Suburbana	1400	(70)
Rural	360	(18)
Total	2000	(100)



el 70% de estos estudiantes proviene de áreas suburbanas, mientras que sólo el 18% proviene de áreas rurales.

GRAFICAS DE BARRA

La gráfica de barra nos proporciona una ilustración sencilla y rápida de datos que pueden dividirse en unas cuantas categorías. Por comparación, la *gráfica de barra (o histograma)* puede acomodar cualquier número de categorías a cualquier nivel de medición y, por lo tanto, se utiliza más ampliamente en la investigación social.

Examinemos la gráfica de barra de la Figura 3.2 que ilustra una distribución de frecuencia de clases sociales. Esta gráfica de barra se construye siguiendo el orden estándar: una línea de base horizontal (o eje x) a lo largo de la cual se marcan los valores de los puntajes o categorías (en este ejemplo, las clases sociales) y una línea vertical (eje y) a lo largo del costado de la figura que representa las frecuencias por cada puntaje o categoría. (En el caso de los datos agrupados, los puntos medios de los intervalos de clase se ordenan a lo largo de la línea base horizontal.) Nótese que las barras rectangulares dan las frecuencias para la amplitud de los valores de los porcentajes. Mientras más alta es la barra, mayor es la frecuencia de ocurrencia.

En la Figura 3.2, las barras rectangulares de la gráfica se han unido para enfatizar los distintos grados de estatus social representados por diferencias de clases sociales. Además, las clases sociales se han trazado sobre la línea de base en orden *ascendente* de baja-baja a alta-alta. Este es el orden convencional para construir gráficas de barra de nivel ordinal y por intervalos.

Sin embargo, al dibujar una gráfica de barra de puntajes nominales, las barras deben estar *separadas,* y no unidas, para evitar implicar continuidad entre las categorías. Es más, las categorías de nivel nominal se pueden ordenar en cualquier forma a lo largo de la línea base horizontal. La Figura 3.3 ilustra tales características de las gráficas de barra de nivel nominal.

FIGURA 3.2 Gráfica de barra de una distribución de clases sociales

Clase social	*f*
Alta-alta	5
Alta-baja	14
Media alta	23
Media baja	45
Baja-alta	38
Baja-baja	25
Total	150

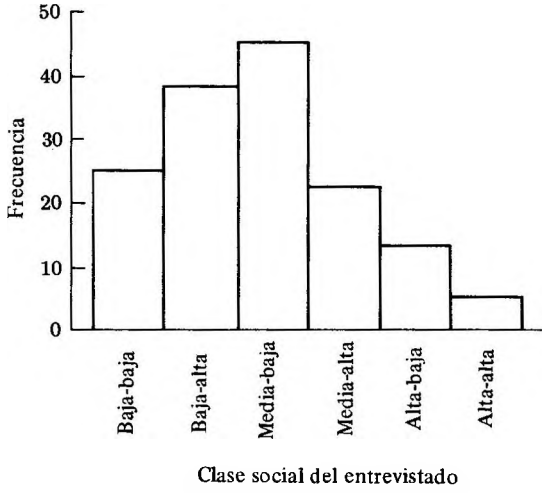
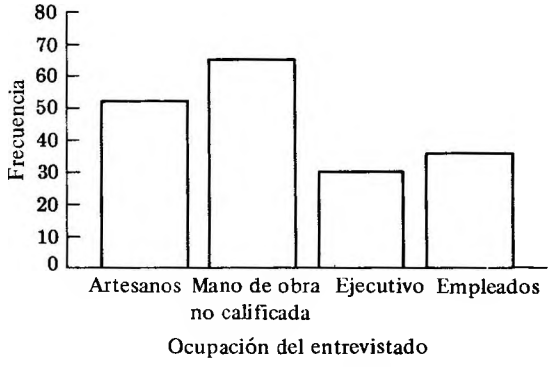


FIGURA 3. 3 Gráfica de barra de una distribución ocupacional

Ocupación	*f*
Artesanos	52
Mano de obra no calificada	65
Ejecutivo	29
Empleados	34
Total	180



POLIGONOS DE FRECUENCIA

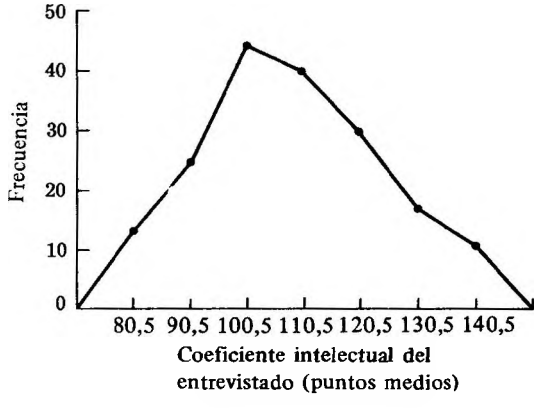
Otro método gráfico que se emplea comúnmente es el *polígono de frecuencia.* Aunque el polígono de frecuencia puede acomodar una amplia variedad de categorías, tiende a enfatizar la *continuidad,* a lo largo de una escala, más que las *diferencias* y es, por tanto, particularmente útil para representar puntajes ordinales y por intervalos. Esto se debe a que las frecuencias se indican por medio de una serie de puntos colocados sobre los valores de los puntajes o los puntos medios de cada intervalo de clase. Los puntos adyacentes se conectan mediante una línea recta que cae sobre la línea base en uno y otro extremo. Como lo muestra la Figura 3.4, la altura de cada punto indica la frecuencia de ocurrencia.

Para graficar frecuencias acumuladas (o porcentajes acumulados), puede construirse un *polígono de frecuencia acumulada.**

Como se ve en la Figura 3.5, las frecuencias acumuladas se ordenan a lo largo de la línea vertical de la gráfica y están indicadas por la altura de los puntos, sobre la línea base horizontal. Sin embargo, a diferencia de un polígono de frecuencia

FIGURA 3.4 Polígono de frecuencia de una distribución de puntajes de coeficiente intelectual

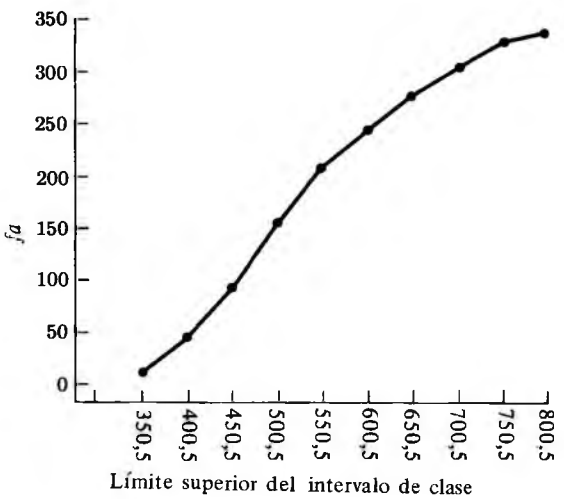
Intervalo de clase	*f*
136-145	11
126-135	16
116-125	29
106-115	40
96-105	44
86-95	25
76-85	13
Total	178



* N. del R. También se suele llamar ojiva.

FIGURA 3.5 Polígono de frecuencia acumulada para los datos de la tabla 2.8

Intervalo de clase	*f*	*fa*
751-800	6	336
701-750	25	330
651-700	31	305
601-650	30	274
551-600	35	244
501-550	55	209
451-500	61	154
401-450	48	93
351-400	33	45
301-350	12	12
	N = 336	



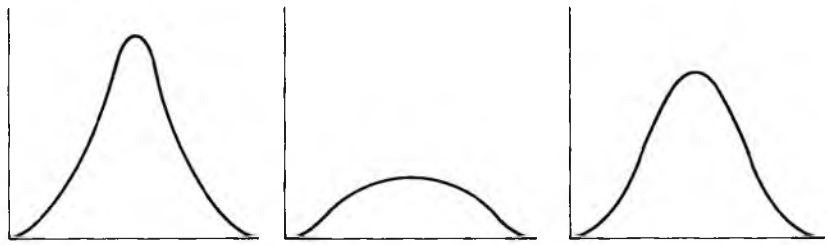
común, la línea recta que conecta todos los puntos del polígono de frecuencia acumulada no tiene que tocar otra vez la línea base horizontal, ya que las frecuencias acumuladas que se están representando son el producto de sumas sucesivas. Ninguna frecuencia acumulada es menor (generalmente es mayor) que la anterior. También, a diferencia de un polígono de frecuencia común, los puntos de una gráfica acumulada se trazan sobre los límites superiores de los intervalos de clase en lugar de sobre los puntos medios. Esto se debe a que la frecuencia acumulada representa el número total de casos *tanto dentro como por debajo* de un intervalo de clase en particular.

CONSTRUCCION DE GRAFICAS DE BARRA Y POLIGONOS DE FRECUENCIA

Las siguientes reglas y procedimientos pueden aplicarse a la construcción de gráficas de barra y polígonos de frecuencia:

1. Como una cuestión de tradición, y para evitar confusiones, el investigador siempre ordena los porcentajes a lo largo de la línea base horizontal y las frecuencias (o el porcentaje de casos) a lo largo de la línea vertical.
2. Toda gráfica debe ir completamente rotulada. La línea base horizontal debe rotularse en relación con las características (por ej., edad del entrevistado), la línea vertical debe rotularse de acuerdo con lo que se está representando (ya sean "frecuencias" o "porcentajes") y los valores numéricos de los puntos a lo largo de la escala. Además, la gráfica debe titularse indicando la naturaleza de los puntajes que se están ilustrando.
3. Al construir una gráfica, la longitud de la línea vertical debe ser como de un 75% de la longitud de la línea base horizontal. Este arreglo representa una manera relativamente estándar de dibujar gráficas y minimiza una fuente de confusión potencial.

FIGURA 3.6 Algunas variaciones de la curtosis entre las distribuciones simétricas



(a) Leptocúrticas (b) Platocúrticas (c) Mesocúrticas

4. El primer punto sobre la línea vertical —aquel punto en el cual se cruza con la línea horizontal— debe empezar siempre en cero, ya que cualquier otro orden podría dar una visión distorsionada de los puntajes.

FORMA DE UNA DISTRIBUCION DE FRECUENCIA

Los métodos gráficos pueden ayudarnos a visualizar la variedad de formas que toman las distribuciones de frecuencia. Algunas distribuciones son *simétricas;* al doblar la curva por el centro se crean dos mitades idénticas. Por lo tanto, tales distribuciones contienen el mismo número de valores extremos en ambas direcciones, alta y baja. Se dice que otras distribuciones están *sesgadas* y tienen más casos extremos en una dirección que en otra.

Existen variaciones considerables entre las distribuciones simétricas. Por ejemplo, pueden diferir marcadamente en términos de su *"puntiagudez"* (o *curtosis*). Algunas distribuciones simétricas, como en la Figura 3.6(a), son bastante picudas o altas (llamadas *leptocúrticas*); otras, como en la Figura 3.6(b), son bastante planas (llamadas *platocúrticas*) y, aun otras, no son ni muy picudas ni muy planas (llamadas *mesocúrticas*). Una clase de distribución simétrica mesocúrtica, como la que se muestra en la Figura 3.6(c), la *curva normal,* tiene especial importancia para la investigación social y se estudiará en detalle en el Capítulo 6.

Existe una variedad de distribuciones asimétricas o sesgadas. Cuando existe sesgo, apilándose los puntajes en una sola dirección, la distribución tendrá una "cola" pronunciada. La posición de esta cola indica dónde están localizados los relativamente pocos puntajes extremos y determina la *dirección* del sesgo.

La distribución (a) en la Figura 3.7 está *negativamente sesgada* (sesgada hacia la izquierda), ya que tiene una cola mucho más larga a la izquierda que a la derecha. Esta distribución indica que la mayoría de los entrevistados recibieron puntajes altos y que sólo unos cuantos obtuvieron puntajes bajos. Si se tratara de una distribución de calificaciones, en un examen final, podríamos afirmar que a la mayoría de los estudiantes les fue bastante bien y a unos cuantos mal.

Miremos ahora la distribución (b) cuya cola está situada a la derecha. Ya que la dirección de la cola indica el sesgo, podemos decir que la distribución está *positivamente sesgada* (sesgada hacia la derecha). ¡Las calificaciones del examen final de los estudiantes de nuestro hipotético grupo serían bastante bajas!

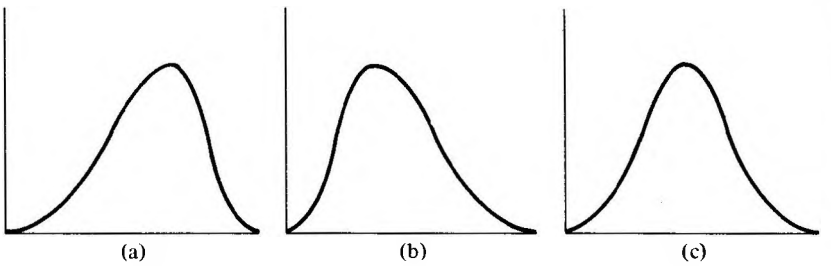


FIGURA 3.7 Tres distribuciones que representan la dirección del sesgo

Examinemos finalmente la distribución (c) que contiene dos colas idénticas. En tal caso, existe el mismo número de puntajes en ambas direcciones. La distribución no está en absoluto sesgada, sino que es perfectamente simétrica. Si se tratara de la distribución de calificaciones en nuestro examen final, tendríamos un gran número de estudiantes más o menos promedio y pocos alumnos que obtuvieran calificaciones altas o bajas.

RESUMEN

Las presentaciones gráficas de datos pueden usarse para aumentar la legibilidad de los hallazgos de la investigación. Nuestro análisis de las presentaciones gráficas incluyó gráficas de sectores, gráficas de barra y polígonos de frecuencia. Las gráficas de sectores nos dan una simple ilustración de los puntajes que pueden dividirse en unas cuantas categorías. Las gráficas de barra se utilizan más ampliamente, ya que pueden acomodar cualquier número de categorías. Los polígonos de frecuencia acomodan también un amplio rango de categorías, pero son especialmente útiles para datos ordinales y por intervalos, ya que enfatizan una continuidad a lo largo de la escala.

Las variaciones en la forma de las distribuciones pueden caracterizarse en términos de simetría o, si contienen más casos extremos en una dirección que en otra, en términos de sesgo positivo o negativo.

4

Medidas de tendencia central

Los investigadores, en muchos campos, han utilizado el término "promedio" para hacer preguntas tales como: ¿Cuál es el ingreso *promedio* que perciben los bachilleres y los profesionales? ¿Cuántos cigarrillos se fuma el adolescente *promedio*? ¿Cuál es el *promedio* de calificaciones de las universitarias? En *promedio*, ¿cuántos accidentes automovilísticos ocurren como resultado directo del alcohol o las drogas?

Una forma útil de describir a un grupo en su totalidad es encontrar un número único que represente lo "promedio" o "típico" de ese conjunto de puntajes. En la investigación social, ese valor se conoce como una *medida de tendencia central*, ya que está generalmente localizada hacia el medio o centro de una distribución en la que la mayoría de los puntajes tienden a concentrarse.

Lo que el lego quiere decir con el término "promedio" resulta a menudo vago y hasta confuso. La concepción del investigador social es mucho más precisa que la de uso popular; se expresa numéricamente como una entre varias clases distintas de mediciones de "promedio" o tendencia central que puede asumir valores numéricos bastante diferentes en el mismo conjunto de puntajes. Sólo trataremos aquí de las tres medidas de tendencia central más conocidas: la *moda,* la *mediana* y la *media*.

LA MODA

Para obtener la moda (Mo), simplemente buscamos el puntaje o categoría que ocurre más frecuentemente en una distribución. La moda puede encontrarse fácilmente por inspección más que por cálculo. Por ejemplo, en el conjunto de datos (1), 2, 3, (1), (1), 6, 5, 4, (1), 4, 4, 3, la moda es 1, ya que es el número que ocurre más que cualquier otro en el conjunto (ocurre 4 veces).

En el caso de una distribución de frecuencia simple en la que los valores de los puntajes y las frecuencias se presentan en columnas separadas, la moda es el valor

TABLA 4.1
Cómo buscar la moda en una distribución de frecuencia simple

	Valor de los puntajes	*f*
	7	2
	6	3
	5	4
Mo ⟶	4	5
	3	4
	2	3
	1	2
	Total	23

que aparece más a menudo en la columna de frecuencia de la tabla. Por lo tanto, en la distribución de frecuencia simple localizada en la Tabla 4.1, Mo=4.

Algunas distribuciones de frecuencia contienen dos o más modas. En el siguiente conjunto de datos, por ejemplo, los puntajes 2 y 6 ocurren *ambos* más frecuentemente: 6,6,7,2,6,1,2,3,2,4. Gráficamente, tales distribuciones tienen dos puntos de frecuencia máxima, sugiriéndonos las dos jorobas del lomo de un camello. Nos referimos a estas distribuciones como *bimodales,* en contraste con la variedad *unimodal* más común, que tiene una sola joroba o punto de máxima frecuencia (ver Figura 4.1)

LA MEDIANA

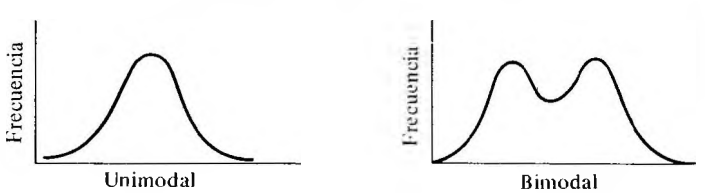
Cuando los puntajes ordinales o por intervalos, se organizan por orden de tamaño, resulta posible localizar la mediana (Mdn), el punto *más cercano al medio* en una distribución. Por lo tanto, se considera la mediana como la medida de tendencia central que corta la distribución en dos partes iguales.

Si tenemos un número impar de casos, entonces la mediana será el caso que cae exactamente en la mitad de la distribución. La posición del valor de la mediana puede localizarse por inspección o por fórmula.

$$\text{Posición de la mediana} = \frac{N + 1}{2}$$

Así, 16 es el valor de la mediana para los puntajes 11,12,13,(16),17,20,25; este es el caso en que divide los números de manera que le quedan 3 números a cada lado. De acuerdo con la fórmula (7 + 1)/2, vemos que la mediana 16 es el cuarto puntaje en la distribución, contando desde cualquiera de los 2 extremos.

FIGURA 4.1 Presentaciones gráficas de distribuciones unimodales y bimodales



Si el número de casos es par, la mediana es siempre aquel *punto* sobre el cual cae el 50% de los casos y bajo el cual cae el otro 50% de los mismos. Para un número par de casos habrá dos casos medios. Para ilustrar, los números 16 y 17 representan los casos medios para los siguientes puntajes: 11,12,13,⑯,⑰, 20,25,26. Por la fórmula (8 + 1)/2 = 4,5, la mediana caerá a mitad de camino entre el cuarto y el quinto caso; el punto más cercano al medio en esta distribución resulta ser 16,5 ya que está a medio camino entre 16 y 17, los puntajes cuarto y quinto del conjunto. De igual forma, la mediana es 9 en los puntajes 2,5,8,10,11,12, nuevamente por estar situado exactamente a medio camino entre los dos casos medios (6 + 1)/2 = 3,5.

Debemos explicar e ilustrar otra circunstancia: tal vez nos pidan que busquemos la mediana de puntajes que contienen varios puntajes medios de idéntico valor numérico. La solución es simple: la mediana es el valor numérico. Por lo tanto, en los puntajes 11,12,13, 16, 16, 16, 25,26,27, el caso mediano es 16, a pesar de que ocurre más de una vez.

Cómo obtener la mediana de una distribución de frecuencia simple

Para encontrar la mediana de puntajes ordenados en forma de distribución de frecuencia simple, comenzamos con el procedimiento que acabamos de ver. En el caso de la Tabla 4.1,

$$\text{Posición de la mediana} = \frac{23 + 1}{2} = \frac{24}{2} = 12$$

La mediana resulta ser el duodécimo puntaje en esta distribución de frecuencia. Para ayudar a localizar este duodécimo puntaje, podríamos construir una distribución de frecuencia acumulada como se muestra en la tercera columna de la Tabla 4.2 (esto puede hacerse mentalmente para un número pequeño de puntajes). Comenzando con el valor más bajo, sumamos frecuencias hasta llegar al duodécimo puntaje

TABLA 4. 2 Cómo encontrar la mediana para una distribución de frecuencia simple

Valores del puntaje		*f*	*fa*
7		2	23
6		3	21
5		4	18
Mdn ⟶ 4		5	14
3		4	9
2		3	5
1		2	2
	Total	23	

en la distribución. En el presente ejemplo, la mediana de los valores de los puntajes es 4.

LA MEDIA

La medida de tendencia central más comúnmente utilizada, la media aritmética $\overline{X}$, puede obtenerse sumando un conjunto de porcentajes y dividiendo entre el número de éstos. Por lo tanto, definimos la media más formalmente como *la suma de un conjunto de puntajes dividido entre el número total de puntajes del conjunto.* Por fórmula,

$$\overline{X} = \frac{\Sigma X}{N}$$

donde

$\overline{X}$ = la media (léase X barra)
Σ = la suma (expresada como la letra mayúscula griega sigma)[1]
X = un puntaje no procesado en un conjunto de datos
N = el número total de puntajes en un conjunto.

Aplicando la fórmula arriba expuesta, encontramos que la media del coeficiente intelectual de los 8 entrevistados listados en la Tabla 4.3 es 108.

TABLA 4.3 Cómo calcular la media: un ejemplo

Entrevistado	*X(C.I.)*	
Leticia	125	
Francisco	92	$\overline{X} = \frac{\Sigma X}{N}$
Sara	72	
Miguel	126	
Rebeca	120	$= \frac{864}{8}$
Rocío	99	
Benjamín	130	
Pablo	100	$= 108$
	$\Sigma X = 864$	

A diferencia de la moda, la media no es siempre el puntaje que ocurre más a menudo. A diferencia de la mediana, no es necesariamente el punto más cercano al medio en una distribución. Entonces, ¿qué significa *media*? ¿cómo puede interpretarse? Como veremos, la media puede considerarse como el "centro de gravedad", el

[1] La letra mayúscula griega sigma (Σ) se encontrará muchas veces en el texto. Indica simplemente que debemos *sumar* lo que sigue. En el presente ejemplo, ΣX indica sumar los porcentajes crudos o no procesados.

punto alrededor del cual las desviaciones positivas y negativas de cualquier distribución se equilibran. Para comprender esta característica de la media, debemos comprender primero el concepto de *desviación*, que indica la distancia entre cualquier puntaje no procesado y la media. Para encontrar la desviación, simplemente le restamos la media a cualquier puntaje no procesado. De acuerdo con la fórmula,

$$x = X - \bar{X}$$

donde

x = el puntaje de desviación (simbolizado siempre por x minúscula)
X = **cualquier puntaje no procesado en la distribución**
$\bar{X}$ = la media

TABLA 4.4 Desviaciones de un conjunto de puntajes no procesados de $\bar{X}$

X	x		
9	+3	+5	
8	+2		
6	0		$\bar{X} = 6$
4	−2	−5	
3	−3		

Como $X = 6$ para el conjunto de puntajes no procesados 9,8,6,4, y 3, el puntaje no procesado 9 se encuentra exactamente 3 unidades de puntajes no procesados *por sobre* la media de 6 (o $X - \bar{X} = 9 - 6 = +3$). De igual forma, el puntaje no procesado 4 está 2 unidades de puntaje no procesado *por debajo* de la media (o $X - \bar{X} = 4 - 6 = -2$). Conclusión: mientras más grande es la desviación x, más grande es la distancia entre ese puntaje no procesado y la media de la distribucion.

Considerando la media como un punto de equilibrio en la distribución, podemos decir ahora que la suma de las desviaciones que caen por encima de la media es igual en valor absoluto (haciendo caso omiso de los signos menos) a la suma de las desviaciones que caen por abajo de la media. Volvamos a un ejemplo anterior, al conjunto de puntajes 9,8,6,4,3 en que $\bar{X} = 6$. Si la media para esta distribución es el "centro de gravedad", pasando por alto los signos menos, la suma de las desviaciones positivas (desviaciones de los puntajes no procesados 8 y 9) debieran igualar la suma de las desviaciones negativas (desviaciones de los puntajes no procesados 4 y 3). Como se indica en la Tabla 4.4, este resulta ser el caso, ya que la suma de las desviaciones por abajo de $\bar{X}$ (−5) es igual a la suma de las desviaciones por encima de $\bar{X}$ (+5).

Tomando otro ejemplo, 4 es la media para los números 1,2,3,5,6 y 7.

Vemos que la suma de las desviaciones por abajo de este puntaje es −6, mientras que la suma de las desviaciones por encima de él es + 6. Volveremos sobre el concepto de la desviación en los Capítulos 5 y 6.

Cómo obtener la media de una distribución de frecuencia simple

La fórmula $\bar{X} = \Sigma X/N$ sirve para obtener la media de un pequeño número de puntajes. Sin embargo, cuando tenemos un mayor número de casos podría ser más práctico, y se gastaría menos tiempo, calcular la media de una distribución de frecuencia por la fórmula

$$\bar{X} = \frac{\Sigma fx}{N}$$

en que

$\bar{X}$ = la media
X = el valor de un puntaje no procesado en la distribución
fX = un puntaje multiplicado por su frecuencia de ocurrencia
ΣfX = la suma de los $fX's$
N = el número total de puntajes

La Tabla 4.5 ilustra el cálculo de la media de una distribución de frecuencia simple.

TABLA 4.5 Cómo obtener $\bar{X}$ de una distribución de frecuencia simple

X	f	fX	
8	2	16	
7	3	21	
6	5	30	
5	6	30	$\bar{X} = \frac{\Sigma fX}{N} = \frac{132}{28} = 4{,}71$
4	4	16	
3	4	12	
2	3	6	
1	1	1	
	$N = 28$	$\Sigma fX = 132$	

COMPARACION DE LA MODA, LA MEDIANA Y LA MEDIA

Llega un momento en que el investigador social escoge una medida de tendencia central para una situación en una investigación particular. ¿Empleará la moda, la mediana o la media? Su decisión involucra varios factores que incluyen:

1. El nivel de medición,
2. la forma de distribución de sus puntajes, y
3. el objetivo de la investigación.

Nivel de medición

Como la moda requiere sólo un conteo de frecuencia, puede aplicarse a cualquier conjunto de datos en el nivel de medición nominal, ordinal o por

intervalos. Por ejemplo, podríamos determinar que la categoría modal en una medición de nivel nominal de afiliaciones religiosas (protestante, católica y judía) es "protestante", ya que el mayor número de nuestros entrevistados se identifican como tales. Del mismo modo, podríamos saber que el mayor número de estudiantes que asisten a una universidad privada tiene un promedio de 2.5 (Mo = 2,5).

La mediana requiere un ordenamiento de categorías de la más alta a la más baja. Es por esto que sólo puede obtenerse a partir de datos ordinales o por intervalos y *no* de datos nominales. Para ilustrar, podríamos encontrar que la mediana de los ingresos anuales entre los dentistas de un pequeño pueblo es $17 000. Este resultado nos da una forma signficativa de examinar la tendencia central de nuestros datos. Por contraste, tendría poco sentido que fuéramos a calcular la mediana para escalas de afiliación religiosa (protestante, católica o judía), sexo (masculino o femenino) o país u origen (Inglaterra, Polonia, Francia o Alemania), cuando no se ha realizado una categorización o ajuste a una escala.

El uso de la media se restringe exclusivamente a los datos por intervalos. Su aplicación a datos ordinales o nominales da un resultado sin significado que generalmente no indica en absoluto la tendencia central. ¿Qué sentido tendría calcular la media para una distribución de afiliación religiosa o de sexo? Aunque es menos obvio, es igualmente inapropiado calcular una media para datos que pueden categorizarse pero no puntuarse.

Forma de la distribución

La forma de una distribución es otro factor que puede influir en la elección de la medida de tendencia central que haga el investigador. En una distribución unimodal perfectamente simétrica, la moda, la mediana y la media serán idénticas, ya que el punto de máxima frecuencia (Mo) es también el puntaje más cercano a la mediana (Mdn), así como el "centro de gravedad" ($\overline{X}$). Como se muestra en la Figura 4.2, las medidas de tendencia central coincidirán en el punto más central, en el "pico" de la distribución simétrica.

Cuando el investigador social trabaja con una distribución simétrica, su elección de la medida de tendencia central se basará principalmente en sus objetivos particu-

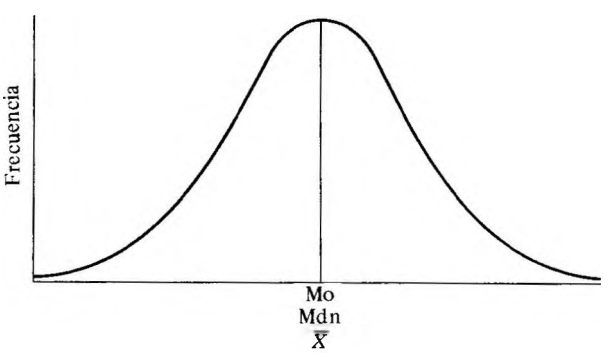
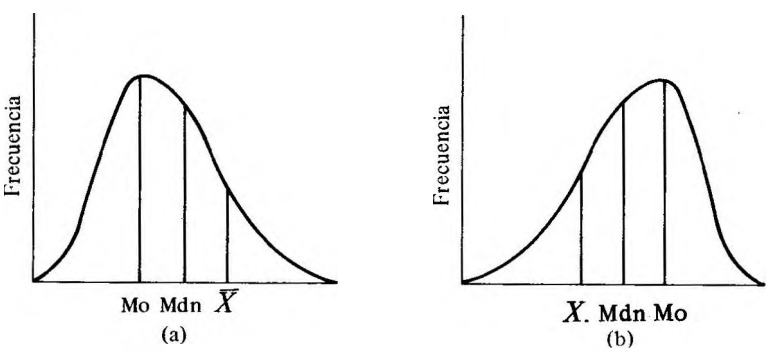


FIGURA 4.2 Una distribución simétrica, unimodal, que demuestra que la moda, la mediana y la media asumen valores idénticos

lares de investigación y en el nivel a que estén medidos sus datos. Sin embargo, cuando trabaje con una distribución sesgada su decisión estará muy influida por la forma de sus datos.

Como lo demuestra la Figura 4.3, la moda, la mediana y la media no coinciden en las distribuciones sesgadas, *a pesar de que sus posiciones relativas permanecen constantes* –alejándose del "pico" y acercándose a la "cola"–, el orden es siempre de moda, a mediana y a media. La moda cae más cerca del "pico" de la curva, ya que este es el punto en que ocurren los puntajes más frecuentes. Por contraste, la media se encuentra más cerca de la "cola", donde están localizados relativamente pocos valores de puntajes extremos. Por este motivo, el puntaje medio en la distribución sesgada positivamente de la Figura 4.3 (a) se encuentra cerca de los valores altos; la media en la distribución sesgada negativamente de la Figura 4.3 (b) cae cerca de los valores bajos.

FIGURA 4.3 Posiciones relativas de medidas de tendencia central en (a) una distribución sesgada positivamente y (b) una distribución sesgada negativamente



Mientras que la media está muy influida por los puntajes extremos en ambas direcciones, los cambios en los valores extremos modifican poco o nada la mediana. Esto se debe a que la media considera todos los puntajes en una distribución, mientras que, por definición, la mediana se entiende sólo con el valor numérico de puntaje que cae en la posición más cercana al medio de la distribución. Como se ilustra más adelante, el cambio del valor de un puntaje extremo de 10, en la distribución A, a 95 en la distribución B no modifica en absoluto el valor de la mediana (Mdn = 7,5), en tanto que la media varía de 7,63 a 18,25:

distribución A: 5 6 6 7 8 9 10 10 Mdn = 7,5 $\bar{X}$ = 7,63
distribución B: 5 6 6 7 8 9 10 95 Mdn = 7,5 $\bar{X}$ = 18,25

En una distribución sesgada, la mediana cae siempre en algún punto entre la media y la moda. Es esta característica la que convierte a la mediana en la medida de tendencia central más deseable para describir una distribución de puntajes sesgada. Para ilustrar esta ventaja de la mediana volvamos a la Tabla 4.6 y examinemos el salario anual "promedio" entre los empleados de una pequeña corporación. Si fuéramos publirrelacionistas contratados por una corporación para darle una imagen

pública favorable, probablemente querríamos calcular la media para demostrar que el empleado "promedio" gana $18 000 y está relativamente bien pagado. Por otra parte, si fuéramos representantes sindicales que buscan elevar los niveles salariales, querríamos, probablemente, emplear la moda para demostrar que el salario "promedio" es de sólo $1 000, una suma atrozmente baja. Finalmente, si fuéramos investigadores sociales buscando informar con exactitud sobre el salario "promedio" entre los empleados de la corporación, sabiamente emplearíamos la mediana ($3 000), ya que cae entre las otras medidas de tendencia central y da, por lo tanto, una visión más equilibrada de la estructura salarial. El método más aceptable sería el de dar a conocer las tres medidas de tendencia central y dejar que el público interpretase los resultados. Desafortunadamente, es cierto que pocos investigadores sociales –publirrelacionistas y los representantes sindicales– informan sobre más de una medida de tendencia central. Es más desafortunado aún el hecho de que algunos informes de investigación no especifican exactamente cuál medida de tendencia central –la moda, la mediana o la media– se utilizó para calcular la cantidad "promedio" o la posición dentro de un grupo de puntajes. Como lo demuestra la ilustración anterior, sería imposible una interpretación razonable de los descubrimientos si no se contara con tal información.

TABLA 4.6 Medidas de tendencia central de una distribución sesgada de salarios anuales

Salario	
$100 000	
25 000	$\bar{X}$ = $18 000
10 000	
5 000	Mdn = $3 000
1 000	
1 000	Mo = $1 000
1 000	
1 000	

Ya se anotó, anteriormente, que algunas distribuciones de frecuencia pueden caracterizarse como bimodales, ya que contienen dos puntos de frecuencia máxima. Para describir apropiadamente las distribuciones bimodales, generalmente es útil identificar *ambas* modas; el uso de la mediana o la media podría oscurecer aspectos importantes de tales distribuciones.

Consideremos la situación del investigador social que dirigió entrevistas con 26 personas de bajos ingresos para determinar cuál era su concepción ideal sobre el tamaño de su familia. A cada entrevistado se le preguntó: "Suponga que usted puede decidir exactamente qué tan grande debe ser su familia, ¿cuántas personas le gustaría ver en su familia ideal, incluyendo a todos los niños y adultos? " Como se muestra en la Tabla 4.7, los resultados de este estudio indicaron una amplia gama de preferencias en cuanto al tamaño de la familia; desde vivir solo (1) hasta vivir con muchas personas (10). Usando la media o la mediana, podríamos concluir que la familia ideal de los entrevistados constaba de seis miembros ($\bar{X}$ = 5,58; Mdn = 6). Sin embargo, sabiendo que la distribución es bimodal, vemos que estaban represen-

tadas, en realidad, dos concepciones ideales sobre el tamaño de la familia dentro del grupo de entrevistados: una con un número bastante grande de personas (Mo = 8), y la otra con sólo unas cuantas personas (Mo = 3).

El Objetivo de la Investigación

Hasta este punto, hemos estudiado la elección de una medida de tendencia central en términos del nivel de medición y de la forma de una distribución de los puntajes. Preguntamos ahora: ¿qué espera hacer el investigador social con su medida de tendencia central? Si busca una medición rápida, sencilla, pero crudamente descriptiva o si está trabajando con una distribución bimodal, empleará generalmente la moda. Sin embargo, en la mayoría de las situaciones que enfrenta el investigador, la moda sólo tiene utilidad como un indicador preliminar de la tendencia central que puede obtenerse rápidamente mediante una breve exploración de los puntajes. Si busca una medición precisa de la tendencia central, la decisión está generalmente entre la mediana y la media.

Para describir una distribución sesgada, el investigador generalmente escoge la mediana ya que (como se anotó anteriormente) tiende a dar un cuadro equilibrado de los puntajes extremos. La mediana se utiliza además como un punto de la distribución donde los puntajes pueden dividirse en dos categorías de acuerdo con preferencias sobre el tamaño familiar —aquéllos que prefieren una familia pequeña contra los que prefieren una familia grande.

Para una medida precisa de las distribuciones simétricas se tiende a preferir la media sobre la mediana, ya que la media puede usarse fácilmente en el análisis estadístico más avanzado, como el que se introduce en los capítulos subsiguientes del texto. Es más, la media es más estable que la mediana, ya que varía menos a través de las distintas muestras tomadas de cualquier población dada. Esta ventaja de la media —aunque quizás no haya sido entendida o apreciada por el estudiante— se hará más manifiesta en el subsiguiente estudio de la función de toma de decisiones de la estadística (ver Capítulo 7).

TABLA 4.7 Concepciones ideales sobre el tamaño de la familia entre 26 entrevistados de bajos ingresos: una distribución bimodal

Tamaño ideal de la familia	*f*
10	1
9	2
8	6
7	3
6	2
5	1
4	2
3	6
2	2
1	1
	$N = 26$

COMO OBTENER LA MODA, LA MEDIANA Y LA MEDIA DE UNA DISTRIBUCION DE FRECUENCIA AGRUPADA

En una distribución de frecuencia agrupada, la moda es el punto medio del intervalo de clase que tiene mayor frecuencia. De acuerdo con esta definición, la moda para la distribución situada en la Tabla 4.8 es 72, ya que éste es el punto medio del intervalo que ocurre más frecuentemente (ocurre 17 veces).

Para encontrar la mediana de los puntajes agrupados en una distribución de frecuencia, debemos (1) encontrar el intervalo de clase que contiene la mediana y (2) interpolar.

TABLA 4.8 Cómo obtener la moda de una distribución de frecuencia agrupada

Intervalo de clase	*Punto medio*	*f*
95–99	97	3
90–94	92	2
85–89	87	4
80–84	82	7
75–79	77	12
70–74	72	17
65–69	67	12
60–64	62	5
55–59	57	5
50–54	52	4
		$N = 71$

Paso 1—para localizar el intervalo mediano, construimos primero una distribución de frecuencia acumulada, como se indica en la tercera columna de la Tabla 4.9. Comenzando con el intervalo que contenga los valores más bajos (las edades menores, 20-29), sumamos las frecuencias hasta llegar al intervalo que contenga el caso que divide a la distribución en dos partes iguales, el puntaje más cercano al medio.

En el presente ejemplo, $N = 100$ y, por lo tanto, buscamos el quincuagésimo caso ($N/2 = 100/2 = 50$). Subiendo desde el intervalo más bajo, vemos que 26 de los casos tienen edades de 39 o menos. Vemos también que el quincuágesimo caso cae dentro del intervalo 40-49, ya que éste es el intervalo de clase cuyas frecuencias acumuladas contienen a 53 o a más de la mitad de los casos. En otras palabras, refiriéndose a las frecuencias acumuladas, los casos vigesimoséptimo hasta el quincuagésimotercero se encuentran dentro del intervalo 40-49. Esta es la mediana del intervalo.

TABLA 4.9 Una distribución de frecuencia agrupada por edades

Intervalo	*f*	*fa*
60–69	15	100
50–59	32	85
40–49	27	53
30–39	16	26
20–29	10	10
	$N = 100$	

Paso 2—Para encontrar el valor exacto de la mediana, aplicamos la fórmula

$$\text{Mediana} = \begin{array}{l}\text{Límite inferior}\\ \text{de la mediana}\\ \text{del intervalo}\end{array} + \left(\frac{\frac{N}{2} - \begin{array}{c}fa \text{ bajo el límite}\\ \text{inferior de la}\\ \text{mediana del intervalo}\end{array}}{f \text{ en la mediana del intervalo}} \right) \begin{array}{l}\text{tamaño del}\\ \text{intervalo}\end{array}$$

Para los datos de la Tabla 4.9, la mediana se determina como sigue:

$$\text{Mediana} = 39{,}5 + \left(\frac{50 - 26}{27}\right) 10$$
$$= 39{,}5 + 8{,}89$$
$$= 48{,}39$$

Para calcular la media de una distribución de frecuencia agrupada, puede utilizarse una versión modificada de la fórmula para una distribución de frecuencia simple (ver Tabla 4.5). Como se muestra abajo, el símbolo X ya no se usa para designar un puntaje, sino que se refiere al *punto medio de un intervalo de clase.* Por lo tanto,

$$\bar{X} = \frac{\Sigma fX}{N}$$

en que

$\bar{X}$ = la media

X = el punto medio de un intervalo de clase

fX = un punto medio multiplicado por el número de casos dentro de su intervalo de clase

N =el número total de puntajes

Podemos ilustrar el cálculo de una media de datos agrupados con referencia a la siguiente distribución:

Intervalo	f
17–19	1
14–16	2
11–13	3
8–10	5
5–7	4
2–4	2
	$N = 17$

PASO 1: Encontrar el punto medio de cada intervalo de clase

Intervalo	X = *punto medio*
17–19	18
14–16	15
11–13	12
8–10	9
5–7	6
2–4	3

PASO 2: Multiplicar cada punto medio por el número de casos dentro de su intervalo y obtener ΣfX

Intervalo	X = *punto medio*	f	fX
17–19	18	1	18
14–16	15	2	30
11–13	12	3	36
8–10	9	5	45
5–7	6	4	24
2–4	3	2	6
		$N = 17$	$\Sigma fX = 159$

PASO 3: Insertar el Resultado del Paso 2 en la Fórmula para $\overline{X}$

$$\overline{X} = \frac{\Sigma fX}{N}$$

$$= \frac{159}{17}$$

$$= 9{,}35$$

RESUMEN

Este capítulo ha presentado las tres medidas de tendencia central más conocidas, medidas de lo que es "promedio" o "típico" en un conjunto de datos. Se definió la moda como la categoría o puntaje que ocurre más a menudo; se consideró la mediana como el punto más cercano al medio en una distribución; la media se consideró como la suma de un conjunto de puntajes dividida entre el número total de puntajes en un conjunto. Se compararon estas medidas de tendencia central considerando el nivel de medición, la forma de su distribución y el objetivo de la investigación. Podemos resumir esas condiciones para elegir entre tres medidas de la siguiente manera:

Moda:

1. Nivel de medición: nominal, ordinal o por intervalos.
2. Forma de la distribución: más apropiada para la bimodal.
3. Objetivo: medida de tendencia central rápida y sencilla pero aproximativa.

Mediana:

1. Nivel de medición: ordinal o por intervalos
2. Forma de la distribución: más apropiada para las altamente sesgadas.
3. Objetivo: medición precisa de la tendencia central, puede utilizarse a veces para operaciones estadísticas más avanzadas o para dividir las distribuciones en dos categorías (por ejemplo, alto contra bajo).

Media:

1. Nivel de medición: por intervalos
2. Forma de la distribución: más apropiada para las simétricas unimodales.
3. Objetivo: medición precisa de la tendencia central, puede utilizarse a menudo para operaciones estadísticas más avanzadas, incluyendo pruebas para tomar decisiones de las que se tratará en los capítulos subsiguientes del texto.

PROBLEMAS

1. Los salarios por hora de siete empleados de una pequeña compañía son \$9, \$8, \$9, \$4, \$1, \$6, y \$3. Encontrar (a) el salario modal por hora, (b) el salario mediano por hora y (c) el salario medio por hora.
2. Supongamos que la pequeña compañía del Problema 1 contrató a otro empleado con un salario de \$1 por hora, dando por resultado los siguientes salarios por hora: \$9, \$8, \$9, \$4, \$1, \$6, \$3 y \$1. Encontrar (a) el salario modal por hora, (b) el salario mediano por hora, (c) el salario medio por hora.
3. Encontrar (a) la moda, (b) la mediana y (c) la media para los puntajes 205, 6, 5, 5, 5, 2 y 1. ¿Qué medida de tendencia central *no* usaría para describir este conjunto de puntajes? ¿Por qué?
4. Seis alumnos de un seminario de sociología fueron interrogados mediante una medición de nivel por intervalos respecto de su actitud hacia los portorriqueños. Sus respuestas en la escala de 1 a 10 (los valores de puntajes más altos indican actitudes más favorables hacia los portorriqueños) fueron como sigue: 5, 2, 6, 3, 1 y 1.

   Buscar (a) la moda (b) la mediana y (c) la media para los anteriores puntajes de actitud. En conjunto, ¿qué tan favorables eran estos estudiantes hacia los portorriqueños?
5. Buscar (a) la moda (b) la mediana y (c) la media para los puntajes 10, 12, 14, 8, 6, 7, 10, 10.
6. Buscar (a) la moda (b) la mediana y (c) la media para los puntajes 3, 3, 4, 3, 1, 6, 5, 6, 6, 4.
7. Encontrar (a) la moda (b) la mediana y (c) la media para los puntajes 8, 8, 7, 9, 10, 5, 6, 8, 8.
8. Buscar (a) la moda (b) la mediana y (c) la media para los puntajes 5, 4, 6, 6, 1, y 3.

9. Buscar (a) la moda (b) la mediana y (c) la media para los puntajes 8, 6, 10, 12, 1, 3, 4, 4.
10. Buscar (a) la moda (b) la mediana y (c) la media para los puntajes 12, 12, 1, 12, 5, 6, 7.
11. ¿Cuál es la desviación de cada uno de los siguientes puntajes de una media de 20,5? (a) $X = 20,5$; (b) $X = 33,0$; (c) $X = 15,0$; (d) $X = 21,0$.
12. ¿Cuál es la desviación de cada uno de los siguientes puntajes de una media de 3,0? (a) $X = 4,0$, (b) $X = 2,5$; (c) $X = 6,3$; (d) $X = 3,0$.
13. ¿Cuál es la desviación de cada uno de los siguientes puntajes de una media de 15? (a) $X = 22,5$; (b) $X = 3$; (c) $X = 15$; (d) $X = 10,5$;
14. Los puntajes de actitudes hacia los portorriqueños, de 31 estudiantes, se ubicaron en la siguiente distribución de frecuencia (los puntajes más altos indican actitudes más favorables hacia los portorriqueños):

Puntaje de actitud	f
7	3
6	4
5	6
4	7
3	5
2	4
1	2
	$N = 31$

Encontrar (a) la moda (b) la mediana y (c) la media.

15. Se pidió, a 31 niños matriculados en el 3er. curso elemental de una escuela urbana, que indicaran el número de sus hermanos y/o hermanas que vivieran en su hogar. Los datos resultantes se ordenaron en forma de distribución de frecuencia como sigue:

Número de hermanos	f
5	6
4	7
3	9
2	5
1	4
	$N = 31$

Encontrar (a) el número modal de hermano (b) el número mediano de hermanos y (c) el número medio de hermanos para este grupo de 31 estudiantes.

16. Encontrar (a) la moda (b) la mediana y (c) la media para la siguiente distribución de frecuencia:

Valores del puntaje	*f*
10	3
9	4
8	6
7	8
6	9
5	7
4	5
3	2
2	1
1	1
	$N = 46$

17. Encontrar (a) la moda (b) la mediana y (c) la media para la siguiente distribución de frecuencia agrupada:

Intervalo de clase	*f*
20–24	2
15–19	4
10–14	8
5–9	5
	$N = 19$

18. Encontrar (a) la moda (b) la mediana y (c) la media para la siguiente distribución de frecuencia agrupada:

Intervalo de clase	*f*
90–99	16
80–89	17
70–79	15
60–69	3
50–59	2
40–49	3
	$N = 56$

19. Encontrar (a) la moda (b) la mediana y (c) la media para la siguiente distribución de frecuencia agrupada:

Intervalo de clase	*f*
17–19	2
14–16	3
11–13	6
8–10	5
5–7	1
	$N = 17$

5

Medidas de dispersión o variabilidad

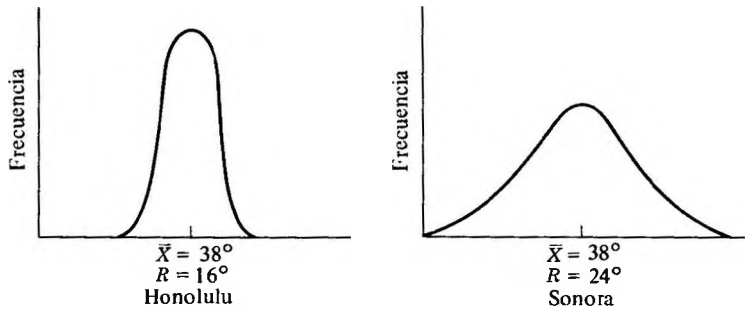
En el Capítulo 4 vimos que la moda, la mediana y la media podían usarse para resumir, en un sólo valor, lo que es "promedio" en una distribución. Sin embargo, cuando se usa cualquier medida de tendencia central, ésta nos da sólo un cuadro incompleto de un conjunto de datos y, por consiguiente, podría conducir tanto a conclusiones erróneas o distorsionadas como a una posible aclaración.

Para ilustrar esta posibilidad, supongamos que Honolulu, Hawaii y Sonora, México tienen la misma temperatura media de 38°C durante el día. ¿Podemos entonces suponer que la temperatura es básicamente igual en ambas localidades? O, ¿no es posible que una ciudad sea más apropiada que la otra para la natación y otras actividades al aire libre? Como se muestra en la Figura 5.1, la temperatura de Honolulu sólo tiene leves variaciones durante el año, fluctuando usualmente entre 33°C y 42°C. Por contraste, la temperatura en Sonora puede diferir, de estación en estación, de una mínima de cerca de 21°C en enero a una máxima de cerca de 45°C en julio y agosto. No es necesario decir que las playas de Sonora no se encuentran atestadas durante todo el año.

Tomemos otro ejemplo: supongamos que se ha encontrado que los ladrones y los profesores de secundaria, en una ciudad determinada, tienen el mismo ingreso anual medio de $ 8 000. ¿Indicaría *necesariamente,* este descubrimiento, que las dos distribuciones de ingresos son iguales? Por el contrario, podría encontrarse que difieren marcadamente en otro aspecto importante —o sea, que los ingresos de los profesores se agrupan estrechamente alrededor de los $ 8 000, mientras que los ingresos de los ladrones son mucho más irregulares, reflejando mayores oportunidades de encarcelamiento, desempleo y pobreza, así como de una riqueza poco usual.

Se puede ver que, además de una medida de tendencia central, necesitamos un índice de cómo están diseminados los puntajes alrededor del centro de la distribución. En una palabra, necesitamos una medida de lo que se conoce comúnmente

FIGURA 5.1 Diferencias de dispersión: La distribución de temperatura en Honolulu y Sonora (números aproximados)



como *dispersión* o *variabilidad.* Volviéndo sobre el ejemplo anterior, podríamos decir que la distribución de temperatura en Sonora, México, tiene *mayor variabilidad* que la distribución de temperatura en Honolulu, Hawaii. Del mismo modo, podemos decir que la distribución de ingresos entre los profesores tiene *menor variabilidad* que la distribución de ingresos entre los ladrones. Este capítulo trata sólo de las medidas de dispersión o variabilidad más conocidas: *el rango,* la *desviación media* y la *desviación estándar.*

EL RANGO

Para lograr una medida de dispersión rápida, pero aproximada, podríamos buscar lo que se conoce como el rango (R), o sea la diferencia entre el puntaje más alto y el más bajo de la distribución. Por ejemplo, si la temperatura más alta de Honolulu, en el año fue de 44°C y la más fría de 28°C, entonces el rango de la temperatura anual en Honolulu sería 16° C (44° – 28° = 16°). Si el día más caluroso en Sonora fue de 47°C y el más frío de 23°C, el rango de la temperatura en Sonora sería 24°C (47° – 23° = 24°C).

La ventaja del rango –su cálculo rápido y fácil– es a la vez su más importante desventaja. Es decir, que el rango depende totalmente de sólo dos valores de puntajes, del caso más grande y el más pequeño, en un determinado conjunto de datos dado. Como resultado, el rango generalmente da sólo un índice no procesado de la dispersión de la distribución. Por ejemplo, R = 98 en los datos 2, 6, 7, 7, 10, 12, 13, 100, (R = 100 – 2 = 98), mientras que R = 12 en los datos 2, 6, 7, 7, 10, 12, 13, 14, (R = 14 – 2 = 12). Por lo tanto, cambiando *un solo* puntaje (de 100 a 14), hicimos que el rango fluctuara bruscamente de 98 a 12. Cualquier medición que esté tan afectada por los puntajes de un sólo entrevistado, no puede darnos una idea precisa con respecto a la dispersión y, en el mejor de los casos, debe considerarse sólo como un índice preliminar o muy aproximado.

LA DESVIACION MEDIA

En el capítulo anterior se definió el concepto de desviación como la distancia entre cualquier porcentaje no procesado y su media. Para encontrar la desviación, se nos dijo que le restáramos la media a cualquier porcentaje no procesado ($x = X - \bar{X}$). Si

deseamos obtener ahora una medida de dispersión que tome en cuenta cada puntaje en una distribución (en vez de sólo dos valores), podríamos tomar la desviación absoluta (o distancia) entre cada puntaje y la media de la distribución ($|x|$), sumar estas desviaciones, y luego dividir esta suma entre el número de puntajes. El resultado sería la desviación media. Por fórmula,

$$DM = \frac{\Sigma|x|}{N}$$

en que

DM = la desviación media
$\Sigma|x|$ = la suma de las desviaciones absolutas (sin tomar en cuenta los signos + y −)
N = el número total de puntajes

Una nota importante: para llegar a $\Sigma|x|$, *debemos* pasar por alto los signos (+) y (−) y sumar valores absolutos. Esto es cierto porque la suma de las desviaciones reales (Σx) —desviaciones que usan signos para mostrar la dirección ya sea por encima o por abajo de la media— es siempre igual a cero. Las desviaciones positivas y negativas se cancelan a sí mismas y, por tanto, no pueden usarse para describir o comparar la dispersión de las distribuciones. Por contraste, la suma de las desviaciones absolutas tiende a agrandarse a medida que aumenta la dispersión o variabilidad de la distribución.

Podemos ilustrar ahora el procedimiento paso a paso para calcular la desviación media, considerando el conjunto de datos 9, 8, 6, 4, 2 y 1.

PASO 1: Buscar la Media para la Distribución

X	
9	
8	$\bar{X} = \frac{\Sigma X}{N}$
6	
4	$= \frac{30}{6}$
2	
1	$= 5$
$\Sigma X = 30$	

PASO 2: Restarle la media a cada puntaje no procesado (crudo) y sumar estas desviaciones (sin considerar sus signos)

X	x
9	+4
8	+3
6	+1
4	−1
2	−3
1	−4
$\Sigma X = 30$	$\Sigma\|x\| = 16$

PASO 3: Dividir $\Sigma|x|$ entre N para controlar el número de casos involucrados

$$DM = \frac{\Sigma|x|}{N}$$
$$= \frac{16}{6}$$
$$= 2{,}67$$

Siguiendo el procedimiento anterior, vemos que para el conjunto de datos 9, 8, 6, 4, 2 y 1, la desviación media es 2,67. Esto indica que, en promedio, los puntajes de esta distribución se desvían de la media por 2,67 unidades.

Para comprender mejor la utilidad de la desviación media, volvamos a las distribuciones de ingresos diarios (a), (b) y (c), tal como están localizadas en la Tabla 5.1. Nótese primero que la media de cada distribución es $ 20. Nótese también que parecen existir importantes diferencias de dispersión entre las distribuciones, diferencias que pueden detectarse con ayuda del rango y la desviación media.

Examinemos primero la distribución de ingresos (a) en la que todos los ingresos son exactamente iguales. Como todos los puntajes de esta distribución toman valores numéricos idénticos (20), podemos decir que la distribución (a) no tiene ninguna dispersión. Todos ganaron la misma cantidad de dinero ese día. Como resultado, el rango es 0 y no hay absolutamente ninguna desviación de la media (DM = 0). Las distribuciones (b) y (c) sí contienen dispersión. Más específicamente, la distribución (b) tiene un rango de 6 y una desviación media de 1,71; la distribución (c) tiene un rango de 30 y una desviación media de 8,57 Podemos afirmar, por lo tanto, que la distribución (b) contiene menor variabilidad que la distribución (c) –los ingresos de la distribución (b) son más parecidos que los ingresos de la distribución (c).

TABLA 5.1 Dispersión en las distribuciones de ingresos diarios que tienen la misma media ($ 20)

Distribución (a)		*Distribución (b)*		*Distribución (c)*	
X	$\|x\|$	X	$\|x\|$	X	$\|x\|$
$20	0	$23	+3	$35	+15
20	0	22	+2	30	+10
20	0	21	+1	25	+5
20	0	20	0	20	0
20	0	19	−1	15	−5
20	0	18	−2	10	−10
20	0	17	−3	5	−15
$\Sigma\|x\|$ = 0		$\Sigma\|x\|$ = 12		$\Sigma\|x\|$ = 60	
$\bar{X}$ = $20		$\bar{X}$ = $20		$\bar{X}$ = $20	
R = $ 0		R = $ 6		R = $30	
DM = $ 0		DM = $ 1,71		DM = $ 8,57	
Ninguna dispersión		Alguna dispersión		Mayor dispersión	

LA DESVIACION ESTANDAR

Por motivos que pronto serán evidentes, la desviación media ya no es utilizada ampliamente por los investigadores sociales; ha sido abandonada como medida de dispersión en favor de una más efectiva, *la desviación estándar.* Sin embargo, como veremos, la desviación media no puede considerarse como una pérdida de tiempo, ya que, por lo menos, nos da una base firme para comprender la naturaleza de la desviación estándar.

En un estudio previo vimos que la desviación media evita el problema de los números negativos, que cancelan a los positivos, pasando por alto los signos (+) y (–) y sumando las desviaciones absolutas de la media. Este procedimiento para crear una medida de variabilidad tiene la notoria desventaja de que tales valores absolutos no son siempre útiles en el análisis estadístico más avanzado (ya que no se pueden manipular algebraicamente con facilidad).

Para superar este problema y obtener una medida de dispersión que sea más tratable, en los procedimientos estadísticos más avanzados, podríamos *elevar al cuadrado las desviaciones reales de la media y sumarlas* (Σx^2). Como lo ilustra la Tabla 5.2, este procedimiento se libraría de los signos –ya que los números elevados al cuadrado son siempre positivos.

Después de sumar las desviaciones de la media elevadas al cuadrado, podríamos *dividir esta suma entre N para controlar el número de puntajes involucrados y obtener lo que se conoce como la media de estas desviaciones cuadráticas.* (*Nota:* Recuérdese que se siguió un procedimiento semejante para llegar a la desviación media cuando dividimos $\Sigma |x|$ entre N). Continuando con la ilustración de la Tabla 5.2, vemos que

$$\frac{\Sigma x^2}{N} = \frac{52}{6} = 8{,}67$$

Surge aún otro problema. Como resultado directo de la elevación al cuadrado de las desviaciones de la media, la unidad de medición ha cambiado, lo que hace que nuestro resultado 8,67 sea bastante difícil de interpretar. Tenemos 8,67 ¿pero 8,67 unidades de qué? Entonces, para regresar a nuestra unidad de medición original, *tomamos la raíz cuadrada de la media de las desviaciones elevadas al cuadrado:*

$$\sqrt{\frac{\Sigma x^2}{N}} = \sqrt{8{,}67} = 2{,}95$$

Definimos ahora la desviación estándar como el resultado de la anterior serie de operaciones, es decir, como *la raíz cuadrada de la media de las desviaciones de la media de una distribución elevadas al cuadrado.* Simbolizada por DE o por la letra minúscula griega sigma σ.

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}}$$

TABLA 5.2 Puntaje de desviaciones cuadráticas para eliminar los números negativos: en el ejemplo se utilizan los datos de la Tabla 5.1.

X	x	x^2
9	+4	16
8	+3	9
6	+1	1
4	−1	1
2	−3	9
1	−4	16
	$\Sigma x = 0$	$\Sigma x^2 = 52$

en que

σ = la desviación estándar
Σx^2 = la suma de las desviaciones de la media elevadas al cuadrado
N = el número total de puntajes

Para resumir, el procedimiento para calcular la desviación estándar no difiere mucho del método que vimos anteriormente para obtener la desviación media. En relación con el presente ejemplo, se desarrollan los siguientes pasos.

PASO 1: Encontrar la media para la distribución

X	
9	$\bar{X} = \frac{\Sigma X}{N}$
8	
6	$= \frac{30}{6}$
4	
2	$= 5$
1	
$\Sigma X = 30$	

PASO 2: Restar la media a cada puntaje no procesado para obtener la desviación

X	x
9	+4
8	+3
6	+1
4	−1
2	−3
1	−4

PASO 3: Elevar cada desviación al cuadrado antes de sumar las desviaciones elevadas al cuadrado

X	x	x^2
9	+4	16
8	+3	9
6	+1	1
4	−1	1
2	−3	9
1	−4	16
		$\Sigma x^2 = 52$

PASO 4: Dividir entre N y encontrar la raíz cuadrada del resultado

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}}$$
$$= \sqrt{\frac{52}{6}}$$
$$= \sqrt{8{,}67}$$
$$= 2{,}95.$$

Podemos decir ahora que la desviación estándar para el conjunto de datos 9, 8, 6, 4, 2 y 1 es 2,95.

La fórmula de los puntajes crudos o no procesados para DE

Hasta ahora se ha utilizado la fórmula $\sqrt{\Sigma x^2/N}$ para calcular la desviación estándar. Existe un método más sencillo para obtener DE —especialmente si hay una calculadora a la mano— un método que no requiere buscar las desviaciones, sino que trabaja directamente con los puntajes no procesados.

La fórmula de los puntajes crudos es

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - \overline{X}^2}$$

en la que

σ = la desviación estándar
ΣX^2 = la suma de los puntajes no procesados elevados al cuadrado (importante: cada puntaje no procesado se eleva al cuadrado *primero* y luego se suman estos puntajes no procesados elevados al cuadrado)
N = el número total de puntajes
$\overline{X}^2$ = la media elevada al cuadrado

El procedimiento paso a paso para calcular DE, por el método de los puntajes no procesados, puede ilustrarse volviendo sobre los datos de la Tabla 5.2.

PASO 1: Elevar cada puntaje no procesado al cuadrado antes de sumar los puntajes no procesados elevados al cuadrado

X	X^2
9	81
8	64
6	36
4	16
2	4
1	1
	$\Sigma X^2 = 202$

PASO 2: Obtener la media y elevarla al cuadrado

X	
9	
8	
6	$\bar{X} = \frac{\Sigma X}{N} = \frac{30}{6} = 5$
4	
2	$\bar{X}^2 = 25$
1	
$\Sigma X = 30$	

PASO 3: "Insertar" los resultados de los pasos 1 y 2 en la fórmula

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - \bar{X}^2}$$
$$= \sqrt{\frac{202}{6} - 25}$$
$$= \sqrt{33{,}67 - 25{,}00}$$
$$= \sqrt{8{,}67}$$
$$= 2{,}95$$

Como se mostró anteriormente, la aplicación de la fórmula de los puntajes no procesados a los datos de la Tabla 5.2 nos da exactamente el mismo resultado que el método original.

Cómo obtener la DE de una distribución de frecuencia simple

Para obtener la desviación estándar de datos ordenados en forma de distribución de frecuencia simple, aplicamos la fórmula

$$\sigma = \sqrt{\frac{\Sigma f X^2}{N} - \bar{X}^2}$$

Para ilustrar paso a paso, calculemos la desviación estándar de la siguiente distribución:

Valor de los puntajes	*f*
7	1
6	2
5	3
4	5
3	2
2	2
1	1
	$N = 16$

PASO 1: Multiplicar cada valor (X) por su f para obtener fX

X	f	fX
7	1	7
6	2	12
5	3	15
4	5	20
3	2	6
2	2	4
1	1	1

PASO 2: Multiplicar cada fX por X para obtener fX^2 (antes de sumar para obtener ΣfX^2)

X	fX	fX^2
7	7	49
6	12	72
5	15	75
4	20	80
3	6	18
2	4	8
1	1	1
		$\Sigma fX^2 = 303$

PASO 3: Obtener la media y elevarla al cuadrado

fX		
7		
12	$\bar{X} = \frac{\Sigma fX}{N}$	
15		
20	$= \frac{65}{16}$	$\bar{X}^2 = 16{,}48$
6		
4		
1	$= 4{,}06$	
$\Sigma fX = 65$		

PASO 4: "Insertar" los resultados de los pasos 1, 2 y 3 en la fórmula

$$\sigma = \sqrt{\frac{\Sigma f X^2}{N} - \overline{X}^2}$$
$$= \sqrt{\frac{3033}{16} - 16{,}48}$$
$$= \sqrt{18{,}94 - 16{,}48}$$
$$= \sqrt{2{,}46}$$
$$= 1{,}57$$

El significado de la desviación estándar

La serie de pasos que se requieren para calcular la desviación estándar puede dejar al estudiante con una sensación de incertidumbre con respecto al significado de su resultado. Por ejemplo, supongamos que encontramos que $\sigma = 4$ en una distribución particular de puntajes. ¿Qué nos indica este número? ¿Qué podemos exactamente decir ahora sobre esa distribución, que no pudimos haber dicho antes?

El siguiente capítulo buscará aclarar el significado completo de la desviación estándar. Por ahora, notemos brevemente que la desviación estándar (como la desviación media que le antecede) representa la "variabilidad promedio" de una distribución, ya que mide el promedio de desviaciones de la media. También entran a escena los procedimientos de elevar al cuadrado y sacar la raíz cuadrada pero, principalmente, con el fin de eliminar los signos (–) y volver a la unidad de medición más cómoda, la unidad del puntaje no procesado.

Notemos también que mientras mayor sea la dispersión alrededor de la media en una distribución, mayor será la desviación estándar. Así, $\sigma = 4{,}5$ indica una mayor variabilidad que $\sigma = 2{,}5$. Por ejemplo, la distribución de la temperatura diaria en Sonora, México, tiene una desviación estándar mayor que la que tiene la distribución de temperatura, en la misma época, en Honolulu, Hawaii.

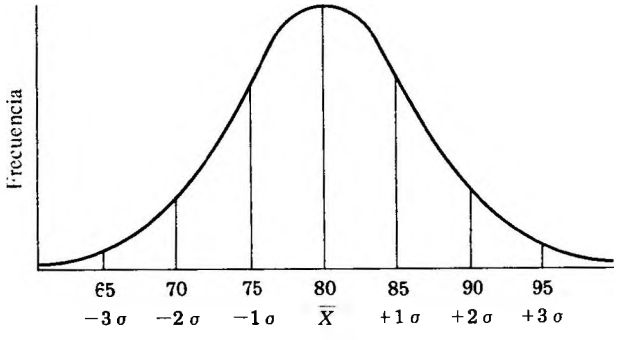
Si deseamos estudiar la distancia entre una mesa y la pared de la sala, podríamos pensar en términos de metros o centímetros como unidades de medición (por ejemplo, "la mesa de la sala está situada a 50 centímetros de esta pared"). Pero, ¿cómo medimos la anchura de la línea base de un polígono de frecuencia que contenga los puntajes de un grupo de entrevistados ordenados de bajo a alto (en orden ascendente)? Como un asunto relacionado, ¿cómo ingeniamos un método para encontrar la distancia entre cualquier puntaje no procesado y su media –un método estandarizado que permita comparaciones entre puntajes no procesados dentro de la misma distribución, así como entre diferentes distribuciones? Si estuviéramos hablando de mesas, podríamos encontrar que una está a 50 cm de la pared de la sala, mientras que la otra está a 100 cm de la pared de la cocina. Tenemos una unidad de medición estándar en el concepto de centímetros y, por lo tanto, podemos hacer tales comparaciones en forma significativa. Pero, ¿qué hay con las comparaciones entre puntajes crudos? Por ejemplo, ¿podemos siempre comparar un 85 en un examen de inglés con un 80 en alemán? ¿Cuál es en realidad la

calificación más alta? Un poco de reflexión nos mostrará que depende de cómo les haya ido a los otros estudiantes en cada clase.

Un método que da una estimación aproximada de la anchura de una línea base es el rango, ya que da la distancia entre los puntajes más alto y más bajo a lo largo de la línea base. Pero el rango no puede utilizarse efectivamente para situar un puntaje en relación con su media, ya que —aparte de sus otras debilidades— la amplitud cubre la anchura completa de la línea base. Por contraste, el tamaño de la desviación estándar es más pequeño que el del rango y usualmente cubre mucho menos que la anchura completa de la línea base.

Tal como medimos un tapete en centímetros o metros, también podríamos medir la línea base en unidades de desviación estándar (en unidades sigma). Por ejemplo, podríamos sumar la desviación estándar al valor de la media para encontrar cuál puntaje no procesado está situado exactamente a una desviación estándar (una distancia sigma) de la media. Por lo tanto, como lo muestra la Figura 5.2, si $\bar{X} = 80$ y DE = 5, entonces el puntaje no procesado 85 está exactamente una desviación estándar *por sobre* la media (80 + 5 = 85), una distancia de $+1\sigma$. Esta dirección es "más" porque todas las desviaciones *sobre* la media son positivas; todas las desviaciones por debajo de la media son "menos" o negativas.

FIGURA 5.2 Trazado de la línea base en unidades de desviación estándar cuando la desviación estándar (σ) es 5 y la media ($\bar{X}$) es 80



Continuamos trazando la línea base sumando el valor de la desviación estándar con el puntaje no procesado 85. Este procedimiento nos da el puntaje no procesado 90, que está exactamente a dos desviaciones estándar sobre la media (85 + 5 = 90). Del mismo modo, le sumamos la desviación estándar al puntaje no procesado y obtenemos 95, lo cual representa el puntaje no procesado que cae exactamente tres desviaciones estándar sobre la media. Para continuar el proceso por abajo de la media, restamos la desviación estándar de la media: restamos 5 de 80, 5 de 75 y 5 de 70 para obtener -1σ, -2σ, y -3σ.

Como se ilustra en la Figura 5.3, el proceso de trazado de la línea base en unidades de desviación estándar es, en muchos aspectos, igual que medir la distancia entre una mesa y la pared en unidades de centímetros. Sin embargo, la analogía se rompe en por lo menos un aspecto importante: mientras los centímetros y los metros son de dimensión constante (1 centímetro siempre es igual a la centésima

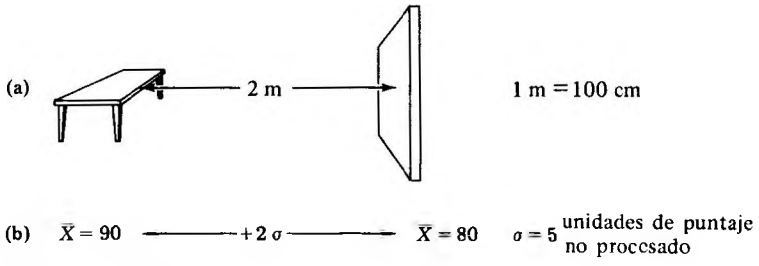


FIGURA 5.3 Medición de la distancia (a) entre una mesa y una pared en unidades de cm y (b) entre un puntaje no procesado y una media en unidades de desviación estándar

parte del metro, 1 metro siempre tendrá 100 cm), el valor de la desviación estándar varía de distribución a distribución. De otro modo, no podríamos utilizar la desviación estándar como se ilustraba anteriormente para comparar distribuciones en cuanto a su variabilidad (por ejemplo, DE = $ 5 000 para la distribución de ingresos de profesores de secundaria; DE = $ 15 000 para la distribución de ingresos de los ladrones). Por este motivo, debemos calcular el tamaño de la desviación estándar para cualquier distribución con la que estemos trabajando. Como resultado, es por lo general más difícil entender la desviación estándar en contraposición con centímetros o metros como unidad de medición. Volveremos sobre este concepto de la desviación estándar en el capítulo siguiente.

COMPARACION DEL RANGO, LA DESVIACION MEDIA Y LA DESVIACION ESTANDAR

El rango se considera meramente como un índice preliminar o aproximado de la variabilidad de una distribución. Es rápida y fácil de obtener, pero no muy confiable, y puede aplicarse a datos ordinales o por intervalos.

El rango tiene un propósito útil en relación con el cálculo de las desviaciones estándar. Como se ilustra en la Figura 5.2, seis desviaciones estándar cubren casi la distancia total entre el puntaje más alto y el más bajo en una distribución (-3σ a $+3\sigma$). Este sólo hecho nos proporciona un método conveniente para la estimación (pero no para el cálculo) de la desviación estándar. Generalmente, el tamaño de la desviación estándar es de aproximadamente un sexto del tamaño del rango. Por ejemplo, si el rango es de 36, entonces podría suponerse que DE cae cerca de 6; si el rango es 6, la DE, estará probablemente cerca de 1.

Esta regla puede revestir de una considerable importancia para el estudiante que desea saber si su resultado está cercano a lo correcto. Para tomar un caso extremo, si $R = 10$ y DE que hemos calculado, es 12, hemos cometido algún error, ya que DE no puede ser mayor que el rango. Una nota de precaución: la regla de un sexto es aplicable cuando tenemos un gran número de puntajes. Para un pequeño número de casos, habrá generalmente un número menor de desviaciones estándar para cubrir el rango de la distribución.

Mientras que el rango se calcula con sólo 2 valores numéricos, tanto la desviación estándar como la desviación media toman en cuenta cada valor en una distribución. Sin embargo, a pesar de su relativa estabilidad, la desviación media ya

no se utiliza ampliamente en la investigación social, ya que no puede emplearse en muchos análisis estadísticos avanzados. Por contraste, la desviación estándar emplea el procedimiento matemáticamente aceptable de despejar los signos en lugar de pasarlos por alto. Como resultado, la desviación estándar se ha convertido en el paso inicial para obtener ciertas medidas estadísticas, especialmente en el contexto de la toma de decisiones en estadística. Analizaremos esta característica de la desviación estándar en detalle en los capítulos subsiguientes, particularmente en los Capítulos 6 y 7.

A pesar de su utilidad como medida confiable de dispersión, la desviación estándar tiene también sus desventajas. Comparada con otras medidas de variabilidad, calcular la desviación estándar tiende a ser difícil y tardado. Sin embargo, esta desventaja está siendo superada más y más por el creciente uso de calculadoras de alta velocidad y computadoras para realizar análisis estadísticos. La desviación estándar (como la desviación media) tiene también la característica de ser una medida de nivel por intervalos y, por lo tanto, no puede usarse con datos nominales u ordinales —datos que frecuentemente les sirven a muchos investigadores sociales.

COMO OBTENER EL RANGO, LA DESVIACION MEDIA Y LA DESVIACION ESTANDAR DE DATOS AGRUPADOS

Ya sea que se trabaje con datos agrupados o no agrupados, el rango es siempre la diferencia entre los puntajes más altos y más bajos. No es necesario ningún método o fórmula especial.

A fin de ilustrar el procedimiento paso a paso para obtener la desviación media para una distribución de frecuencia agrupada, consideremos la siguiente distribución de frecuencia agrupada:

Intervalo de clase	*f*
17–19	1
14–16	2
11–13	3
8–10	5
5–7	4
2–4	2
	$N = 17$

PASO 1: Encontrar el punto medio de cada intervalo de clase

Intervalo	X = *punto medio*
17–19	18
14–16	15
11–13	12
8–10	9
5–7	6
2–4	3

PASO 2: Determinar la media de la distribución

X = *punto medio*	f	fX	
18	1	18	
15	2	30	$\bar{X} = \frac{\Sigma fX}{N}$
12	3	36	
9	5	45	$= \frac{159}{17}$
6	4	24	
3	2	6	$= 9{,}35$
		$\Sigma fX = 159$	

PASO 3: Encontrar la desviación, de cada punto medio, de la media

X = *punto medio*	$X - \bar{X} = \lvert x \rvert$
18	8,65
15	5,65
12	2,65
9	,35
6	3,35
3	6,35

PASO 4: Multiplicar cada puntaje de desviación por la frecuencia en el respectivo intervalo de clase y sumar estos productos

Intervalo	f	$\lvert x \rvert$	$f\lvert x \rvert$
17–19	1	8,65	8,65
14–16	2	5,65	11,30
11–13	3	2,65	7,95
8–10	5	,35	1,75
5–7	4	3,35	13,40
2–4	2	6,35	12,70
	$N = 17$		$\Sigma f\lvert x \rvert = 55{,}75$

PASO 5: Dividir entre N

$$\text{DM} = \frac{\Sigma f|x|}{N}$$

$$= \frac{55{,}75}{17}$$

$$= 3{,}28$$

Llegamos a una desviación media de 3,28.

Una fórmula de puntajes no procesados puede usarse para calcular la desviación estándar para una distribución de frecuencia agrupada. En términos de fórmula,

$$\sigma = \sqrt{\frac{\Sigma fX^2}{N} - \bar{X}^2}$$

en que

σ = la desviación estándar
f = la frecuencia en un intervalo de clase
X = el punto medio de un intervalo de clase
N = el número total de puntajes
$\overline{X}^2$ = la media elevada al cuadrado

El procedimiento paso a paso para encontrar la desviación estándar puede ilustrarse con referencia a los datos agrupados:

Intervalo de clase	f
17–19	1
14–16	2
11–13	3
8–10	5
5–7	4
2–4	2

PASO 1: Multiplicar cada punto medio por la frecuencia en el intervalo de clase y sumar estos productos

Intervalo de clase	f	*Punto medio* (X)	fX
17–19	1	18	18
14–16	2	15	30
11–13	3	12	36
8–10	5	9	45
5–7	4	6	24
2–4	2	3	6
	$N = 17$		$\Sigma fX = 159$

PASO 2: Obtener la media y elevarla al cuadrado

$$\overline{X} = \frac{\Sigma fX}{N}$$

$$= \frac{159}{17} \qquad \overline{X}^2 = 87{,}42$$

$$= 9{,}35$$

PASO 3: Multiplicar cada punto medio por fX y sumar estos productos

Intervalo de clase	f	*Punto medio* (X)	fX	fX^2
17–19	1	18	18	324
14–16	2	15	30	450
11–13	3	12	36	432
8–10	5	9	45	405
5–7	4	6	24	144
2–4	2	3	6	18
				$\Sigma fX^2 = 1773$

PASO 4: "Insertar" los resultados de los pasos 2 y 3 en la fórmula

$$\sigma = \sqrt{\frac{\Sigma f X^2}{N} - \overline{X}^2}$$
$$= \sqrt{\frac{1773}{17} - 87{,}42}$$
$$= \sqrt{104{,}29 - 87{,}42}$$
$$= \sqrt{16{,}87}$$
$$= 4{,}11$$

La desviación estándar resulta ser 4,11.

RESUMEN

En el presente capítulo nos han presentado el rango, la desviación media y la desviación estándar (tres medidas de dispersión o cómo los puntajes se encuentran dispersos alrededor del centro de una distribución). Se ha considerado el rango como un indicador rápido, pero muy general, de dispersión o variabilidad, que puede encontrarse fácilmente tomando la diferencia entre los puntajes más alto y más bajo en una distribución. La desviación media (la suma de las desviaciones absolutas dividida entre N) se trató como una medida de dispersión matemáticamente inadecuada, pero como una base sólida para comprender la desviación estándar, la raíz cuadrada del promedio de las desviaciones de la media elevadas al cuadrado. En la desviación estándar tenemos una medida de dispersión confiable, a nivel de intervalos, que puede utilizarse para operaciones estadísticas descriptivas y en toma de decisiones más avanzadas. El sentido completo de la desviación estándar se analizará en el subsiguiente estudio de la curva normal y de las generalizaciones de muestras a poblaciones.

PROBLEMAS

1. Los puntajes de examen obtenidos por un grupo de 5 estudiantes son 7, 5, 3, 2 y 1 sobre una escala de 10 puntos. Para este conjunto de puntajes, buscar (a) el rango (b) la desviación media y (c) la desviación estándar.
2. Sobre una escala diseñada para medir actitudes hacia la segregación racial, dos grupos universitarios lograron los siguientes puntajes:

Grupo A	*Grupo B*
4	3
6	3
2	2
1	1
1	4
1	2

Comparar la variabilidad de actitudes hacia la segregación racial entre los miembros de los grupos A y B calculando (a) el rango de los puntajes para cada grupo (b) la desviación media de los puntajes para cada grupo y (c) la desviación estándar de los puntajes para cada grupo. ¿Cuál grupo tiene mayor variabilidad de puntajes de actitud?

3. Para el conjunto de puntajes 3, 5, 5, 4, 1 hallar (a) el rango, (b) la desviación media y (c) la desviación estándar.
4. Para el conjunto de puntajes 1, 6, 6, 3, 7, 4, 10, calcular la desviación estándar.
5. Calcular la desviación estándar para el conjunto de puntajes 12, 12, 10, 9, 8.
6. Hallar la desviación estándar para la siguiente distribución de frecuencia de puntajes:

X	f
5	3
4	5
3	6
2	2
1	2
	$N = 18$

7. Hallar la desviación estándar para la siguiente distribución de frecuencia de puntajes:

X	f
7	2
6	3
5	5
4	7
3	4
2	3
1	1
	$N = 25$

8. Hallar la desviación estándar para la siguiente distribución de frecuencia de puntajes:

X	f
10	2
9	5
8	8
7	7
6	4
5	3
	$N = 29$

9. Hallar (a) el rango (b) la desviación media y (c) la desviación estándar para la siguiente distribución de frecuencia agrupada de puntajes:

Intervalo de clase	*f*
90–99	6
80–89	8
70–79	4
60–69	3
50–59	2
	$N = 23$

10. Hallar (a) el rango (b) la desviación media y (c) la desviación estándar para la siguiente distribución de frecuencia agrupada de puntajes:

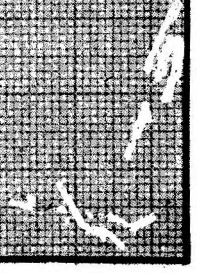
Intervalo de clase	*f*
17–19	2
14–16	3
11–13	6
8–10	5
5–7	1

11. Hallar (a) el rango (b) la desviación media y (c) la desviación estándar para la siguiente distribución de frecuencia agrupada de puntajes:

Intervalo de clase	*f*
20–24	2
15–19	4
10–14	8
5–9	5
	$N = 19$

PARTE II

De la descripción a la toma de decisiones



6

La curva normal

En los capítulos anteriores vimos que las distribuciones de frecuencia pueden tomar una variedad de formas. Algunas son perfectamente simétricas o libres de sesgo; otras son sesgadas ya sea negativa o positivamente y algunas otras, incluso, tienen más de una "joroba", etc. Dentro de esta gran diversidad existe una distribución de frecuencia con la cual muchos de nosotros ya estamos familiarizados, aunque sea sólo por las calificaciones que nos dan los instructores de acuerdo a la "curva". Esta distribución, que se conoce comúnmente como la *curva normal,* es un modelo teórico o ideal que se obtuvo de una ecuación matemática más que de una investigación y recolección de datos real.[1] Sin embargo, la utilidad de la curva normal, para el investigador social, puede verse en sus aplicaciones a las situaciones reales de investigación.

Como veremos en el presente capítulo, por ejemplo, la curva normal puede utilizarse para describir distribuciones de puntajes, para interpretar la desviación estándar y para hacer un informe de probabilidades. En los capítulos siguientes veremos que la curva normal es un ingrediente esencial en la toma de decisiones en estadística, por medio de la cual el investigador social generaliza sus resultados de muestras a poblaciones. Antes de proceder a un estudio de las técnicas de la toma de decisiones es necesario lograr primero una comprensión de las propiedades de la curva normal.

[1] La curva normal puede construirse con la fórmula

$$Y = \frac{N}{\sigma\sqrt{2\pi}} e^{-(X-\bar{X})^2/2\sigma^2}$$

donde

Y = la ordenada para un valor dado de X (frecuencia con que ocurre)
$\pi = 3{,}1416$
$e = 2{,}7183$

CARACTERISTICAS DE LA CURVA NORMAL

¿Cómo puede caracterizarse la curva normal? y ¿cuáles son las propiedades que la distinguen de otras distribuciones? Como lo indica la Figura 6.1, la curva normal es un tipo de curva uniforme y simétrica cuya forma recuerda a muchos una campana y por tanto se conoce como la "curva en forma de campana". Tal vez el rasgo más sobresaliente de la curva normal es su *simetría:* si doblamos la curva en su punto más alto al centro, crearíamos dos mitades iguales, cada una fiel imagen de la otra.

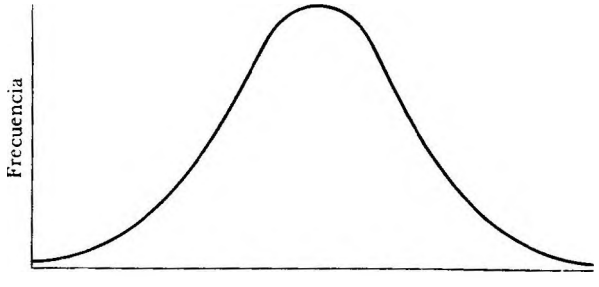
Además, la curva normal es *unimodal,* ya que sólo tiene un pico o punto de máxima frecuencia –aquel punto en la mitad de la curva en el cual coinciden la media, la mediana y la moda– (el alumno recordará que la media, la mediana y la moda ocurren en distintos puntos en una distribución sesgada, ver Capítulo 3). Desde el pico central redondeado de la distribución normal, la curva cae gradualmente en ambas colas, extendiéndose indefinidamente en una y otra dirección y acercándose más y más a la línea de base sin alcanzarla realmente.

CURVAS NORMALES: EL MODELO Y EL MUNDO REAL

Podríamos preguntarnos: ¿hasta qué punto se asemejan o aproximan las distribuciones de datos reales (esto es, los datos recogidos por los investigadores sociales en el curso de una investigación) a la forma de la curva normal? Imaginemos, con fines ilustrativos, que todos los fenómenos sociales, psicológicos y físicos estuvieran distribuidos normalmente, ¿cómo sería este mundo hipotético?

En lo concerniente a las características físicas de los humanos, la mayoría de los adultos caería dentro del campo de los 1,60 y 1,80 m de estatura, siendo muy pocos muy bajos (menos de 1,60 m) o muy altos (más de 1,90 m). Como lo muestra la Figura 6.2, el Coeficiente Intelectual (C.I.) sería igualmente predecible –la mayor proporción de puntajes de C.I. caerían entre 90 y 110; veríamos una caída gradual de los puntajes en una y otra cola con unos pocos "genios" que marcarían más de 140; igualmente, pocos marcarían menos de 60. De igual manera, relativamente pocos individuos se catalogarían como extremistas políticos, ya sea de derecha o izquierda, mientras que a la mayoría se les consideraría políticamente moderados o neutrales. Finalmente, hasta el patrón del uso resultante del flujo de tráfico en las entradas se

FIGURA 6.1 La forma de la curva normal



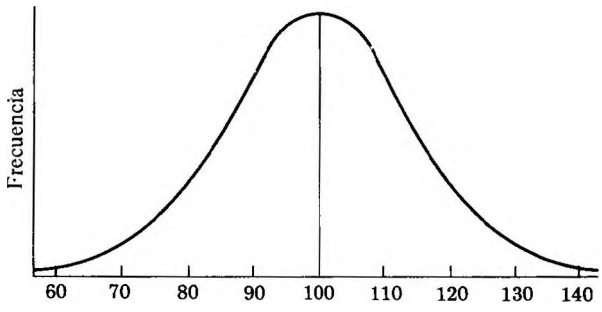


FIGURA 6.2 Distribución hipotética de puntajes de coeficiente intelectual

asemejaría a la distribución normal —el mayor uso ocurriría en el centro de la entrada, mientras que a uno y otro lado ocurrirían cantidades gradualmente decrecientes.

Hasta este punto, algunos lectores habrán notado que el mundo hipotético de la curva normal no difiere radicalmente del "mundo real" en que vivimos actualmente. De hecho, fenómenos tales como la estatura, el coeficiente intelectual, la orientación política y el uso en las entradas parecen aproximarse a la distribución normal teórica. Debido a que muchos fenómenos poseen esta característica, ya que ocurre frecuentemente en la naturaleza (y por otros motivos que luego conoceremos), los investigadores, en muchos campos, han hecho extensivo el uso de la curva normal aplicándola a los datos que recogen y analizan.

Pero debería anotarse también que algunos fenómenos, tanto en las ciencias sociales como en otros campos, simplemente no se ajustan a la noción teórica de la distribución normal. Muchas distribuciones son sesgadas; otras tienen más de un pico; algunas son simétricas pero no tienen forma de campana. Como un ejemplo concreto, consideremos la distribución de la riqueza en el mundo. Es muy bien sabido que los "desposeídos" superan en número a los "pudientes". Así, como lo muestra la Figura 6.3, la distribución de la riqueza (como lo indica el ingreso per cápita) está aparentemente muy sesgada, de tal manera que una pequeña proporción de la población mundial recibe una gran proporción del ingreso mundial. Del mismo modo, los especialistas en población nos dicen que los Estados Unidos se han

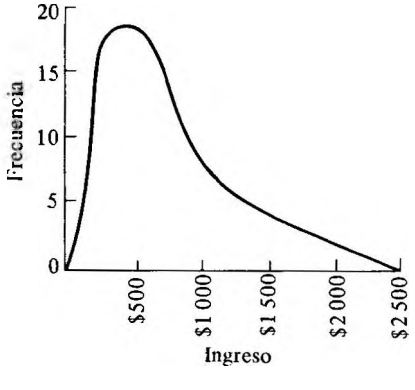


FIGURA 6.3 La distribución del ingreso per cápita entre las naciones del mundo (en dólares americanos)

convertido recientemente en una tierra de jóvenes y ancianos. Desde el punto de vista económico, esta distribución de edad representa una carga para una fuerza de trabajo relativamente pequeña, compuesta por ciudadanos de "mediana edad", que está manteniendo a un número desproporcionadamente grande de personas no productivas, tanto jubilados como jóvenes en edad escolar.

Cuando tenemos buenos motivos para suponer alejamientos radicales de la normalidad —como en el caso de la edad y el ingreso— la curva normal no puede usarse como un modelo de los datos que hemos obtenido. Por tanto, no puede aplicársele, a voluntad, a todas las distribuciones con que se encuentre el investigador, sino que debe usarse con una buena dosis de discreción. Afortunadamente, los estadísticos saben que muchos fenómenos de interés para el investigador social toman la forma de la curva normal.

EL AREA BAJO LA CURVA NORMAL

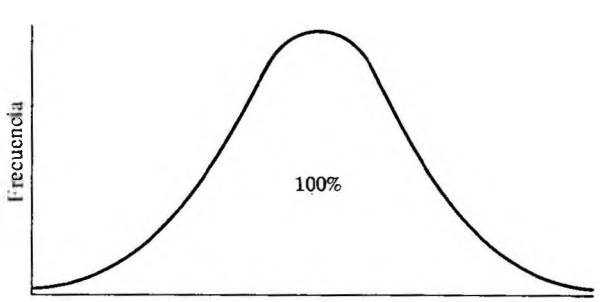
Para poder emplear la curva normal en la resolución de problemas, debemos familiarizarnos con el área bajo la curva normal: *aquella área que está entre la curva y la línea base y que contiene el 100 por ciento, o todos los casos, en una distribución normal dada.* La Figura 6.4 ilustra esta característica.

Podríamos encerrar una porción de esta área total dibujando líneas a partir de dos puntos cualesquiera en la línea base hasta la curva. Por ejemplo, usando la media como punto de partida, podríamos dibujar una línea en $\bar{X}$ y otra en el punto que está a 1 DE (una distancia sigma)* sobre $\bar{X}$. Como lo ilustra la Figura 6.5, esta porción sombreada de la curva normal incluye 34,13% de la frecuencia total.

De igual manera, podemos decir que el 47,72% de los casos, bajo la curva normal, están entre $\bar{X}$ y 2 DE_s arriba de la $\bar{X}$ y que el 49,87% están entre $\bar{X}$ y 3 DE_s arriba de la $\bar{X}$ (ver Figura 6.6).

Como veremos, *una proporción constante del área total, bajo la curva normal, estará entre la media y cualquier distancia dada de X, medida en unidades* DE. Esto es cierto a pesar de la media y la DE de la distribución en particular, y se aplica universalmente a todos los datos normales distribuidos. Así, el área bajo la curva normal entre $\bar{X}$ y el punto 1 DE arriba de la $\bar{X}$ incluye *siempre* el 34,13% del total de casos, así estemos estudiando la distribución de estatura, inteligencia, orientación

FIGURA 6.4 Area bajo la curva normal



* N. del R. Debemos anotar que el término "distancia sigma" se refiere a la misma "desviación estándar" pero "poblacional". Las mayúsculas "DE", en el capítulo anterior, indican una "desviación estándar muestral".

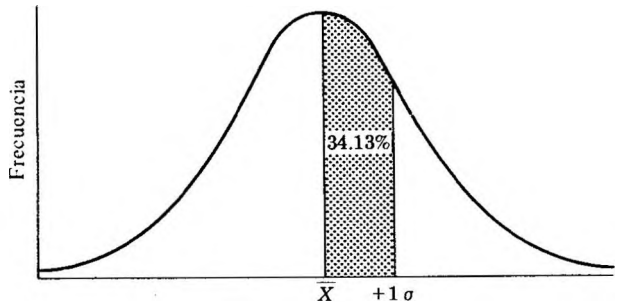


FIGURA 6.5 El porcentaje del área total bajo la curva normal entre $\bar{X}$ y el punto uno de desviación estándar arriba de la $\bar{X}$.

política o el patrón de uso en una entrada. El requisito básico, en cada caso, es sólo que estemos trabajando con una distribución *normal* de puntajes.

La naturaleza simétrica de la curva normal nos lleva a otra importante conclusión; a saber, que cualquier distancia sigma dada arriba de la media *contiene una proporción idéntica de casos que la misma distancia sigma por abajo de la media.* Así, si el 34,13% del área está entre la media y 1 DE por arriba de la $\bar{X}$, entonces el 34,13% del área total está entre la media y 1 DE por abajo de $\bar{X}$; si el 47,72% está entre la media y 2 DE_s por arriba de la $\bar{X}$, entonces el 47,72% está entre la media y 2 DE_s *por abajo de* $\bar{X}$; si el 49,87% está entre la media y 3 DE_s *por arriba de* $\bar{X}$, entonces el 49,87% está también entre la media y 3 DE_s *por abajo* de $\bar{X}$. En otras palabras, como se ilustra en la Figura 6.7, el 68,26% del área total de la curva normal (34,13% + 34,13%) caen entre -1σ y $+1\sigma$ de la media; el 95,44% del área (47,72% + 47,72%) caen entre -2σ y $+2\sigma$ de la media; el 99.74%, o casi todos los casos (49,87% + 49,87%) caen entre -3σ y $+3\sigma$ de la media. Puede decirse, entonces que 6 DE_s incluyen prácticamente todos los casos (más del 99%) bajo cualquier distribución normal.

ACLARANDO LA DESVIACION ESTANDAR: UNA ILUSTRACION

Una importante función de la curva normal es la interpretación y aclaración del significado de la desviación estándar. Para comprender cómo se realiza esta función,

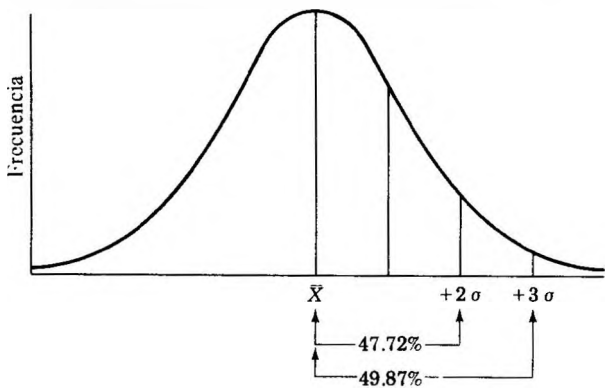


FIGURA 6.6 El porcentaje del área bajo la curva normal entre $\bar{X}$ y los puntos uno y dos de desviaciones estándar a partir de $\bar{X}$.

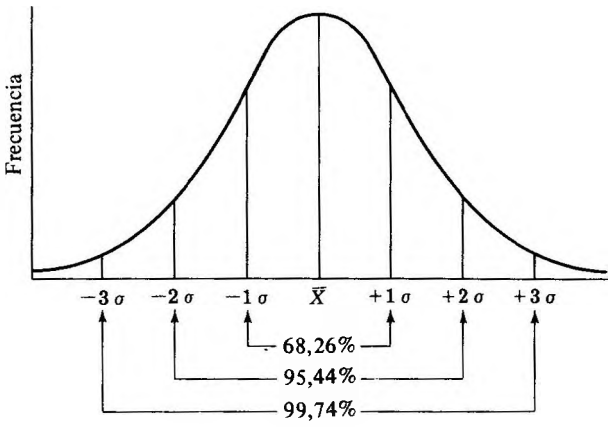


FIGURA 6.7 El porcentaje del área total bajo la curva normal entre -1σ y $+1\sigma$, -2σ y $+2\sigma$, y -3σ y $+3\sigma$

examinemos lo que nos dicen los antropólogos sobre las diferencias de sexo en cuanto al coeficiente intelectual. A pesar de las pretensiones de los chauvinistas, existen evidencias de que tanto los hombres como las mujeres tienen puntajes medios de coeficiente intelectual de aproximadamente 100. Digamos también que estos puntajes de coeficiente intelectual difieren marcadamente en términos de la variabilidad alrededor de la media. En particular, supongamos que los coeficientes intelectuales masculinos tienen mayor *heterogeneidad* que los femeninos, esto es, la distribución de los coeficientes intelectuales masculinos presenta un porcentaje mucho mayor de puntajes extremos que representan tanto a individuos muy inteligentes como a otros muy tontos, en tanto que la distribución de coeficientes femeninos tiene un mayor porcentaje localizado cerca del promedio, hallándose al centro el punto de máxima frecuencia.

Como la desviación estándar es una medida de variación, estas diferencias de sexo en la variabilidad deberían reflejarse en el valor de las DE en cada distribución de puntajes de coeficiente intelectual. Así, podríamos encontrar que la DE para los coeficientes intelectuales masculinos es 10, mientras que para los femeninos es de 5.

Conociendo la desviación estándar de cada conjunto de puntajes de coeficiente intelectual, y suponiendo que cada conjunto está distribuido normalmente, podría-

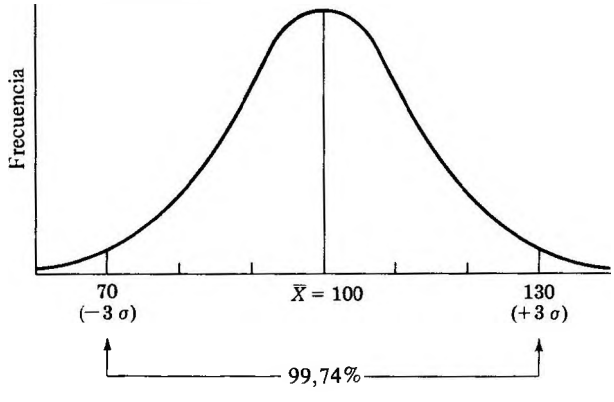


FIGURA 6.8 Una distribución de puntajes de coeficientes intelectuales masculinos

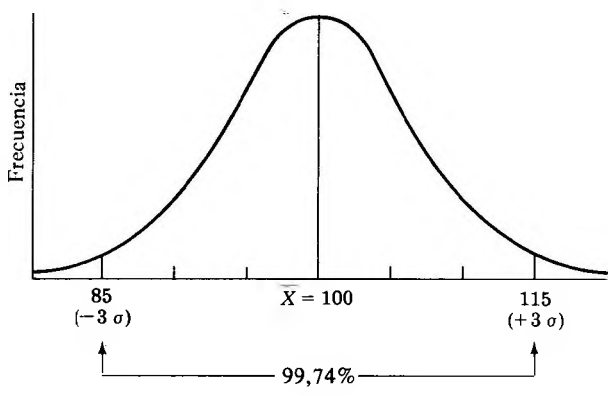


FIGURA 6.9 Una distribución de puntajes de coeficientes intelectuales femeninos

mos estimar y comparar el porcentaje de hombres y mujeres que tienen cualquier extensión de puntajes de coeficiente intelectual.

Por ejemplo, midiendo la línea base de la distribución de coeficientes intelectuales masculinos en unidades DE, sabremos que el 68,26% de los puntajes de coeficientes intelectuales masculinos cae entre -1σ y $+1\sigma$ de la media. De igual manera, como la desviación estándar siempre está dada en unidades de puntaje crudas* y $\sigma = 10$, sabremos también que éstos son puntos de la distribución en los que se localizan los coeficientes 110 y 90 ($\bar{X} - \sigma = X$: $100 - 10 = 90$ y $100 + 10 = 110$). Así, el 68,25% de los hombres tendrían puntajes de coeficiente intelectual que fluctúan entre 90 y 110.

Alejándonos de la $\bar{X}$, y más allá de estos puntos, encontraríamos, como se ilustra en la Figura 6.8, que el 99,74% de estos casos, o prácticamente todos los hombres, tienen puntajes de coeficiente intelectual entre 70 y 130 (entre -3σ y $+3\sigma$).

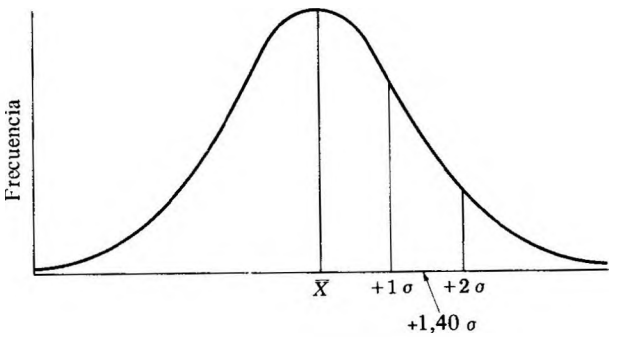
Del mismo modo, mirando ahora la distribución de los puntajes de coeficientes intelectuales femeninos como se grafican en la Figura 6.9, vemos que el 99,74% de estos casos caerían entre los puntajes 85 y 115 (entre -3σ y $+3\sigma$). Entonces, en contraste con los hombres, la distribución de puntajes de coeficientes intelectuales femeninos podría considerarse relativamente *homogénea,* teniendo una proporción menor de puntajes extremos en una y otra dirección. Esta diferencia se refleja en el tamaño comparativo de cada DE, y en los coeficientes intelectuales que oscilan entre -3σ y $+3\sigma$ de la media.

EL USO DE LA TABLA B

Al estudiar la distribución normal sólo hemos analizado aquellas distancias de la media que son múltiplos exactos de la desviación estándar. Es decir, las DE 1, 2 o 3 ya sea por arriba o por abajo de la media. Por lo tanto, surge ahora la pregunta: ¿qué debemos hacer para determinar el porcentaje de casos para las distancias entre dos ordenadas cualesquiera? Supongamos, por ejemplo, que desea-

* N. del E. Recordemos que también se llaman "no procesadas".

FIGURA 6.10 La posición de un puntaje crudo que está a 1,40 DE_s por arriba de $\overline{X}$



mos determinar el porcentaje de la frecuencia total que cae entre la media y un porcentaje crudo que está localizado a 1,40 DE_s por arriba de la media. Como lo ilustra la Figura 6.10, un puntaje crudo a 1,40 DE_s por arriba de la media es obviamente más grande que 1 DE, pero menor que 2 DE_s a partir de la media. Así, sabemos que esta distancia de la media incluiría más del 34,13%, pero menos del 47,72% del área total bajo la curva normal.

Para determinar el porcentaje *exacto* dentro de este intervalo, debemos emplear la tabla B al final del texto que da el porcentaje bajo la curva normal entre la media y varias distancias sigma de ella. Estas distancias sigma (de 0,0 a 5,0) se encuentran en la columna del lado izquierdo de la Tabla B y se les ha asignado un lugar decimal. El segundo lugar decimal se ha dado en la hilera superior o primera de la tabla.

Nótese que la simetría de la curva normal permite dar porcentajes para un sólo lado de la media que constituye sólo la mitad de la curva (50%). Los valores en la Tabla B representan uno y otro lado. A continuación se reproduce una parte de la misma.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	00.00	00.40	00.80	01.20	01.60	01.99	02.39	02.79	03.19	03.59
0.1	03.98	04.38	04.78	05.17	05.57	05.96	06.36	06.75	07.14	07.53
0.2	07.93	08.32	08.71	09.10	09.48	09.87	10.26	10.64	11.03	11.41
0.3	11.79	12.17	12.55	12.93	13.31	13.68	14.06	14.43	14.80	15.17
0.4	15.54	15.91	16.28	16.64	17.00	17.36	17.72	18.08	18.44	18.79

Cuando aprendamos a usar y entender la Tabla B, podremos intentar localizar primero el porcentaje de casos entre una distancia sigma de 1,0 y la media (pues ya sabemos que el 34,13% del área total cae entre estos puntos sobre la línea base). Observando la Tabla B nos damos cuenta, ciertamente, de que ésta nos indica que exactamente el 34,13% del área total oscila entre la media y una distancia sigma de 1,00. Igualmente, vemos que la distancia sigma 2,00 incluye exactamente el 47,72% del área total bajo la curva, mientras que la distancia sigma 2,01 contiene el 47,78% de esta área total.

* N. del E. Las Tablas de esta edición (Apéndice B) han sido fotografiadas fielmente del original en inglés; en el resto del texto se ha eliminado el tradicional punto decimal y puesto, en cambio, la coma decimal.

LOS PUNTAJES ESTANDAR Y LA CURVA NORMAL

De este modo, estamos preparados para encontrar el porcentaje del área total, bajo la curva normal, en relación con cualquier distancia sigma de la media dada. Sin embargo, queda por lo menos una importante pregunta más por contestar: ¿cómo determinamos la distancia sigma de cualquier puntaje crudo? es decir, ¿cómo nos las arreglamos por traducir nuestro puntaje crudo –que recogimos originalmente de nuestros entrevistados– a unidades de desviación estándar? Si deseáramos convertir centímetros a metros, simplemente dividiríamos el número de centímetros entre 100 ya que hay 100 en un metro. Igualmente, si estuviéramos convirtiendo minutos en horas, dividiríamos el número de minutos entre 60, ya que hay 60 minutos en cada hora. Exactamente de la misma manera, podemos convertir cualquier puntaje crudo en unidades DE dividiendo la distancia entre éste y la media entre la DE. Para ilustrar imaginemos un puntaje crudo de 6 en una distribución donde la media es 3 y la DE es 2. Tomando la diferencia entre el puntaje crudo y la media, y obteniendo un puntaje de desviación (6-3), vemos que una puntuación de 6 está a 3 unidades de puntaje crudo por arriba de la media. En otras palabras, la distancia sigma de un puntaje crudo de 6 es 1,5 *en esta distribución en particular.* Debemos hacer notar que siempre hay 100 centímetros en 1 metro y 60 minutos en una hora, sin importar la situación de medición. La desviación estándar no comparte la constancia que marca a estas otras medias estándares, sino que cambia de una distribución a otra. Es por esto que debemos conocer la desviación estándar de una distribución, ya sea que la calculemos, la estimemos o la sepamos de otra persona, antes de poder convertir cualquier puntaje particular a unidades de desviación estándar.

El proceso que acabamos de ilustrar –de encontrar la distancia sigma de X– da un valor que se llama *puntaje z* o *estándar,* que indica *la dirección y el grado en que cualquier puntaje crudo se desvía de la media de una distribución en una escala de unidades* DE (nótese que la columna al lado izquierdo de la Tabla B, al final del libro, lleva el título "z". Así, un puntaje z de +1,4 indica que el puntaje crudo se encuentra a 1,4 DE (casi $1\frac{1}{2}$ DE) *por arriba* de la media, mientras que un puntaje z de –2,1 significa que el puntaje cae un poco más de 2 DE_s *por abajo* de la media (ver Figura 6.11).

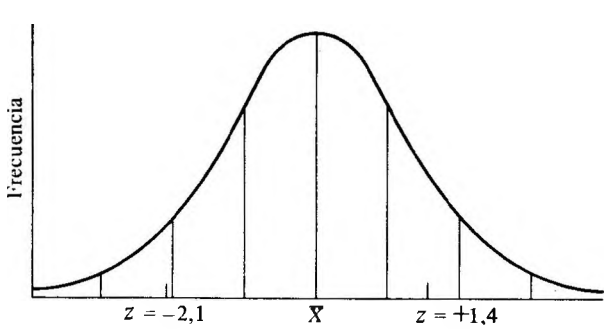


FIGURA 6.11 La posición de $z = -2,1$ y $z = +1,4$ en una distribución normal

Obtenemos un puntaje z encontrando el puntaje de desviación ($x = X - \overline{X}$) (que da la distancia entre el puntaje no crudo y la media) y luego dividiéndola entre σ.

Calculado por fórmula,

$$z = \frac{X - \overline{X}}{\sigma} \quad \text{o} \quad \frac{x}{\sigma}$$

donde

x = el puntaje de desviación
σ = la desviación estándar de una distribución
z = un puntaje estándar

Ejemplo 1

Estamos estudiando la distribución del ingreso anual en una ciudad en la cual el ingreso medio anual es de $ 5 000 y la desviación estándar es $ 1 500. Suponiendo que la distribución del ingreso anual está normalmente distribuida, podemos convertir el puntaje crudo de esta distribución, $ 7 000, en un puntaje estándar, de la siguiente manera:

$$z = \frac{7000 - 5000}{1500} = +1{,}33$$

Así, un ingreso anual de $ 7 000 está a 1,33 desviaciones estándar por arriba del ingreso medio anual de $ 5 000 (ver Figura 6.12).

Ejemplo 2

Estamos trabajando con una distribución de puntajes normal que representa la conformidad de un grupo de presuntos inquilinos con la vivienda pública (los puntajes más altos indican mayor satisfacción con la vivienda pública). Digamos que esta distribución tiene un media de 10 y una desviación estándar de 2. Para determinar a cuántas desviaciones estándar está un puntaje de 3 de la media de 10. obtenemos la diferencia entre este puntaje y la media, esto es,

$$\begin{aligned} x &= X - \overline{X} \\ &= 3 - 10 \\ &= -7 \end{aligned}$$

Dividimos entonces entre la desviación estándar

$$\begin{aligned} z &= \frac{x}{\sigma} \\ &= -\tfrac{7}{2} \\ &= -3{,}5 \end{aligned}$$

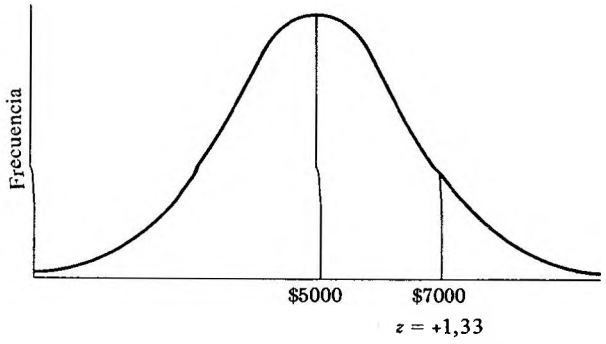


FIGURA 6.12 La posición de $z = 1{,}33$ para un puntaje crudo de $ 7 000

Entonces, como se ve en la Figura 6.13, un puntaje crudo de 3 cae a **3,5** desviaciones estándar por abajo de la media en esta distribución de frecuencias.

Nota: si conocemos un puntaje z y buscamos obtener su equivalente en puntajes crudos, usamos la fórmula

$$X = z\sigma + \overline{X}$$

Para el presente ejemplo,

$$\begin{aligned} X &= (-3{,}5)(2) + 10 \\ &= -7 + 10 \\ &= 3 \end{aligned}$$

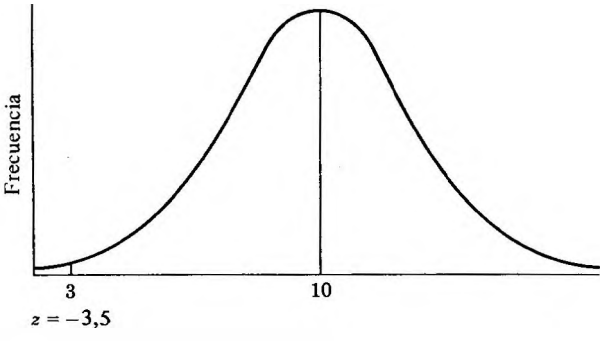


FIGURA 6.13 La posición de $z = -3{,}5$ para el puntaje crudo 3

PROBABILIDAD Y LA CURVA NORMAL

Como veremos ahora, la curva normal puede usarse conjuntamente con los puntajes z y la Tabla B para determinar la probabilidad de obtener cualquier puntaje crudo en una distribución. En el presente contexto, el término *probabilidad* se refiere a la frecuencia relativa de ocurrencia de cualquier resultado o evento dado; esto es, *la probabilidad asociada con cualquier evento es el número de veces en que dicho evento puede ocurrir, en relación con el número total de eventos.* En forma de proposición,

$$\text{La probabilidad de cualquier resultado o evento} = \frac{\text{número de veces en que el resultado o evento puede ocurrir}}{\text{número total de resultados o eventos}}$$

Así, la probabilidad de sacar una sola carta (digamos el as de espadas) de una baraja de 52 cartas es 1 en 52, ya que el resultado del "as de espadas" sólo puede ocurrir una vez entre el número total de tales resultados, 52 cartas. La probabilidad de caer en "cara" una moneda "imparcial o perfectamente equilibrada" que se lanza al aire sólo una vez, es 1 en 2, ya que "cara" ocurre una vez entre el número total de posibles resultados, que es 2. Igualmente, si se nos dijera que abriéramos un libro de 100 páginas en cualquier página dada (digamos, en la página 23) la probabilidad de abrir el libro "al azar" en la página deseada en un solo intento es 1 en 100.

En el presente contexto, la curva normal es una distribución en la cual es posible determinar probabilidades asociadas con varios puntos a lo largo de su línea base. Como se hizo notar anteriormente, la curva normal es una *distribución de frecuencia* en la cual la frecuencia total bajo la curva es igual a 100% ; contiene un área central que rodea la media, donde los puntajes ocurren con mayor frecuencia, y áreas más pequeñas hacia uno y otro lado, donde hay un aplanamiento gradual y por tanto una menor proporción de puntajes extremadamente altos y bajos. Entonces, en términos de probabilidad, podemos decir que la probabilidad disminuye a medida que viajamos a lo largo de la línea base alejándonos de la media en una y otra dirección. Por tanto, decir que el 68,26% de la frecuencia total bajo la curva normal cae entre -1σ y $+1\sigma$ de la media, es decir, que la probabilidad de que cualquier puntaje crudo caiga dentro de este intervalo, es de 68 en 100 aproximadamente. De igual manera, decir que el 95,44% de la frecuencia total bajo la curva normal cae entre -2σ y $+2\sigma$ de la media es decir, también, que la probabilidad de que cualquier puntaje crudo caiga dentro de este intervalo es de 95 en 100 aproximadamente, y así sucesivamente.

Este es precisamente el mismo concepto de probabilidad o *frecuencia relativa* que vimos operar al sacar una sola carta de una baraja completa, al lanzar una moneda al aire o al abrir un libro en una página determinada. Nótese, sin embargo, que las probabilidades asociadas con áreas bajo la curva normal se dan siempre en relación con el 100% que constituye toda el área bajo la curva (por ejemplo, 68 en 100, 95 en 100, 99 en 100 y así sucesivamente). Por este motivo, y para dar una forma estándar de ver la probabilidad a través de este libro, estaremos tratando *la probabilidad como el número de veces entre 100 en que puede ocurrir cualquier evento dado.* Así, la probabilidad de sacar el as de espadas de un conjunto de naipes barajado es 1,92 en 100 ($\frac{1}{52}$) y de caer "cara" al lanzar la moneda al aire es 50 en 100 ($\frac{1}{2}$). Es más, nótese que la probabilidad se expresa usualmente en decimales como una proporción (P). Por ejemplo, podemos decir que $P = 0{,}50$ ($\frac{50}{100}$) de caer "cara" al lanzar sólo una vez la moneda. Igualmente, podemos decir que $P = 0{,}68$ ($\frac{68}{100}$) y que cualquier puntaje crudo caerá entre -1σ y $+1\sigma$ bajo la curva normal.

Expresada como proporción, *la probabilidad siempre oscila entre 0 y 1.* La probabilidad de un evento es 0 cuando estamos absolutamente seguros de que no ocurrirá; la probabilidad de un evento es 1 cuando estamos absolutamente seguros de

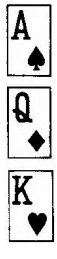
que ocurrirá. ¡Los investigadores sociales nunca, no están, absolutamente seguros de nada! Como resultado, podríamos esperar frecuentemente encontrar probabilidades iguales a 0,60, 0,25 o 0,05, pero casi nunca esperaríamos reducir la probabilidad a 0 o aumentarla a 1.

Otra característica importante de la probabilidad es la *regla de la suma,* que afirma que *la probabilidad de obtener un resultado cualquiera entre varios diferentes es igual a la suma de sus distintas probabilidades.* Supongamos, por ejemplo, que deseamos encontrar la probabilidad de sacar *ya sea* el as de espadas, la reina de diamantes, *o* el rey de corazones de un conjunto de naipes bien barajado de 52 cartas en el primer intento. Sumando sus probabilidades separadas ($\frac{1}{52} + \frac{1}{52} + \frac{1}{52}$), vemos que la probabilidad de obtener cualquiera de estas cartas, en un solo intento, es igual a $\frac{3}{52}$ ($P = 0{,}06$). En otras palabras, tenemos 6 oportunidades entre 100 de obtener ya sea el as de espadas, la reina de diamantes o el rey de corazones a la primera tentativa (ver Figura 6.14).

La regla de la suma siempre supone que los resultados *se excluyen mutuamente,* esto es, no pueden ocurrir simultáneamente dos resultados. Por ejemplo, ninguna carta de una baraja de 52 cartas puede ser espada, diamante *y* corazón al mismo tiempo. Igualmente, una moneda que se lanza sólo una vez no puede, de ninguna manera, caer sobre su "cara" y su "cruz" al mismo tiempo.

Suponiendo que los resultados se excluyesen mutuamente, podemos decir que la probabilidad asociada con todos los posibles resultados de un evento siempre es igual a 1. Esto indica que debe ocurrir algún resultado. Si no es "cara", entonces será "cruz"; si no es un as, entonces será un rey, reina, sota, diez, etc. Al lanzar una moneda la probabilidad de caer "cruz" es igual a $\frac{1}{2}$ ($P = 0{,}50$). Por supuesto, la probabilidad de caer "cruz" también es $\frac{1}{2}$ ($P = 0{,}50$). Sumando las probabilidades de todos los resultados posibles, vemos que la probabilidad de caer "cara" o "cruz" es igual a 1 ($\frac{1}{2} + \frac{1}{2} = 1$).

Otra propiedad importante de la probabilidad ocurre en la *regla de la multiplicación* que se centra en el problema de obtener dos o más resultados en orden sucesivo, uno después del otro. La regla de la multiplicación afirma que *la probabili-*



Probabilidad de sacar el as de espadas $\frac{1}{52}$

Probabilidad de sacar la reina de diamantes $\frac{1}{52}$

Probabilidad de sacar el rey de corazones + $\frac{1}{52}$

Probabilidad de sacar ya sea el as de espadas, la reina de diamantes o el rey de corazones $\frac{3}{52}$ ($P = 0{,}06$)

FIGURA 6.14 La probabilidad de obtener ya sea el as de espadas, la reina de diamantes o el rey de corazones en un solo intento de una baraja de 52 cartas: una ilustración de la regla de la suma

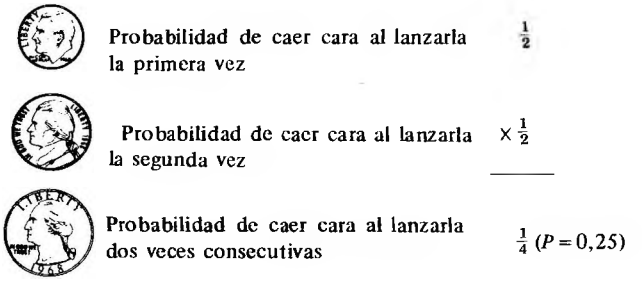


FIGURA 6.15 La probabilidad de sacar "caras" en dos lanzamientos sucesivos de una moneda: una ilustración de la regla de la multiplicación

dad de obtener una combinación de resultados que se excluyan mutuamente, es igual al producto de sus probabilidades por separado. En lugar de "ya sea. . . o. . .", la regla de la multiplicación establece el "primero, segundo, tercero".

Por ejemplo, ¿cuál es la probabilidad de sacar "caras" al lanzar dos veces consecutivas una moneda? Como estos resultados son independientes uno del otro, el resultado, al lanzar la moneda por primera vez, no influye en el resultado que se obtiene la segunda vez. En el primer lanzamiento de la moneda, la probabilidad de obtener "caras" es igual a $\frac{1}{2}$ ($P = 0{,}50$); en el segundo, la probabilidad de obtener "caras" también es igual a $\frac{1}{2}$ ($P = 0{,}50$). Por lo tanto, la probabilidad de caer "caras" al lanzar dos veces consectivas la moneda es igual a $(\frac{1}{2})\ (\frac{1}{2}) = \frac{1}{4}$ (o $P = 0{,}25$). Ver Figura 6.15).

Para aplicar la anterior concepción de probabilidad, en relación con la distribución normal, volvamos a un ejemplo anterior. Se nos pidió que convirtiéramos un puntaje crudo de una distribución del ingreso anual de una ciudad, que supusimos se aproximaba a la curva normal en su puntaje z equivalente. Esta distribución de ingreso tenía una media de \$ 5 000 con un DE de \$ 1 500.

Aplicando la fórmula del puntaje z, vimos anteriormente que un ingreso anual de \$ 7 000 estaba a 1.33 DE por arriba de la media de \$ 5 000, esto es,

$$z = \frac{7000 - 5000}{1500} = +1{,}33$$

Determinemos ahora la probabilidad de obtener un puntaje que esté entre \$ 5 000 la media y \$ 7 000. En otras palabras, ¿cuál es la probabilidad de elegir al azar, en una sola tentativa, a una persona de esta ciudad cuyo ingreso anual fluctúe entre \$ 5 000 y \$ 7 000? El problema se ilustra gráficamente en la Figura 6.16 (nos estamos refiriendo al área sombreada bajo la curva) y puede resolverse en dos pasos, utilizando la fórmula del puntaje z y la Tabla B al final del libro.

PASO 1: Convertir el puntaje crudo (\$ 7 000) en un puntaje z

$$z = \frac{X - \overline{X}}{\sigma}$$

$$= \frac{7\,000 - 5\,000}{1\,500} = +1{,}33$$

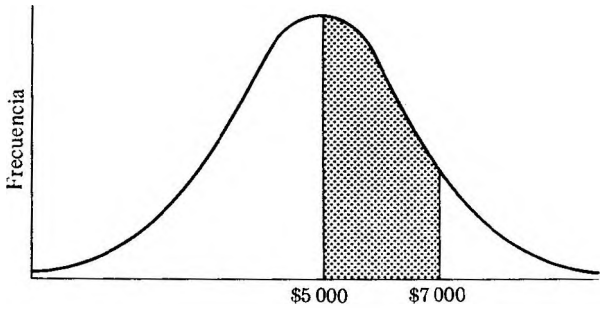


FIGURA 6.16 La porción del área total bajo la curva normal para la cual buscamos la probabilidad de ocurrencia

Así, un puntaje crudo $ 7 000 se encuentra a 1,33 DE_s sobre la media.

PASO 2: Usando la Tabla B, buscar el porcentaje de la frecuencia total bajo la curva que cae entre el puntaje z ($z = +1{,}33$) y la media.

En la Tabla B, vemos que el 40,82% (41%) de la población total de esta ciudad gana entre $ 5 000 y $ 7 000 (ver la Figura 6.17). Así, recorriendo 2 decimales hacia la izquierda, vemos que la probabilidad (redondeando) es de 41 de 100: $P = 0{,}41$ de que obtuviéramos un individuo cuyo ingreso anual esté entre esta cifras.

En el ejemplo anterior se nos pidió determinar la probabilidad asociada con la distancia entre la media y una cierta distancia sigma de ella. Sin embargo, puede que muchas veces deseemos encontrar el porcentaje del área que está en un determinado puntaje crudo o *más allá* de él hacia una u otra cola de la distribución, o bien encontrar la probabilidad para obtener estos puntajes. Por ejemplo, en el presente caso, podríamos desear conocer la probabilidad para obtener un ingreso anual de $ 7 000 o *más*.

Este problema puede ilustrarse gráficamente, como se muestra en la Figura 6.18 (nos estamos refiriendo al área sombreada bajo la curva). En este caso, seguiríamos los pasos 1 y 2 descritos anteriormente, obteniendo así el puntaje z y encontrando el porcentaje bajo la curva normal entre $5 000 y una $z = 1{,}33$ (de la Tabla B). Sin embargo, en el presente caso debemos dar un paso más adelante y *restar* el

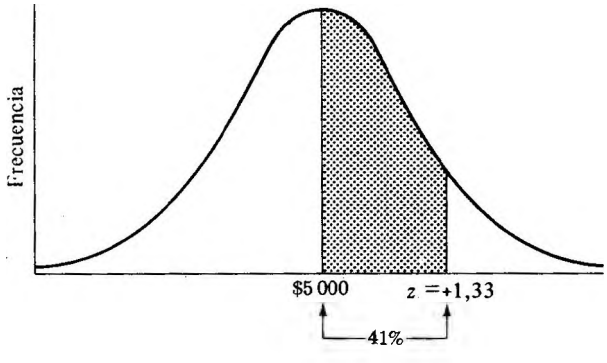
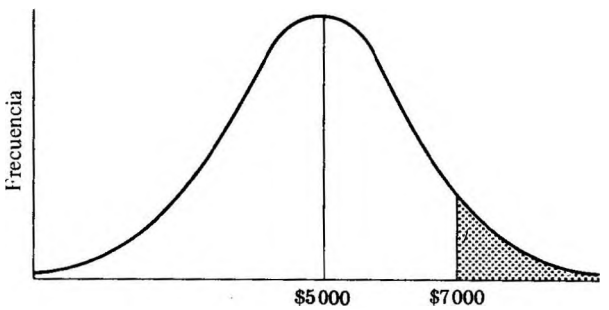


FIGURA 6.17 El porcentaje del área total bajo la curva normal entre X = $ 5 000 y z = 1,33

FIGURA 6.18 La porción del área total bajo la curva normal para la cual buscamos determinar la probabilidad de que ocurra.



porcentaje obtenido en la Tabla B de 50% —el porcentaje del área total localizado a uno y otro lado de $\bar{X}$. Esto resulta cierto ya que *la tabla B siempre se refiere al porcentaje del área entre un puntaje z y la media, nunca al porcentaje de área en un puntaje z o más allá de éste.*

Por lo tanto, restando 40,82% de 50% vemos que ligeramente más del 9% (9,18%) caen *en* $ 7 000 o *más allá.* En términos de probabilidad, podemos decir (recorriendo 2 decimales hacia la izquierda) que hay sólo un poco más de 9 oportunidades, entre 100 (P = 0,09), de que encontremos un individuo en esta ciudad cuyo ingreso sea de $ 7 000 o más.

Ya se anotó que cualquier distancia sigma dada por arriba de la media contiene una proporción idéntica de casos que la misma distancia sigma por abajo de la media. Por este motivo, nuestro procedimiento para encontrar probabilidades asociadas con puntos abajo de $\bar{X}$ es idéntico al que se siguió en los ejemplos anteriores.

Por ejemplo, el porcentaje de frecuencia total entre el puntaje z −1,33 ($ 3 000) y la media es idéntico al porcentaje entre el puntaje z +1,33 ($ 7 000) y la media. Por lo tanto, sabemos que un individuo cuyo ingreso fluctúe entre $ 3 000 y $ 5 000 obtiene P = 0,41. Igualmente, el porcentaje de frecuencia total en −1,33 ($ 5 000 menos) o mayor es igual que en +1,33 ($ 7 000 o más) o más allá. Así, sabemos que hay una P = 0,09 de que encontremos que alguien de la ciudad tiene un ingreso anual de $ 3 000 o menor.

Podemos usar la regla de la suma para encontrar la probabilidad de obtener más de una sola porción del área bajo la curva normal. Por ejemplo, ya hemos determinade que P = 0,09 es para ingresos de $ 3 000 o menos, y para ingresos de $ 7 000 o más. Para encontrar la probabilidad de obtener *ya sea* $ 3 000 o menos, o $ 7 000 o más; simplemente sumamos sus probabilidades por separado como sigue:

$$P = 0{,}09 + 0{,}09$$
$$= 0{,}18$$

De manera semejante, podemos buscar la probabilidad de hablar a alguien cuyo ingreso oscile entre $ 3 000 y $7 000, sumando las probabilidades asociadas con los puntajes z de 1,33 a uno y otro lado de la media. Por lo tanto,

$$P = 0{,}41 + 0{,}41$$
$$= 0{,}82$$

Nótese que 0.82 + 0.18 es igual a 1, lo que representa todos los posibles eventos bajo la curva normal.

La aplicación de la regla de la multiplicación a la curva normal puede ilustrarse buscando la probabilidad de obtener cuatro individuos cuyos ingresos sean de $ 7 000 o más. Sabemos ya que P = 0,09 asociada con la búsqueda de un individuo cuyo ingreso sea de por lo menos $ 7 000. Por lo tanto,

$$P = (0{,}09)\,(0{,}09)\,(0{,}09)\,(0{,}09)$$
$$= (0{,}09)^4$$
$$= 0{,}00007$$

Aplicando la regla de la multiplicación vemos que la probabilidad de obtener cuatro individuos con ingresos de $ 7 000 o más, es de 7 oportunidades entre 100 000.

RESUMEN

Este capítulo trató de relacionar las propiedades de la distribución normal teórica con los problemas del "mundo real" en la investigación social. Así, se demostró que el área bajo la curva normal puede ser empleada para interpretar la desviación estándar y hacer afirmaciones de probabilidad. La importancia de la distribución normal se hará más evidente en los subsiguientes capítulos del texto.

PROBLEMAS

1. En cualquier distribución normal de puntajes, ¿qué porcentaje del área total cae (a) entre -1 DE y $+1$ DE, (b) entre -2 DE_s y $+2$ DE_s (c) entre -3 DE_s y $+3$ DE_s?
2. Dada una distribución normal de puntajes crudos en la cual X = 7,5 y DE = 1,3, expresar cada uno de los siguientes puntajes crudos como puntaje z: **(a)(b)(c)(d)(e)(f)(g)**
3. Dada una distribución normal de ingreso diario en la cual X = $ 10,50 y DE = $ 1,80, expresar cada uno de los siguientes ingresos como puntaje z: (a)(b)(c)(d) (f)(g)
4. Para el Problema 3, de la distribución de ingreso, determinar (a) el porcentaje de entrevistados que tienen un ingreso diario de $ 15,00 o más, (b) la probabilidad de localizar un entrevistado cuyo ingreso diario sea de $ 15.00 o más; (c) el porcentaje de entrevistados que ganan entre $ 10,00 y $ 10,50; (d) la probabilidad de localizar un entrevistado cuyo ingreso fluctúe entre $ 10.00 y $ 10,50; (e) la probabilidad de localizar un entrevistado cuyo ingreso sea de $ 10,00 o menos; (f) la probabilidad de localizar un entrevistado cuyo ingreso sea *ya* de $ 10,00 o menos *o* de $ 11,00 o más; (g) la probabilidad de localizar dos entrevistados cuyo ingreso sea $ 10,00 o menos.

5. Dada una distribución normal de puntajes crudos en la cual $\overline{X}$ = 80 y DE= 7,5, determinar (a) el porcentaje de entrevistados que obtuvieron puntajes de 60 o menos; (b) la probabilidad de localizar a un entrevistado que haya obtenido un puntaje de 60 o menos; (c) el porcentaje de entrevistados que obtuvieron puntajes entre 80 y 90; (d) la probabilidad de localizar un entrevistado que haya obtenido puntajes entre 80 y 90; (e) el porcentaje de entrevistados que lograron puntajes de 85 o más; (f) la probabilidad de localizar a un entrevistado que haya obtenido un puntaje de 85 o más; (g) la probabilidad de localizar a un entrevistado que haya obtenido puntajes *sea ya* de 70 o menos *o* de 90 o más; (h) la probabilidad de obtener tres entrevistados que hayan logrado puntajes de 90 o más.

7

Muestras y poblaciones

El investigador social generalmente busca sacar conclusiones acerca de grandes números de individuos. Por ejemplo, podría desear estudiar a los 350 000 000 de ciudadanos de Latinoamérica, a los 1 000 miembros de un determinado sindicato de trabajadores, a los 10 000 indígenas que viven en los pueblos del sur de México o a los 45 000 estudiantes inscritos en determinada universidad.

Hasta este punto, hemos estado suponiendo que el investigador social investiga la totalidad del grupo que intenta comprender. Este grupo, conocido como *población o universo,* consiste en un conjunto de individuos que comparten por lo menos una característica, sea una ciudadanía común, la calidad de ser miembros de una asociación voluntaria o de una raza, la matrícula en una misma universidad, o similares. Así, podríamos hablar de la población de Colombia o de México, del número de miembros de un sindicato de trabajadores, de la población de indígenas residentes en un pueblo sureño o de la cantidad de estudiantes universitarios.

Como el investigador social trabaja con limitaciones de tiempo, energía y recursos económicos, rara vez estudia a todos y cada uno de los miembros de la población en que está interesado. En cambio, el investigador analiza sólo una *muestra:* un número pequeño de individuos tomado de alguna población. A través del proceso de muestreo, el investigador social busca generalizar de su muestra (grupo pequeño) a la totalidad de la población de donde la obtuvo (grupo mayor).

El proceso de muestreo es una parte integral de la vida diaria. ¿De qué otra forma obtendríamos información acerca de los demás si no haciendo muestreos a nuestro alrededor? Por ejemplo, podríamos discutir informalmente sobre temas políticos con otros estudiantes para averiguar cuáles son, en general, sus opiniones políticas; podríamos intentar determinar de qué manera nuestros compañeros de curso estudian para cierto examen poniéndonos en contacto, anticipadamente, con sólo algunos miembros de la clase; incluso podríamos invertir en el mercado de valores

después de descubrir que una pequeña muestra de nuestros compañeros ha ganado dinero de una manera similar.

METODOS DE MUESTREO

Los métodos de muestreo del investigador social son generalmente más cuidadosos y sistemáticos que los de la vida diaria. Su preocupación central es asegurarse de que los miembros de su muestra sean lo suficientemente representativos de la población entera como para permitir hacer generalizaciones precisas acerca de ella. Para hacer tales inferencias, el investigador escoge un método de muestreo apropiado para ver si todos y cada uno de los miembros de la muestra tienen igual oportunidad de ser integrados en ella. Si a cada miembro de la población se le da igual oportunidad de ser escogido para la muestra, se está utilizando un método *aleatorio;* de no ser así, el método empleado viene a ser *no aleatorio.*

Muestras no aleatorias

El método de muestreo no aleatorio más usual es el muestreo por accidente y es el que menos difiere con nuestros procedimientos diarios de muestreo, ya que se basa exclusivamente en lo que es conveniente para el investigador. Es decir, el investigador simplemente incluye los casos más convenientes en su muestra y excluye de ella los casos inconvenientes. La mayoría de los estudiantes podrá recordar al menos algunas ocasiones en que el maestro que está realizando una investigación les ha pedido a todos los alumnos de su clase que participen en un experimento o llenen un cuestionario. La popularidad de esta forma de muestreo por accidente en psicología ha ocasionado que algunos detractores vean a la psicología como "la ciencia del estudiante universitario" de 2o semestre debido a que muchos de ellos son sujetos de investigación.

Otro tipo no aleatorio es el muestreo *por cuota.* En este procedimiento de muestreo, las diversas características de una población, tales como edad, sexo, clase social o raza, son muestreadas de acuerdo con el porcentaje que ocupan dentro de la población. Supongamos, por ejemplo, que se nos pidiera sacar una muestra por cuota de los estudiantes que asisten a una universidad donde el 42% son mujeres y el 58% son hombres. Usando este método, se da a los entrevistadores una cuota de estudiantes para localizar, de manera que sólo el 42% de la muestra consista de mujeres y el 58% de hombres. Se incluyen en la muestra los mismos porcentajes que están representados en la población. Si el tamaño total de la muestra es 200, entonces se seleccionan 84 estudiantes del sexo femenino y 116 del sexo masculino.

Una tercera variedad de muestra no aleatoria se conoce como muestreo *intencional* o de *juicio.* La idea básica que involucra este tipo de muestra es que la lógica, el sentido común o el sano juicio, pueden usarse para seleccionar una muestra que sea representativa de una población. Por ejemplo, para sacar una muestra de juicio de revistas

que reflejen los valores de la clase media, podríamos, a un nivel intuitivo, escoger Visión, Vanidades, ya que los artículos que aparecen en estas revistas *parecen* reflejar lo que la mayoría de los latinoamericanos de la clase media desean (por ejemplo, el nivel de vida del norteamericano, el éxito económico y similares). De manera semejante, los distritos estatales que tradicionalmente han votado por los candidatos ganadores para cargos públicos podrían ser encuestados en un intento por predecir el resultado de determinadas elecciones.

Muestras aleatorias

Como se anotó anteriormente, el muestreo aleatorio le da a todos y cada uno de los miembros de la población igual oportunidad de ser seleccionados para la muestra. Esta característica del muestreo aleatorio indica que cada miembro de la poblacion debe ser identificado antes de obtener dicha muestra aleatoria, requisito que generalmente se llena obteniendo una lista que incluya a todos y cada uno de los miembros de la población. Si pensamos un poco veremos que la obtención de una lista completa de los miembros de la población no es siempre una tarea fácil, especialmente si se está estudiando una población grande y diversa. Para tomar un ejemplo relativamente fácil, ¿dónde podríamos conseguir una lista *completa* de los estudiantes inscritos en una universidad importante? Aquellos investigadores sociales que lo han intentado darán fe de su dificultad. Para una tarea más laboriosa, tratemos de encontrar una lista de todos los residentes de una gran ciudad. ¿Cómo podemos asegurarnos de identificarlos a todos, incluso a aquellos residentes que no desean ser identificados?

El tipo básico de muestra aleatoria, el *muestreo aleatorio simple,* puede obtenerse mediante un proceso no muy distinto de la técnica, actualmente conocida, de poner todos los nombres en diferentes pedazos de papel y luego sacar sólo algunos nombres de un sombrero con los ojos vendados. Este procedimiento le da, idealmente, igual oportunidad a todos los miembros de la población de ser seleccionados para la muestra ya que se incluye sólo un pedazo de papel por persona. Por varios motivos (incluyendo el hecho de que el investigador necesitaría un sombrero extremadamente grande) el investigador social que intenta tomar una muestra aleatoria generalmente no saca nombres de sombreros. En cambio, usa una *tabla de números aleatorios* tal como la tabla H localizada al final del texto. Hemos reproducido a continuación una porción de una tabla de números aleatorios.

	Número de columna																			
Número de fila	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	2	3	1	5	7	5	4	8	5	9	0	1	8	3	7	2	5	9	9	3
2	6	2	4	9	7	0	8	8	6	9	5	2	3	0	3	6	7	4	4	0
3	*0*	*4*	*5*	5	5	0	*4*	*3*	*1*	0	5	3	7	4	3	5	0	8	9	0
4	1	1	8	3	7	4	4	1	0	9	6	2	2	1	3	4	3	1	4	8
5	1	6	0	3	5	0	3	2	4	0	4	3	6	2	2	2	3	5	0	0

Una tabla de números aleatorios se construye en forma tal que genere series de números sin ningún patrón u orden determinado. Como resultado, el proceso de usar una tabla de números aleatorios produce una muestra imparcial semejante a aquélla que se logra poniendo pedazos de papel en un sombrero y sacando nombres con los ojos vendados.

Para obtener una muestra aleatoria simple por medio de una tabla de números aleatorios, el investigador social obtiene primero su lista de la población y le asigna un número de identificación único a todos y cada uno de sus miembros. Por ejemplo, si está realizando una investigación acerca de los 500 estudiantes inscritos en la materia de "Introducción a la Sociología" podría obtener una lista de ellos con el profesor y asignarle a cada alumno un número de 001 a 500. Habiendo preparado la lista, procede a sacar los miembros de su muestra de una tabla de números aleatorios. Digamos que el investigador busca sacar una muestra de 50 estudiantes para representar a los 500 miembros de la población del curso. Podría entrar a la tabla de números aleatorios en cualquier número (con los ojos cerrados, por ejemplo) y moverse en cualquier dirección tomando números apropiados hasta que haya seleccionado los 50 miembros para la muestra. Mirando una porción de la anterior tabla de números aleatorios, podríamos comenzar arbitrariamente en la intersección de la columna 1 y la fila 3 moviéndonos de izquierda a derecha para tomar cada número que aparezca entre 001 y 500. Los primeros números que aparecen en la columna 1 y la fila 3 son 0, 4 y 5. Por lo tanto, el alumno número 045 es el primer miembro de la población que se elegirá para la muestra. Continuando de izquierda a derecha vemos que 4, 3 y 1 aparecen enseguida, de manera que se selecciona el alumno número 431. Se continúa con este proceso hasta que se hayan tomado todos los 50 miembros para la muestra. Una nota para el estudiante: al usar la tabla de números aleatorios, pase siempre por alto los números que aparezcan por segunda vez o que estén más arriba de lo necesario.

Todos los métodos de muestreo aleatorio son en realidad variaciones del procedimiento de muestreo simple que se acaba de ilustrar. Por ejemplo, con el muestreo *sistemático* no se requiere tabla de números aleatorios, ya que se hace el muestreo con una lista de miembros de la población por intervalos fijos. Entonces, empleando el muestreo sistemático se incluye cada *enésimo* miembro de una población, en una muestra de ella. Para ilustrar, al sacar una muestra de la población de 10 000 amas de casa de cierta colonia podríamos organizar una lista de amas de casa, tomar cada *décimo* nombre de la lista y presentar una lista de 1 000 amas de casa.

La ventaja del muestreo sistemático es que no se requiere una tabla de números aleatorios. Como resultado, este método es siempre menos demorado que el procedimiento aleatorio simple, especialmente para sacar muestras de grandes poblaciones. Por el contrario, al tomar una muestra sistemática se presume que la posición en una lista de miembros de una población no influye en la aleatoriedad. Si esta presunción no se toma seriamente, el resultado puede ser que se seleccionen más de una vez

ciertos miembros de la población, mientras que otros definitivamente no se seleccionan. Esto puede suceder, por ejemplo, cuando se muestrean sistemáticamente casas de una lista en la que las casas de esquina (que son generalmente más caras que las demás casas de la cuadra) ocupan una posición fija o cuando se sacan muestras de los nombres de un directorio telefónico por intervalos fijos, de manera que los nombres asociados a ciertos lazos étnicos no se seleccionan.

Otra variación del muestreo aleatorio simple es el muestreo *estratificado;* involucra la división de la población en subgrupos o *estratos* más homogéneos de los que se toman entonces muestras aleatorias simples. Supongamos, por ejemplo, que deseamos estudiar la aceptación de varios métodos de control de la natalidad entre la población de cierta ciudad. Como las actitudes hacia el control de la natalidad varían según la religión y el estatus socioeconómico, podríamos estratificar nuestra población sobre estas variables, formando así subgrupos más homogéneos con respecto a la aceptación del control de la natalidad. Más específicamente, digamos que podríamos identificar a los miembros de la población, católicos, protestantes y judíos, así como a los de clase alta, media y baja. Nuestro procedimiento de estratificación podría dar los siguientes subgrupos o estratos:

Protestantes de clase alta
Protestantes de clase media
Protestantes de clase baja
Católicos de clase alta
Católicos de clase media
Católicos de clase baja
Judíos de clase alta
Judíos de clase media
Judíos de clase baja

Habiendo identificado nuestros estratos, procedemos a tomar una muestra aleatoria simple de cada subgrupo o estrato (por ejemplo, de protestantes de clase baja, de católicos de clase media, etc.) hasta que hayamos muestreado la población entera. O sea que, para los efectos del muestreo, cada estrato se trata como una población completa y se aplica el muestreo aleatorio simple. Específicamente se le da a cada miembro de un estrato un número de identificación, se pone en lista y se saca una muestra por medio de una tabla de números aleatorios. Como paso final del procedimiento, los miembros seleccionados de cada subgrupo o estrato se combinan para lograr tener una muestra de toda la población.

La estratificación se basa en la idea de que un grupo homogéneo requiere una muestra más pequeña que un grupo heterogéneo. Por ejemplo, el estudio de los individuos que caminan por la esquina de una calle céntrica requiere, probablemente, una muestra más grande que el estudio de los individuos de clase media que viven en un suburbio. Se pueden encontrar generalmente caminando por el centro individuos

que tienen cualquier combinación de características. Por contraste, las personas de la clase media que viven en un suburbio son generalmente más parecidos entre sí en lo que se refiere a educación, ingresos, orientación política, tamaño de la familia, actitud hacia el trabajo, para mencionar sólo algunas características.

A primera instancia, las muestras aleatorias estratificadas tienen una asombrosa semejanza con el método no aleatorio por cuotas tal como se explicó anteriormente, ya que ambos procedimientos requieren usualmente que se incluyan las características de la muestra en las proporciones exactas en que contribuyen a la población. Por lo tanto, si el 32% de nuestra muestra se compone de protestantes de la clase media, entonces exactamente el 32% de nuestra muestra debe sacarse de protestantes de clase media; del mismo modo, si el 11% de nuestra población consiste de judíos de clase baja, entonces el 11% de nuestra muestra debe constituirse de manera semejante y así sucesivamente. Surge una excepción en el contexto del muestreo estratificado cuando un estrato en particular está desproporcionadamente bien representado en la muestra, posibilitando un subanálisis más intensivo de ese grupo.

Tal evento puede surgir, por ejemplo, cuando los indígenas, quienes constituyen una pequeña proporción de una población dada, son "sobre-muestreados" en un esfuerzo por examinar más de cerca sus características.

A pesar de sus semejanzas superficiales, las muestras por cuotas y estratificadas son esencialmente diferentes. Mientras los miembros de las muestras por cuotas se toman por cualquier método que escoje el investigador, los miembros de las muestras estratificadas se seleccionan siempre sobre una base aleatoria, generalmente por medio de una tabla de números aleatorios aplicada a una lista completa de miembros de la población.

Antes de dejar el tema de los métodos de muestreo, examinemos la naturaleza de una forma de muestreo aleatorio especialmente popular que se conoce como el método de *cúmulos*. Tales muestras se usan ampliamente para reducir los costos de las grandes encuestas en que los entrevistadores deben ser enviados a localidades dispersas, ya que se requieren muchos viajes. Empleando el método de cúmulos se desarrollan por lo menos dos niveles de muestreo:

1. La *unidad primaria de muestreo* o cúmulo, que es aquella área bien delineada en la que se considera que están incluidas características que se encuentran en toda la población (por ejemplo, un estado, una región de empadronamiento, una cuadra de una ciudad, etc.), y
2. Los miembros de la muestra dentro de cada cúmulo.

Imaginemos, con fines ilustrativos, que quisiéramos entrevistar a una muestra representativa de individuos que viven en una gran área de nuestra ciudad. Extraer una muestra aleatoria simple, sistemática o estratificada de entrevistados diseminados sobre una amplia área implicaría una buena cantidad de viajes, sin mencionar tiempo y dinero. Sin embargo, por medio del muestreo por cúmulos limitaríamos nuestras

entrevistas a aquellos individuos situados dentro de relativamente pocos cúmulos. Por ejemplo, podríamos empezar tratando al primer cuadro de la ciudad como nuestra unidad primaria de muestreo o cúmulo. Podríamos proceder entonces a obtener una lista de todas las cuadras dentro del área, por lo cual tomamos una muestra aleatoria simple de cuadras. Habiendo tomado nuestra muestra de cuadras, podríamos seleccionar a los entrevistados individuales (o familias) en cada cuadra por el mismo método aleatorio simple. Más específicamente, todos los individuos (o familias) en cada una de las cuadras seleccionadas se ponen en una lista y se escoge una muestra de entrevistados de cada cuadro con ayuda de una tabla de números aleatorios. Utilizando el método de cúmulos, cualquier entrevistador dado localiza una de las cuadras seleccionadas y hace contacto con más de un entrevistado que vive allí.

A una escala mucho más amplia, se puede aplicar el mismo procedimiento de cúmulos a encuestas nacionales, tratando a las ciudades, estados o pueblos, como unidades primarias de muestreo para ser seleccionadas inicialmente y entrevistando a una muestra aleatoria simple de cada una de las ciudades, estados o pueblos escogidos. De esta manera, los entrevistadores no necesitan cubrir todos y cada uno de éstos, sino sólo un número mucho menor de tales áreas que han sido seleccionadas aleatoriamente para ser incluidas.

ERROR DE MUESTREO

A través del resto del texto seremos cuidadosos en distinguir entre las características de las muestras que estudiamos realmente y las poblaciones a las cuales esperamos generalizar. Para hacer esta distinción, en nuestros procedimientos estadísticos, no podemos, por tanto, seguir usando los mismos símbolos para representar la media y la desviación estándar tanto de la muestra como de la población. En su lugar debemos emplear diferentes símbolos, dependiendo de si nos estamos refiriendo a características de la muestra o de la población. En relación con la media, simbolizaremos siempre a la media de una *muestra* como $\bar{X}$ y a la media de una *población* como μ. En relación con la desviación estándar, simbolizaremos a la desviación estándar de una *muestra* como s y a la desviación estándar de su *población* como σ.

Normalmente, el investigador social trata de obtener una muestra que sea representativa de la población en la que está interesado. Como las muestras aleatorias le dan a todos y a cada uno de los miembros de la población la misma oportunidad de ser seleccionados para la muestra, son, a la larga, más representativas de las características poblacionales que sus contrapartes no aleatorias. Sin embargo, como se explicó brevemente en el Capítulo 1, *siempre* podemos esperar, por mera casualidad, que haya alguna diferencia entre una muestra, aleatoria o de otro tipo, y la población de la que se ha extraído. $\bar{X}$ casi nunca será exactamente igual a μ y s rara vez será exactamente igual a σ. Esta diferencia, conocida como *error de muestreo,* resulta sin importar qué tan bien se haya diseñado y realizado el plan de muestreo

TABLA 7.1 Una población y tres muestras aleatorias de calificaciones de exámenes finales

Población			*Muestra A*	*Muestra B*	*Muestra C*
70	80	93	96	40	72
86	85	90	99	86	96
56	52	67	56	56	49
40	78	57	52	67	56
89	49	48	303	249	273
99	72	30			
96	94	1 431	$\bar{X}$ = 75.75	$\bar{X}$ = 62.25	$\bar{X}$ = 68.25
	μ = 71.55				

con las mejores intenciones del investigador y donde no ocurre ningún fraude ni se han cometido errores.

Para ilustrar la operación del error de muestreo miremos ahora la Tabla 7.1, que contiene una población de 20 calificaciones de exámenes finales y 3 muestras, A, B y C, extraídas aleatoriamente de esta población (cada una se tomó con la ayuda de una tabla de números aleatorios). Como se esperaba, la media de la población (μ = 71,55) no es aritméticamente idéntica con ninguna de las tres medias muestrales; de manera similar, existen diferencias entre las mismas medias muestrales.

DISTRIBUCION MUESTRAL DE MEDIAS

Dada la presencia del error de muestreo, el estudiante se preguntará cómo es posible generalizar *siempre* a partir de una muestra a una población. Para llegar a una respuesta razonable, consideremos el trabajo de un hipotético investigador social que estudia la audición de radio entre el millón de residentes de una ciudad. Para ahorrar tiempo y dinero entrevista a sólo una muestra tomada aleatoriamente del total de la población de residentes. Extrae 500 residentes por medio de una tabla de números aleatorios y le pregunta a cada miembro de la muestra: ¿cuántos minutos escucha usted la radio diariamente?" y encuentra que el tiempo empleado en escucharla va desde 0 a 240 minutos. Como se ve en la Figura 7.1, el tiempo medio empleado en escuchar la radio en una muestra de 500 residentes es de 101,55 minutos.

Resulta que nuestro hipotético investigador social es levemente excéntrico y tiene una notable inclinación a extraer muestras de poblaciones. Es tan intenso su entusiasmo por el muestreo que continúa extrayendo muchas muestras adicionales de 500 residentes cada una y calculando el tiempo de audición de radio de los miembros de cada muestra. Este procedimiento continúa hasta que nuestro excéntrico investigador ha extraído 98 muestras de 500 residentes *cada una.* En el proceso de extraer 98 muestras aleatorias estudia, de hecho, a 49 000 entrevistados (500 x 98 = 49 000).

Supongamos, como se muestra en la Figura 7.2, que la población total de nuestra ciudad en estudio tiene un tiempo promedio de 99,75 minutos de audición de radio. Como lo ilustra también la Figura 7.2, supongamos que las muestras tomadas por

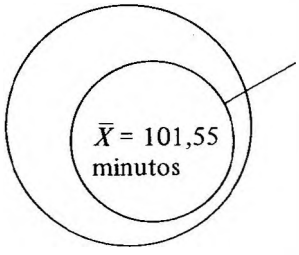


FIGURA 7.1 El tiempo promedio de audición para una muestra aleatoria tomada de una población hipotética.

nuestro excéntrico investigador social producen medias que van desde 89 hasta 111 minutos. De acuerdo con nuestro estudio previo, esto podría suceder fácilmente, simplemente con base en el error de muestreo.

Las distribuciones de frecuencia de los *puntajes crudos* pueden obtenerse tanto de muestras como de poblaciones. De modo semejante podemos construir una *distribución muestral de medias,* una distribución de frecuencia de un gran número de *medias* de muestras aleatorias que se han extraído de la misma población. La Tabla 7.2 presenta las 98 medias muestrales recogidas por nuestro excéntrico investigador social en forma de distribución muestral. Como cuando se trabaja con una distribución de puntajes crudos, las medias de la Tabla 7.2 se han ordenado en forma decreciente (de alta a baja) y la frecuencia con que ocurren se ha indicado en una columna adyacente.

Características de una distribución muestral de medias

Hasta este punto, no nos hemos enfrentado directamente al problema de generalizar

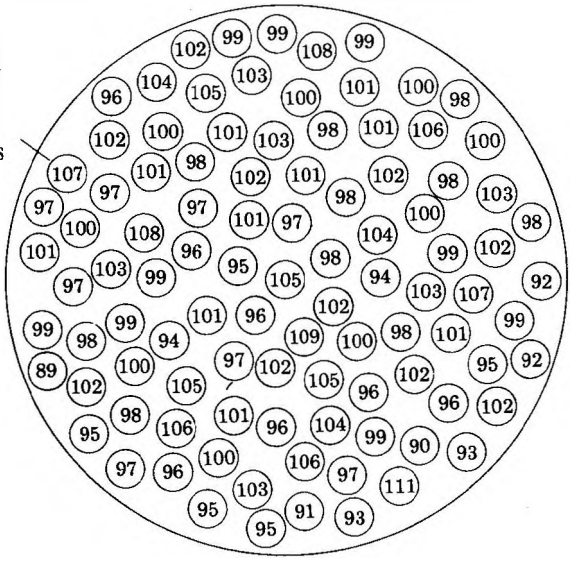


FIGURA 7.2 El tiempo promedio de audición en 98 muestras aleatorias tomadas de una población hipotética en la que μ = 99,75 minutos.

TABLA 7.2 Distribución muestral de medias (audición de radio) para 98 muestras aleatorias.

Media	*f*
111 min	1
110	1
109	1
108	2
107	2
106	3
105	4
104	5
103	6
102	8
101	9
100	9
99	9
98	8
97	7
96	6
95	5
94	4
93	3
92	2
91	1
90	1
89 min	1
	$N = 98$

de muestras a poblaciones. El modelo teórico conocido como distribución muestral de medias (como lo ilustran las 98 medias muestrales obtenidas por nuestro excéntrico investigador social) tiene ciertas propiedades que le otorgan un importante papel en el proceso de muestreo. Antes de dirigirnos hacia el procedimiento para hacer generalizaciones de muestras a poblaciones, debemos examinar primero las características de una distribución muestral de medias:

1. *La distribución muestral de medias se aproxima a una curva normal.* Como lo ilustra gráficamente la Figura 7.3 (a), al arreglar las medias muestrales de la Tabla 7.2, en un polígono de frecuencia, obtenemos la forma de una distribución normal. Esto es cierto para todas las distribuciones muestrales de medias sin importar la forma de la distribución de puntajes crudos de la población de la cual se extraen las medias.[1]
2. *La media de una distribución muestral de medias ("la media de medias") es igual a la verdadera media de la población.* Si tomamos un gran número de medias de muestras aleatorias de la misma población y encontramos la media de todas las medias muestrales tendremos el valor de la verdadera media de la población. Por lo tanto, como se ve en la Figura 7.3, la media de la

[1] Esto supone que hemos extraído grandes muestras aleatorias, de igual tamaño, de una población dada de puntajes crudos.

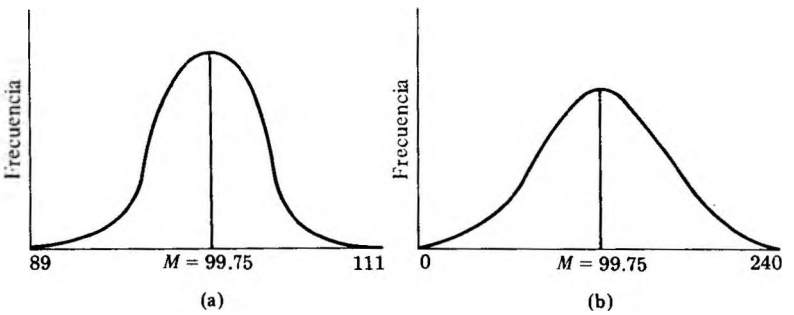


FIGURA 7.3 Polígonos de frecuencia de (a) la distribución muestral de medias de la Tabla 7.2 y (b) de la población de la que se extrajeron estas medias.

distribución muestral de medias (a) es la misma que la media de la población de la que se sacó (b). Pueden considerarse como valores intercambiables.

3. *La desviación estándar de una distribución muestral de medias es menor que la desviación estándar de la población.*

Como lo ilustra la Figura 7.3, la dispersión de la distribución muestral es siempre menor que la dispersión de la población total. Esto es cierto porque tomamos datos medios (más que el rango de puntajes crudos que componen esas medias), eliminando así los valores de puntajes crudos extremos. Por ejemplo, el puntaje de desviación media 100 puede obtenerse de los puntajes crudos 60, 90, 110 y 140. (60 + 90 + 110 + 140 = 400/4 = 100). Graficando los puntajes crudos, incluimos valores entre 60 y 140. Graficando el puntaje de la media, sin embargo, reducimos obviamente la ocurrencia de tales valores extremos de los puntajes a un valor único de 100. Como resultado, esperamos obtener una desviación estándar menor cuando se tomen en conjunto y se grafique un determinado número de puntajes de medias.

La distribución muestral de medias como una curva normal

Como se indicó en el Capítulo 6, si definimos la probabilidad en términos de frecuencia de ocurrencia, entonces la curva normal puede considerarse como una distribución de probabilidad (podemos decir que la probabilidad disminuye a medida que viajamos por la línea base alejándonos de la media en una u otra dirección).

Con esta idea, podemos encontrar la probabilidad de obtener varios puntajes crudos en una distribución, dadas una cierta media y su desviación estándar. Por ejemplo, para encontrar la probabilidad asociada con la obtención de alguien que tenga un ingreso anual entre $5 000 y $7 000, en una población con un ingreso medio de $5 000 y una desviación estándar de $1 500, convertimos el puntaje crudo $7 000 en un puntaje z (+1,33) y vamos a la Tabla B al final del texto para obtener el porcentaje de la frecuencia total que cae entre el puntaje z 1,33 y la media. Esta área contiene el 40,82% de los puntajes crudos. Así, P = 0,41 redondeado, para que

encontremos un individuo cuyo ingreso anual oscile entre $5 000 y $7 000. Si queremos saber la probabilidad que existe de encontrar a alguien cuyo ingreso sea de $7 000 o más, debemos ir un paso más allá y restar el porcentaje obtenido en la Tabla B de 50% —el porcentaje del área que está a uno y otro lado de la media. Restando 40,82% de 50%, vemos que el 9,18% cae en o más allá de $7 000. Por lo tanto, moviéndonos 2 lugares decimales hacia la izquierda, podemos decir que tenemos P = 0,09 (9 oportunidades entre 100) de encontrar un individuo cuyo ingreso sea de ~~$80 000 o más.~~

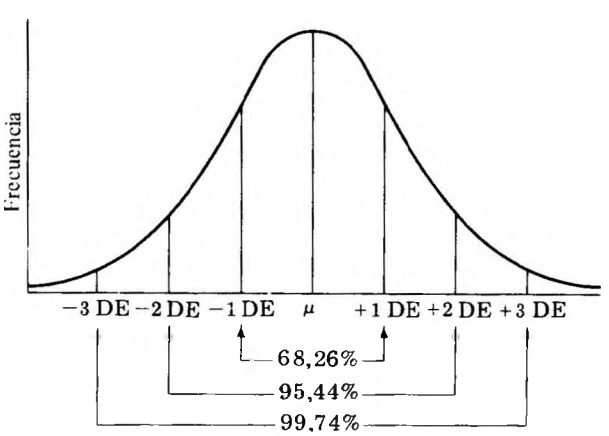
En el presente contexto no nos interesa ya obtener probabilidades asociadas con la distribución de *puntajes crudos.* En lugar de esto nos encontramos trabajando con una distribución de *medias muestrales* que se han extraído de la población total de puntajes y deseamos hacer afirmaciones de probabilidad acerca de esas medias muestrales.

Como lo ilustra la Figura 7.4, ya que la distribución muestral de medias toma la forma de la curva normal, podemos decir que la probabilidad disminuye a medida que nos alejamos de la media de medias (la verdadera media de la población). Esto tiene sentido porque, como recordará el estudiante, la distribución muestral es producto de diferencias casuales entre las medias muestrales (error de muestreo). Por este motivo esperamos que por casualidad, y sólo por casualidad, la mayoría de las medias muestrales caigan cerca del valor de la verdadera media de la población, mientras que relativamente pocas medias muestrales caigan lejos de ella.

La Figura 7.4 indica que cerca del 68% de las medias muestrales en una distribución muestral fluctúan entre — 1DE y + 1DE de la media de medias (verdadera media poblacional). En términos de probabilidad, podemos decir que P = 0,68 de cualquier media muestral dada que caiga dentro de este intervalo. De igual manera, podemos decir que la probabilidad de que cualquier media muestral caiga entre —2DE y + 2DE de la media de medias es de cerca de 0,95 (95 oportunidades entre 100) y así sucesivamente.

Dado que la distribución muestral toma la forma de la curva normal, podemos

FIGURA 7.4 La distribución muestral de medias como una distribución de probabilidad



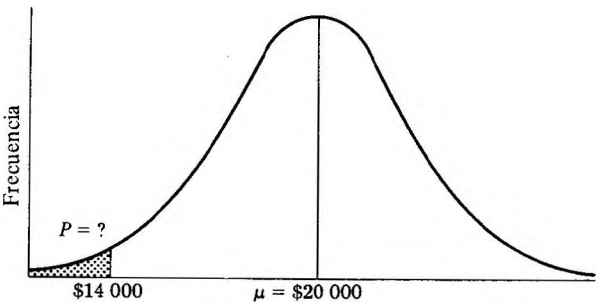


FIGURA 7.5 La probabilidad asociada con la obtención de una media muestral de $14 000 o menos, si la verdadera media poblacional es de 20 000 y la desviación estándar es de $2 600.

usar también los puntajes *z* y la Tabla B para obtener la probabilidad de cualquier media muestral y no sólo aquellas que son múltiplos exactos de la desviación estándar. Dada una media de medias y la desviación estándar de la distribución muestral, el proceso es idéntico al que se usó en el capítulo anterior para una distribución de puntajes crudos. Sólo se han cambiado los nombres.

Imaginemos, por ejemplo, que cierta universidad sostiene que sus ex-alumnos tienen un ingreso anual promedio (μ) de $20 000. Tenemos motivos para dudar de la legitimidad de esta pretensión y decidimos ensayarla en una muestra aleatoria de 100 ex alumnos. En el proceso obtenemos una media muestral de sólo $14 000. Preguntamos ahora: ¿qué tan probable sería que obtuviéramos una media de $14 000 o al menos de que la verdadera media poblacional fuera realmente $20 000? ¿Ha dicho la universidad la verdad? O, ¿es este sólo un intento de hacer publicidad entre el público para incrementar las inscripciones o donaciones? La Figura 7.5 ilustra el área para la cual buscamos una solución.

Supongamos que sabemos que la desviación estándar de la distribución muestral es $2 600. Siguiendo el procedimiento estándar, convertimos la media muestral en un puntaje *z*, como sigue:

$$Z = \frac{\overline{X} - M}{\sigma_{\overline{X}}} = \frac{14\,000 - 20\,000}{2600} = -2{,}31$$

donde

$\overline{X}$ = una media muestral en la distribución

$M = \mu =$ la media de medias (igual a la pretensión de la universidad sobre la verdadera media de la población)

$\sigma_{\overline{X}}$ = la desviación estándar de la distribución muestral de medias

El resultado del procedimiento anterior nos dirá que una media muestral de $14 000 yace exactamente en 2,31 desviaciones estándar por abajo de la supuesta media poblacional verdadera, $20 000. Recurriendo a la Tabla B, al final del texto, vemos que el 48,96% de las medias muestrales caen entre $14,000 y $20,000. Restando del 50% obtenemos el porcentaje de la distribución que representa medias muestrales de $14 000 o menos si es que la verdadera media poblacional es de $20 000.

Esta cifra es 1,04% (50% – 48,96% = 1,04%). Por lo tanto, la probabilidad es 0,01 redondeando (1 oportunidad entre 100) de obtener una media muestral de $14 00 o menos, cuando la verdadera media poblacional es $20 000. Con una probabilidad tan pequeña de equivocarnos, podemos decir, con cierta confianza, que la verdadera media de la población *no* es realmente $20 000. Es dudoso que el informe de la universidad sobre el ingreso anual de sus exalumnos represente algo más que mala publicidad.

ERROR ESTANDAR DE LA MEDIA

Hasta ahora hemos hecho de cuenta que el investigador social tiene efectivamente información de primera mano acerca de la distribución muestral de las medias. Hemos actuado como si él, al igual que el investigador excéntrico, hubiera recogido realmente datos sobre un gran número de medias muestrales que se extrajeron aleatoriamente de alguna población. Si así fuera, sería una tarea bastante simple hacer generalizaciones acerca de la población, ya que la media de medias toma un valor que es igual al de la verdadera media poblacional.

En la práctica real, el investigador social rara vez recoge datos sobre más de una o dos muestras de las que aún espera generalizar a una población completa. Extraer una distribución muestral de medias requiere el mismo esfuerzo que tomaría estudiar a todos y cada uno de los miembros de la población. Como resultado, el investigador social no tiene un conocimiento real sobre la media de medias o la desviación estándar de la distribución muestral. Sin embargo, sí tiene un buen método para *estimar* la desviación estándar de la distribución muestral de medias sobre la base de los datos recogidos en una sola muestra. Esta estimación se conoce como el *error estándar de la media* y se simboliza por $\sigma_{\bar{x}}$[2]. Por fórmula,

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{N-1}}$$

donde

$\sigma_{\bar{X}}$ = el error estándar de la media (una estimación de la desviación estándar de una distribución muestral de medias)
s = la desviación estándar de una *muestra*
N = el número total de puntajes en una *muestra*

[2]En muchos textos, el error estándar de la media, basado en la desviación estándar poblacional y simbolizado por σ_X, se distingue del error estándar de la media estimado, basado en la desviación estándar de la muestra y simbolizado por $s_{\bar{x}}$. Sin embargo, si no se mide la población entera no se conoce el valor de la desviación estándar de la población y por lo tanto debe estimarse. Con el fin de simplificar, hemos elegido, por tanto, pasar por alto la anterior distinción e introducir en su lugar una fórmula única para el error estándar de la media, simbolizado por $\sigma_{\bar{x}}$ y basado en los datos de la muestra.

Para ilustrar, si la desviación estándar de una muestra de diez entrevistados es 2,5, entonces

$$\sigma_{\bar{Y}} = \frac{2,5}{\sqrt{10 - 1}}$$
$$= \frac{2,5}{3,0}$$
$$= 0,83$$

Como se anotó arriba, el investigador social que sólo estudia una o dos muestras no puede conocer la media de medias, cuyo valor es igual al de la verdadera media de la población. Sólo tiene la media muestral que ha obtenido, que difiere de la verdadera media poblacional como resultado del error de muestreo. Pero, ¿no hemos caído en un círculo vicioso? ¿Cómo es posible estimar la verdadera media poblacional a partir de una sola media muestral, especialmente a la vista de tales diferencias inevitables entre muestras y poblaciones?

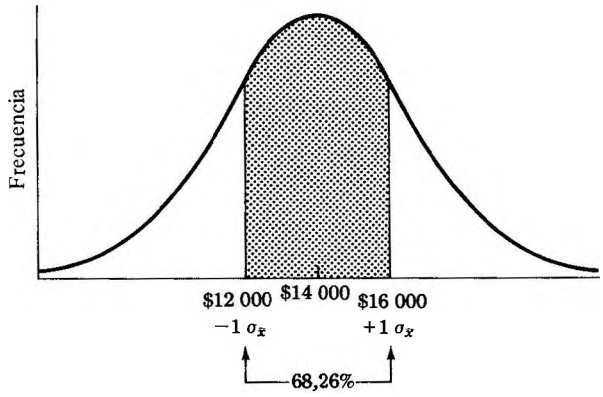
De hecho, hemos recorrido una distancia considerable desde nuestra posición original. Habiendo tratado la naturaleza de la distribución muestral de medias, estamos preparados ahora para estimar el valor de una media poblacional. Con la ayuda del error estándar de la media, podemos encontrar *el rango de valores de la media dentro del cual es probable que fluctúe nuestra verdadera media poblacional. Podemos también estimar la probabilidad de que nuestra media poblacional caiga realmente dentro de ese rango de valores medios.* Este es el concepto del *intervalo de confianza.*

INTERVALOS DE CONFIANZA

Para explorar el procedimiento que se sigue para encontrar un intervalo de confianza, ampliemos un ejemplo anterior. Supongamos que la muestra aleatoria (de un investigador) de 100 exalumnos de cierta universidad marca un ingreso anual medio de $ 14 000. Como sus datos provienen solamente de una muestra aleatoria, y no de la población total de exalumnos, no podemos estar seguros de que el ingreso medio reportado sea realmente un reflejo de esta población de exalumnos universitarios. Como ya hemos visto, el error de muestreo es, después de todo, el producto inevitable de sacar muestras de poblaciones.

Sin embargo, sí sabemos que el 68,26% de todas las medias muestrales aleatorias, en la distribución muestral de medias, caerán entre −1 DE y +1 DE de la verdadera media poblacional. Estimando la desviación estándar de la distribución muestral ($\sigma_{\bar{x}}$ = $2 000) y usando nuestra media muestral $14 000 como una estimación de la media poblacional, podemos establecer el rango dentro del cual hay 68 oportunidades entre 100 (redondeando) de que la verdadera media poblacional caiga. Este rango de ingresos medios, conocido como el *intervalo de confianza del 68%* se ilustra gráficamente en la Figura 7.6.

FIGURA 7.6 Un intervalo de Confianza del 68% cuando $\sigma_{\bar{x}}$ = \$2 000 y $\overline{X}$ = \$14 000



El intervalo de confianza del 68% puede obtenerse de la siguiente manera:

$$\text{intervalo de confianza del 68\%} = \overline{X} + \sigma_{\bar{x}}$$

donde

$\overline{X}$ = una media muestral
$\sigma_{\bar{x}}$ = el error estándar de la media

Aplicando la fórmula anterior a nuestro problema:

$$\text{el intervalo de confianza del 68\%} = \$14{,}000 \pm \$2\,000$$
$$= 12\,000 \longleftrightarrow \$16\,000$$

Por lo tanto, el investigador social informa que *tiene un 68% de confianza* en que el ingreso poblacional medio estre estos exalumnos universitarios sea de \$ 14 000, más o menos \$ 2 000. En otras palabras, hay 68 oportunidades entre 100 (P = 0,68) de que la verdadera media poblacional caiga realmente dentro de un rango entre \$ 12 000 y \$ 16 000 (\$ 14 000 − \$ 2 000 = \$ 12 000; \$ 14 000 + \$ 2 000 = \$ 16 000). Esta estimación se hace a pesar del error de muestreo, aunque dentro de un margen de error (más o menos \$20 000) y a un nivel de confianza específico (del 68%).

Pueden construirse intervalos de confianza para cualquier nivel de probabilidad. La mayoría de los investigadores sociales no están suficientemente seguros para estimar una media poblacional sabiendo que sólo hay 68 oportunidades entre 100 de estar en lo correcto (68 de cada 100 medias muestrales caen dentro del intervalo entre \$ 12 000 y \$ 16 000). Como resultado, se ha convertido en una cuestión convencional utilizar un intervalo de confianza *más amplio*, menos preciso, que tiene *mejores probabilidades* de hacer una estimación exacta de la media poblacional. Tal modelo se encuentra en el *intervalo de confianza del 95%*, por medio del cual se estima la media poblacional sabiendo que hay 95 oportunidades entre 100 de estar en lo cierto; hay 5 oportunidades entre 100 de equivocarse

(95 de cada 100 medias muestrales caen dentro del intervalo). Sin embargo, incluso usando el intervalo de confianza del 95%, debe tenerse en mente el hecho de que la media muestral del investigador podría ser una de esas cinco medias muestrales que caen fuera del intervalo establecido. En la toma de decisiones, en estadística, nunca se está completamente seguro.

¿Cómo hacemos para encontrar el intervalo de confianza del 95%? Sabemos ya que el 95,44% de las medias muestrales en una distribución muestral se encuentran entre − 2 DE y + 2 DE de la media de medias. Mirando la Tabla B podemos afirmar que 1.96 desviaciones estándar en ambas direcciones cubren exactamente el 95% de las medias muestrales (47,50% a cada lado de la media de medias). Para encontrar el intervalo de confianza del 95%, debemos multiplicar primero el error estándar de la media por 1,96 (el intervalo está a 1,96 unidades de $\sigma_{\bar{x}}$ en una y otra dirección de la media). Por lo tanto,

el intervalo de confianza del 95% = $\overline{X} \pm (1{,}96)\sigma_{\bar{X}}$

donde

$\overline{X}$ = una media muestral
$\sigma_{\bar{X}}$ = el error estándar de la media

Si aplicamos el intervalo de confianza del 95% a nuestra estimación del ingreso medio entre los exalumnos universitarios, vemos que:

el intervalo de confianza del 95% = \$14 000 ± (1,96) \$2 000
= \$14 000 ± \$3 920
= \$10 080 ⟷ \$17 920

Conclusión: Tenemos un 95% de confianza en que la verdadera media poblacional cae entre los \$ 10 080 y los \$ 17 920.

Resumamos el procedimiento paso a paso para obtener el intervalo de confianza del 95% en la siguiente muestra aleatoria de datos crudos.

X
1
5
2
3
4
1
2
2
4
3

PASO 1: Encontrar la media de la muestra

X	
1	
5	
2	
3	$\bar{X} = \frac{\Sigma X}{N}$
4	
1	$= \frac{27}{10}$
2	
2	
4	$= 2{,}7$
3	
$\Sigma X = 27$	

PASO 2: Obtener la desviación estándar de la muestra

X	X^2	
1	1	
5	25	
2	4	$s = \sqrt{\frac{\Sigma X^2}{N} - \bar{X}^2}$
3	9	
4	16	$= \sqrt{\frac{89}{10} - (2{,}7)^2}$
1	1	$= \sqrt{8{,}9 - 7{,}29}$
2	4	
2	4	$= \sqrt{1{,}61}$
4	16	$= 1{,}27$
3	9	
	$\Sigma X^2 = 89$	

PASO 3: Obtener el error estándar de la media

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{N-1}}$$

$$= \frac{1{,}27}{\sqrt{10-1}}$$

$$= \frac{1{,}27}{3}$$

$$= 0{,}42$$

PASO 4: Multiplicar el error estándar de la media por 1.96

El intervalo de confianza del 95% $= \overline{X} \pm (1{,}96)\sigma_X$

$= 2.7 \pm (1{,}96)\,(0{,}42)$

$= 2{,}7 \pm 0{,}82$

PASO 5: Sumar y restar este producto de la media muestral para encontrar el rango de puntajes promedio dentro de los cuales cae la media poblacional:

$$\text{el intervalo de confianza del 95\%} = 2{,}7 \pm 0{,}82$$
$$= 1{,}88 \longleftrightarrow 3{,}52$$

Podemos tener un 95% de confianza de que la verdadera media poblacional está entre 1,88 y 3,52.[3]

Un intervalo de confianza aún más riguroso es el *intervalo de confianza* del 99%. En la Tabla B, al final del texto, vemos que el puntaje z 2,58 representa el 49,50% del área a cada lado de la curva. Doblar esta cantidad produce el 99% del área bajo la curva; el 99% de las medias muestrales cae dentro de ese intervalo. En términos de probabilidad, 99 de cada 100 medias muestrales se encuentran entre −2,58 DE y +2,58 DE de la media. A la inversa, sólo 1 de cada 100 medias cae fuera del intervalo. Por fórmula, el intervalo de confianza del 99% $= \overline{X} \pm (2{,}58)\sigma_{\bar{x}}$

donde

$\overline{X}$ = una media muestral
σ_X = el error estándar de la media

Con respecto a nuestra estimación del ingreso medio entre exalumnos universitarios:

$$\text{el intervalo de confianza del 99\%} = \$14\,000 \pm (2{,}58)\ \$2\,000$$
$$= \$14\,000 \pm \$5\,160$$
$$= \$8\,840 \longleftrightarrow \$19\,160$$

Hemos determinado, con un 99% de confianza, que la verdadera media poblacional cae en algún sitio entre $ 8 840 y $ 19 160.

El estudiante deberá notar que el intervalo de confianza del 99% consiste en una banda más amplia ($ 8 840 a $ 19 160) que el intervalo de confianza del 95% de $ 10 080 a $ 17 920). El intervalo del 99% abarca más del área total bajo la curva normal y, por lo tanto, a un mayor número de medias muestrales. Esta banda más amplia de puntajes promedio nos da mayor confianza en que hemos estimado la verdadera media poblacional con exactitud. Una sola media muestral de cada 100 se encuentra fuera del intervalo. Por otra parte, al aumentar nuestra confianza del 95 al 99 por ciento, hemos sacrificado también un grado de precisión al señalar la media poblacional. Manteniendo constante el tamaño de la muestra, el investigador social

[3] Para propósitos ilustrativos empleamos una muestra pequeña. En la práctica, el investigador que utilice dicho procedimiento para encontrar un intervalo de confianza deberá trabajar por lo menos con 30 casos para hallar la condición de normalidad en la distribución muestral de medias (véase la discusión de la razón t Capítulo 8).

debe escoger entre una mayor precisión o una mayor confianza de estar en lo correcto.

Para resumir el procedimiento que se sigue paso a paso para encontrar el intervalo de confianza del 99%, reexaminemos la muestra aleatoria de puntajes:

X
1
5
2
3
4
1
2
2
4
3

PASO 1: Encontrar la media de la muestra

1	
5	
2	
3	$\bar{X} = \frac{\Sigma X}{N}$
4	
1	$= \frac{27}{10}$
2	
2	$= 2{,}7$
4	
3	
$\Sigma X = 27$	

PASO 2: Obtener la desviación estándar de la muestra

X	X^2	
1	1	
5	25	
2	4	$s = \sqrt{\frac{\Sigma X^2}{N} - \bar{X}^2}$
3	9	
4	16	$= \sqrt{\frac{89}{10} - (2{,}7)^2}$
1	1	
2	4	$= \sqrt{8{,}9 - 7{,}29}$
2	4	
4	16	$= \sqrt{1{,}61}$
3	9	
	$\Sigma X^2 = 89$	$= 1{,}27$

PASO 3: Obtener el error estándar de la media

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{N-1}}$$
$$= \frac{1{,}27}{\sqrt{10-1}}$$
$$= \frac{1{,}27}{3}$$
$$= 0{,}42$$

PASO 4: Multiplicar el error estándar de la media por 2,58

el intervalo de confianza del 99% $= \bar{X} \pm (2{,}58)\,\sigma_{\bar{X}}$
$= 2{,}7 \pm (2{,}58)\,(0{,}42)$
$= 2{,}7 \pm 1{,}08$

PASO 5: Sumar y restar este producto de la media muestral para encontrar el rango de puntajes promedio dentro del cual cae la media poblacional

el intervalo de confianza del 99% $= 2{,}7 \pm 1{,}08$
$= 1{,}62 \longleftrightarrow 3{,}78$

Tenemos un 99% de confianza en que la verdadera media poblacional cae entre 1,62 y 3,78.

ESTIMACION DE PROPORCIONES

Hasta aquí, nos hemos centrado en los procedimientos para estimar medias poblacionales. El investigador social a menudo busca presentar una estimación de una *proporción* poblacional estrictamente con base en la proporción que obtiene en una muestra aleatoria. Una circunstancia conocida es la del encuestador, cuyos datos sugieren que una cierta proporción de los votos irán hacia un determinado tema o candidato político para un cargo público. Cuando un encuestador informa que el 45% de la votación será a favor de cierto candidato, lo hace sabiéndolo con una precisión menor de 100%. En general, tiene una confianza de 95 o 99% de que su proporción estimada cae dentro de la extensión del rango (por ejemplo, entre 40 y 50 por ciento).

Estimamos las proporciones por medio del procedimiento que acabamos de usar para estimar medias. Todos los estadísticos –incluyendo las medias y las proporciones– tienen sus distribuciones muestrales. Tal como encontramos anteriormente, el

error estándar de la media, podemos buscar ahora el *error estándar de la proporción.* Por fórmula,

$$\sigma_P = \sqrt{\frac{P(1-P)}{N}}$$

donde

σ_P = el error estándar de la proporción (una estimación de la desviación estándar de la distribución muestral de proporciones)
P = una proporción muestral
N = el número total en la muestra

Con fines ilustrativos, digamos que el 45 por ciento de una muestra aleatoria de 100 estudiantes universitarios informa que éstos están a favor de la legalización de las drogas. El error estándar de la proporción sería

$$\begin{aligned}\sigma_P &= \sqrt{\frac{0{,}45(0{,}55)}{100}} \\ &= \sqrt{\frac{0{,}2475}{100}} \\ &= \sqrt{0{,}0025} \\ &= 0{,}05\end{aligned}$$

Para encontrar el intervalo de confianza del 95 por ciento multiplicamos el error estándar de la proporción por 1,96 y sumamos y restamos este producto a la proporción muestral:

$$\text{el intervalo de confianza del 95\%} = P \pm (1{,}96)\,\sigma_P$$

donde

P = una proporción muestral
σ_P = el error estándar de la proporción

Si buscamos la proporción de estudiantes universitarios que están a favor de la legalización de las drogas,

$$\begin{aligned}\text{el intervalo de confianza del 95\%} &= 0{,}45 \pm (1{,}96)\ 0{,}05 \\ &= 0{,}45 \pm 0{,}098 \\ &= 0{,}35 \longleftrightarrow 0{,}55\end{aligned}$$

Tenemos un 95 por ciento de confianza en que la verdadera proporción poblacional no es ni menor a 0,35 ni mayor de 0,55. Más específicamente, entre el 35 y el 55 por ciento de esta población de estudiantes universitarios están a favor de la legalización de todas las drogas. Existe un 5 por ciento de probabilidad de que nos equivoquemos; 5 veces entre 100, tales intervalos de confianza no contendrán la verdadera proporción poblacional.

Resumamos el procedimiento para estimar una proporción por medio del intervalo de confianza del 95%. Supongamos que la proporción muestral para la cual haremos nuestra estimación resulta ser 0,40 (40 por ciento de los 100 casos caen dentro de esta categoría).

PASO 1: Obtener el error estándar de la proporción

$$\sigma_P = \sqrt{\frac{P(1-P)}{N}}$$
$$= \sqrt{\frac{0{,}40(0{,}60)}{100}}$$
$$= \sqrt{\frac{0{,}24}{100}}$$
$$= \sqrt{0{,}0024}$$
$$= 0{,}049$$

PASO 2: Multiplicar el error estándar de la proporción por 1,96 el intervalo de confianza del 95% $= P \pm (1{,}96)\sigma_P$
$= 0{,}40 \pm (1{,}96)\ (0{,}049)$
$= 0{,}40 \pm 0{,}096$

PASO 3: Sumar y restar este producto de la proporción muestral para encontrar el rango de proporciones dentro de la que cae la proporción poblacional

el intervalo de confianza del 95% $= 0{,}40 \pm 0{,}096$
$= 0{,}30 \longleftrightarrow 0{,}50$

Podemos decir, con un 95% de confianza, que la verdadera proporción poblacional fluctúa entre 0,30 y 0,50.

RESUMEN

Este capítulo ha explorado los procedimientos y conceptos claves relacionados con la generalización de muestras a poblaciones. Se presentaron los métodos aleatorios y no aleatorios de muestreo. Se señaló que el error de muestreo –la diferencia inevitable entre muestras y poblaciones– ocurre a pesar de un plan de muestreo bien diseñado y ejecutado. Como resultado del error de muestreo podemos estudiar las

características de la distribución muestral de medias, una distribución que forma una curva normal y cuya desviación estándar puede estimarse con la ayuda del error estándar de la media. Armados con tal información, podemos construir intervalos de confianza para las medias (o las proporciones) dentro de las cuales tenemos confianza (95 por ciento o 99 por ciento) de que caiga la verdadera media (o proporción) poblacional. De esta manera podemos hacer generalizaciones de una muestra a una población.

PROBLEMAS

1. Encontrar el error estándar de la media con la siguiente muestra de 30 puntajes:

3	5
3	3
2	3
1	2
5	2
4	3
5	2
1	4
6	6
3	1
2	1
1	3
1	4
2	3
3	4

2. Con la media muestral del Problema 1 buscar (a) el intervalo de confianza del 95% y (b) el intervalo de confianza del 99%.

3. Buscar el error estándar de la media con la siguiente muestra de 34 puntajes:

10	1
4	8
10	7
5	5
5	6
6	10
7	6
3	8
5	7
4	7
4	6
5	5
6	5
6	4
7	3

5	4
8	5

4. Con la media muestral del Problema 3 encontrar (a) el intervalo de confianza del 95% y (b) el intervalo de confianza del 99%.

5. Hallar el error estándar de la media con la siguiente muestra de 32 puntajes:

4	4
2	3
5	6
6	6
1	7
1	1
7	5
8	7
7	8
8	8
8	4
2	5
6	3
5	2
6	6
4	5

6. Con la media muestral del Problema 5 buscar (a) el intervalo de confianza del 95% y (b) el intervalo de confianza del 99% .
7. Para estimar la proporción de estudiantes de una determinada universidad que favorecen la abolición de grupos políticos, un investigador social entrevistó una muestra aleatoria de 50 estudiantes de la población universitaria. Encontró que el 57 por ciento de la muestra estaba a favor de deshacerse de los grupos políticos (proporción muestral = 0,57). Con esta información (a) buscar el error estándar de la proporción y (b) construir un intervalo de confianza del 95% .
8. Dados el tamaño muestral de 150 y una proporción muestral de 0,32 (a) buscar el error estándar de la proporción y (b) construir un intervalo de confianza del 95%.
9. Dados el tamaño muestral de 200 y una proporción muestral de 0,25 (a) buscar el error estándar de la proporción y (b) construir un intervalo de confianza del 95%.

PARTE III
La toma de decisiones

8

Comprobación de diferencias entre medias

En el Capítulo 7 vimos que una media poblacional o una proporción puede estimarse a partir de la información que obtenemos de una sola muestra. Por ejemplo, podríamos estimar el nivel de anomia en una ciudad, en particular la proporción de personas ancianas que están en una situación económica mala o la actitud media hacia la segregación racial entre una población de negros norteamericanos.

Aunque el enfoque descriptivo y de recolección de datos de la estimación de medias y proporciones tiene una importancia obvia, *no* constituye el objetivo fundamental de la toma de decisiones o de la actividad de la investigación social. Muy por el contrario, la mayoría de los investigadores sociales se interesan en la tarea de *contrastar las hipótesis* que existen acerca de las diferencias entre dos o más muestras.

Cuando comprueban diferencias entre las muestras, los investigadores sociales se hacen preguntas tales como: ¿Difieren los alemanes de los norteamericanos con respecto a la obediencia a la autoridad? ¿Quién presenta una tasa de suicidios más alta, los católicos o los protestantes? ¿Qué efecto producen los entrevistadores negros frente a los blancos sobre la honestidad de los entrevistados negros? ¿Las personas políticamente conservadoras disciplinan más severamente a sus niños que las personas políticamente liberales? (ver Capítulo 1). Nótese que cada pregunta de investigación implica hacer una *comparación* entre dos grupos: conservadores frente a liberales, entrevistadores negros frente a entrevistadores blancos; protestantes frente a católicos; alemanes frente a norteamericanos.

LA HIPOTESIS NULA: NINGUNA DIFERENCIA ENTRE LAS MEDIAS

En el análisis estadístico se ha vuelto convencional empezar con la comprobación de la *hipótesis nula* —la hipótesis que sustenta que dos muestras han sido extraídas de la

misma población. De acuerdo con la hipótesis nula, cualquier diferencia observada entre las muestras se considera como un hecho casual resultante únicamente del error de muestreo. Por lo tanto, la diferencia que existe entre dos medias muestrales no representa una diferencia real entre sus medias poblacionales.

En el presente contexto, la hipótesis nula puede simbolizarse como

$$\mu_1 = \mu_2$$

donde

μ_1 = la media de la primera población
μ_2 = la media de la segunda población

Examinemos las hipótesis nulas para las preguntas de investigación planteadas anteriormente:

1. Los alemanes no son ni más ni menos obedientes a la autoridad que los norteamericanos.
2. Los protestantes presentan la misma tasa de suicidios que los católicos.
3. Los entrevistados negros son igualmente sinceros, sean entrevistados por blancos o por negros.
4. Las personas políticamente conservadoras disciplinan a sus niños en el mismo grado que las personas políticamente liberales.

Debe notarse que la hipótesis nula no niega la posibilidad de obtener diferencias entre medias *muestrales.* Al contrario, busca explicar tales diferencias entre las medias muestrales atribuyéndolas a la operación del error de muestreo. Por ejemplo, de acuerdo con la hipótesis nula, si encontramos que una *muestra* aleatoria de mujeres dentistas ganan menos dinero ($\bar{X}$ = $12 000) que una *muestra* aleatoria de hombres dentistas ($\bar{X}$ = $15 000), no concluimos, sobre esa base, que la *población* de mujeres dentistas gana menos dinero que la *población* de hombres dentistas. En lugar de esto tratamos la diferencia muestral obtenida ($15 000 – $12 000 = $3 000) como producto del error de muestreo —la diferencia que resulta inevitablemente del proceso de muestrear de una población dada. Como veremos más tarde, este aspecto de la hipótesis nula proporciona un importante vínculo con la teoría del muestreo.

LA HIPOTESIS DE INVESTIGACION: ALGUNA DIFERENCIA ENTRE LAS MEDIAS

La hipótesis nula se expone generalmente (aunque no necesariamente) con la esperanza de rechazarla. Esto tiene sentido, ya que la mayoría de los investigadores sociales busca establecer relaciones entre variables. Esto es, están frecuentemente más interesados en encontrar diferencias que en determinar que las diferencias no existen. Para

ilustrar, ¿quién se molestaría en estudiar a los católicos y a los protestantes con la esperanza de que sus tasas de suicidio *no* difieran? Las diferencias que existen entre los grupos —ya sea que se esperen en terrenos teóricos o empíricos— proporcionan a menudo la razón fundamental sobre la cual se realiza el estudio.

Si rechazamos la hipótesis nula, si encontramos que nuestra hipótesis, de que no existe ninguna diferencia entre las medias, no se sostiene, aceptamos automáticamente la *hipótesis de investigación* (hipótesis alterna) que plantea que sí existe una verdadera diferencia poblacional. Este es un resultado frecuentemente esperado en la investigación social. La hipótesis de investigación establece que las dos muestras se han tomado de la población teniendo medias diferentes. Afirma que la diferencia obtenida entre medias muestrales es demasiado grande como para ser explicada por el error de muestreo.

La hipótesis de investigación para diferencias entre medias se simboliza como

$$\mu_1 \neq \mu_2$$

donde

μ_1 = la media de la primera población
μ_2 = la media de la segunda población (el signo $\neq$ se lee: "no es igual")

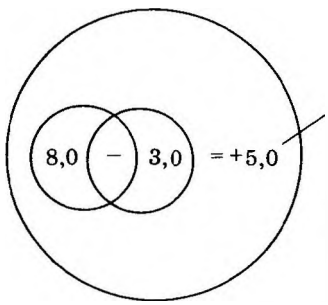
Podemos especificar las siguientes hipótesis de investigación para las preguntas planteadas anteriormente:

1. Los alemanes difieren de los americanos con respecto a la obediencia a la autoridad.
2. Los protestantes no tiene la misma tasa de suicidio que los católicos.
3. La honestidad de los entrevistados negros difiere, dependiendo de si los entrevistan blancos o negros.
4. Las personas políticamente liberales difieren de las políticamente conservadoras con respecto a sus métodos en la crianza de los niños.

DISTRIBUCION MUESTRAL DE DIFERENCIAS DE MEDIAS

En el capítulo anterior vimos que las 98 medias de las 98 muestras extraídas por nuestro investigador social excéntrico podían representarse en forma de distribución muestral de medias. De manera semejante, imaginemos ahora que el mismo investigador social excéntrico toma al mismo tiempo *no una*, sino *dos* muestras aleatorias de una población dada de personas. Supongamos, por ejemplo, que toma una muestra de 500 personas políticamente liberales y otra de 500 personas políticamente conservadoras. Para comprobar la hipótesis de investigación de que los liberales son menos estrictos como padres, que los conservadores, él interroga entonces a todos los miembros de la muestra acerca de sus métodos de crianza (por ejemplo: ¿Castiga usted siempre a sus niños? ¿Les pega usted? Si es así, ¿qué tan frecuentemente?).

FIGURA 8.1 La diferencia media en permisibilidad entre muestras de liberales y conservadores tomada de una población hipotética



De las respuestas a tales preguntas se obtiene una medida de permisibilidad* en la crianza de los niños que puede utilizarse para comparar las muestras liberal y conservadora. Los puntajes de esta medida van desde 1 (no rígido) hasta 10 (muy rígido). Como se ilustra gráficamente en la Figura 8.1, nuestro investigador social excéntrico encuentra que su muestra de liberales es ~~menos~~ rígida ($\bar{X}$ = 8,0) que su muestra de conservadores ($\bar{X}$ = 3,0).

Podríamos preguntarnos: A la luz del error de muestreo, ¿podemos esperar que una diferencia entre 8,0 y 3,0 (8,0 − 3,0 = +5,0) se dé estrictamente con base en el azar y solamente por el azar? , ¿debemos aceptar la hipótesis nula de que no existe ninguna diferencia poblacional? , ¿esta diferencia muestral obtenida de +5,0 es lo suficientemente amplia para indicar la verdadera diferencia poblacional que se muestra entre los conservadores y los liberales con respecto a sus prácticas de crianza de los niños?

En el Capítulo 2 se nos presentaron las distribuciones de frecuencia de puntajes crudos de una población dada. En el Capítulo 7 vimos que era posible construir una distribución muestral de puntajes promedio, una distribución de frecuencia de medias muestrales. Al dirigirnos al asunto que tenemos entre manos, debemos llevar la idea de la distribución de frecuencia un paso más adelante y examinar la naturaleza de una *distribución muestral de diferencias*, esto es, una distribución de frecuencia de un gran número de *diferencias* entre medias muestrales aleatorias que se han extraído de una población dada.

Para ilustrar la distribución muestral de diferencias, volvamos sobre el trabajo de nuestro investigador social excéntrico cuya pasión por la extracción de muestras aleatorias lo ha llevado una vez más a continuar el proceso de muestreo más allá de los límites ordinarios. En lugar de tomar una sola muestra de 500 liberales y una sola muestra de 500 conservadores, toma 70 *pares* de tales muestras (70 muestras que contienen 500 conservadores y 70 muestras con 500 liberales *cada una*). O sea que, cada vez que extrae aleatoriamente 500 conservadores, extrae también 500 liberales.

Habiendo tomado sus muestras, nuestro investigador social excéntrico interroga a todos y cada uno de los miembros de la muestra (1 000 × 70 = 70 000 personas)

* N. del E. Término utilizado para denotar la cualidad de mostrarse poco estricto con los hijos.

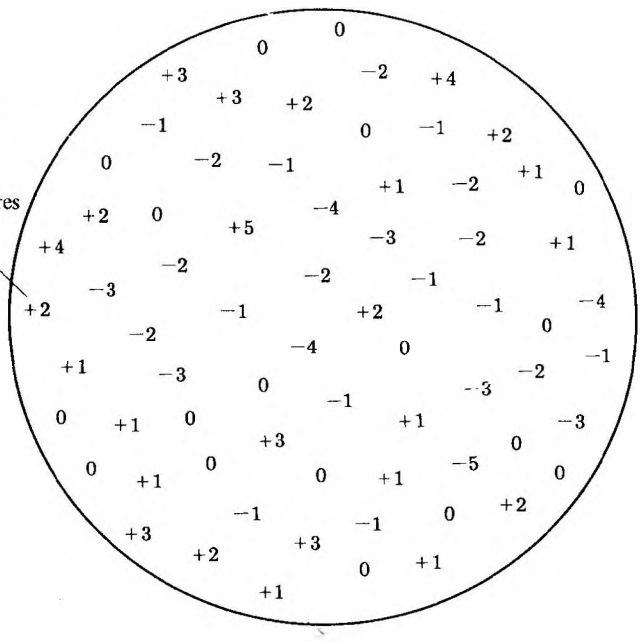


FIGURA 8.2 Setenta puntajes de diferencia entre medias que representan diferencias de permisibilidad entre muestras liberales y conservadoras tomadas aleatoriamente de una población hipotética

acerca de sus métodos de crianza de los niños y presenta un puntaje medio de permisibilidad para cada una de las muestras liberales y conservadoras. Además, obtiene un dato de diferencia entre las medias restando el puntaje medio conservador del puntaje medio liberal por cada par de muestras. Por ejemplo, si el puntaje medio de permisibilidad de los liberales es de 7,0 y el puntaje medio de los conservadores es de 6,0, entonces el puntaje de diferencia sería +1,0; igualmente, si el puntaje medio liberal es de 5,0 y el puntaje medio conservador es de 8,0, la diferencia sería −3,0. Obviamente, mientras mayor es el puntaje de diferencia, más difieren las dos muestras con respecto a la característica que se está investigando. Nótese que siempre restamos la segunda media muestral de la primera (en el presente caso restamos los puntajes medios conservadores de los puntajes medios de los liberales). Los 70 puntajes de diferencia entre las medias obtenidas por nuestro investigador social excéntrico se ilustran en la Figura 8.2.

Supongamos que sabemos que las poblaciones de conservadores y liberales realmente no difieren en absoluto con respecto a la permisibilidad en los métodos de crianza de los niños. Digamos que $\mu = 5{,}0$ en ambas poblaciones. Si suponemos que la hipótesis nula es correcta y que los liberales y los conservadores son idénticos en este aspecto, podemos usar las 70 diferencias entre las medias obtenidas por nuestro excéntrico investigador social para ilustrar la distribución muestral de diferencias. Esto es cierto porque la distribución muestral de diferencias supone que todos los pares de muestras difieren sólo en virtud del error de muestreo y no en función de verdaderas diferencias poblacionales.

TABLA 8.1 Distribución muestral de diferencias para 70 pares de muestras aleatorias

Diferencia entre medias[a]	*f*
+5	1
+4	2
+3	5
+2	7
+1	10
0	18
−1	10
−2	8
−3	5
−4	3
−5	1
	$N = 70$

[a]Estos puntajes de diferencia incluyen valores fraccionarios (por ejemplo, −5 incluye los valores desde −5,0 hasta +5,9).

Las 70 diferencias medias de la Figura 8.2 se han ordenado como una distribución muestral de diferencias de medias en la Tabla 8.1. Como los puntajes de otros tipos de distribuciones de frecuencia, éstos se han ordenado en forma decreciente mientras que la frecuencia en que ocurre se indica en una columna adyacente.

Para describir mejor las propiedades claves de una distribución muestral de diferencias, los datos de la Tabla 8.1 se han presentado gráficamente en la Figura 8.3. Tal como allí se ilustra, vemos que *la distribución muestral de diferencias entre medias muestrales se aproxima a una curva normal cuya media ("media de diferencias") es cero.*[1] Esto es lógico porque las diferencias positivas y negativas de las medias de la distribución tienden a cancelarse unas a otras (por cada valor negativo tiende a haber un valor positivo a igual distancia de la media).

Como curva normal, la mayoría de las diferencias entre medias muestrales de esta distribución cae cerca de cero —su punto más cercano al centro; hay relativamente pocas diferencias entre medias con valores extremos en una u otra dirección de la media de diferencias. Esto es de esperarse ya que la distribución de diferencias completa es un producto del error de muestreo más que de diferencias poblacionales reales entre conservadores y liberales. En otras palabras, si la diferencia media real entre las poblaciones de conservadores y liberales es cero, esperamos también que la media de la distribución muestral de diferencias sea cero.

CONTRASTACION DE LAS HIPOTESIS CON LA DISTRIBUCION DE DIFERENCIAS

En capítulos anteriores aprendimos a hacer afirmaciones de probabilidad con respecto a la frecuencia con que ocurren tanto los puntajes crudos como las medias muestrales. En el presente caso buscamos hacer afirmaciones de probabilidad acerca de los puntajes de diferencia en la distribución muestral de diferencias entre medias. Como se señaló anteriormente, esta distribución muestral toma la forma de la curva normal y, por lo tanto, puede considerarse como una distribución de probabilidad.

[1] Esto supone que hemos extraído grandes muestras aleatorias de una población dada de puntajes crudos.

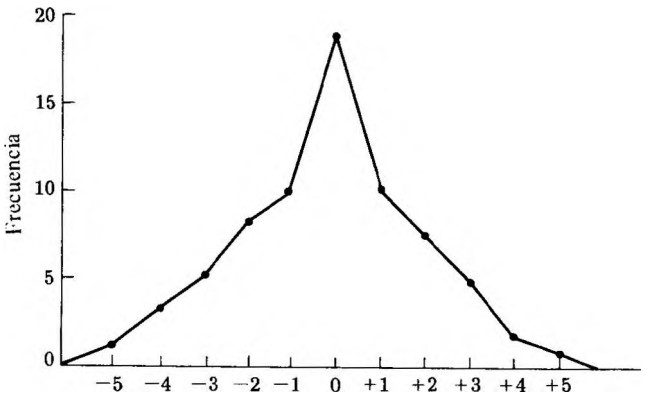


FIGURA 8.3 Polígono de frecuencia de la distribución muestral de diferencias de la Tabla 8.1

Podemos decir que la probabilidad disminuye a medida que nos alejamos más y más de la media de diferencias (cero). Más específicamente, como se ilustra en la Figura 8.4, vemos que el 68,26 por ciento de las diferencias entre medias caen entre −1 DE y +1 DE de cero. En términos de probabilidad, esto indica que P = 0,68 de que cualquier diferencia entre medias muestrales caiga dentro de este intervalo. De manera similar, podemos decir que la probabilidad es aproximadamente 0,95 (95 oportunidades entre 100) de que cualquier diferencia entre medias muestrales caiga entre −2 DE y + 2 DE de una diferencia media de cero, y así sucesivamente.

La distribución muestral de diferencias proporciona una base sólida para comprobar hipótesis acerca de la diferencia de media entre dos muestras aleatorias. Supongamos, por ejemplo, que una muestra de 100 liberales tiene un puntaje medio de permisibilidad de 7, mientras que una muestra de 100 conservadores tiene un puntaje medio de permisibilidad de 2. El razonamiento es así: si nuestra diferencia entre medias obtenida de 5 (7 − 2 = 5) está tan lejos de una diferencia de cero que sólo tiene una pequeña *probabilidad* de ocurrir en la distribución muestral de diferencias, rechazamos la hipótesis nula, que como antes dijimos es la hipótesis que establece que la diferencia obtenida es un resultado del error de muestreo. Si por

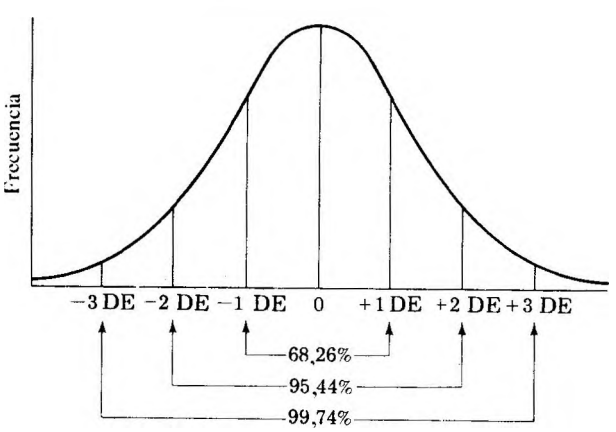


FIGURA 8.4 La distribución muestral de diferencias como una distribución de probabilidad

otra parte nuestra diferencia de medias muestrales cae tan cerca de cero que la *probabilidad* de que ocurra es grande, ~~debemos aceptar la hipótesis nula~~ y tratar nuestra diferencia obtenida como un resultado del error de muestreo.

Por lo tanto, buscamos determinar qué tan lejos está muestra diferencia, entre las medias, obtenida (en este caso 5) de una diferencia media de cero. Al hacerlo debemos convertir primero nuestra diferencia obtenida a unidades de desviación estándar.

Recordemos que convertimos los *puntajes crudos** a unidades de desviación estándar por la fórmula.

$$z = \frac{X - \bar{X}}{\sigma}$$

donde

X = un puntaje crudo
$\bar{X}$ = la media de la distribución de puntajes crudos
σ = la desviación estándar de la distribución de puntajes crudos

Igualmente, convertimos los *puntajes medios* de una distribución de medias muestrales a unidades de desviación estándar por la fórmula

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}}$$

donde

$\bar{X}$ = una media muestral
μ = la media poblacional (media de medias)
$\sigma_{\bar{x}}$ = el error estándar de la media (estimación de la desviación estándar de la distribución de medias)

En el presente contexto buscamos, de un modo similar, traducir nuestra diferencia entre medias muestrales (+5) a unidades de desviación estándar por la fórmula

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{\sigma_{\text{dif}}}$$

donde

* N. de E. "no procesados."

$\bar{X}_1$ = La media de la primera muestra
$\bar{X}_2$ = la media de la segunda muestra
"0" = cero, el valor de la media de la distribución muestral de diferencias (suponemos que $\mu_1 - \mu_2 = 0$)
σ_{dif} = la desviación estándar de la distribución muestral de diferencias

Debido a que siempre se supone que el valor de la media de la distribución de diferencias es cero, podemos desprendernos de él. en la fórmula del puntaje z, sin alterar nuestro resultado. Por lo tanto,

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{dif}}$$

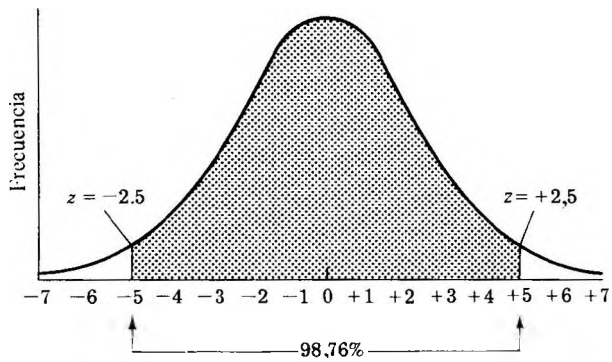
Con respecto a la permisibilidad que existe entre los liberales y los conservadores, debemos traducir primero nuestra diferencia entre medias obtenidas a su puntaje z equivalente . Si la desviación estándar de la distribución muestral de diferencias (σ_{dif}) es 2, obtenemos el siguiente puntaje z:

$$z = \frac{7 - 2}{2}$$
$$= \frac{5}{2}$$
$$= +2{,}5$$

Así, una diferencia de medias de 5 entre los liberales y los conservadores cae a 2,5 desviaciones estándar de una diferencia media de cero en la distribución de diferencias.

Nos preguntamos: *¿Qué probabilidad hay de que una diferencia de 5 o más, entre medias muestrales, pueda suceder estrictamente con base en el error de muestreo?* Acudiendo a la Tabla B, al final del texto, vemos que z = 2,5 representa el 49,38 por ciento de la distribución *en una u otra dirección* de la media de cero. O sea que el 98,76 por ciento (49,38% + 49,38+ = 98,76%) de las diferencias entre medias muestrales están entre cero y una diferencia media de 5 *en ambas direcciones* de cero, más *y* menos (ver Figura 8.5). En términos de probabilidad, esto indica que P = 0,99 (99 oportunidades entre 100) de que una diferencia entre medias caiga entre −5 y +5. Restando de 100 por ciento (100% − 98,76% = 1,24%), encontramos que P = 0,01 (redondeado) de que una diferencia media de 5 (o mayor de 5) entre las muestras, pueda ocurrir estrictamente con base en el error de muestreo. Esto es, que una diferencia media de 5 o más ocurre por error de muestreo (y por lo tanto aparece en la distribución muestral) *sólo una vez* en cada 100 diferencias entre medias. Sabiendo esto, ¿no pensaríamos en rechazar la hipótesis nula y aceptar la hipótesis de investigación de que una diferencia poblacional existe realmente entre conservadores y liberales con respecto a la permisibilidad en la crianza de los niños?

FIGURA 8.5 Representación gráfica del porcentaje del área total en la distribución de diferencias entre $z = -2{,}5$ y $z = +2{,}5$



Una oportunidad entre 100 representa una probabilidad bastante buena ¿no es verdad?

Dada la situación anterior, la mayoría de nosotros elegiría rechazar la hipótesis nula a pesar de que nos podríamos equivocar al hacerlo (no olvidemos que aún queda 1 oportunidad entre 100). Sin embargo, la decisión no es siempre tan clara. Supongamos, por ejemplo, que nos enteramos de que nuestra diferencia media sucede por error de muestreo 10 ($P = 0{,}10$), 15 ($P = 0{,}15$), o 20 ($P = 0{,}20$) veces de 100. ¿Rechazamos aún la hipótesis nula? o ¿"vamos a lo seguro" y atribuimos nuestra diferencia obtenida al error de muestreo?

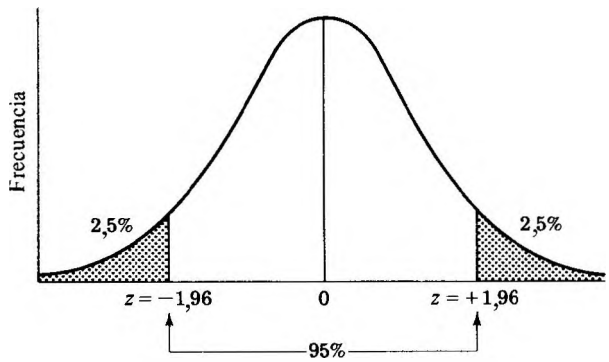
Necesitamos un punto de referencia consistente para decidir si una diferencia entre dos medias muestrales es tan grande que ya no puede atribuírsele al error de muestreo. Necesitamos un método para determinar cuánto es *estadísticamente significativo* nuestro resultado.

NIVELES DE CONFIANZA

Para establecer si nuestra diferencia muestral obtenida es estadísticamente significativa —resultado de una diferencia poblacional real y no sólo del error de muestreo— se acostumbra establecer un *nivel de confianza* (también conocido como *nivel de significancia*), nivel de probabilidad en el cual se puede rechazar a la hipótesis nula y se puede aceptar con confianza la hipótesis de investigación. Por lo tanto, decidimos rechazar la hipótesis nula si la probabilidad es muy pequeña (por ejemplo, sólo 5 oportunidades entre 100) de que la diferencia muestral sea un producto del error de muestreo.

Es un asunto convencional utilizar el *nivel de confianza de 0,05*. O sea que estamos dispuestos a rechazar la hipótesis nula si una diferencia muestral obtenida ocurre casualmente sólo 5 veces o menos entre 100 (5 por ciento). El nivel de confianza de 0,05 se ha representado gráficamente en la Figura 8.6. Como se muestra allí, el nivel de confianza de 0,05 se encuentra en las pequeñas áreas de las "colas" de la distribución de diferencias de medias. Estas son las áreas bajo la curva que representan una distancia de más o menos 1,96 desviaciones estándar de una diferencia media de cero.

FIGURA 8.6
Representación gráfica del nivel de confianza de 0,05



Para comprender mejor por qué este punto en particular de la distribución muestral representa el nivel de confianza de 0,05 podríamos volver a la Tabla B, al final del texto, para determinar el porcentaje de frecuencia total asociado con 1,96 desviaciones estándar de la media. Vemos que 1,96 desviaciones estándar *en una u otra* dirección representan el 2,5% de las diferencias entre medias muestrales (50% − 47,5% = 2,5%). En otras palabras, el 95 por ciento de las diferencias muestrales cae entre −1.96 DE y + 1.96 DE de una diferencia media de cero; sólo el 5 por ciento cae en este punto o más allá de él (2,5% + 2,5% = 5%).

Los niveles de confianza pueden establecerse para cualquier grado de probabilidad. Por ejemplo, un nivel de confianza más estricto es el *nivel de confianza de 0,01*, por medio del cual se rechaza la hipótesis nula si solamente hay 1 oportunidad entre 100 de que la diferencia muestral obtenida pueda ocurrir por error de muestreo (1 por ciento). El nivel de confianza de 0,01 está representado por el área que está a 2,58 desviaciones estándar en ambas direcciones de una diferencia de media de cero.

Los niveles de confianza no nos dan una afirmación *absoluta* acerca de la corrección de la hipótesis nula. Siempre que decidamos rechazar la hipótesis nula a un cierto nivel de confianza, nos abriremos a la posibilidad de tomar la decisión equivocada. Rechazar la hipótesis nula cuando se debería aceptar se conoce como el *error alpha (o error tipo I)*. La probabilidad de cometer el error alpha sólo puede surgir cuando rechazamos la hipótesis nula y varía de acuerdo con el nivel de confianza que escojamos. Por ejemplo, si rechazamos la hipótesis nula al nivel de confianza de 0,05 y concluimos que los conservadores realmente difieren de los liberales en términos de sus métodos de crianza de los niños, entonces hay 5 oportunidades entre 100 de que nos equivoquemos. En otras palabras, P = 0,05 de que hayamos cometido el error alpha y de que los conservadores no difieran realmente de los liberales. Igualmente, si escogemos el nivel de confianza de 0,01 sólo existe una oportunidad entre 100 (P = 0,01) de tomar la decisión equivocada con respecto a la diferencia entre liberales y conservadores. Obviamente, mientras más riguroso sea nuestro nivel de confianza (mientras más cerca de la cola se encuentre), menos probabilidades tendremos de cometer el error alpha. Tomando un ejemplo extremo, establecer un nivel de confianza de 0,001 produce un riesgo de que el error alpha ocurra solamente una vez entre mil.

Sin embargo, mientras más cerca de la cola de la curva caiga nuestro nivel de confianza, mayor será el riesgo de cometer otra clase de error, conocido como el *error beta (o error tipo II),* error en el que se cae al aceptar la hipótesis nula cuando debió haber sido rechazada. El error beta indica que nuestra hipótesis de investigación puede ser aún correcta, a pesar de la decisión de rechazarla y de aceptar la hipótesis nula. Un método para reducir el riesgo de cometer el error beta es aumentar el tamaño de las muestras de manera que sea más probable que quede representada una diferencia poblacional real.

Nunca podemos estar seguros de que no hemos tomado una decisión equivocada con respecto a la hipótesis nula, ya que examinamos solamente una muestra y no la población entera. Mientras no tengamos conocimiento de los verdaderos valores poblacionales, correremos el riesgo de cometer un error tipo I o tipo II, dependiendo de nuestra decisión. Este es el riesgo de la toma de decisiones estadísticas que el investigador social debe estar dispuesto a asumir.

EL ERROR ESTANDAR DE LA DIFERENCIA

Nunca podemos tener conocimientos de fuentes directas acerca de la desviación estándar de la distribución de diferencias de medias y, al igual que en el caso de la distribución muestral de medias (Capítulo 7), resultaría un esfuerzo mayor el extraer realmente un gran número de pares de muestras para poder calcularla. Sin embargo, esta desviación estándar desempeña un importante papel en el método que se sigue para contrastar hipótesis acerca de las diferencias entre las medias y, por lo tanto, no puede pasarse por alto.

Afortunadamente, tenemos un método sencillo por medio del cual puede estimarse con exactitud la desviación estándar de la distribución de diferencias con base en las dos muestras que hemos extraído realmente. A esta estimación de la desviación estándar de la distribución muestral de diferencias la llamaremos *error estándar de la diferencia,* el cual se simboliza con σ_{dif}, por fórmula,

$$\sigma_{\text{dif}} = \sqrt{\sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2}$$

donde

σ_{dif} = el error estándar de la diferencia
$\sigma_{\bar{X}_1}$ = el error estándar de la primera media muestral
$\sigma_{\bar{X}_2}$ = el error estándar de la segunda media muestral

Supongamos, con fines ilustrativos, que hemos obtenido los siguientes datos de una muestra de 50 liberales y una muestra de 50 conservadores:

Liberales (N = 50)	*Conservadores (N = 50)*
$\bar{X} = 7.0$	$\bar{X} = 6.0$
$s = 2{,}0$	$s = 1{,}5$

Para calcular el error estándar de la diferencia, debemos encontrar primero el error estándar para cada media muestral. Recordemos que esto se hace como sigue, a partir de la desviación estándar para cada muestra (ver Capítulo 7):

$$\sigma_{\bar{X}_1} = \frac{s_1}{\sqrt{N_1 - 1}} \qquad \sigma_{\bar{X}_2} = \frac{s_2}{\sqrt{N_2 - 1}}$$

$$= \frac{2{,}0}{\sqrt{50 - 1}} \qquad = \frac{1{,}5}{\sqrt{50 - 1}}$$

$$= \frac{2{,}0}{7{,}0} \qquad = \frac{1{,}5}{7{,}0}$$

$$= 0{,}29 \qquad = 0{,}21$$

Una vez que conocemos $\sigma_{\bar{X}}$ para cada media muestral, podemos obtener σ_{dif} como sigue:

$$\sigma_{\text{dif}} = \sqrt{\sigma_{\bar{X}_1}{}^2 + \sigma_{\bar{X}_2}{}^2}$$

$$= \sqrt{0{,}29^2 + 0{,}21^2}$$

$$= \sqrt{0{,}08 + 0{,}04}$$

$$= \sqrt{0{,}12}$$

$$= 0{,}35$$

El error estándar de la diferencia (nuestra estimación de la desviación estándar de la distribución de diferencias) resulta ser 0,35. Si estamos comprobando la diferencia entre los liberales (X = 7,0) y los conservadores (X = 6,0) con respecto a la permisibilidad, usaríamos nuestro resultado para convertir la diferencia entre medias muestrales obtenida a su puntaje z equivalente:

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\text{dif}}}$$

$$= \frac{7 - 6}{0{,}35}$$

$$= \frac{1}{0{,}35}$$

$$= 2{,}86$$

Remitiéndonos a la Tabla B, al final del libro, vemos que un puntaje z de 2,86 equivale exactamente al 49,79 por ciento de las diferencias de medias *a uno u otro* lado o al 99,58 por ciento de las diferencias de medias a *ambos* lados de una diferencia de media de cero (49,79% + 49,79% = 99,58%). Si restamos esta suma de 100 por ciento encontramos que menos del 1% (0,42%) de los puntajes de diferencias de medias tienen un valor de 1 o mayor de 1. Por lo tanto, P es menor a 0,01 de obtener una diferencia de media de 1 con base en el error de muestreo. Podemos rechazar la hipótesis nula ya sea al nivel de confianza de 0,05 o de 0,01, cualquiera que sea el que hayamos establecido para nuestro estudio.

Una Ilustración

Para proporcionar una ilustración minuciosa del procedimiento anterior, para comprobar una diferencia entre dos medias muestrales, supongamos que quisimos contrastar la hipótesis nula al nivel de confianza de 0,05 que planteaba que las mujeres no son ni más ni menos etnocéntricas que los hombres ($\mu_1 = \mu_2$). Nuestra hipótesis de investigación establece que las mujeres difieren de los hombres con respecto al etnocentrismo[2] ($\mu_1 \neq \mu_2$). Para comprobar esta hipótesis, digamos que le dimos una medida de etnocentrismo (por ejemplo, la escala de etnocentrismo) a una muestra aleatoria de 35 mujeres y a una muestra aleatoria de 35 hombres y obtuvimos los siguientes puntajes de etnocentrismo para cada muestra (X = datos que van desde 1, representando bajo etnocentrismo, hasta 5, representando alto etnocentrismo):

Hombres ($N = 35$)		*Mujeres* ($N = 35$)	
X_1	X^2	X_2	X^2
1	1	1	1
1	1	1	1
1	1	1	1
1	1	2	4
2	4	1	1
1	1	1	1
1	1	1	1
3	9	3	9
3	9	1	1
1	1	2	4
2	4	4	16
1	1	1	1
2	4	1	1
1	1	1	1
1	1	1	1
1	1	5	25
1	1	1	1
2	4	2	4
4	16	2	4
5	25	1	1
1	1	1	1
1	1	1	1
2	4	1	1
1	1	2	4
2	4	3	9
1	1	1	1
2	4	1	1
1	1	1	1
1	1	2	4
1	1	2	4
1	1	2	4
3	9	1	1
3	9	1	1
1	1	1	1
4	16	1	1
$\Sigma X = 60$	$\Sigma X^2 = 142$	$\Sigma X = 54$	$\Sigma X^2 = 114$

[2] "Etnocentrismo" se refiere a la tendencia a evaluar a todos los grupos de personas usando nuestras propias normas culturales.

PASO 1: Encontrar la media para cada muestra

$$\overline{X}_1 = \frac{\Sigma X_1}{N} \qquad \overline{X}_2 = \frac{\Sigma X_2}{N}$$
$$= \frac{60}{35} \qquad = \frac{54}{35}$$
$$= 1{,}71 \qquad = 1{,}54$$

PASO 2: Encontrar la desviación estándar para cada muestra

$$s_1 = \sqrt{\frac{\Sigma X^2}{N} - \overline{X}^2} \qquad s_2 = \sqrt{\frac{\Sigma X^2}{N} - \overline{X}^2}$$
$$= \sqrt{\frac{142}{35} - 2{,}92} \qquad = \sqrt{\frac{114}{35} - 2{,}37}$$
$$= \sqrt{4{,}06 - 2{,}92} \qquad = \sqrt{3.26 - 2{,}37}$$
$$= \sqrt{1{,}14} \qquad = \sqrt{0{,}89}$$
$$= 1{,}07 \qquad = 0{,}94$$

PASO 3: Encontrar el error estándar de cada media

$$\sigma_{\overline{X}_1} = \frac{s_1}{\sqrt{N-1}} \qquad \sigma_{\overline{X}_2} = \frac{s_2}{\sqrt{N-1}}$$
$$= \frac{1{,}07}{\sqrt{34}} \qquad = \frac{0{,}94}{\sqrt{34}}$$
$$= \frac{1{,}07}{5{,}83} \qquad = \frac{0{,}94}{5{,}83}$$
$$= 0{,}18 \qquad = 0{,}16$$

PASO 4: Encontrar el error estándar de la diferencia

$$\sigma_{\text{dif}} = \sqrt{\sigma_{\overline{X}_1}^2 + \sigma_{\overline{X}_2}^2}$$
$$= \sqrt{(0{,}18)^2 + (0{,}16)^2}$$
$$= \sqrt{0{,}03 + 0{,}03}$$
$$= \sqrt{0{,}06}$$
$$= 0{,}25$$

PASO 5: Convertir la diferencia entre medias muestrales a unidades de error estándar de la diferencia

$$z = \frac{\overline{X}_1 - \overline{X}_2}{\sigma_{\text{dif}}}$$
$$= \frac{1{,}71 - 1{,}54}{0{,}25}$$
$$= \frac{0{,}17}{0{,}25}$$
$$= 0{,}68$$

PASO 6: Encontrar el porcentaje del área total bajo la curva normal entre z y una diferencia media de cero (ver Tabla B)

25,17%
+ 25,17%
50,34%

PASO 7: Restar de 100% para encontrar el porcentaje del área total asociado con la diferencia entre medias muestrales obtenida

100,00%
− 50,34%
49,66%

Del resultado del Paso 7 vemos que P = 0,50 (redondeado) de obtener una diferencia media de 0,17 (1,71 − 1,54) por error de muestreo. Como resultado debemos aceptar la hipótesis nula y rechazar la hipótesis de investigación al nivel de confianza de 0,05. La probabilidad de que ocurra nuestra diferencia entre medias obtenida entre hombres y mujeres es mayor a 5 de 100. Para ser exactos, ¡es igual a 50 de 100! Conclusión: Los datos de nuestra muestra no indican que las mujeres sean ni más ni menos etnocéntricas que los hombres.

COMPARACIONES ENTRE MUESTRAS PEQUEÑAS

Los investigadores sociales trabajan frecuentemente con muestras que contienen un pequeño número de entrevistados o casos (por ejemplo, menos de 30). Mientras que puede ser conveniente, si no necesario, obtener resultados basados en muestras de pequeño tamaño, éstos pueden ser seriamente engañosos si se interpretan de acuerdo al área señalada bajo la curva normal en la Tabla B. Esto resulta cierto ya que la distribución muestral de diferencias toma la forma de la curva normal sólo si las muestras que van a constituirla son grandes. Un investigador social que trabaja con 5, 10 o 20 entrevistados en cada muestra no puede encontrarse con esta suposición. Como resultado no puede usar puntajes z basados en la distribución normal.

Para compensar estadísticamente este alejamiento de la normalidad, en la distribución de diferencias, obtenemos en su lugar lo que se conoce comúnmente como la razón t. Al igual que el puntaje z, la razón t puede usarse para convertir una diferencia entre medias muestrales a unidades de error estándar de la diferencia. También de la misma manera en que se llega al puntaje z obtenemos una razón t, tomando la diferencia entre nuestras medias muestrales y dividiéndolas por nuestro error estándar de la diferencia. Por fórmula,

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sigma_{\text{dif}}}$$

donde

$\overline{X}_1$ = la media de la primera muestra
$\overline{X}_2$ = la media de la segunda muestra
σ_{dif} = el error estándar de la diferencia

Como se muestra arriba, la fórmula de la razón *t* es idéntica a la fórmula para el puntaje *z* que aprendimos anteriormente. Sin embargo, a diferencia de un puntaje *z*, la razón *t* debe interpretarse con referencia a los *grados de libertad*[3] (gl), que varían directamente con el tamaño de la muestra y van a determinar la forma de la distribución muestral de diferencias. Mientras mayor sea el tamaño de la muestra, mayores serán nuestros grados de libertad. Mientras mayores sean nuestros grados de libertad, más se acercará la distribución de diferencias a una aproximación de la curva normal. Con infinitos grados de libertad, nuestra razón *t* se convierte en puntaje *z* y de ese modo podemos emplear la Tabla B para interpretar nuestro resultado.

Pero, ¿qué sucede cuando trabajamos con muestras pequeñas? ¿cómo sorteamos el asunto para encontrar grados de libertad e interpretar nuestra razón *t*? Para una razón *t* que representa dos medias muestrales, el número de grados de libertad puede encontrarse por la fórmula

$$gl = N_1 + N_2 - 2$$

donde

N_1 = el tamaño de la primera muestra
N_2 = el tamaño de la segunda muestra

Por lo tanto, si estamos comparando una muestra de 6 liberales y 8 conservadores, nuestros grados de libertad serán $6 + 8 - 2 = 12$.

Podemos interpretar cualquier razón *t* que obtengamos con la ayuda de la Tabla C, al final del libro, y del número de grados de libertad que hemos calculado. La Tabla C proporciona los valores de *t* que se requieren para rechazar la hipótesis nula a los niveles de confianza de 0,05 y 0,01 para varios grados de libertad. Volviendo a la Tabla C, vemos una columna marcada gl (grados de libertad) y una lista de valores *t* para cada grado de libertad a los niveles de confianza de 0,05 y 0,01. Como veremos, estos valores *t* pueden usarse para interpretar la razón *t* que hemos calculado.

Una ilustración de una comparación entre muestras pequeñas

Para ilustrar el uso de la razón de *t*, de los grados de libertad y de la Tabla C para comprobar una diferencia de medias entre muestras pequeñas, pensemos en la siguiente situación de investigación: Un investigador social busca comprobar la

[3] Grados de libertad se refiere técnicamente a la libertad de variación entre un conjunto de puntaje. Si tenemos una muestra de 6 puntajes, entonces 5 son libres de variar mientras que sólo uno es de valor fijo. Por lo tanto, en una sola muestra de 6 entrevistados, $gl = N - 1$ o 5.

hipótesis de que el comportamiento caritativo varía según si la donación se hace anónimamente o si se da a conocer la identidad del donante. Por lo tanto,

Hipótesis nula: ($\mu_1 = \mu_2$) *El grado de comportamiento caritativo no difiere si la donación es anónima o no.*

Hipótesis de investigación: ($\mu_1 \neq \mu_2$) *El grado de comportamiento caritativo difiere si la donación se hace anónimamente o no.*

Para probar esta hipótesis el investigador estipula el nivel de confianza de 0,05; esto es, escoge inicialmente rechazar la hipótesis nula sólo si resulta que hay 5 oportunidades entre 100 de que la diferencia entre medias muestrales obtenida sea producto del error de muestreo. Habiendo establecido este criterio de significancia, él obtiene dos muestras aleatorias de donantes potenciales. A todos los miembros de ambas muestras les pide donaciones en dinero para distribuirlo entre los sobrevivientes de un gran terremoto. A los 6 miembros de la primera muestra les asegura el anonimato completo; a los 6 miembros de la segunda muestra les promete colocar los nombres de los donantes en un lugar público visible. Por tanto, tenemos las condiciones experimentales de *anonimato* contra *identidad conocida.*

A continuación se enumeran las cantidades de dinero donadas por los miembros de ambas muestras:

Anonimato (N = 6)		*Identidad conocida (N = 6)*	
X_1	X_1^2	X_2	X_2^2
\$1	1	\$3	9
2	4	5	25
1	1	5	25
1	1	5	25
2	4	4	16
1	1	5	25
$\Sigma X_1 = 8$	$\Sigma X_1^2 = 12$	$\Sigma X_2 = 27$	$\Sigma X_2^2 = 125$

Vemos que los 6 miembros de la muestra que quedó en el anonimato dieron \$8 mientras que los 6 miembros de la muestra de identidad conocida dieron \$27. El siguiente procedimiento puede usarse paso a paso para probar la significancia estadística de la diferencia obtenida.

PASO 1: Encontrar la media de cada muestra

$$\overline{X}_1 = \frac{\Sigma X_1}{N} \qquad \overline{X}_2 = \frac{\Sigma X_2}{N}$$

$$= \frac{8}{6} \qquad = \frac{27}{6}$$

$$= \$1{,}33 \qquad = \$4{,}50$$

PASO 2: Encontrar la desviación estándar de cada muestra

$$s_1 = \sqrt{\frac{\Sigma X_1^2}{N_1} - \overline{X}_1^2} \qquad s_2 = \sqrt{\frac{\Sigma X_2^2}{N} - \overline{X}_2^2}$$

$$= \sqrt{\frac{12}{6} - (1{,}33)^2} \qquad = \sqrt{\frac{125}{6} - (4{,}50)^2}$$

$$= \sqrt{2{,}00 - 1{,}77} \qquad = \sqrt{20{,}83 - 20{,}25}$$

$$= \sqrt{0{,}23} \qquad = \sqrt{58}$$

$$= 0{,}48 \qquad = 0{,}76$$

PASO 3: Encontrar el error estándar de cada media

$$\sigma_{\overline{X}_1} = \frac{s_1}{\sqrt{N_1 - 1}} \qquad \sigma_{\overline{X}_2} = \frac{s_2}{\sqrt{N_2 - 1}}$$

$$= \frac{0{,}48}{\sqrt{5}} \qquad = \frac{0{,}76}{\sqrt{5}}$$

$$= \frac{0{,}48}{2{,}24} \qquad = \frac{0{,}76}{2{,}24}$$

$$= 0{,}21 \qquad = 0{,}34$$

PASO 4: Encontrar el error estándar de la diferencia

$$\sigma_{\text{dif}} = \sqrt{\sigma_{\overline{X}_1}{}^2 + \sigma_{\overline{X}_2}{}^2}$$

$$= \sqrt{(0{,}21)^2 + (0{,}34)^2}$$

$$= \sqrt{0{,}04 + 0{,}12}$$

$$= \sqrt{0{,}16}$$

$$= 0{,}40$$

PASO 5: Convertir la diferencia entre medias muestrales a unidades de error estándar de la diferencia

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sigma_{\text{dif}}}$$

$$= \frac{1{,}33 - 4{,}50}{0{,}40}$$

$$= -\frac{3{,}17}{0{,}40}$$

$$= -7{,}93$$

PASO 6: Buscar el número de grados de libertad

$$\text{gl} = N_1 + N_2 - 2$$

$$= 6 + 6 - 2$$

$$= 10$$

PASO 7: Comparar la razón *t* obtenida con la razón *t* apropiada de la Tabla C

razón *t* obtenida = 7,93
razón *t* de la tabla = 2,228
gl = 10
P = 0,05

Como se ve en el Paso 7, para poder rechazar la hipótesis nula al nivel de confianza de 0,05 con 10 grados de libertad, nuestra razón *t* calculada debe ser 2,228 o más. En el presente caso hemos obtenido una razón *t* de 7,93. Por lo tanto, rechazamos la hipótesis nula y aceptamos la hipótesis de investigación. El grado de comportamiento caritativo realmente varía de acuerdo a si la donación se hace anónimamente o bien si se da a conocer la identidad del donante. Más específicamente, la condición de **"identidad conocida" produce significativamente más caridad (X_2 = \$4,50) que la condición de "anonimato" ($\bar{X}_1$ = \$1,33).**

COMPARACIONES ENTRE MUESTRAS DE DISTINTO TAMAÑO

Hasta ahora hemos trabajado con muestras que contienen exactamente el mismo número de entrevistados o casos. Por ejemplo, en la ilustración anterior cada muestra contenía 6 entrevistados. Sin embargo, cuando realmente salimos a realizar la investigación encontramos que, con frecuencia, nuestras muestras difieren en tamaño. Así podemos tener una muestra de 50 liberales y 64 conservadores, una muestra de 15 hombres y 22 mujeres. Para hacer comparaciones entre muestras de distinto tamaño debemos encontrar una forma de dar el *peso* apropiado a la influencia relativa de cada muestra. En el caso de $\bar{X}$ esto se hace automáticamente, ya que siempre dividimos ΣX entre N. Este no es el caso para el error estándar de la diferencia: cada desviación estándar de la muestra en que se basa σ_{dif} contribuye igualmente a la fórmula que aprendimos anteriormente, aunque existan diferencias grandes e importantes en el tamaño de las muestras.

Este problema puede superarse utilizando una fórmula para el error estándar de la diferencia, en la cual la influencia relativa de cada desviación estándar puede ser ponderada en términos del tamaño de su muestra. Tal fórmula se presenta a continuación:

$$\sigma_{dif} = \sqrt{\left(\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$$

donde

s_1 = la desviación estándar de la primera muestra
s_2 = la desviación estándar de la segunda muestra
N_1 = el número total en la primera muestra
N_2 = el número total en la segunda muestra

Para ilustrar el procedimiento que se sigue para comparar muestras de distinto tamaño, pensemos en la hipótesis de que los niños negros y blancos de cierto barrio difieren respecto a la tendencia hacia la criminalidad. En este caso,

Hipótesis nula: *Los niños negros y blancos no difieren respecto a su tendencia hacia la criminalidad.* ($\mu_1 = \mu_2$)

Hipótesis de investigación: *Los niños negros y blancos difieren respecto a su tendencia hacia la criminalidad.* ($\mu_1 \neq \mu_2$)

Para comprobar este hecho en el nivel de confianza de 0,05, imaginemos que cierto investigador administró una medida de "tendencia hacia la criminalidad" a una muestra aleatoria de 4 blancos y a una muestra aleatoria de 7 negros. Resultaron los siguientes puntajes de "tendencia hacia la criminalidad" (los datos van desde 1, que representa poca tendencia hacia la criminalidad, hasta 5, que representa una fuerte tendencia hacia la criminalidad):

Blancos (N = 4)		*Negros (N = 7)*	
X_1	X_1^2	X_2	X_2^2
1	1	4	16
2	4	1	1
1	1	1	1
3	9	1	1
$\Sigma X_1 = 7$	$\Sigma X_1^2 = 15$	2	4
		2	4
		1	1
		$\Sigma X_2 = 12$	$\Sigma X_2^2 = 28$

El procedimiento detallado para comprobar la hipótesis anterior puede ilustrarse como sigue:

PASO 1: Encontrar la media de cada muestra

$$\bar{X}_1 = \frac{\Sigma X_1}{N_1} \qquad \bar{X}_2 = \frac{\Sigma X_2}{N_2}$$

$$= \frac{7}{4} \qquad = \frac{12}{7}$$

$$= 1.75 \qquad = 1.71$$

PASO 2: Encontrar la desviación estándar de cada muestra

$$s_1 = \sqrt{\frac{\Sigma X_1^2}{N_1} - \bar{X}_1^2} \qquad s_2 = \sqrt{\frac{\Sigma X_2^2}{N_2} - \bar{X}_2^2}$$

$$= \sqrt{\frac{15}{4} - 3.06} \qquad = \sqrt{\frac{28}{7} - 2.92}$$

$$= \sqrt{3,75 - 3,06} \qquad = \sqrt{4,00 - 2,92}$$
$$= \sqrt{0,69} \qquad = \sqrt{1,08}$$
$$= 0,83 \qquad = 1,04$$

PASO 3: Encontrar el error estándar de la diferencia

$$\sigma_{dif} = \sqrt{\left(\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$$
$$= \sqrt{\left(\frac{4(0,83)^2 + 7(1.04)^2}{4 + 7 - 2}\right)\left(\frac{1}{4} + \frac{1}{7}\right)}$$
$$= \sqrt{\left(\frac{2.76 + 7,56}{9}\right)(0,25 + 0,14)}$$
$$= \sqrt{\left(\frac{10,32}{9}\right)(0,39)}$$
$$= \sqrt{(1,15)(0,39)}$$
$$= \sqrt{0,45}$$
$$= 0,67$$

PASO 4: Convertir la diferencia entre medias muestrales a unidades de error estándar de la diferencia

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sigma_{dif}}$$
$$= \frac{1,75 - 1,71}{0,67}$$
$$= \frac{0,04}{0,67}$$
$$= 0,06$$

PASO 5: Buscar el número de grados de libertad

$$gl = N_1 + N_2 - 2$$
$$= 4 + 7 - 2$$
$$= 9$$

PASO 6: Comparar la razón *t* obtenida, con la razón *t* apropiada de la Tabla C

razón *t* obtenida = 0,06
razón *t* de la tabla = 2,262
gl = 9
P = 0,05

Como se indica en el Paso 6, para rechazar la hipótesis nula. al nivel de confianza de

0,05 con 9 grados de libertad, nuestra razón *t* obtenida tendría que ser 2,262 o más. Como hemos calculado una razón *t* de sólo 0,06 debemos aceptar la hipótesis nula y rechazar la hipótesis de investigación. Nuestros resultados no respaldan el concepto de que los niños negros y blancos difieren respecto a su tendencia hacia la criminalidad.

COMPARACION DE LA MISMA MUESTRA MEDIDA DOS VECES

Hasta aquí hemos analizado las comparaciones que se hacen entre dos muestras que se han extraído independientemente (por ejemplo, hombres contra mujeres, negros contra blancos o liberales contra conservadores). Antes de dejar este tema presentaremos ahora una última variación de la comparación entre dos medias a la que nos referimos como un diseño de *antes-después* o de *panel:* es el caso de *una sola* muestra medida en dos puntos diferentes en el tiempo (tiempo 1 contra tiempo 2). Por ejemplo, un encuestador puede tratar de medir las reacciones que experimenta una sola muestra de niños tanto antes como después de ver cierto programa de televisión. Del mismo modo podríamos desear medir las diferencias de actitudes hacia un determinado candidato a un cargo público antes y después de su campaña.

Para dar una ilustración paso a paso de una comparación de *antes-después,* supongamos que varios individuos han sido obligados por el gobierno a reubicar sus hogares debido a la construcción de una carretera. Como investigadores sociales, nos interesa determinar el impacto que la reubicación residencial forzada tiene sobre los sentimientos de buena vecindad (esto es, sentimientos positivos hacia los vecinos del barrio, *pre*-reubicación, contra los sentimientos hacia los vecinos del barrio, *post*-reubicación). En este caso, entonces, μ_1 es el puntaje medio de buena vecindad en el tiempo 1 (*antes* de la reubicación) y μ^2 es el puntaje medio de buena vecindad en el tiempo 2 (*después* de la reubicación). Por lo tanto,

Hipótesis nula: ($\mu_1 = \mu_2$) *El grado de buena vecindad no difiere antes ni después de la reubicación.*

Hipótesis de investigación: ($\mu_1 \neq \mu_2$) *El grado de buena vecindad difiere antes y después de la reubicación.*

Para probar el impacto que causa la reubicación forzada sobre la buena vecindad, entrevistamos una muestra aleatoria de 6 individuos tanto antes como después de que se les obligó a mudarse. Nuestras entrevistas producen los siguientes puntajes de buena vecindad (los puntajes más altos de 1 a 4 indican mayor grado de buena vecindad):

Entrevistado	*Antes de mudarse* X_1	*Después de mudarse* X_2	*Diferencia* $X_1 - X_2 = D$	*(Diferencia)*^{2} D^2
Rosalba	2	1	1	1
Raúl	1	2	−1	1
Carolina	3	1	2	4
Lilia	3	1	2	4
Alberto	1	2	−1	1
Mario	4	1	3	9
	$\Sigma X_1 = 14$	$\Sigma X_2 = 8$		$\Sigma D^2 = 20$

Como se mostró anteriormente, hacer una comparación antes-después, concentra nuestra atención en la *diferencia* que hay entre el tiempo 1 y el tiempo 2; esto se refleja en la fórmula para obtener la desviación estándar (para la distribución de puntajes de diferencias antes-después:

$$s = \sqrt{\frac{\Sigma D^2}{N} - (\overline{X}_1 - \overline{X}_2)^2}$$

donde:

s = la desviación estándar de la distribución de puntajes de diferencias antes-después
D = el puntaje crudo "después", restado del puntaje crudo "antes"
N = el número de casos o entrevistados en la muestra

PASO 1: Encontrar la media para cada punto en el tiempo

$$\overline{X}_1 = \frac{\Sigma X_1}{N} \qquad \overline{X}_2 = \frac{\Sigma X_2}{N}$$
$$= \frac{14}{6} \qquad = \frac{8}{6}$$
$$= 2{,}33 \qquad = 1{,}33$$

PASO 2: Encontrar la desviación estándar para la diferencia entre el tiempo 1 y el tiempo 2

$$s = \sqrt{\frac{\Sigma D^2}{N} - (\overline{X}_1 - \overline{X}_2)^2}$$
$$= \sqrt{\frac{20}{6} - (2{,}33 - 1{,}33)^2}$$
$$= \sqrt{\frac{20}{6} - 1{,}00}$$
$$= \sqrt{3{,}33 - 1{,}00}$$
$$= \sqrt{2{,}33}$$
$$= 1{,}53$$

PASO 3: Encontrar el error estándar de la diferencia

$$\sigma_{dif} = \frac{s}{\sqrt{N-1}}$$
$$= \frac{1{,}53}{\sqrt{6-1}}$$
$$= \frac{1{,}53}{2{,}24}$$
$$= 0{,}68$$

PASO 4: Convertir la diferencia entre medias muestrales a unidades de error estándar de la diferencia

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sigma_{dif}}$$
$$= \frac{2{,}33 - 1{,}33}{0{,}68}$$
$$= \frac{1{,}00}{0{,}68}$$
$$= 1{,}47$$

PASO 5: Encontrar el número de grados de libertad

$$gl = N - 1$$
$$= 6 - 1$$
$$= 5$$

Nota: N se refiere al número total de casos, no al número de puntajes, para los cuales hay 2 por caso o entrevistado.

PASO 6: Comparar la razón t obtenida con la razón apropiada de la Tabla C

razón t obtenida = 1,47
razón t de la Tabla C = 2,571
gl = 5
P = 0,05

Para poder rechazar la hipótesis nula al nivel de confianza de 0,05 con 5 grados de libertad, debemos obtener una razón t calculada de 2,571. Ya que nuestra razón t es de sólo 1,47 —menor al valor requerido por la tabla— aceptamos la hipótesis nula y rechazamos la hipótesis de investigación. La diferencia muestral obtenida en lo que respecta a la buena vecindad antes y después de la reubicación era, en realidad, un resultado del error de muestreo.

REQUISITOS PARA EL USO DEL PUNTAJE z Y LA RAZON t

Como veremos a través del resto de este texto, cada prueba estadística debe utilizarse sólo si el investigador social ha tomado en cuenta por lo menos ciertos re-

quisitos, condiciones o suposiciones. El empleo inadecuado de una prueba puede confundir un problema y conducir al investigador a conclusiones erróneas. Como resultado, se deben tener muy presentes los siguientes requisitos al pensar en las características del puntaje *z* o la razón *t* como una prueba de significancia:

1. Una comparación entre dos medias:el puntaje *z* y la razón *t* se emplean para poder hacer comparaciones entre dos medias de muestras independientes o de una sola muestra ordenadas en un diseño de panel "antes-después."
2. Datos por invervalos:la suposición consiste en que tenemos puntajes al nivel de medición por intervalos. Por lo tanto, no podemos usar el puntaje *z* o la razón *t* para datos colocados por grados o datos que sólo pueden categorizarse al nivel nominal de medición (ver Capítulo 1).
3. Muestreo aleatorio:debemos haber extraído nuestras muestras sobre una base aleatoria de una población de puntajes.
4. Una distribución normal:la razón *t* para muestras pequeñas requiere que la característica de la muestra que hayamos medido esté normalmente distribuida en la población fundamental (el puntaje *z* para grandes muestras no se ve muy afectado si no se cumple esta condición). A menudo, no podemos estar 100 por ciento seguros de que existe normalidad. Al no tener motivos para creer otra cosa, muchos investigadores suponen pragmáticamente que su característica muestral está normalmente distribuida. Sin embargo, si el investigador tiene motivos para sospechar que no se puede suponer normalidad, estará más acertado si considera que la razón *t* puede ser una prueba inapropiada (ver Capítulo 6).

RESUMEN

Este capítulo se ha concentrado en la comprobación de hipótesis acerca de las diferencias entre medias muestrales. Se describió e ilustró la distribución muestral de las diferencias entre medias como una distribución de probabilidad relacionada con este propósito. Con ayuda de esta distribución, y del error estándar de la diferencia, podría hacerse una afirmación de probabilidad y, sobre esa base, rechazar o aceptar una hipótesis nula a un nivel de confianza específico. Además, vimos que la razón *t* (y los grados de libertad) podrían usarse para comprobar hipótesis acerca de diferencias entre muestras pequeñas, entre muestras de distinto tamaño y para una sola muestra medida en dos puntos en el tiempo. La propiedad de la razón *t* depende de ciertos requisitos tales como (1) hacer una comparación entre dos medias, (2) los datos por intervalos, (3) el muestreo aleatorio y (4) una distribución normal.

PROBLEMAS

1. Los investigadores sociales buscaban comprobar la hipótesis de que la prensa clandestina no está ni más ni menos orientada, hacia cuestiones sexuales, que la

prensa de la clase media. Empleando un "índice de sexualidad", recogieron datos de una muestra aleatoria de 40 artículos publicados en revistas de la clase media y de 40 artículos de revistas clandestinas. Mientras que la muestra de clase media tenía un puntaje medio de sexualidad de 3,0 y una desviación estándar de 1,5, la muestra clandestina tenía un puntaje medio de sexualidad de 4,0 y una desviación estándar de 2,0 (los puntajes medios más altos indican mayor sexualidad). Usando los datos anteriores, comprobar la hipótesis nula de que no existe ninguna diferencia con respecto a la sexualidad entre la prensa de clase media y la prensa clandestina. ¿Qué indican sus resultados?

2. Dos grupos de estudiantes tuvieron exámenes finales de estadística. Sólo se dio a un grupo la preparación formal para el examen, el otro grupo leyó el texto requerido, pero nunca asistió a clases. El primer grupo (que asistió a clases) logró calificaciones de 2, 2, 3 y 4 en el examen; el segundo grupo (que nunca asistió a clases) obtuvo calificaciones de examen de 1, 1, 2 y 3. Comprobar la hipótesis nula de que no existe ninguna diferencia en cuanto a calificaciones de examen entre los estudiantes que no asistieron a clases y los que asistieron. ¿Qué indican sus resultados? (*Nota:* Los exámenes se calificaron de 1 a 10; las calificaciones más altas representaban mejores conocimientos de estadística).
3. Comprobar la significancia de la diferencia entre las medias de las siguientes muestras aleatorias de puntajes:

Muestra 1	*Muestra 2*
8	1
3	5
1	8
7	3
7	2
6	1
8	2

4. Comprobar la significancia de la diferencia entre las medias de las siguientes muestras aleatorias de puntajes:

Muestra 1	*Muestra 2*
6	6
6	5
8	7
7	7
5	3
4	3
8	5
7	6
7	3

5. Comprobar la significancia de la diferencia entre las medias de las siguientes muestras aleatorias de puntajes

Muestra 1	*Muestra 2*
15	10
18	11
12	12
17	10
19	10

6. Comprobar la significancia de la diferencia entre las medias de las siguientes muestras aleatorias de puntajes

Muestra 1	*Muestra 2*
1	2
1	2
2	4
3	2
3	2

7. Comprobar la significancia de la diferencia entre medias de los siguientes muestras aleatorias de puntajes:

Muestra 1	*Muestra 2*
5	10
7	7
7	9
3	9
6	7
5	8
4	
6	
7	

8. Comprobar la significancia de la diferencia entre las medias de las siguientes muestras aleatorias de puntajes:

Muestra 1	*Muestra 2*
3	7
6	8
4	8
2	9
1	9
	6
	5

9. Comprobar la significancia de la diferencia entre las medias de las siguientes muestras aleatorias de puntajes:

Muestra 1	*Muestra 2*
10	10
4	10
1	8
2	7
4	
8	
3	
5	

10. Tanto antes como después de ver una película diseñada para reducir los prejuicios contra los grupos minoritarios, se interrogó a seis estudiantes acerca de sus actitudes hacia los judíos. Sobre los siguientes datos comprobar la hipótesis de que no hubo diferencia en las actitudes hacia los judíos entre estos estudiantes antes y después de ver la película (los puntajes más altos indican actitudes más favorables hacia los judíos):

Estudiante	*Antes*	*Después*
A	2	4
B	2	5
C	4	3
D	6	8
E	7	9
F	5	8

11. Comprobar la significancia de la diferencia "antes-después" entre las medias en la siguiente muestra aleatoria de puntajes:

Entrevistado	*Antes*	*Después*
A	7	3
B	6	4
C	5	2
D	4	3

12. Comprobar la significancia de la diferencia "antes-después" entre las medias en la siguiente muestra aleatoria de puntajes:

Entrevistado	*Antes*	*Después*
A	6	3
B	7	4
C	10	9
D	9	7
E	8	5

9

Análisis de varianza

Negros contra blancos, hombres contra mujeres y liberales contra conservadores representan el tipo de comparaciones entre dos muestras que ocupó nuestra atención en el capítulo anterior. No obstante, la realidad social no siempre puede rebanarse convenientemente en dos grupos; los entrevistados no siempre se dividen en forma tan simple.

Como resultado, el investigador social busca frecuentemente hacer comparaciones entre tres, cuatro, cinco o más muestras o grupos. Como ejemplo diremos que puede estudiar la influencia de la identidad racial (negra, blanca u oriental) en la discriminación laboral, el grado de privación económica (grave, moderada o leve) en la delincuencia juvenil, o la clase social subjetiva (alta, media, trabajadora o baja) en la motivación para la realización.

El estudiante se preguntará si usamos una *serie* de razones *t* para hacer comparaciones entre tres o más medias muestrales. Supóngase por ejemplo, que queremos comprobar la influencia de la clase social en la motivación para la realización. ¿Por qué no comparar por pares todas las posibles combinaciones de clases sociales y tener una razón *t* para cada comparación? Usando este método, cuatro muestras generan seis pares de combinaciones para las cuales se deben calcular seis razones *t*:

1. clase alta contra clase media;
2. clase alta contra clase trabajadora;
3. clase alta contra clase baja;
4. clase media contra clase trabajadora;
5. clase media contra clase baja;
6. clase trabajadora contra clase baja.

El procedimiento de calcular una serie de razones *t* no sólo implica una gran cantidad de trabajo, sino que también tiene una limitación estadística. Esto se debe a que aumenta la probabilidad de cometer el error alpha: error de rechazar la hipótesis nula cuando debe ser aceptada. Recordemos que el investigador social generalmente está dispuesto a aceptar un riesgo del 5 por ciento de cometer el error alpha (el nivel de confianza de 0,05). Por lo tanto, espera que *por mera casualidad* 5 de cada 100 diferencias entre medias muestrales serán lo suficientemente grandes como para considerarlas significativas. Sin embargo, mientras más pruebas estadísticas realicemos, más probable será que obtengamos resultados estadísticamente significativos por error de muestreo (más que por una verdadera diferencia poblacional) y que por ello cometamos el error alpha. Cuando llevamos a cabo un gran número de estas pruebas, la interpretación de nuestro resultado se vuelve problemática. Para tomar un ejemplo extremo: ¿cómo interpretaríamos una razón *t* significativa de entre 1 000 comparaciones en un determinado estudio? Sabemos que podemos esperar que por lo menos algunas grandes diferencias entre medias ocurran simplemente con base en el error de muestreo.

Para superar este problema y aclarar la interpretación de nuestro resultado, necesitamos una prueba estadística que mantenga el error alpha a un nivel constante, haciendo una decisión global *única* acerca de si existe una diferencia significativa entre las tres o más medias muestrales que buscamos comparar. Tal prueba se conoce como el *análisis de varianza.*

LA LOGICA DEL ANALISIS DE VARIANZA

Para realizar un análisis de varianza, tratamos la *variación* total en un conjunto de puntajes como si se pudiera dividir en dos componentes: la distancia entre los puntajes crudos y su media de grupo, conocida como la *variación dentro de los grupos* y la distancia entre las medias de los grupos, conocida como *variación entre grupos.*

Para examinar la variación dentro de los grupos, representamos gráficamente, en la Figura 9.1, los datos de motivación para la realización de los miembros de cuatro clases sociales –(1) baja, (2) trabajadora, (3) media y (4) alta– donde X_1, X_2, X_3 y X_4 representan cualquier puntaje crudo de su respectivo grupo y $\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$ y $\bar{X}_4$ constituyen las medias de dichos grupos. En términos simbólicos, vemos que la variación dentro de los grupos se refiere a la distancia entre X_1 y $\bar{X}_1$, entre X_2 y $\bar{X}_2$, entre X_3 y $\bar{X}_3$, y entre X_4 y $\bar{X}_4$.

También podemos visualizar la variación entre grupos. Con la ayuda de la Figura 9.2 vemos que el grado de motivación para la realización está en función de

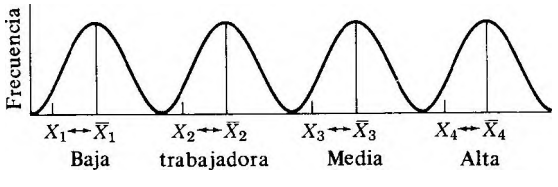


FIGURA 9.1 Representación gráfica de la variación dentro de cuatro grupos de clases sociales.

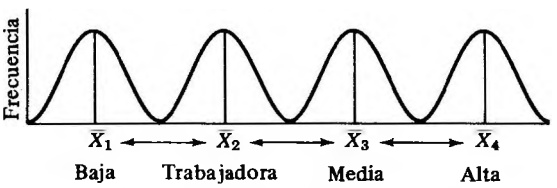


FIGURA 9.2 Representación gráfica de la variación entre cuatro grupos de clases sociales.

la clase social: el grupo de clase alta (X_4) tiene una mayor motivación para la realización que el grupo de clase media (X_3), el cual tiene a su vez mayor motivación que el grupos de clase trabajadora (X_2), cuya motivación también es mayor que la del grupo de clase baja (X_1).

La diferencia entre variación *dentro* de los grupos y variación *entre* grupos no es privativa del análisis de varianza. Aunque no se nombró como tal, encontramos una distinción semejante en la forma de la razón *t*, en la cual se comparó una diferencia *entre* $\bar{X}_1$ y $\bar{X}_2$ con el error estándar de la diferencia (σ_{dif}), estimación combinada de las diferencias *dentro* de cada grupo. Por lo tanto,

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{dif}} \quad \begin{matrix} \longleftarrow \text{variación entre grupos} \\ \longleftarrow \text{variación dentro de los grupos} \end{matrix}$$

De igual manera, el análisis de varianza produce una *razón F*, cuyo numerador representa la variación entre los grupos que se comparan y cuyo denominador contiene una estimación de la variación dentro de estos grupos. Como veremos, la razón *F* indica la magnitud de la diferencia entre los grupos *en relación* con la magnitud de la variación dentro de cada grupo. Como sucedió con la razón *t*, mientras mayor sea la razón *F* (mientras mayor sea la variación entre los grupos en relación con la variación dentro de ellos), mayor será la probabilidad de rechazar la hipótesis nula y aceptar la hipótesis de investigación.

LAS SUMAS DE CUADRADOS

El concepto de la *suma de cuadrados* está en el centro del análisis de varianza y representa el paso inicial para medir la variación total, así como la variación entre los grupos y dentro de ellos. Saber que sólo el rótulo "suma de los cuadrados" es nuevo para nosotros, puede ser una agradable sorpresa. El concepto mismo se presentó en el Capítulo 5 como un paso importante en el procedimiento para obtener la desviación estándar. En ese contexto aprendimos a encontrar la suma de los cuadrados elevando al cuadrado las desviaciones de la media de una distribución y sumando estos puntajes de desviación (Σx^2). Este procedimiento eliminaba los signos menos pero seguía proporcionando una sólida base matemática para la desviación estándar.

Cuando se aplica a una situación en la que se están comparando grupos, existe más de un tipo de suma de cuadrados, aunque cada tipo representa la *suma de desviaciones de la media elevadas al cuadrado*. En correspondencia con la distinción

entre la variación total y sus dos componentes, tenemos la suma total de cuadrados (SC_{total}), la suma de cuadrados entre grupos (SC_{ent}), y la suma de cuadrados dentro de los grupos (SC_{dentro}).

Un ejemplo de investigación

Consideremos una situación de investigación en la que se podría calcular cada tipo de suma de cuadrados. Supóngase que buscamos determinar la influencia de la orientación política en los métodos de crianza de los niños. En el capítulo anterior abordamos este problema mediante una comparación entre liberales y conservadores. Por contraste, ahora queremos hacer comparaciones que representen *varios* puntos en la escala política. Por ejemplo, podríamos comparar la permisibilidad, en la crianza de los niños, de conservadores, liberales, radicales y moderados. En tal caso,

Hipótesis Nula: ($\mu_1 = \mu_2 = \mu_3 = \mu_4$) *Los conservadores, liberales, radicales y moderados no difieren entre sí respecto a la permisibilidad en la crianza de los niños.*

Hipótesis de Investigación: ($\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$) *Los conservadores, liberales, radicales y moderados, difieren entre sí respecto a la permisibilidad en la crianza de los niños.*

Imaginemos que realmente hemos entrevistado muestras aleatorias de cuatro conservadores, cuatro liberales, cuatro radicales y cuatro moderados, para determinar sus métodos de crianza de los niños. Imaginemos además que hemos obtenido los puntajes de permisibilidad que se ven en la Tabla 9.1 (los puntajes van desde 1, que representa poca permisibilidad, hasta 5, que representa mucha permisibilidad).

La suma de cuadrados dentro de los grupos

La suma de cuadrados dentro de los grupos nos da la *suma de las desviaciones de cada puntaje crudo con su media muestral elevadas al cuadrado*. Por lo tanto, la suma de cuadrados dentro de los grupos puede obtenerse por la simple *combinación* de las sumas de cuadrados dentro de cada muestra. Por fórmula,

$$SC_{dentro} = \Sigma x_1^2 + \Sigma x_2^2 + \Sigma x_3^2 + \Sigma x_4^2$$

donde

x = un puntaje de desviación $(X - \bar{X})$

Aplicando la fórmula SC_{dentro} a los datos de la Tabla 9.1, vemos que

$$\begin{aligned} SC_{dentro} &= 1{,}00 + 2{,}00 + 0{,}74 + 2{,}74 \\ &= 6{,}48 \end{aligned}$$

Suma de cuadrados entre los grupos

La *suma de cuadrados entre los grupos* representa la *suma de las desviaciones de cada media muestral de la media total elevadas al cuadrado*. En consecuencia,

debemos determinar la diferencia entre cada media muestral y la media total ($X - \bar{X}_{total}$), elevar al cuadrado este puntaje de diferencia, multiplicar por el número de puntajes en la muestra y sumar estas cantidades. La fórmula de definición para la suma de cuadrados entre los grupos es

$$SC_{ent} = \Sigma(\bar{X} - \bar{X}_{total})^2 N$$

donde

$\bar{X}$ = cualquier media muestral

$\bar{X}_{total}$ = la media total (la media de todos los puntajes crudos de la totalidad de las muestras combinadas)

N = el número de puntajes de cualquier muestra

SC_{ent} = la suma de cuadrados entre los grupos

El procedimiento para encontrar la suma de cuadrados entre los grupos para los datos de la Tabla 9.1 puede resumirse como sigue:

TABLA 9.1 Puntajes de permisibilidad en la crianza de los niños para muestras de conservadores, moderados, liberales y radicales

Conservadores ($N = 4$)			*Moderados* ($N = 4$)		
X_1	x	x^2	X_2	x	x^2
1	−0,50	0,25	1	−1	1
2	0,50	0,25	3	1	1
1	−0,50	0,25	2	0	0
2	0,50	0,25	2	0	0
$\Sigma X_1 = 6$		$\Sigma x^2 = 1,00$	$\Sigma X_2 = 8$		$\Sigma x^2 = 2,00$
$\bar{X}_1 = \frac{6}{4} = 1,5$			$\bar{X}_2 = \frac{8}{4} = 2,0$		

Liberales ($N = 4$)			*Radicales* ($N = 4$)		
X_3	x	x^2	X_4	x	x^2
1	−0,75	0,56	3	1,25	1,56
2	0,25	0,06	2	0,25	0,06
2	0,25	0,06	1	−0,75	0,56
2	0,25	0,06	1	−0,75	0,56
$\Sigma X_3 = 7$		$\Sigma x^2 = 0,74$	$\Sigma X_4 = 7$		$\Sigma x^2 = 2,74$
$\bar{X}_3 = \frac{7}{4} = 1,75$			$\bar{X}_4 = \frac{7}{4} = 1,75$		

$$\bar{X}_{total} = 1,75$$

$$\begin{aligned} SC_{ent} &= (1,50 - 1,75)^2 4 + (2,0 - 1,75)^2 4 \\ &\quad + (1,75 - 1,75)^2 4 + (1,75 - 1,75)^2 4 \\ &= (-0,25)^2 4 + (0,25)^2 4 + (0)4 + (0)4 \\ &= (0,06)4 + (0,06)4 + (0)4 + (0)4 \\ &= 0,24 + 0,24 \\ &= 0,48 \end{aligned}$$

La suma total de cuadrados

Puede demostrarse que la *suma total de cuadrados, la suma de las desviaciones de cada puntaje crudo de la media total del estudio elevadas al cuadrado,* es igual a una combinación de sus componentes dentro y entre los grupos. La suma total de cuadrados para los datos de la Tabla 9.1 se puede encontrar como sigue:

$$\begin{aligned} SC_{total} &= SC_{ent} + SC_{dentro} \\ &= 0{,}48 + 6{,}48 \\ &= 6{,}96 \end{aligned}$$

La suma total de cuadrados también se puede definir en términos de la ecuación

$$SC_{total} = \Sigma(X - \bar{X}_{total})^2$$

donde

X = un puntaje crudo en cualquier muestra

$\bar{X}_{total}$ = la media total (la media de todos los puntajes crudos de todas las muestras combinadas)

SC_{total} = la suma total de cuadrados

Utilizando la fórmula anterior, restamos la media total ($\bar{X}_{total}$) de cada puntaje crudo del estudio (X), elevamos al cuadrado, los puntajes de desviación que resulten y los sumamos.
Para los datos de la Tabla 9.1,

$$\begin{aligned} SC_{total} &= (1 - 1{,}75)^2 + (2 - 1{,}75)^2 + (1 - 1{,}75)^2 + (2 - 1{,}75)^2 \\ &\quad + (1 - 1{,}75)^2 + (3 - 1{,}75)^2 + (2 - 1{,}75)^2 \\ &\quad + (2 - 1{,}75)^2 + (1 - 1{,}75)^2 + (2 - 1{,}75)^2 \\ &\quad + (2 - 1{,}75)^2 + (2 - 1{,}75)^2 + (3 - 1{,}75)^2 \\ &\quad + (2 - 1{,}75)^2 + (1 - 1{,}75)^2 + (1 - 1{,}75)^2 \\ &= (-0{,}75)^2 + (0{,}25)^2 + (-0{,}75)^2 + (0{,}25)^2 + (-0{,}75)^2 \\ &\quad + (1{,}25)^2 + (0{,}25^2 + (0{,}25)^2 + (-0{,}75)^2 + (0{,}25)^2 \\ &\quad + (0{,}25)^2 + (0{,}25)^2 + (0{,}25)^2 + (0{,}75)^2 + \\ &\quad + (-0{,}75)^2 \\ &= 0{,}56 + 0{,}06 + 0{,}56 + 0{,}06 + 0{,}56 + 1{,}56 + 0{,}06 \\ &\quad + 0{,}06 + 0{,}56 + 0{,}06 + 0{,}06 + 0{,}06 + 1{,}56 + 0{,}06 \\ &\quad + 0{,}56 + 0{,}56 \\ &= 6{,}96 \end{aligned}$$

Cómo calcular sumas de cuadrados

Las fórmulas de definición para las sumas de cuadrados, dentro de los grupos, entre los grupos y totales, en la forma en que se presentaron anteriormente, se basan en el

manejo de puntajes de desviación, requisito difícil y demorado. Afortunadamente, podemos usar en su lugar las fórmulas de cálculo que se indican más adelante, las cuales son mucho más simples para obtener un resultado en forma de razón F, que es idéntica (exceptuando los errores de redondeo) a la que obtuvimos con las fórmulas de definición mucho más largas.

Los puntajes crudos de la Tabla 9.1 se han colocado en la Tabla 9.2 con el fin de ilustrar el uso de las fórmulas de cálculo de la suma de cuadrados.

La fórmula para calcular la suma total de cuadrados es la siguiente:

$$SC_{total} = \Sigma X^2{}_{total} - \frac{(\Sigma X_{total})^2}{N_{total}}$$

donde

N_{total} = el número total de puntajes en todas las muestras combinadas.

Desarrollando esta fórmula para los datos de la Tabla 9.2,

$$SC_{total} = (10 + 18 + 13 + 15) - \frac{(6 + 8 + 7 + 7)^2}{4 + 4 + 4 + 4}$$

$$= 56 - \frac{(28)^2}{16}$$

$$= 56 - \frac{784}{16}$$

$$= 56 - 49$$

$$= 7$$

TABLA 9.2 Puntajes de permisibilidad en la crianza de los niños para muestras de conservadores, liberales, radicales y moderados.

Conservadores ($N = 4$)		*Moderados* ($N = 4$)	
X_1	X^2	X_2	X^2
1	1	1	1
2	4	3	9
1	1	2	4
2	4	2	4
$\Sigma X = 6$	$\Sigma X^2 = 10$	$\Sigma X = 8$	$\Sigma X^2 = 18$
$\bar{X}_1 = \frac{6}{4} = 1{,}5$		$\bar{X}_2 = \frac{8}{4} = 2{,}0$	
Liberales ($N = 4$)		*Radicales* ($N = 4$)	
X_3	X^2	X_4	X^2
1	1	3	9
2	4	2	4
2	4	1	1
2	4	1	1
$\Sigma X = 7$	$\Sigma X^2 = 13$	$\Sigma X = 7$	$\Sigma X^2 = 15$
$\bar{X}_3 = \frac{7}{4} = 1{,}75$		$\bar{X}_4 = \frac{7}{4} = 1{,}75$	
		$\bar{X}_{total} = 1{,}75$	

La suma de cuadrados entre los grupos puede obtenerse por medio de la siguiente fórmula:

$$SC_{ent} = \left[\sum \frac{(\Sigma X)^2}{N}\right] - \frac{(\Sigma X_{total})^2}{N_{total}}$$

donde

N = el número total de puntajes en cualquier muestra
N_{total} = el número total de puntajes en todas las muestras combinadas

Por ejemplo, en la Tabla 9.2,

$$\begin{aligned} SC_{ent} &= \frac{(6)^2}{4} + \frac{(8)^2}{4} + \frac{(7)^2}{4} + \frac{(7)^2}{4} - \frac{(28)^2}{16} \\ &= \frac{36}{4} + \frac{64}{4} + \frac{49}{4} + \frac{49}{4} - \frac{784}{16} \\ &= 9{,}0 + 16 + 12{,}25 + 12{,}25 - 49{,}0 \\ &= 49{,}5 - 49{,}0 \\ &= 0{,}50 \end{aligned}$$

En virtud de que la suma de cuadrados dentro de los grupos es más lenta para calcularse, podemos sacar ventaja del hecho de que la suma total de los cuadrados es igual a una combinación de sus dos componentes. Por lo tanto,

$$SC_{dentro} = SC_{total} - SC_{ent}$$

En el presente caso,

$$\begin{aligned} SC_{dentro} &= 7{,}00 - 0{,}50 \\ &= 6{,}50 \end{aligned}$$

La siguiente fórmula para la suma de cuadrados dentro de los grupos puede servir como verificación de errores de cálculo:

$$SC_{dentro} = \sum\left[(\Sigma X^2) - \frac{(\Sigma X)^2}{N}\right]$$

donde

X = un puntaje crudo en cualquier muestra

N = el número total de puntajes en cualquier muestra

Sustituyendo los datos de la Tabla 9.2,

$$SC_{dentro} = \left[10 - \frac{(6)^2}{4}\right] + \left[18 - \frac{(8)^2}{4}\right] +$$

$$+\left[13-\frac{(7)^2}{4}\right]+\left[15-\frac{(7)^2}{4}\right]$$

$$=\left(10-\frac{36}{4}\right)+\left(18-\frac{64}{4}\right)$$

$$+\left(13-\frac{49}{4}\right)+\left(15-\frac{49}{4}\right)$$

$$=(10-9{,}0)+(18-16{,}0)+(13-12{,}25)$$
$$+(15-12{,}25)$$
$$=1{,}0+2{,}0+0{,}75+2{,}75$$
$$=6{,}50$$

LA MEDIA CUADRATICA

Como es de esperarse de una medida de variación, el valor de las sumas de los cuadrados tiende a crecer a medida que la variación aumenta. Por ejemplo, SC = 10,9 probablemente indica mayor variación que *SC* = 1,3. Sin embargo, la suma de los cuadrados también crece con el aumento de la magnitud de la muestra, la manera que $N = 200$ producirá un SC mayor que $N = 20$. Como resultado, la suma de los cuadrados no puede considerarse una medida "pura" de variación totalmente satisfactoria, a no ser, por supuesto, que podamos encontrar una forma de controlar el número de puntajes involucrados.

Afortunadamente existe tal método en una medida de variación conocida como la *media cuadrática* (o *varianza*), que obtenemos dividiendo SC_{ent} o SC_{dentro} mediante los grados de libertad apropiados (en el Capítulo 5 dividimos igualmente Σx^2 por N como un paso hacia la obtención de la desviación estándar). Por lo tanto,

$$\mu C_{ent} = \frac{SC_{ent}}{gl_{ent}}$$

donde

μC_{ent} = la media cuadrática entre los grupos
SC_{ent} = la suma de cuadrados entre los grupos
gl_{ent} = los grados de libertad entre los grupos

y

$$\mu C_{dentro} = \frac{SC_{dentro}}{gl_{dentro}}$$

donde

μC_{dentro} = la media cuadrática dentro de los grupos

SC_{dentro} = la suma de cuadrados dentro de los grupos

gl_{dentro} = los grados de libertad dentro de los grupos

Pero aún debemos obtener los grados de libertad apropiados. Para la media cuadrática entre los grupos,

$$gl_{ent} = k - 1$$

donde

k = el número de muestras

Para encontrar la media cuadrática dentro de los grupos,

$$gl_{dentro} = N_{total} - k$$

donde

N_{total} = el número total de puntajes en todas las muestras combinadas

k = el número de muestras

Ilustrando con los datos de la Tabla 9.2, para los cuales $SC_{ent} = 0{,}50$ y $SC_{dentro} = 6{,}50$, calculamos nuestros grados de libertad como sigue:

$$gl_{ent} = 4 - 1 = 3$$

y

$$gl_{dentro} = 16 - 4 = 12$$

Ahora estamos preparados para obtener las medias cuadráticas

$$\mu C_{ent} = \frac{0{,}50}{3} = 0{,}17$$

y

$$\mu C_{dentro} = \frac{6{,}50}{12} = 0{,}54$$

RAZON O COCIENTE *F*

Como se anotó anteriormente, el análisis de varianza produce una razón F en la que se comparan la variación entre los grupos y la variación dentro de los grupos. Ahora

estamos en condiciones de especificar el grado de cada tipo de variación tal como se midió por las medias cuadráticas. Por lo tanto, la razón *F* puede considerarse como un indicador de la magnitud de la media cuadrática entre los grupos en relación con el tamaño de la media cuadrática dentro de los grupos, o

$$F = \frac{\mu C_{ent}}{\mu C_{dentro}}$$

Para la Tabla 9.2,

$$F = \frac{0{,}17}{0{,}54}$$
$$= 0{,}31$$

Habiendo obtenido una razón *F* debemos determinar ahora si es lo suficientemente grande para rechazar la hipótesis nula y aceptar la hipótesis de investigación. ¿Difieren los conservadores, los liberales, los radicales y los moderados con respecto a la permisibilidad en la crianza de los niños? Mientras mayor sea nuestra razón *F* calculada (mientras mayor sea la MC_{ent} y menor la MC_{dentro}), más probabilidades tendremos de obtener un resultado estadísticamente significativo.

Pero, ¿cómo reconocer exactamente una razón *F* significativa? Recordemos que, en el Capítulo 8, la razón *t* obtenida con los grados de libertad apropiados, se comparaba con una tabla de razones *t* para el nivel de confianza de 0,05, etc. Igualmente, ahora debemos interpretar la razón *F* que hemos calculado, con la ayuda de la Tabla D al final del libro. La Tabla D contiene una lista de razones *F* significativas —razones *F* que debemos obtener para poder rechazar la hipótesis nula a los niveles de confianza de 0,05 y 0,01. Al igual que en caso de la razón *t* el valor exacto de *F* que debemos obtener depende de sus grados de libertad asociados. Por lo tanto, nuestro uso de la Tabla D se inicia buscando los dos valores gl, los grados de libertad ~~entre~~ los grupos y los grados de libertad ~~dentro~~ de los grupos. Los grados de libertad asociados con el numerador (gl_{ent}) se han indicado en la parte superior de la página, mientras que los grados de libertad asociados con el denominador (gl_{dentro}) se han colocado al lado izquierdo de la tabla. El cuerpo de la Tabla D presenta razones *F* significativas a los niveles de confianza de 0,05 y 0,01.

Para los datos de la Tabla 9.2, hemos encontrado que $gl_{ent} = 3$ y $gl_{dentro} = 12$. Así, en la Tabla D vamos hacia la columna marcada gl = 3 y desde ese punto continuamos hacia abajo hasta llegar a la columna marcada gl = 12. Mediante este procedimiento encontramos que una razón *F* significativa al nivel de confianza de 0,05 debe ser por lo menos 3,49 y al nivel de confianza de 0,01 debe ser igual o mayor que 5,95. La razón *F* que hemos calculado es de sólo 0,31. Como resultado, no tenemos más alternativa que *aceptar* la hipótesis nula y atribuir nuestra diferencia entre medias muestrales, sobre la permisibilidad en la crianza de los niños, al error de muestreo más que a una diferencia real en las poblaciones de conservadores, liberales, radicales y moderados.

TABLA 9.3 Tabla de resumen del análisis de varianza para los datos de la Tabla 9.2.

Fuente de la variación	*gl*	*SC*	*MC*	*F*
Entre grupos	3	0,50	0,17	0,31
Dentro de los grupos	12	6,50	0,54	

Los resultados de nuestro análisis de varianza se pueden colocar en una "tabla de resumen" como la que se muestra en la Tabla 9.3. Se ha convertido en un procedimiento estándar resumir de esta manera un análisis de varianza.

Una ilustración

Para ilustrar paso a paso un análisis de varianza, supongamos que deseamos comprobar la hipótesis de que el coeficiente intelectual (C.I.) varía según la clase social. Por lo tanto,

Hipótesis Nula: Las clases alta, media y baja, no difieren respecto al coeficiente intelectual. ($\mu_1 = \mu_2 = \mu_3$)

Hipótesis de Investigación: Las clases alta, media y baja, difieren respecto al coeficiente intelectual. ($\mu_1 \neq \mu_2 \neq \mu_3$)

Digamos que, para investigar esta hipótesis, establecemos el nivel de confianza de 0,05 como criterio significativo. Imaginemos que podemos medir el C.I. de los miembros de tres muestras de clases sociales: alta, media y baja. Se supone que resultan los siguientes puntajes de C.I.:

Alta (N = 5)		*Media (N = 5)*	
X_1	X^2	X_2	X^2
130	16 900	120	14 400
125	15 625	115	13 225
130	16 900	115	13 225
120	14 400	110	12 100
122	14 884	112	12 544
$\Sigma X = 627$	$\Sigma X^2 = 78\,709$	$\Sigma X = 572$	$\Sigma X^2 = 65\,494$
$\bar{X}_1 = 125{,}4$		$\bar{X}_2 = 114{,}4$	

Baja (N = 5)	
X_3	X^2
110	12 100
100	10 000
90	8 100
100	10 000
85	7 225
$\Sigma X = 485$	$\Sigma X^2 = 47\,425$
$\bar{X}_3 = 97{,}0$	

El procedimiento, paso por paso, para verificar la significancia estadística de la diferencia obtenida entre las medias es como sigue.

PASO 1: Encontrar la media de cada muestra

$$\bar{X}_1 = \frac{\Sigma X_1}{N} \qquad \bar{X}_2 = \frac{\Sigma X_2}{N} \qquad \bar{X}_3 = \frac{\Sigma X_3}{N}$$

$$= \frac{627}{5} \qquad = \frac{572}{5} \qquad = \frac{485}{5}$$

$$= 125{,}4 \qquad = 114{,}4 \qquad = 97{,}0$$

Nótese que las diferencias entre las medias existen, siendo la tendencia que los puntajes de C.I. aumenten de la clase baja a la media o a la alta.

PASO 2: Encontrar la suma total de cuadrados

$$SC_{total} = \Sigma X^2{}_{total} - \frac{(\Sigma X_{total})^2}{N_{total}}$$

$$= (78709 + 65494 + 47425) - \frac{(627 + 572 + 485)^2}{15}$$

$$= 191628 - \frac{(1684)^2}{15}$$

$$= 191628 - \frac{2835856}{15}$$

$$= 191628 - 189057{,}07$$

$$= 2570{,}93$$

PASO 3: Encontrar la suma de cuadrados entre los grupos

$$SC_{ent} = \left[\sum \frac{(\Sigma X)^2}{N}\right] - \frac{(\Sigma X_{total})^2}{N_{total}}$$

$$= \frac{(627)^2}{5} + \frac{(572)^2}{5} + \frac{(485)^2}{5} - \frac{(1684)^2}{15}$$

$$= \frac{393129}{5} + \frac{327184}{5} + \frac{235225}{5} - \frac{2835856}{15}$$

$$= 78625{,}8 + 65436{,}8 + 47045{,}0 - 189057{,}07$$

$$= 191107{,}60 - 189057{,}07$$

$$= 2050{,}53$$

PASO 4: Encontrar la suma de los cuadrados dentro de los grupos

$$SC_{dentro} = SC_{total} - SC_{ent}$$

$$= 2570{,}93 - 2050{,}53$$

$$= 520{,}40$$

$$SC_{dentro} = \sum \left[(\Sigma X^2) - \frac{(\Sigma X)^2}{N} \right]$$

$$= \left[78709 - \frac{(627)^2}{5}\right] + \left[65494 - \frac{(572)^2}{5}\right] + \left[47425 - \frac{(485)^2}{5}\right]$$

$$= \left[78709 - \frac{393129}{5}\right] + \left[65494 - \frac{327184}{5}\right] + \left[47425 - \frac{235225}{5}\right]$$

$$= [78709 - 78625{,}8] + [65494 - 65436{,}8] + [47425 - 47045{,}0]$$

$$= 83{,}2 + 57{,}2 + 380{,}0$$

$$= 520{,}40$$

PASO 5: Encontrar los grados de libertad entre los grupos

$$gl_{ent} = K - 1$$
$$= 3 - 1$$
$$= 2$$

PASO 6: Encontrar los grados de libertad dentro de los grupos

$$gl_{dentro} = N_{total} - K$$
$$= 15 - 3$$
$$= 12$$

PASO 7: Encontrar la media cuadrática entre grupos

$$\mu C_{ent} = \frac{SC_{ent}}{gl_{ent}}$$
$$= \frac{2050{,}53}{2}$$
$$= 1025{,}27$$

PASO 8: Buscar la media cuadrática dentro de los grupos

$$\mu C_{dentro} = \frac{SC_{dentro}}{gl_{dentro}}$$
$$= \frac{520{,}40}{12}$$
$$= 43{,}37$$

PASO 9: Obtener la razón F

$$F = \frac{\mu C_{Ent}}{\mu C_{dentro}}$$

$$= \frac{1025{,}27}{43{,}37}$$

$$= 23{,}64$$

PASO 10: Comparar la razón F obtenida con la razón F correspondiente en la Tabla D

razón F obtenida = 23,64
razón F de la tabla = 3,88
gl = 2/12
P = 0,05

Como muestra el Paso 10, para rechazar la hipótesis nula al nivel de confianza de 0,05 con 2/12 grados de libertad, la razón calculada F debe ser al menos 3,88 Debido a que obtuvimos una razón F de 23,64, podemos rechazar la hipótesis nula y aceptar la hipótesis de investigación. Específicamente, concluimos que las clases baja, media y alta, realmente difieren respecto al C.I.

UNA COMPARACION MULTIPLE DE MEDIAS

Una razón F significativa nos informa de una *diferencia global* entre los grupos que se están estudiando. Si estuviéramos investigando una diferencia entre sólo dos medias muestrales, no se necesitaría ningún análisis adicional para interpretar nuestro resultado: en tal caso, la diferencia obtenida es estadísticamente significativa o no, dependiendo de la magnitud de nuestra razón F. Sin embargo, cuando encontramos una F significativa para las diferencias entre tres o más medias, puede ser importante determinar exactamente dónde están las diferencias significativas. Por ejemplo, en la ilustración anterior, descubrimos diferencias de C.I. estadísticamente significativas entre tres clases sociales. Considérense las posibilidades que presenta esta razón F significativa: $\bar{X}_1$ (alta) puede diferir significativamente de $\bar{X}_2$ (media); $\bar{X}_1$ (alta) puede diferir significativamente de $\bar{X}_3$ (baja); o $\bar{X}_2$ puede diferir significativamente de $\bar{X}_3$ (baja).

Como se explicó anteriormente en este capítulo, obtener una razón t para cada comparación $-\bar{X}_1$ contra $\bar{X}_2$; $\bar{X}_1$ contra $\bar{X}_3$; $\bar{X}_2$ contra $\bar{X}_3$ – implicaría una gran cantidad de trabajo y también aumentaría la probabilidad del error alpha. Afortunadamente se han desarrollado muchas otras pruebas estadísticas para hacer comparaciones múltiples después de una razón F significativa, con el fin de señalar dónde se encuentran las diferencias significativas entre medias. Presentaremos la DSH de Tukey –diferencia significativa honesta (honestly significant difference HSD– una de las más útiles pruebas de comparación múltiple.

La DSH de Tukey se usa sólo después de haber obtenido una razón F significativa. Por el método de Tukey comparamos la diferencia entre dos puntajes medios cualquiera con la DSH. Una diferencia entre medias es estadísticamente significativa sólo si es igual o mayor que la DSH. Por fórmula,

$$\text{DSH} = q\alpha\sqrt{\frac{MC_{\text{dentro}}}{n}}$$

donde

$q\alpha$ = un valor de la tabla a un nivel de confianza dado para el número máximo de medias que se estén comparando

μC_{dentro} = la media cuadrática dentro de los grupos (que se obtuvo del análisis de varianza)

n = el número de entrevistados en cada grupo (supone el mismo número en cada grupo)

A diferencia de la razón *t*, la DSH toma en cuenta que la probabilidad del error alpha se incrementa a medida que aumenta el número de medias que se esté comparando. Dependiendo del valor de $q\alpha$, mientras mayor sea el número de medias, más "conservadora" se volverá la DSH en cuanto al rechazo de la hipótesis nula. Como resultado, se obtendrán menos diferencias significativas con la DSH que con la razón *t*. Además, una diferencia entre medias será posiblemente más significativa en una comparación múltiple, entre tres medias, que en una comparación múltiple entre cuatro o cinco medias.

Para ilustrar el uso de la DSH, regresemos a un ejemplo anterior en el cual se encontró que las clases sociales diferían en relación con el C.I. Más específicamente, obtuvimos una razón *F* significativa ($F = 23{,}64$) para las siguientes diferencias entre las muestras de clase alta, media y baja:

$\bar{X}_1$ (alta) = 125,4
$\bar{X}_2$ (media) = 114,4
$\bar{X}_3$ (baja) = 97,0

PASO 1: Construir una tabla de diferencias entre medias ordenadas. Para los presentes datos, el orden jerárquico de las medias (de menor a mayor) es 97,0, 114,4 y 125,4. Estos puntajes medios se colocan en forma de tabla de manera que la diferencia entre cada par de medias se muestran dentro de una tabla. Así, la diferencia entre $\bar{X}_1$ (alta) y $\bar{X}_3$ (baja) es 28,40; la diferencia entre $\bar{X}_1$ (alta y $\bar{X}_2$ (media) es 11,0; y la diferencia entre $\bar{X}_2$ (media) y $\bar{X}_3$ (baja) es 17,4.

	$\bar{X}_3 = 97{,}0$	$\bar{X}_2 = 114{,}4$	$\bar{X}_1 = 125{,}4$
$\bar{X}_3$	—	17,4	28,4
$\bar{X}_2$	—	—	11,0
$\bar{X}_1$	—	—	—

PASO 2: Encontrar $q\alpha$ en la Tabla I. Para encontrar $q\alpha$ en la Tabla I, al final del libro, debemos tener (a) los grados de libertad (gl) para μC_{dentro}, (b) el mayor

número de medias (k), y (c) un nivel de confianza, bien sea 0,01 o 0,05. Del análisis de varianza sabemos ya que gl = 12. Por lo tanto, seguimos la columna de la izquierda de la Tabla I hasta llegar a los 12 grados de libertad. Posteriormente, ya que estamos comparando por pares tres puntajes medios, nos movemos a través de la Tabla I hasta un número máximo de medias (k) igual a 3. Suponiendo un nivel de confianza de 0,05 encontramos que $q0{,}05 = 3{,}77$.

PASO 3: Encontrar la DSH

$$\begin{aligned} \text{DSH} &= q0{,}05\sqrt{\frac{\mu C_{\text{dentro}}}{n}} \\ &= 3{,}77\sqrt{\frac{43{,}37}{5}} \\ &= 3{,}77\sqrt{8{,}67} \\ &= 3{,}77(2{,}94) \\ &= 11{,}08 \end{aligned}$$

PASO 4: Comparar DSH con la tabla de las diferencias entre medias. Para que se la considere estadísticamente significativa, cualquier diferencia entre medias que obtengamos debe ser igual o mayor que la DSH. Refiriéndonos a nuestra anterior tabla de diferencias entre medias, vemos que la diferencia de C.I. de 28,4 entre $\overline{X}_1$ (clase alta) y $\overline{X}_3$ (clase baja) y la diferencia de C.I. de 17,4 entre $\overline{X}_2$ (clase media) y $\overline{X}_3$ (clase baja) son mayores que la DSH = 11,08. Como resultado, concluimos que estas diferencias entre las medias son estadísticamente significativas al nivel de confianza de 0,05. Sólo la diferencia de 11,0 entre $\overline{X}_2$ y $\overline{X}_1$ no es igual ni mayor que la DSH y, por lo tanto, no es estadísticamente significativa.

REQUISITOS PARA EL USO DE LA RAZON F

El análisis de varianza deberá hacerse sólo después de que el investigador haya tomado en cuenta los siguientes requisitos:

1. Una comparación entre tres o más medias independientes: la razón F se emplea usualmente para comparar tres o más medias de muestras independientes. No se puede comprobar una sola muestra colocada en un diseño de panel. Sin embargo, es posible obtener una razón F en lugar de una razón t cuando se hacen comparaciones entre dos muestras. Para el caso de dos muestras $F = t^2$ y se obtienen resultados idénticos.
2. Los datos de intervalo: para realizar un análisis de varianza suponemos que hemos logrado el nivel de medición por intervalos. Preferentemente, no se usarán datos categorizados o colocados por rango.
3. El muestreo aleatorio: debimos haber tomado nuestras muestras aleatoriamente de una población de puntajes.
4. Una distribución normal: suponemos que la característica muestral que medimos está distribuida normalmente en la población original.

RESUMEN

El análisis de varianza puede usarse para hacer comparaciones entre tres o más medias muestrales. Esta prueba origina una razón F cuyo numerador representa la variación entre los grupos y cuyo denominador contiene una estimación de la variación dentro de los grupos. La suma de cuadrados representa el paso inicial para medir la variación. Sin embargo, está muy afectada por la magnitud de la muestra. Para superar este problema dividimos SC_{ent} o SC_{dentro} entre los grados de libertad correspondientes para obtener la media cuadrática. F indica el tamaño de la media cuadrática entre los grupos con respecto al tamaño de la media cuadrática dentro de los grupos. Interpretamos nuestra razón F calculada comparándola con la razón F correspondiente en la Tabla D. Sobre esa base decidimos si rechazamos o aceptamos nuestra hipótesis nula. Después de obtener una F significativa podemos determinar exactamente dónde están las diferencias significativas aplicando el método de Tukey para la comparación múltiple de medias.

PROBLEMAS

1. Comprobar, en las siguientes muestras aleatorias de clases sociales, la hipótesis nula de que la sociabilidad no varía según la clase social. (*Nota*: Los puntajes más altos indican mayor sociabilidad.)

Baja	*Trabajadora*	*Media*	*Alta*
8	7	6	5
4	3	5	2
7	2	5	1
8	8	4	3

2. Comprobar la significancia de las diferencias entre las medias de las siguientes muestras aleatorias de puntajes:

Muestra 1	*Muestra 2*	*Muestra 3*
2	5	8
1	4	9
3	3	7
3	4	8

3. Comprobar la significancia de las diferencias entre las medias de las siguientes muestras aleatorias de puntajes:

Muestra 1	*Muestra 2*	*Muestra 3*
12	6	3
6	5	2
8	7	5
7	5	3
6	1	1

4. Comprobar la significancia de las diferencias entre las medias de las siguientes muestras aleatorias de puntajes:

Muestra 1	*Muestra 2*	*Muestra 3*
5	4	3
5	3	5
4	2	1
3	2	3
6	1	3

5. Realizar una comparación múltiple de medias siguiendo el método de Tukey para determinar exactamente dónde ocurren las diferencias significativas del problema anterior.
6. Comprobar la significancia de las diferencias entre las medias de las siguientes muestras aleatorias de puntajes:

Muestra 1	*Muestra 2*	*Muestra 3*	*Muestra 4*
1	3	4	6
1	2	4	6
3	2	2	5
4	1	2	5
2	5	3	4
1	5	3	6

7. Realizar una comparación múltiple de medias según el método de Tukey para determinar exactamente dónde ocurren las diferencias significativas del Problema 6.

10

Chi cuadrada y otras pruebas no paramétricas

Como se indicó en los Capítulos 8 y 9, debemos exigir bastante del investigador social que emplea una razón *t* o un análisis de varianza para hacer comparaciones entre sus muestras. Cada una de estas pruebas de significancia tiene una lista de requisitos que incluye la suposición de que la característica que se estudia está distribuida normalmente en una determinada población. Además, cada prueba exige el nivel de medición por intervalos, de manera que se le pueda asignar un puntaje a cada miembro de la muestra. Cuando una prueba de significancia, tal como la razón o cociente *t* o el análisis de varianza, requiere de (1) normalidad y (2) de una medida de nivel por intervalos, a la cual nos referimos como una *prueba paramétrica.*[1]

¿Qué sucede con el investigador social que no puede emplear una prueba paramétrica, esto es, que, o no puede suponer honestamente la normalidad o cuyos datos no se sujetan a una medida de nivel por intervalos? Supongamos, por ejemplo, que está trabajando con una distribución sesgada, tal como el ingreso anual, o con datos que han sido categorizados y contados (nivel nominal) o colocados por rangos (nivel ordinal). ¿Cómo se las arregla este investigador para hacer comparaciones entre las muestras sin violar los requisitos de una prueba determinada?

Afortunadamente, los estadísticos han desarrollado varias pruebas *no paramétricas* de significancia –pruebas cuya lista de requisitos no incluye una distribución normal o el nivel de medición por intervalos. Para comprender la importante posición de las pruebas no paramétricas en la investigación social, debemos entender también el concepto estadístico de potencia. *La potencia de una prueba* es la probabilidad de rechazar la hipótesis nula cuando ésta es realmente falsa y debe ser rechazada.

La potencia varía de una prueba a otra. Las pruebas más poderosas –aquéllas que más probablemente rechazarán la hipótesis nula cuando ésta sea falsa– son las

[1] Esta designación se basa en el término "parámetro", que se refiere a cualquier característica de una población.

pruebas que tienen los requisitos más fuertes o los más difíciles de satisfacer. Generalmente, estas son pruebas paramétricas tales como t o F las cuales suponen que se han logrado datos por intervalos y que las características en estudio se hallan distribuidas normalmente en sus poblaciones. En contraste, las alternativas no paramétricas tienen exigencias menos estrictas y constituyen pruebas de significancia menos poderosas que sus contrapartes paramétricas. Como resultado, suponiendo que la hipótesis nula sea falsa (y se mantengan constantes otros factores tales como el tamaño de la muestra), será más probable que un investigador rechace la hipótesis nula mediante el uso apropiado de F o t que de una alternativa no paramétrica.

Es natural que los investigadores sociales ansíen rechazar la hipótesis nula cuando ésta es falsa. Como resultado, muchos de ellos preferirían emplear idealmente pruebas de significancia paramétricas. Sin embargo, como ya se anotó, frecuentemente no es posible satisfacer los requisitos de las pruebas paramétricas. En primer lugar, muchos de los datos de la investigación social están al nivel de medición ordinal o nominal. En segundo lugar, no siempre podemos estar seguros de que las características que se estudian están de hecho distribuidas normalmente en la población.

No es posible conocer la potencia de una prueba estadística cuando se han violado sus requisitos. Por lo tanto, los resultados de una prueba paramétrica cuyos requisitos no se han llenado carecen de interpretación significativa. Bajo tales condiciones, muchos investigadores sociales recurren sabiamente a las pruebas de significancia no paramétricas.

Este capítulo presenta algunas de las pruebas de significancia más conocidas: la chi cuadrada, la prueba de la mediana, el análisis de varianza en una dirección de Kruskal-Wallis y el análisis de varianza en dos direcciones de Friedman.

CHI CUADRADA COMO UNA PRUEBA DE SIGNIFICANCIA

La prueba de significancia no paramétrica más popular en la investigación social se conoce como *chi cuadrada* (χ^2). Como veremos, la prueba χ^2 se usa para hacer comparaciones entre dos o más muestras.

Como en el caso de la razón t y el análisis de varianza, hay una distribución muestral para chi cuadrada que se puede usar para estimar la probabilidad de obtener por mera casualidad un valor de chi cuadrada significativo más que por diferencias poblacionales reales. Sin embargo, a diferencia de las anteriores pruebas de significancia, chi cuadrada se emplea para hacer comparaciones entre *frecuencias* más que entre puntajes medios. Como resultado la hipótesis nula para la prueba chi cuadrada establece que las poblaciones no difieren con respecto a la frecuencia de ocurrencia de una característica dada, en tanto que la hipótesis de investigación dice que las diferencias muestrales reflejan diferencias poblacionales reales en cuanto a la frecuencia relativa de una característica dada.

Con el fin de ilustrar el uso de chi cuadrada para los datos de frecuencia (o para proporciones que pueden reducirse a frecuencias), imaginemos que se nos ha

pedido investigar una vez más la relación entre la orientación política y la permisibilidad en la crianza de los niños. Más que *llevar una cuenta* de los liberales y los conservadores, en términos de su grado de permisibilidad, podríamos *categorizar* los miembros de nuestra muestra estrictamente sobre la base de *uno u otro;* esto es, podríamos decidir que *o* son rígidos *o* que no lo son. Por lo tanto,

Hipótesis Nula: *La frecuencia relativa de los liberales que no son rígidos es la misma que la de los conservadores que son rígidos.*

Hipótesis de Investigación: *La frecuencia relativa de los liberales que no son rígidos no es la misma que la de los conservadores que son rígidos.*

CALCULO DE CHI CUADRADA

La prueba de significancia chi cuadrada tiene que ver esencialmente con la distinción entre las frecuencias esperadas y las frecuencias obtenidas. Las *frecuencias esperadas* (f_e) se refieren a los términos de la hipótesis nula, de acuerdo con la cual se espera que la frecuencia relativa (o proporción) sea la misma de un grupo a otro. Por ejemplo, si se espera que el 50% de los liberales no sea rígido, entonces también esperamos que el 50% de los conservadores tampoco lo sea. En contraste, las *frecuencias obtenidas* (f_o) se refieren a los resultados que obtenemos realmente al realizar un estudio y, por lo tanto, pueden variar o no de un grupo a otro. *Sólo si la diferencia entre las frecuencias esperadas y obtenidas es lo suficientemente grande, rechazamos la hipótesis nula y decidimos que existe una diferencia poblacional verdadera.*

Continuando con el mismo ejemplo, supóngase que fuéramos a extraer muestras aleatorias de 20 liberales y 20 conservadores, quienes podrían categorizar como no rígidos o como rígidos respecto a los métodos de crianza de los niños. La Tabla 10.1 muestra las frecuencias obtenidas que podrían resultar.

Los datos de la Tabla 10.1 indican que 5 de 20 liberales y 10 de 20 conservadores usaron métodos no rígidos de crianza de los niños. Estos resultados se pueden volver a escribir en una tabla 2 × 2 (2 renglones por 2 columnas), en la que se presentan las frecuencias obtenidas para cada casilla y entre paréntesis se muestran sus frecuencias esperadas (ver Tabla 10.2). Nótese que estas frecuencias esperadas se basan en la operación de la simple casualidad, suponiendo por tanto que la hipótesis nula es correcta. Nótese también que los totales marginales de la Tabla 10.2 (que se obtienen sumando las frecuencias por casilla en una u otra dirección) están dados para los renglones (15 y 25) y las columnas (20 y 20). El número total ($N = 40$) puede obtenerse sumando los marginales de renglón o de columna.

Habiéndose dado las frecuencias obtenidas y esperadas para el problema por resolver, ahora podemos obtener el valor de chi cuadrada por la fórmula

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

TABLA 10.1 Frecuencias obtenidas en un estudio de permisibilidad según la orientación política

Métodos de crianza de los niños	*Orientación política*	
	Liberales f_o	*Conservadores* f_o
Rígidos	5	10
No rígidos	15	10
Total	20	20

TABLA 10.2 Los datos de la Tabla 10.1 colocados en una Tabla 2 × 2

	Liberales	*conservadores*	
No rígidos	5 (7,5)	10 (7,5)	15
Rígidos	15 (12,5)	10 (12,5)	25
	20	20	$N = 40$

Frecuencia obtenida — *Frecuencia esperada* — *Un total marginal*

donde

f_o = la frecuencia obtenida en cualquier casilla
f_e = la frecuencia esperada en cualquier casilla
χ^2 = chi cuadrada

De acuerdo con la fórmula para χ^2 debemos restar cada frecuencia esperada de su correspondiente frecuencia obtenida, elevar al cuadrado la diferencia, dividir entre la frecuencia esperada apropiada y sumar estos cocientes para obtener el valor de chi cuadrada.

Los datos de la Tabla 10.2 pueden usarse para ilustrar el procedimiento anterior:

$$\chi^2 = \frac{(5 - 7{,}5)^2}{7{,}5} + \frac{(10 - 7{,}5)^2}{7{,}5} + \frac{(15 - 12{,}5)^2}{12{,}5} + \frac{(10 - 12{,}5)^2}{12{,}5}$$

$$= \frac{(-2{,}5)^2}{7{,}5} + \frac{(2{,}5)^2}{7{,}5} + \frac{(2{,}5)^2}{12{,}5} + \frac{(-2{,}5)^2}{12{,}5}$$

$$= \frac{6,25}{7,5} + \frac{6,25}{7,5} + \frac{6,25}{12,5} + \frac{6,25}{12,5}$$
$$= 0,83 + 0,83 + 0,50 + 0,50$$
$$= 2,66$$

Así encontramos que $\chi^2 = 2,66$. Para interpretar este valor de chi cuadrada, debemos determinar aún el número apropiado de grados de libertad. Esto puede hacerse por medio de tablas, teniendo cualquier número de renglones y columnas y empleando la fórmula

$$gl = (r - 1)(c - 1)$$

donde

r = el número de renglones en la tabla de frecuencias obtenidas
c = el número de columnas en la tabla de frecuencias obtenidas
gl = los grados de libertad

Puesto que las frecuencias obtenidas en la Tabla 10.2 forman dos renglones y dos columnas (2 × 2),

$$gl = (2 - 1)(2 - 1)$$
$$= (1)(1)$$
$$= 1$$

Consultando la Tabla E al final del texto, encontramos una lista de valores de chi cuadrada que son significativos a los niveles de confianza de 0,05 y 0,01. Para el nivel de confianza de 0,05 vemos que el valor de chi cuadrada con 1 grado de libertad es de 3,84. Este es el valor que debemos igualar o exceder antes de poder rechazar la hipótesis nula. Ya que la χ^2 que hemos calculado es de sólo 2,66 y, por consiguiente, menor que el valor de la tabla, debemos aceptar la hipótesis nula y rechazar la hipótesis de investigación. Las frecuencias obtenidas no difieren lo suficiente de las frecuencias al azar esperadas para indicar que existen diferencias poblacionales reales.

COMO BUSCAR LAS FRECUENCIAS ESPERADAS

Las frecuencias esperadas para cada casilla deben reflejar la operación del azar bajo los términos de la hipótesis nula. Si las frecuencias esperadas deben indicar "semejanza" a través de todas las muestras, deben ser proporcionales a sus totales marginales tanto para los renglones como para las columnas.

Para obtener la frecuencia esperada para cualquier casilla, simplemente multiplicamos los totales marginales de columna y de renglón para una casilla determinada y dividimos el producto entre N. Por lo tanto,

$$f_e = \frac{\text{(total marginal de renglón) (total marginal de columna)}}{N}$$

Para la casilla superior izquierda en la Tabla 10.2 (liberales no rígidos),

$$f_e = \frac{(20)(15)}{40}$$
$$= \frac{300}{40}$$
$$= 7{,}5$$

Igualmente, para la casilla superior derecha en la Tabla 10.2 (conservadores no rígidos),

$$f_e = \frac{(20)(15)}{40}$$
$$= \frac{300}{40}$$
$$= 7{,}5$$

Para la casilla inferior de la izquierda en la Tabla 10.2 (liberales rígidos),

$$f_e = \frac{(20)(25)}{40}$$
$$= \frac{500}{40}$$
$$= 12{,}5$$

Para la casilla inferior derecha en la Tabla 10.2 (conservadores rígidos),

$$f_e = \frac{(20)(25)}{40}$$
$$= \frac{500}{40}$$
$$= 12{,}5$$

Como veremos, el método anterior para determinar f_e puede aplicarse a cualquier problema de chi cuadrada para los cuales las frecuencias esperadas deben obtenerse.

Una ilustración

Para resumir el procedimiento paso a paso para obtener chi cuadrada, supongamos que queremos estudiar el uso de la mariguana en estudiantes de bachillerato en relación a sus planes de ingreso a la universidad. Podríamos especificar nuestra hipótesis como sigue:

Hipótesis Nula: *La proporción de fumadores de mariguana entre los estudiantes de bachillerato orientados hacia la universidad es igual a la de los estudiantes que no piensan asistir a la universidad.*

Hipótesis de Investigación: *La proporción de fumadores de mariguana entre los estudiantes de bachillerato orientados hacia la universidad no es igual a la de los estudiantes que no piensan asistir a la universidad.*

Para verificar esta hipótesis al nivel de confianza de 0,05, digamos que debemos entrevistar a dos muestras aleatorias de la población de una escuela de bachillerato acerca del uso de la mariguana: una muestra de 21 estudiantes que van a ingresar a la universidad y una muestra de 15 estudiantes que no planean extender su educación más allá del bachillerato. Supóngase que resultaran los datos de la Tabla 10.3.

TABLA 10.3 Uso de la mariguana entre estudiantes orientados y no orientados hacia la universidad

Uso de la mariguana	*Orientación hacia la Universidad*	
	Universidad f_o	*No universidad* f_o
Fumadores	15	5
No fumadores	6	10
Total	21	15

Como se muestra en la Tabla, 15 de 21 estudiantes orientados hacia la universidad, pero sólo 5 de 15 no orientados hacia ella, eran fumadores de mariguana. Para averiguar si esta es una diferencia significativa entre los estudiantes de bachillerato orientados hacia la universidad y los estudiantes no orientados hacia ésta, desarrollamos el siguiente procedimiento paso a paso:

PASO 1: Reordenar los datos en forma de Tabla 2 X 2

	Universidad	*No universidad*	
Fumadores	15 ( )	5 ( )	20
No fumadores	6 ( )	10 ( )	16
	21	15	$N = 36$

PASO 2: Obtener la frecuencia esperada para cada casilla

15 (11,67)	5 (8,33)	20
6 (9,33)	10 (6,67)	16
21	15	$N = 36$

$$\text{(superior izquierda) } f_e = \frac{(21)(20)}{36} = \frac{420}{36} = 11{,}67$$

$$\text{(superior derecha) } f_e = \frac{(15)(20)}{36} = \frac{300}{36} = 8.33$$

$$\text{(inferior izquierda) } f_e = \frac{(21)(16)}{36} = \frac{336}{36} = 9{,}33$$

$$\text{(inferior derecha) } f_e = \frac{(15)(16)}{36} = \frac{240}{36} = 6{,}67$$

PASO 3: Restar las frecuencias esperadas de las frecuencias obtenidas

	$f_o - f_e$
(superior izquierda)	$15 - 11{,}67 = 3{,}33$
(superior derecha)	$5 - 8{,}33 = -3{,}33$
(inferior izquierda)	$6 - 9{,}33 = -3{,}33$
(inferior derecha)	$10 - 6{,}67 = 3{,}33$

PASO 4: Elevar al cuadrado esta diferencia

	$(f_o - f_e)^2$
(superior izquierda)	$(3{,}33)^2 = 11{,}09$
(superior derecha)	$(-3{,}33)^2 = 11{,}09$
(inferior izquierda)	$(-3{,}33)^2 = 11{,}09$
(inferior derecha)	$(3{,}33)^2 = 11{,}09$

PASO 5: Dividir entre la frecuencia esperada

	$\frac{(f_o - f_e)^2}{f_e}$
(superior izquierda)	$\frac{11{,}09}{11{,}67} = 0{,}95$
(superior derecha)	$\frac{11{,}09}{8{,}33} = 1{,}33$

$$\text{(inferior izquierda)} \quad \frac{11{,}09}{9{,}33} = 1{,}19$$
$$\text{(inferior derecha)} \quad \frac{11{,}09}{6{,}67} = 1{,}66$$

PASO 6: Sumar estos cocientes para obtener el valor de chi cuadrada

$$\Sigma \frac{(f_o - f_e)^2}{f_e}$$
$$0{,}95$$
$$1{,}33$$
$$1{,}19$$
$$1{,}66$$
$$\chi^2 = 5{,}13$$

PASO 7: Encontrar los grados de libertad

$$\begin{aligned} gl &= (r - 1)(c - 1) \\ &= (2 - 1)(2 - 1) \\ &= (1)(1) \\ &= 1 \end{aligned}$$

PASO 8: Comparar el valor de chi cuadrada obtenido con el valor de chi cuadrada correspondiente en la Tabla E

$$\text{obtenido } \chi^2 = 5{,}13$$
$$\text{de la tabla } \chi^2 = 3{,}84$$
$$gl = 1$$
$$P = 0{,}05$$

Como se indica en el Paso 8, para rechazar la hipótesis nula, al nivel de confianza de 0,05 con 1 grado de libertad, nuestro valor de chi cuadrada calculado tendría que ser de 3,84 o más. Como hemos obtenido un valor de chi cuadrada de 5,13, podemos rechazar la hipótesis nula y aceptar la hipótesis de investigación. Nuestros resultados sugieren que la proporción de fumadores de mariguana es mayor entre los estudiantes de bachillerato que van a ingresar a la universidad que entre los estudiantes cuyos planes no incluyen el ingreso a la universidad.

El procedimiento que se acaba de ilustrar paso a paso, para la obtención de chi cuadrada, se puede resumir en forma de tabla:

	f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
(superior izquierda)	15	11,67	3,33	11,09	0,95
(superior derecha)	5	8,33	−3,33	11,09	1,33
(inferior izquierda)	6	9,33	−3,33	11,09	1,19
(inferior derecha)	10	6,67	3,33	11,09	1,66
					$\chi^2 = 5{,}13$

UNA FORMULA 2 × 2 PARA CALCULAR CHI CUADRADA

Podemos evitar el largo proceso de calcular las frecuencias esperadas para un problema de chi cuadrada de 2 × 2 (2 renglones por 2 columnas) usando la siguiente fórmula de cálculo:

$$\chi^2 = \frac{N(AD - BC)^2}{(A + B)(C + D)(A + C)(B + D)}$$

donde:

A = la frecuencia obtenida en la casilla superior izquierda
B = la frecuencia obtenida en la casilla superior derecha
C = la frecuencia obtenida en la casilla inferior izquierda
D = la frecuencia obtenida en la casilla inferior derecha
N = el número total en todas las casillas

Graficamos las casillas A, B, C y D y sus totales marginales en una tabla 2 × 2 como sigue:

A	B	$A + B$
C	D	$C + D$
$A + C$	$B + D$	N

Para ilustrar el uso de la fórmula para calcular chi cuadrada, regresamos a los datos de la Tabla 10.3 (uso de la mariguana según la orientación hacia la universidad) para los cuales ya se ha obtenido un valor χ^2 de 5,13. Podemos colocar, las frecuencias obtenidas para la fórmula de cálculo, de la manera siguiente:

15 A	5 B
C 6	D 10

Aplicando la fórmula de cálculo,

$$\chi^2 = \frac{36[(15)(10) - (5)(6)]^2}{(15 + 5)(6 + 10)(15 + 6)(5 + 10)}$$

$$= \frac{36(150 - 30)^2}{(20)(16)(21)(15)}$$

$$= \frac{36(120)^2}{100800}$$

$$= \frac{36(14400)}{100800}$$

$$= \frac{518400}{100800}$$

$$= 5{,}14$$

CORRECCIONES PARA PEQUEÑAS FRECUENCIAS ESPERADAS

Si las frecuencias esperadas en un problema de chi cuadrada 2 × 2 son muy pequeñas (menos de 10 en una casilla), las fórmulas que hemos aprendido hasta aquí pueden producir un valor de chi cuadrada inflado. Nótese que esto es cierto sólo para las frecuencias *esperadas* y no para las frecuencias obtenidas realmente en el curso de la investigación, las cuales pueden ser de cualquier tamaño.

Para reducir la sobreestimación de chi cuadrada y obtener un resultado más conservador, aplicamos lo que se conoce como la *corrección de Yates* a la situación 2 × 2. Usando la corrección de Yates, la diferencia entre las frecuencias obtenidas y esperadas se reduce en 0,50. Ya que χ^2 depende de la magnitud de esa diferencia, también reducimos el tamaño de nuestro valor calculado para chi cuadrada. La fórmula de chi cuadrada corregida para pequeñas frecuencias esperadas es la siguiente:

$$\chi^2 = \Sigma \frac{(|f_o - f_e| - 0{,}50)^2}{f_e}$$

En la fórmula anterior corregida, las líneas rectas que encierran $f_o - f_e$ indican que debemos reducir el valor absoluto (ignorando los signos menos) de cada $f_o - f_e$ en 0,50.

Apliquemos a los datos de la Tabla 10.3 la fórmula corregida:

$$\chi^2 = \frac{(|15 - 11{,}67| - 0{,}50)^2}{11{,}67} + \frac{(|5 - 8{,}33| - 0{,}50)^2}{8{,}33}$$
$$+ \frac{(|6 - 9{,}33| - 0{,}50)^2}{9{,}33} + \frac{(|10 - 6{,}67| - 0{,}50)^2}{6{,}67}$$

$$= \frac{(3{,}33 - 0{,}50)^2}{11{,}67} + \frac{(3{,}33 - 0{,}50)^2}{8{,}33}$$
$$+ \frac{(3{,}33 - 0{,}50)^2}{9{,}33} + \frac{(3{,}33 - 0{,}50)^2}{6{,}67}$$

$$= \frac{(2{,}83)^2}{11{,}67} + \frac{(2{,}83)^2}{8{,}33} + \frac{(2{,}83)^2}{9{,}33} + \frac{(2{,}83)^2}{6{,}67}$$

$$= \frac{8{,}01}{11{,}67} + \frac{8{,}01}{8{,}33} + \frac{8{,}01}{9{,}33} + \frac{8{,}01}{6{,}67}$$

$$= 0{,}69 + 0{,}96 + 0{,}86 + 1{,}20$$

$$= 3{,}71$$

El procedimiento para aplicar la fórmula de chi cuadrada corregida se puede resumir en forma de tabla:

f_o	f_e	$\lvert f_o - f_e \rvert$	$\lvert f_o - f_e \rvert - 0{,}50$
15	11,67	3,33	2,83
5	8,33	3,33	2,83
6	9,33	3,33	2,83
10	6,67	3,33	2,83

$(\lvert f_o - f_e \rvert - 0{,}50)^2$	$\dfrac{(\lvert f_o - f_e \rvert - 0{,}50)^2}{f_e}$
8,01	0,69
8,01	0,96
8,01	0,86
8,01	1,20
	$\chi^2 = 3{,}71$

Como se muestra arriba, la corrección de Yates produce un valor de chi cuadrada menor ($\chi^2 = 3{,}71$) que el que se obtenía mediante la fórmula no corregida ($\chi^2 = 5{,}13$). En el presente ejemplo, nuestra decisión con respecto a la hipótesis nula dependería de si hemos usado o no la corrección de Yates. Con la fórmula corregida, aceptamos la hipótesis nula; sin ella, la rechazamos.

La corrección de Yates también se puede aplicar a la fórmula para calcular una chi cuadrada 2 × 2 como sigue:

$$\chi^2 = \frac{N(|AD - BC| - N/2)^2}{(A + B)(C + D)(A + C)(B + D)}$$

Regresando a los datos de la Tabla 10.3,

$$\chi^2 = \frac{36[|(15)(10) - (5)(6)| - 36/2]^2}{(15 + 5)(6 + 10)(15 + 6)(5 + 10)}$$

$$= \frac{36(|150 - 30| - 18)^2}{(20)(15)(21)(15)}$$

$$= \frac{36(120 - 18)^2}{100800}$$

$$= \frac{36(102)^2}{100800}$$

$$= \frac{36(10404)}{100800}$$

$$= \frac{374544}{100800}$$

$$= 3{,}71$$

COMPARANDO VARIOS GRUPOS

Hasta aquí, hemos limitado nuestras ilustraciones al problema 2 × 2 ampliamente usado. Sin embargo, deberá enfatizarse que chi cuadrada se calcula frecuentemente para tablas mayores que 2 × 2, tablas en que se han de comparar varios grupos o categorías. El procedimiento paso a paso para comparar varios grupos es esencialmente igual a su contraparte 2 × 2. Ejemplifiquemos con un problema 3 × 3 (3 renglones por 3 columnas), aunque se podría usar cualquier número de renglones y columnas.

Imagínese una vez más que estuviéramos investigando la relación entre la orientación política y los métodos de crianza de los niños. Sin embargo, en esta ocasión digamos que pudimos presentar tres muestras aleatorias: 32 conservadores, 30 moderados, y 27 liberales. Supóngase, además que fuéramos a categorizar los métodos de crianza de los niños, de los miembros de nuestra muestra, como no rígidos, moderados o autoritarios. Por lo tanto,

Hipótesis Nula: *La frecuencia relativa de los métodos no rígidos, moderados y autoritarios de crianza de los niños es igual para liberales, moderados y conservadores.*

Hipótesis de Investigación: *La frecuencia relativa de los métodos no rígidos, moderados y autoritarios de crianza de los niños no es igual para liberales, moderados y conservadores.*

Digamos que generamos las diferencias muestrales, en cuanto a métodos de crianza de los niños, que se muestran en la Tabla 10.4. Allí vemos que 7 de 32 conservadores, 9 de 30 moderados y 14 de 27 liberales pueden considerarse no rígidos en sus prácticas de crianza de los niños.

TABLA 10.4 Crianza de los niños según la orientación política: un problema 3 × 3

Método de crianza de los niños	*Orientación política*		
	Conservador f_o	*Moderado* f_o	*Liberal* f_o
No rígido	7	9	14
Moderado	10	10	8
Autoritario	15	11	5
Total	32	30	27

Debe tenerse en cuenta que la corrección de Yates y la fórmula 2 × 2 para calcular χ^2 sólo se aplican al problema 2 × 2 y por lo tanto no pueden utilizarse para comparar varios grupos, como en la presente situación 3 × 3. Para determinar si hay o no una diferencia significativa en la Tabla 10.4, debemos aplicar la fórmula original χ^2 que se presentó anteriormente:

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

La anterior fórmula para chi cuadrada puede aplicársele al problema 3 X 3 en el siguiente procedimiento paso a paso:

PASO 1: Reordenar los datos en forma de una Tabla 3 X 3

Métodos de crianza de los niños	*Orientación política: Conservadores*	*Moderados*	*Liberales*	
No rígidos	7	9	14	30
Moderados	10	10	8 *Frecuencia obtenida*	28
Autoritarios	15	11	5	31
	32	30	27	$N = 89$

Total marginal

PASO 2: Obtener la frecuencia esperada para cada casilla

7 (10,79)	9 (10,11)	14 (9,10)	30
10 (10,07)	10 (9,44)	8 (8,49)	28
15 (11,14)	11 (10,45)	5 (9,40)	31
32	30	27	$N = 89$

(superior izquierda) $f_e = \frac{(30)(32)}{89}$

$= \frac{960}{89}$

$= (10,79)$

28 (central izquierda) $f_e = \frac{(28)(32)}{89}$

$= \frac{896}{89}$

31 $= 10,07$

(inferior izquierda) $f_e = \frac{(31)(32)}{89}$

$= \frac{992}{89}$

$= 11,14$

(central superior) $f_e = \frac{(30)(30)}{89} = \frac{900}{89} = 10{,}11$

(central central) $f_e = \frac{(28)(30)}{89} = \frac{840}{89} = 9{,}44$

(central inferior) $f_e = \frac{(31)(30)}{89} = \frac{930}{89} = 10{,}45$

(superior derecha) $f_e = \frac{(30)(27)}{89} = \frac{810}{89} = 9{,}10$

(central derecha) $f_e = \frac{(28)(27)}{89} = \frac{756}{89} = 8{,}49$

(inferior derecha) $f_e = \frac{(31)(27)}{89} = \frac{837}{89} = 9{,}40$

PASO 3: Restar las frecuencias esperadas de las frecuencias obtenidas

	$f_o - f_e$
(superior izquierda)	$7 - 10{,}79 = -3{,}79$
(central izquierda)	$10 - 10{,}07 = -0{,}07$
(inferior izquierda)	$15 - 11{,}14 = 3{,}86$
(superior central)	$9 - 10{,}11 = -1{,}11$
(central central)	$10 - 9{,}44 = 0{,}56$
(inferior central)	$11 - 10{,}45 = 0{,}55$
(superior derecha)	$14 - 9{,}10 = 4{,}90$
(central derecha)	$8 - 8{,}49 = -0{,}49$
(inferior derecha)	$5 - 9{,}40 = -4{,}40$

PASO 4: Elevar al cuadrado esta diferencia

	$(f_o - f_e)^2$
(superior izquierda)	$(-3{,}79)^2 = 14{,}36$
(central izquierda)	$(-0{,}07)^2 = 0{,}01$
(inferior izquierda)	$(3{,}86)^2 = 14{,}90$
(superior central)	$(-1{,}11)^2 = 1{,}23$
(central central)	$(0{,}56)^2 = 0{,}31$
(inferior central)	$(0{,}55)^2 = 0{,}30$
(superior derecha)	$(4{,}90)^2 = 24{,}01$
(central derecha)	$(-0{,}49)^2 = 0{,}24$
(inferior derecha)	$(-4{,}40)^2 = 19{,}36$

PASO 5: Dividir entre la frecuencia esperada

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

(superior izquierda) $\frac{14,36}{10,79} = 1,33$

(central izquierda) $\frac{0,01}{10,07} = 0,00$

(inferior izquierda) $\frac{14,90}{11,14} = 1,34$

(superior central) $\frac{1,23}{10,11} = 0,12$

(central central) $\frac{0,31}{9,44} = 0,03$

(inferior central) $\frac{0,30}{10,45} = 0,03$

(superior derecha) $\frac{24,01}{9,10} = 2,64$

(central derecha) $\frac{0,24}{8,49} = 0,03$

(inferior derecha) $\frac{19,36}{9,40} = 2,06$

PASO 6: Sumar estos cocientes para obtener el valor de chi cuadrada

$$\Sigma \frac{(f_o - f_e)^2}{f_e}$$

1,33
0,00
1,34
0,12
0,03
0,03
2,64
0,03
2,06

$\chi^2 = 7,58$

PASO 7: Encontrar el número de grados de libertad

$$\begin{aligned} gl &= (r - 1)(c - 1) \\ &= (3 - 1)(3 - 1) \\ &= (2)(2) \\ &= 4 \end{aligned}$$

PASO 8: Comparar el valor de chi cuadrada obtenido con el valor de chi cuadrada correspondiente en la Tabla E

$$\chi^2 \text{ obtenido} = 7{,}58$$
$$\chi^2 \text{ en la tabla} = 9{,}49$$
$$gl = 4$$
$$P = 0{,}05$$

Por lo tanto, necesitamos un valor de chi cuadrada de por lo menos 9,49 para rechazar la hipótesis nula. Dado que nuestra χ^2 obtenida es de sólo 7,58, debemos aceptar la hipótesis nula y atribuir nuestras diferencias muestrales a la operación de la simple casualidad. No hemos descubierto evidencias estadísticamente significativas que indiquen que la frecuencia relativa de los métodos de crianza de los niños difiere para los liberales, los moderados y los conservadores.

REQUISITOS PARA EL USO DE CHI CUADRADA

A pesar del hecho de que las pruebas no paramétricas no suponen una distribución normal en la población, también tienen una serie de requisitos que el investigador social debe tomar en cuenta si ha de hacer una selección inteligente entre las pruebas de significancia. El estudiante notará, sin embargo, que los requisitos para el uso de las pruebas no paramétricas son generalmente más fáciles de satisfacer que aquéllos para el uso de sus contrapartes paramétricas, tales como la razón t o el análisis de varianza. Teniendo esto en mente, veamos algunos de los requisitos más importantes para el uso de la prueba de significancia chi cuadrada:

1. Una comparación entre dos o más muestras: como se describió e ilustró en el presente capítulo, la prueba chi cuadrada se emplea para hacer comparaciones entre dos o más muestras *independientes.* Esto requiere que tengamos por lo menos una tabla 2 × 2 (por lo menos 2 renglones y 2 columnas). La suposición de independencia indica que chi cuadrada no puede aplicarse a una sola muestra colocada en un diseño de panel antes/después. Deben obtenerse por lo menos dos muestras de entrevistados.
2. Los datos nominales: sólo se requieren las frecuencias.
3. El muestreo aleatorio: debimos haber extraído nuestras muestras aleatoriamente de una población determinada.
4. Las frecuencias esperadas por casilla no deben ser demasiado pequeñas: el tamaño exacto de f_e depende de la naturaleza del problema. Para un problema 2 × 2, ninguna frecuencia esperada deberá ser menor que 5. Además, la fórmula corregida de Yates deberá usarse para un problema 2 × 2 en el cual una frecuencia esperada por casilla es menor que 10. Para una situación en la cual se están comparando varios grupos (digamos un problema 3 × 3 o 4 × 5), no existe ninguna regla rápida y rígida respecto al

mínimo de frecuencias por casilla, aunque deberemos tener cuidado de ver que pocas casillas contengan menos de 5 casos. En cualquier evento, las frecuencias esperadas para todas las casillas combinadas (Σf_e) deben ser siempre iguales a las frecuencias obtenidas para todas las casillas combinadas (Σf_o).

LA PRUEBA DE LA MEDIANA

Se puede aplicar chi cuadrada a cualquier número de muestras independientes medidas al nivel nominal. Para datos ordinales, *la prueba de la mediana* es un procedimiento no paramétrico simple para determinar la probabilidad de que dos muestras aleatorias hayan sido tomadas de poblaciones con las mismas medianas.

A fin de ilustrar el procedimiento para realizar la prueba de la mediana, supóngase que un investigador quisiera estudiar las reacciones masculinas y femeninas ante una situación socialmente embarazosa. Para crear la turbación el investigador pidió a 15 hombres y 12 mujeres, quienes poseían una habilidad escasamente "promedio" para el canto, que interpretaran individualmente varias canciones, tales como "El amor es una cosa esplendorosa", ante un auditorio de "expertos". A continuación se muestra el número de minutos que cada sujeto estuvo dispuesto a continuar cantando (un menor periodo de tiempo indica supuestamente mayor turbación):

Número de minutos cantados			
Hombres	*Mujeres*	*Hombres*	*Mujeres*
15	12		
18	7	11	9
15	15	10	11
17	16	8	14
17	6	14	9
16	8	9	
10	10	18	
13	6	16	

PASO 1: Encontrar la mediana de las dos muestras combinadas. Por fórmula,

$$\text{Posición de la mediana} = \frac{N + 1}{2}$$

$$= \frac{27 + 1}{2}$$

$$= 14\text{o.}$$

La mediana es el decimocuarto puntaje contando de uno u otro extremo de la distribución arreglada por tamaños.

Para encontrar la mediana, ordenamos todos los puntajes para hombres y

mujeres en orden consecutivo (sin importar de qué muestra provienen) y localizamos su mediana combinada:

18
18
17
17
16
16
16
15
15
15
14
14
13
12 ← Mediana (el decimocuarto puntaje de uno u otro extremo)
11
11
10
10
10
9
9
9
8
8
7
6
6

PASO 2: Contar el número en cada muestra que cae por encima de la mediana y por abajo de ella (Mdn = 12)

	Hombres f	*Mujeres* f
Sobre la mediana	10	3
Abajo de la mediana	5	9
	$N = 27$	

Como se vio anteriormente, el número que representa el tiempo de canto arriba y abajo de la mediana de cada muestra de hombres y mujeres se representa en una tabla de frecuencia 2 × 2. En el presente ejemplo, 10 de los 15 hombres, pero sólo 3 de las 12 mujeres, continuaron cantando por un periodo de tiempo mayor que el tiempo mediano de canto para la totalidad del grupo.

PASO 3: Realizar una prueba de significancia chi cuadrada. Si no existen diferencias de sexo respecto al tiempo de canto (y, por lo tanto, de turbación social), esperaríamos que la misma mediana se dividiera dentro de cada muestra, de manera que la mitad de los hombres y la mitad de las mujeres cayeran sobre la mediana. Para determinar si las diferencias de sexo obtenidas son estadísticamente significativas o sólo un producto del error de muestreo, realizamos el análisis de χ^2.

	Hombres	*Mujeres*
Sobre la mediana	10 (*A*)	3 (*B*)
Abajo de la mediana	5 (*C*)	9 (*D*)
	$N = 27$	

$$\chi^2 = \frac{N(|AD - BC| - N/2)^2}{(A + B)(C + D)(A + C)(B + D)}$$

$$= \frac{27[|(10)(9) - (3)(5)| - \frac{27}{2}]^2}{(10 + 3)(5 + 9)(10 + 5)(3 + 9)}$$

$$= \frac{27(75 - 13{,}5)^2}{32760}$$

$$= \frac{102120{,}75}{32760}$$

$$= 3{,}12$$

Al buscar en la Tabla E, al final del texto, encontramos que χ^2 debe ser igual o mayor que 3,84 (gl = 1) para poder considerarlo significativo al nivel 0,05. Como nuestra χ^2 obtenida es de 3,12, no podemos rechazar la hipótesis nula. No hay evidencias suficientes para concluir, con base en nuestros resultados, que los hombres difieren de las mujeres respecto a sus reacciones ante una situación socialmente embarazosa.

Requisitos para el uso de la prueba de la mediana

Las siguientes condiciones deben cumplirse para poder aplicar adecuadamente la prueba de la mediana a un problema de investigación.

1. Una comparación entre dos o más medianas independientes: la prueba de la mediana se emplea para hacer comparaciones entre dos o más medianas de muestras independientes.
2. Los datos ordinales: para realizar la prueba de la mediana, suponemos por lo menos el nivel ordinal de medición. Los datos nominales no se pueden usar.
3. El muestreo aleatorio: debimos haber extraído nuestras muestras sobre una base aleatoria de una población dada.

EL ANALISIS DE VARIANZA EN DOS DIRECCIONES POR RANGOS DE FRIEDMAN

En el Capítulo 8 presentamos una variación de la razón t que se podía usar para comparar la misma muestra medida dos veces. Por ejemplo, en el diseño antes/después podría medirse el grado de hostilidad en una muestra de niños antes y después de mirar un violento programa de televisión.

El análisis de varianza en dos direcciones por rangos de Friedman (χ_r^2) constituye un enfoque no paramétrico para verificar las diferencias en una sola muestra de entrevistados a quienes se ha medido al menos bajo dos condiciones.

Por fórmula,

$$\chi_r^2 = \frac{12}{Nk(k+1)} \Sigma (\Sigma R_i)^2 - 3N(k+1)$$

donde

k = el número de mediciones (representa usualmente las condiciones bajo las cuales se estudia a los entrevistados)
N = el número total de entrevistados
ΣR_i = la suma de los rangos para una medición cualquiera (usualmente representa una condición cualquiera en estudio)

Una ilustración

Para ilustrar la aplicación del análisis de varianza en dos direcciones de Friedman, supóngase que deseamos comprobar la hipótesis de que la hostilidad de los niños varía según el nivel de violencia en sus programas de televisión. Con el fin de estudiar la influencia de la violencia televisada, imaginemos que podemos exponer una muestra aleatoria de diez niños a tres distintos niveles de violencia en un programa que es esencialmente igual en todos los demás aspectos. Digamos también que hemos obtenido los siguientes puntajes de hostilidad de estos 10 niños bajo cada condición como espectador de televisión (los puntajes van desde 20 hasta 60; los puntajes más altos representan mayor hostilidad):

PASO 1: Colocar por grados los puntajes de cada entrevistado a través de todas las condiciones (en cada renglón). Para realizar el análisis de varianza en dos direcciones de Friedman, trabajamos directamente con los rangos para cada entrevistado sobre todas las mediciones.[2] Como se muestra arriba, el nivel de hostilidad del niño A

[2] En este ejemplo no hubo empates entre rangos. En caso de rangos empatados (por ejemplo, si el nivel de hostilidad del niño A hubiera sido el mismo para dos o más niveles de violencia) sígase el procedimiento para tratar con rangos empatados como se presentan, en relación con el coeficiente de correlación del orden de los rangos, en el Capítulo 11.

	Condición como espectador		
Niño	*Violencia baja*	*Violencia mediana*	*Violencia alta*
A	23	30	32
B	41	45	43
C	36	35	39
D	28	29	35
E	39	41	47
F	25	28	27
G	38	46	51
H	40	47	49
I	45	46	42
J	29	34	38

aumentó de 23 a 30 y a 32 a medida que el nivel de violencia televisada, al que estaba expuesto, aumentaba de baja a mediana y a alta. Por rango, el puntaje de hostilidad del niño A fue mayor (1) a una violencia alta, un poco menor (2) a una violencia mediana y menor (3) a una violencia baja. Continuando hacia abajo, vemos que la hostilidad del niño B fue mayor (1) a una violencia mediana, un poco menor (2) a una violencia alta y menor (3) a una violencia baja. La del niño C fue mayor (1) a una violencia alta, un poco menor (2) a una violencia baja y menor (3) a una violencia mediana. El orden de los rangos de los tres puntajes de hostilidad de cada niño se muestra a continuación:

Niño	*Violencia baja*	*Rango*	*Violencia mediana*	*Rango*	*Violencia alta*	*Rango*
A	23	3	30	2	32	1
B	41	3	45	1	43	2
C	36	2	35	3	39	1
D	28	3	29	2	35	1
E	39	3	41	2	47	1
F	25	3	28	1	27	2
G	38	3	46	2	51	1
H	40	3	47	2	49	1
I	45	2	46	1	42	3
J	29	3	34	2	38	1

PASO 2: Sumar los rangos bajo cada condición (para cada columna). Si la hipótesis nula es correcta —y no ocurren diferencias significativas entre las condiciones— podemos esperar que las sumas de los rangos a través de las condiciones sean iguales entre sí (menos el error de muestreo). En el presente ejemplo hay tres condiciones: violencia televisada baja, mediana y alta. Los rangos para cada una de estas condiciones se suman como sigue:

Niño	*Rango (baja)*	*Rango (mediana)*	*Rango (alta)*
A	3	2	1
B	3	1	2
C	2	3	1
D	3	2	1
E	3	2	1
F	3	1	2
G	3	2	1
H	3	2	1
I	2	1	3
J	3	2	1
	$\Sigma R = 28$	$\Sigma R = 18$	$\Sigma R = 14$

PASO 3: Reemplazar en la fórmula para obtener χ_r^2

$$\chi_r^2 = \frac{12}{Nk(k+1)} \Sigma (\Sigma R_i)^2 - 3N(k+1)$$

$$= \frac{12}{(10)(3)(3+1)} (28^2 + 18^2 + 14^2) - 3(10)(3+1)$$

$$= \frac{12}{120} (784 + 324 + 196) - 120$$

$$= 0{,}10(1304) - 120$$

$$= 130{,}4 - 120$$

$$= 10{,}4$$

PASO 4: Encontrar el número de grados de libertad

$$\text{gl} = k - 1$$
$$= 3 - 1$$
$$= 2$$

PASO 5: Comparar χ_r^2 con el valor correspondiente de chi cuadrada en la Tabla E

χ_r^2 obtenido = 10,4
χ^2 de la tabla = 5,99
gl = 2
P = 0,05

χ_r^2 es en realidad un valor de chi cuadrada derivado de la suma de los rangos para todas las condiciones. Como resultado, podemos comparar nuestro χ_r^2 obtenido con el correspondiente χ^2 en la Tabla E. Con gl = 2 necesitamos un valor de chi cuadrada de por lo menos 5,99 a fin de rechazar la hipótesis nula. Ya que nuestro χ_r^2 obtenido es de 10,4, rechazamos la hipótesis nula y aceptamos la hipótesis de investigación. Hemos descubierto evidencias de que la violencia televisada sí induce

a la hostilidad en los niños. Hay diferencias significativas en la hostilidad según el nivel de violencia.

Requisitos para el uso del análisis de varianza en dos direcciones por rangos de Friedman

Para aplicar el análisis de varianza en dos direcciones de Friedman, deben cumplirse las siguientes condiciones:

1. Una comparación de una sola muestra medida bajo dos o más condiciones: el procedimiento de Friedman no se puede aplicar para contrastar diferencias entre muestras independientes, sino que supone que la misma muestra de entrevistados se ha medido por lo menos dos veces (o que los miembros de dos o más muestras se han comparado sobre variables apropiadas).
2. Los datos ordinales: sólo se requieren datos que puedan colocarse por rangos.
3. El número de entrevistados no debe ser demasiado pequeño: el requisito mínimo exacto para N depende del número de condiciones (k) a las que se va a exponer a los entrevistados. Por ejemplo, N debe ser igual o mayor que 10 cuando $k = 3$; en tanto que N debe ser igual o mayor que 5 cuando $k = 4$.

ANALISIS DE VARIANZA EN UNA DIRECCION POR RANGOS DE KRUSKAL-WALLIS

El análisis de varianza en una dirección de Kruskal-Wallis es una alternativa no paramétrica para el análisis de varianza (razón **F**) que puede usarse para comparar varias muestras independientes, pero que sólo requiere datos de nivel ordinal. Para aplicar el procedimiento de Kruskal-Wallis buscamos el estadístico **H** como sigue:

$$H = \frac{12}{N(N+1)} \sum \left[\frac{(\Sigma R_i)^2}{n} \right] - 3(N+1)$$

donde

N = el número total de casos c entrevistados
n = el número de casos en una muestra dada
ER_i = la suma de los rangos para una muestra dada.

Una ilustración

A fin de ilustrar el procedimiento para aplicar el análisis de varianza en una dirección por rangos, pensemos en la posible influencia de la edad sobre la capacidad de un individuo para encontrar empleo. Supóngase que estudiamos este problema tomando muestras aleatorias de adultos seniles, de edad mediana y jóvenes a quienes

se da un cierto número de días para encontrar empleo. Digamos que se obtuvieron los siguientes resultados:

Número de días antes de encontrar empleo		
Adultos seniles	*Adultos de edad mediana*	*Adultos jóvenes*
(n = 7) 63	(n = 8) 33	(n = 6) 25
20	42	31
43	27	6
58	28	14
57	51	18
71	64	13
45	12	
	30	

PASO 1: Ordenar por rango el grupo total de puntajes y encontrar la suma de los rangos para cada muestra. Todos los puntajes deben clasificarse por orden de menor a mayor (al puntaje *más pequeño* se le *debe* asignar un rango de 1; de 2 al que le sigue, y así sucesivamente). En este ejemplo, los puntajes se han ordenado desde 1 (que representa 6 días) hasta 21 (que representa 71 días).[3]

X_1	*Rango*	X_2	*Rango*	X_3	*Rango*
63	19	33	12	25	7
20	6	42	13	31	11
43	14	27	8	6	1
58	18	28	9	14	4
57	17	51	16	18	5
71	21	64	20	13	3
45	15	12	2		$\Sigma R_3 = 31$
	$\Sigma R_1 = 110$	30	10		
			$\Sigma R_2 = 90$		

PASO 2: Reemplazar en la fórmula para obtener H

$$H = \frac{12}{N(N+1)} \sum \left[\frac{(\Sigma R_i)^2}{n} \right] - 3(N+1)$$

$$= \left(\frac{12}{21(21+1)} \right) \left(\frac{110^2}{7} + \frac{90^2}{8} + \frac{31^2}{6} \right) - 3(21+1)$$

$$= \left(\frac{12}{462} \right) \left(\frac{12100}{7} + \frac{8100}{8} + \frac{961}{6} \right) - 66$$

$$= (0{,}03)(1728{,}57 + 1012{,}50 + 160{,}17) - 66$$

$$= (0{,}03)(2901{,}24) - 66$$

$$= 87{,}04 - 66$$

$$= 21{,}04$$

[3] En este ejemplo no hubo empates entre rangos. En caso de rangos empatados (por ejemplo, si dos personas demoran exactamente 24 días en encontrar trabajo) sígase el procedimiento para tratar rangos empatados como se presentan, en relación con el coeficiente de correlación de orden de los rangos, en el Capítulo 11.

PASO 3: Encontrar el número de grados de libertad

$$\begin{aligned} gl &= k - 1 \\ &= 3 - 1 \\ &= 2 \end{aligned}$$

PASO 4: Comparar H con el valor de chi cuadrada correspondiente en la Tabla E

$$\begin{aligned} H &= 21{,}04 \\ \chi^2 \text{ de la tabla} &= 5{,}991 \\ gl &= 2 \\ P &= 0{,}05 \end{aligned}$$

Para rechazar la hipótesis nula al nivel de confianza de 0,05 con 2 grados de libertad, nuestro H calculado tendría que ser 5,991 o más. Como hemos obtenido un H igual a 21,04, podemos rechazar la hipótesis nula y aceptar la hipótesis de investigación. Nuestros resultados indican que hay diferencias significativas, según la edad, en la cantidad de tiempo necesario para encontrar un empleo.

Requisitos para el uso del análisis de varianza en una dirección de Kruskal-Wallis

Para aplicar el análisis de varianza en una dirección por rangos debemos considerar los siguientes requisitos:

1. Una comparación de tres o más muestras independientes: el análisis de varianza en una dirección no se puede aplicar para contrastar diferencias dentro de una sola muestra de entrevistados que se midió más de una vez.
2. Los datos ordinales: sólo se requieren datos que puedan colocarse por rangos.
3. Cada muestra debe contener por lo menos 6 casos: cuando hay más de 5 entrevistados en cada grupo, la significancia de H puede determinarse por medio del valor correspondiente de chi cuadrada en la Tabla E. Para comprobar las diferencias entre muestras más pequeñas, recomendamos al lector las tablas especiales de Siegel (1956).

RESUMEN

Los estadísticos han desarrollado varias pruebas de significancia no paramétricas —pruebas cuyos requisitos no incluyen una distribución normal ni el nivel de medición por intervalos. La más conocida de ellas, la chi cuadrada, se emplea para hacer comparaciones entre frecuencias más que entre puntajes medios. Cuando la

diferencia entre las frecuencias esperadas y las frecuencias obtenidas es lo suficientemente grande rechazamos la hipótesis nula y aceptamos la validez de una diferencia poblacional real. Este es el requisito para que un valor de chi cuadrada sea significativo. Otros procedimientos no paramétricos incluyen: la prueba de la mediana para determinar si existe una diferencia significativa entre las medianas de dos muestras, el análisis de varianza en dos direcciones de Friedman para comparar la misma muestra medida por lo menos dos veces, y el análisis de varianza en una dirección por rangos de Kruskal-Wallis para comparar varias muestras independientes.

PROBLEMAS

1. Se entrevistaron muestras aleatorias de hombres y mujeres para determinar si fumaban cigarrillos o no. Se encontró que de 29 hombres 15 eran fumadores y que de 30 mujeres 20 eran fumadoras. Comprobar la hipótesis nula de que la frecuencia relativa de los hombres fumadores es la misma que la de las mujeres fumadoras. ¿Qué indican sus resultados?
2. Dos grupos de estudiantes presentaron exámenes finales de estadística. Sólo se dio preparación formal para el examen a un grupo; el otro leyó el texto requerido pero nunca asistió a clases. Mientras que 22 de los 30 miembros del primer grupo (que asistió a clases) aprobaron el examen, sólo 10 de los 28 miembros del segundo grupo (que no asistió a clases) lo aprobaron. Comprobar la hipótesis nula de que la frecuencia relativa de los "asistentes" que pasan el examen final es la misma que la de los "no asistentes" que lo pasan. ¿Qué indican sus resultados?
3. Realizar una prueba de significancia chi cuadrada aplicando la corrección de Yates al siguiente problema 2 × 2:

16	8
7	11

4. Realizar una prueba de significancia chi cuadrada aplicando la corrección de Yates al siguiente problema 2 × 2:

8	12
10	5

5. Realizar una prueba de significancia chi cuadrada aplicando la corrección de Yates al siguiente problema 2 × 2:

20	14
5	10

6. Realizar una prueba de significancia chi cuadrada para el siguiente problema 3 × 3:

20	17	5
15	16	16
4	14	18

7. Realizar una prueba de significancia chi cuadrada para el siguiente problema 4 × 2:

25	6
19	10
15	15
8	20

8. Realizar una prueba de significancia chi cuadrada para el siguiente problema 2 × 3:

8	10	15
12	10	9

9. Se pidió a dos muestras de estudiantes que leyeran y luego evaluaran un cuento corto escrito por un autor nuevo. A la mitad de ellos se les dijo que el autor era una mujer, mientras que a la otra mitad se le dijo que el autor era un hombre. Se obtuvo la siguiente evaluación: (los puntajes más altos indican evaluaciones más favorables)

X_1 *(Se les dijo que el autor era una mujer)*	X_2 *(Se les dijo que el autor era un hombre)*
6	6
5	8
1	8
1	2
3	5
4	6
3	3
6	8
5	6
5	8
1	2
3	2
5	6
6	8
6	4
3	3

Aplicando la prueba de la mediana, determinar si existe una diferencia significativa entre las medianas de estos grupos. ¿Se vieron influenciadas las evaluaciones del cuento corto por el sexo que se atribuyó al autor?

10. Aplicando la prueba de la mediana, determinar si existe una diferencia significativa entre las medianas de las siguientes muestras de puntajes:

X_1		X_2	
7		4	
8	9	7	3
7	5	3	2
6	9	2	2
7	8	3	6
7	9	4	4
8	7	7	5
9	9	4	4
7	9	5	4
6		6	4
9		2	3

11. La "armonía e identificación de grupo" entre una muestra de 14 niños se midió antes y después de que participaron en una tarea escolar cooperativa preparada para que dependieran más unos de otros en la obtención de una calificación en el curso. Se consiguieron los siguientes puntajes de identificación de grupo (los puntajes más altos indican mayor armonía de grupo):

Estudiante	*Tiempo 1 (Antes de la tarea cooperativa)*	*Tiempo 2 (Después de la tarea cooperativa)*
A	62	75
B	51	53
C	60	62
D	43	51
E	49	52
F	45	46
G	73	62
H	66	68
I	57	55
J	63	69
K	43	45
L	46	45
M	67	68
N	61	67

Aplicando el análisis de varianza en dos direcciones por rangos de Friedman, determinar si existe una diferencia significativa entre el Tiempo 1 y el Tiempo 2 en cuanto a la armonía de grupo.

12. Aplicando el análisis de varianza en dos direcciones por rangos de Friedman, determinar si existe una diferencia significativa entre los puntajes de los tiempos 1, 2 y 3 de la siguiente muestra de 11 entrevistados:

Entrevistado	*Tiempo 1*	*Tiempo 2*	*Tiempo 3*
A	60	62	64
B	53	54	50
C	59	65	71
D	65	66	68
E	55	63	61
F	71	74	76
G	57	58	63
H	77	76	79
I	63	65	70
J	54	59	62
K	63	62	65

13. Los investigadores probaron la alineación política entre muestras de estudiantes que se especializan en artes liberales, ingeniería y bellas artes. Se obtuvieron los siguientes resultados por muestra (los puntajes más altos indican mayor alineación):

X_1 *(Artes liberales)*	X_2 *(Ingeniería)*	X_3 *(Bellas artes)*
100	101	97
110	90	98

X_1 (*Artes liberales*)	X_2 (*Ingeniería*)	X_3 (*Bellas artes*)
95	92	99
93	100	100
106	90	104
102	96	103
	92	

Aplicando el análisis de varianza en una dirección de Kruskal-Wallis, determinar si existe una diferencia significativa según la especialización universitaria con respecto al nivel de alienación política.

14. Aplicando el análisis de varianza en una dirección de Kruskal-Wallis, determinar si existe una diferencia significativa entre las siguientes muestras de puntajes:

X_1	X_2	X_3
125	100	95
100	99	90
122	105	86
127	103	96
115	116	88
129	98	89
130		

11

Correlación

Características tales como la orientación política, la inteligencia y la clase social *varían* de un entrevistado a otro y, por lo tanto, nos referimos a ellas como *variables*. En capítulos anteriores nos hemos preocupado por establecer la presencia o ausencia de una relación entre dos variables cualesquiera que ahora llamaremos X y Y por ejemplo, entre la orientación política (X) y los métodos de crianza de los niños (Y); entre la clase social (X) y la inteligencia (Y); o entre la orientación a estudios universitarios (X) y el uso de la mariguana (Y). Anteriormente, y con ayuda de la razón t, del análisis de varianza o de la chi cuadrada, tratamos de descubrir si una diferencia entre dos o más muestras podía considerarse estadísticamente significativa —reflejo de una diferencia poblacional real— y no como simple producto del error de muestreo.

LA FUERZA DE LA CORRELACION

El descubrimiento de la existencia de una relación no dice mucho acerca del grado de asociación o *correlación* entre dos variables. Muchas relaciones son estadísticamente significativas; pocas expresan una correlación *perfecta* o exacta. Para ilustrar, sabemos que la estatura y el peso están asociados, ya que mientras más alta es una persona su peso tiende a aumentar. Sin embargo, hay numerosas excepciones a la regla. Algunas personas altas pesan muy poco, mientras que algunas personas bajas pesan mucho. Del mismo modo, una relación entre la orientación a estudios universitarios y el uso de la mariguana no impide la posibilidad de encontrar muchos estudiantes que van a ingresar a la universidad que no fuman o bien muchos fumadores entre aquéllos que no piensan asistir a ella.

Las correlaciones realmente varían respecto a su *fuerza*. Podemos visualizar diferencias en la fuerza de la correlación por medio de un *diagrama de dispersión*,

una gráfica que muestra la forma en que los puntajes de dos variables cualesquiera X y Y están dispersas en toda la escala de los posibles valores de los puntajes. En el arreglo convencional, un diagrama de dispersión se construye de manera que la variable X se sitúa a lo largo de la línea base horizontal, mientras que la variable Y se mide sobre la línea vertical.

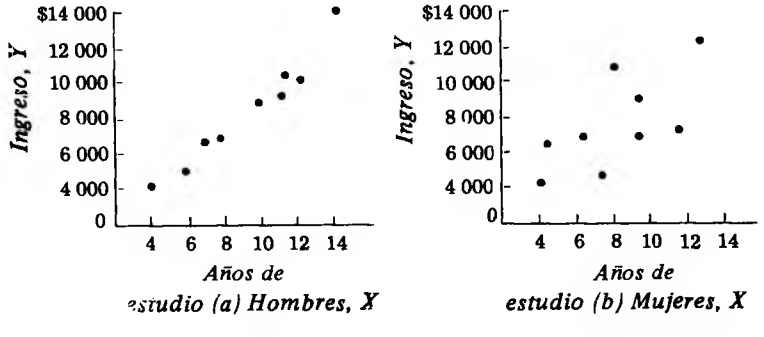
Observando la Figura 11.1 encontramos dos diagramas de dispersión, cada uno de los cuales representa la relación entre los años de estudio (X) y el ingreso (Y). La Figura 11.1(a) grafica esta relación respecto a los hombres, mientras que la Figura 11.1(b) representa la relación respecto a las mujeres. Nótese que todos y cada uno de los puntos en estos diagramas de dispersión grafican *dos* puntajes, estudios e ingreso, obtenidos de *un* entrevistado. Por ejemplo, en la Figura 11,1(a) vemos que un hombre con 4 años de estudio ganaba $ 4 000, mientras que un hombre con 13 años de estudio ganaba $ 10 000.

Podemos decir que la fuerza de la correlación entre X y Y aumenta a medida que los puntos de un diagrama de dispersión forman al estrecharse más una línea recta que baja por el centro de la gráfica. Por lo tanto, la Figura 11.1(a) (hombres) representa una correlación más fuerte que la Figura 11.1(b) (mujeres), aunque ambos diagramas de dispersión indican que el ingreso tiende a aumentar con un mayor estudio. Tales datos respaldarían ciertamente la imagen de que el ingreso de las mujeres (en relación con el de los hombres) está menos relacionado con el nivel de estudios a que llegan.

DIRECCION DE LA CORRELACION

A menudo se puede describir a la correlación como positiva o negativa respecto a la dirección. Una *correlación positiva* indica que los entrevistados que obtienen puntajes *altos* sobre la variable X también tienden a obtener puntajes *altos* sobre la variable Y. Recíprocamente, los entrevistados que obtienen puntajes *bajos* sobre X también tienden a obtener puntajes *bajos* sobre Y. La correlación positiva puede ilustrarse mediante la relación entre estudios e ingreso. Como hemos visto anteriormente, los entrevistados que completan muchos años de estudio tienden a percibir ingresos anuales elevados, en tanto que aquéllos que completan sólo unos cuantos años de estudio tienden a ganar muy poco anualmente.

FIGURA 11.1 Diagramas de dispersión que representan diferencias en la fuerza de la relación entre la preparación y el ingreso para hombres y mujeres



Existe una *correlación negativa,* si los entrevistados que obtienen puntajes *altos* sobre la variable *X* tienden a obtener puntajes *bajos* sobre la variable *Y*. A la inversa, los entrevistados que logran puntajes *bajos* sobre *X* tienden a lograr puntajes *altos* sobre *Y*. La relación entre los estudios y el ingreso *no* representaría una correlación negativa puesto que los entrevistados que completan muchos años de estudio *no* tienden a percibir ingresos anuales bajos. Un ejemplo de correlación negativa más adecuado es la relación entre los estudios y el prejuicio contra los grupos minoritarios. El prejuicio tiende a disminuir a medida que aumenta el nivel educativo. Por lo tanto, los individuos con pocos estudios formales tienden a mantener fuertes prejuicios, en tanto que los individuos con muchos años de estudio tienden a tener pocos prejuicios.

CORRELACION CURVILINEA

Una correlación positiva o negativa representa un tipo de relación *lineal.* Representados gráficamente, los puntos de un diagrama de dispersión tienden a formar una línea recta a través del centro de la gráfica. Si existe una correlación positiva, entonces los puntos del diagrama de dispersión se agruparán alrededor de la línea recta imaginaria que se indica en la Figura 11.2(a). Por el contrario, si una correlación negativa está presente, los puntos del diagrama de dispersión rodearán la línea imaginaria como se muestra en la Figura 11.2(b).

En su mayoría los investigadores sociales buscan establecer una correlación lineal, ya sea positiva o negativa. Sin embargo, es importante hacer notar que no se puede considerar que todas las relaciones entre *X* y *Y* forman una línea recta. Existen muchas correlaciones *curvilíneas* que indican que una variable aumenta a medida que la otra se incrementa hasta que la relación misma se invierte, de manera que una variable decrece finalmente mientras que la otra sigue acrecentándose. O sea que una relación entre *X* y *Y* que comienza como positiva se vuelve negativa; una relación que comienza como negativa se vuelve positiva. Para ilustrar una correlación curvilínea, estúdiese la relación entre el número de hijos (tamaño de la familia) y el estatus socioeconómico. Como se muestra en la Figura 11.3, los puntos del diagrama de dispersión tienden a formar una curva en forma de *U* más que una línea

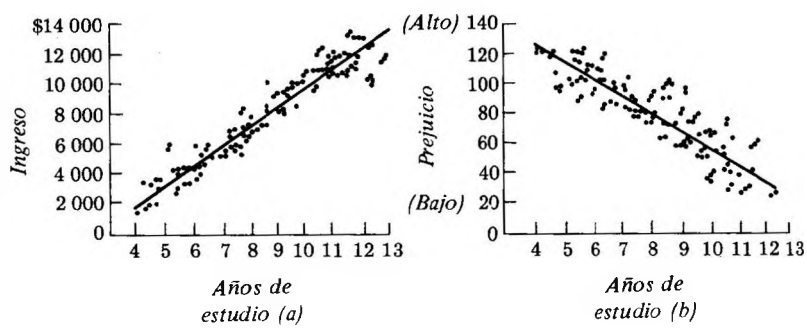


FIGURA 11.2 Diagramas de dispersión que representan (a) una correlación positiva entre la preparación y el ingreso y (b) una correlación negativa entre la preparación y el prejuicio

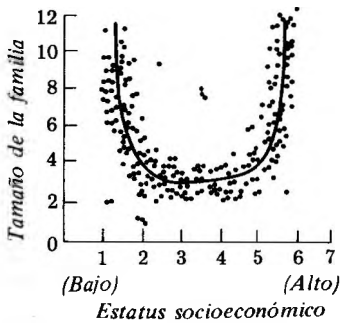


FIGURA 11.3 La relación entre el estatus socioeconómico (X) y el tamaño de la familia (Y): una correlación curvilínea

recta. Así, las familias de clase media tienen un número pequeño de hijos: el tamaño de la familia (Y) aumenta a medida que el estatus socioeconómico (X) se vuelve más alto y más bajo.

EL COEFICIENTE DE CORRELACION

El procedimiento para encontrar la correlación curvilínea se encuentra fuera del ámbito de este texto. En cambio, volvemos nuestra atención hacia los *coeficientes de correlación,* que expresan numéricamente tanto la fuerza como la dirección de la correlación lineal en línea recta. Tales coeficientes de correlación se encuentran generalmente entre −1,00 y +1,00 como sigue:

−1,00	⟵ correlación negativa perfecta
⋮	
−0,95	⟵ correlación negativa fuerte
⋮	
−0,50	⟵ correlación negativa moderada
⋮	
−0,10	⟵ correlación negativa débil
⋮	
0,00	⟵ ninguna correlación
⋮	
+0,10	⟵ correlación positiva débil
⋮	
+0,50	⟵ correlación positiva moderada
⋮	
+0,95	⟵ correlación positiva fuerte
⋮	
+1,00	⟵ correlación positiva perfecta

Vemos entonces que valores numéricos negativos como −1,00, −0,95, −0,50 y −0,10 significan una correlación negativa, en tanto que valores numéricos positivos como +1,00, +0,95, +0,50 y +0,10 indican una correlación positiva. Con respecto al grado de asociación, mientras más cerca esté de 1,00, en una u otra dirección, mayor es la fuerza de la correlación. En vista de que la fuerza de una correlación es independiente de su dirección, podemos decir que −0,10 y +0,10 son iguales en

cuanto a fuerza (ambas son muy débiles) y que −0,95 y +0,95 también tienen igual fuerza (ambas son muy fuertes).

UN COEFICIENTE DE CORRELACION PARA DATOS POR INTERVALOS

Con la ayuda del coeficiente de correlación de Pearson (r), podemos determinar la fuerza y la dirección de la relación entre las variables X y Y, las cuales han sido medidas al nivel por intervalos. La r de Pearson refleja hasta qué punto cada miembro de la muestra obtiene el mismo puntaje z sobre dos variables X y Y. En el caso de una correlación positiva, los dos puntajes z de un entrevistado tienen el mismo signo, ya sea positivo o negativo, y están situados aproximadamente a la misma distancia de la media de cada distribución de puntajes. Así, si el individuo A logra un puntaje por encima de la media en X, también lo hace en Y; si el individuo B logra un puntaje por debajo de la media en X, también lo hace en Y. En el caso de una correlación negativa, los puntajes z de un entrevistado tienen signos opuestos, indicando que son equidistantes de sus medias pero que caen en lados opuestos a ellas. Si el individuo A logra un puntaje sobre la media en X, en Y lo obtiene por debajo de la media si el individuo B obtiene un puntaje por debajo de la media en X, en Y lo logra por encima de ella. La interpretación de la correlación positiva y negativa por el puntaje z se ha ilustrado en la Figura 11.4.

Ahora podemos definir la r de Pearson como *la media de los productos del puntaje z para las variables X y Y*. Por fórmula,

$$r = \frac{\Sigma(z_X z_Y)}{N}$$

donde:

r = el coeficiente de correlación de Pearson

z_X = el puntaje z de un individuo en la variable X, igual a $\dfrac{X - \overline{X}}{s_X}$

z_Y = el puntaje z de un individuo en la variable Y, igual a $\dfrac{Y - \overline{Y}}{s_Y}$

N = el número total de pares de puntajes X y Y

A fin de ilustrar la aplicación de la r de Pearson, utilicemos la fórmula anterior para obtener un coeficiente de correlación para la relación entre el número de años de estudio que completó el padre (X) y el número de años de estudio que completó su hijo (Y). Los datos de la Tabla 11.1 representan esta relación en una muestra aleatoria de siete entrevistados.

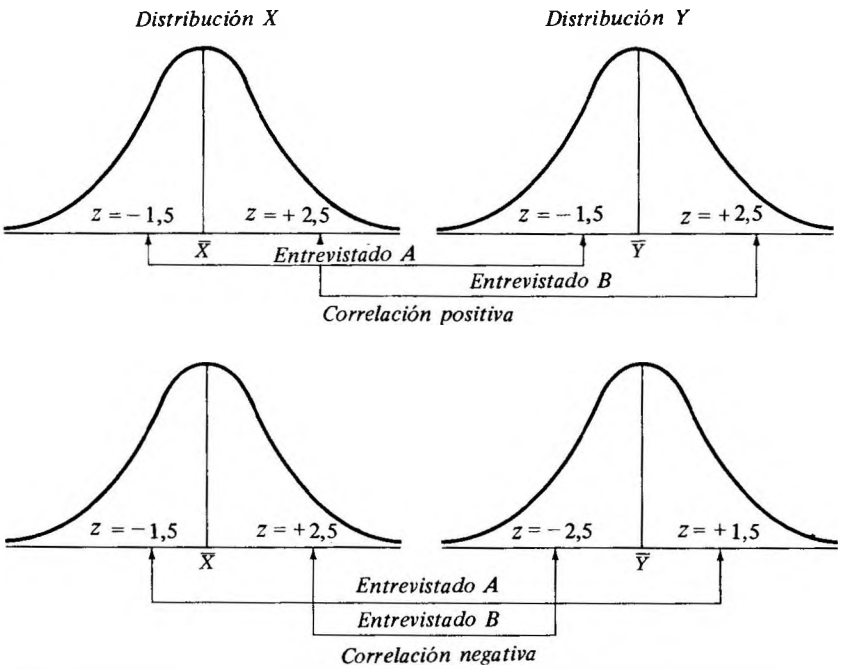


FIGURA 11.4 Una interpretación de la correlación positiva contra la negativa por el puntaje *z*

Para aplicar la fórmula para la r de Pearson debemos encontrar primero X, Y, s_X y s_Y como sigue:

TABLA 11.1 Relación entre el nivel educativo del entrevistado y la preparación del padre

	Años de estudio	
Niño	*Padres (X)*	*Niños (Y)*
A	12	12
B	10	8
C	6	6
D	16	11
E	8	10
F	9	8
G	12	11

Para cada muestra ahora encontramos los puntajes z y los puntajes z-producto para las variables X y Y.

X	X^2	Y	Y^2	
12	144	12	144	
10	100	8	64	
6	36	6	36	
16	256	11	121	$\bar{X} = \frac{\Sigma X}{N}$ $\bar{Y} = \frac{\Sigma Y}{N}$
8	64	10	100	
9	81	8	64	
12	144	11	121	$= \frac{73}{7}$ $= \frac{66}{7}$
$\Sigma X = 73$	$\Sigma X^2 = 825$	$\Sigma Y = 66$	$\Sigma Y^2 = 650$	$= 10{,}43$ $= 9{,}43$

$$s_X = \sqrt{\frac{\Sigma X^2}{N} - \bar{X}^2} \qquad s_Y = \sqrt{\frac{\Sigma Y^2}{N} - \bar{Y}^2}$$

$$= \sqrt{\frac{825}{7} - (10{,}43)^2} \qquad = \sqrt{\frac{650}{7} - (9{,}43)^2}$$

$$= \sqrt{117{,}86 - 108{,}78} \qquad = \sqrt{92{,}86 - 88{,}92}$$

$$= \sqrt{9{,}08} \qquad = \sqrt{3{,}94}$$

$$= 3{,}01 \qquad = 1{,}98$$

Para ilustrar el procedimiento para obtener z_X, z_Y, y $z_X z_Y$, examinemos las respuestas X y Y del miembro A de la muestra. Ya sabemos que $X = 10{,}43$ y $s_X = 3{,}01$. Puesto que $X - \bar{X} = 12 - 10{,}43 = 1{,}57$ para el miembro A de la muestra, encontramos que su $z_X = 1{,}57/3{,}01 = +0{,}52$. En otras palabras, los 12 años de

	X	$X - \bar{X}$	$\frac{X - \bar{X}}{s_X}$	Y	$Y - \bar{Y}$	$\frac{Y - \bar{Y}}{s_Y}$	$z_x z_y$
A	12	1,57	0,52	12	2,57	1,30	0,68
B	10	−0,43	−0,14	8	−1,43	−0,72	0,10
C	6	−4,43	−1,47	6	−3,43	−1,73	2,54
D	16	5,57	1,85	11	1,57	0,79	1,46
E	8	−2,43	−0,81	10	0,57	0,29	−0,24
F	9	−1,43	−0,48	8	−1,43	−0,72	0,34
G	12	1,57	0,52	11	1,57	0,79	0,41
							$\Sigma(z_x z_y) = 5{,}29$

educación de A caen aproximadamente media desviación estándar por encima de la media de la distribución. Igualmente sabemos que $\bar{Y} = 9{,}43$ y $s_Y = 1{,}98$. Ya que $Y - \bar{Y} = 12 - 9{,}43 = 2{,}57$ para el miembro A de la muestra, encontramos que su $z_Y = 2{,}57/1{,}98 = +1{,}30$. En otras palabras, los 12 años de educación de A caen aproximadamente una y un tercio desviaciones estándar por encima de la media de esta distribución. Para obtener $z_X z_Y$ para A, multiplicamos su puntaje z +0,52 por su puntaje z +1,30 (0,52 × 1,30 = 0,68). Como se muestra en la columna de la derecha anterior, la suma de estos puntajes productos z es 5,29.

Sustituyendo en la fórmula de Pearson,

$$r = \frac{\Sigma(z_X z_Y)}{N}$$

$$= \frac{5{,}29}{7}$$

$$= +{,}75$$

En el ejemplo anterior, la r de Pearson es igual a +0,75, lo que indica una correlación positiva bastante fuerte entre el nivel educativo que alcanzan los niños y

el de sus padres. Es decir, los entrevistados cuyos padres alcanzaron un alto nivel educativo también tienden a lograrlo; los entrevistados cuyos padres lograron un nivel educativo bajo también tienden a tener un bajo nivel de educación.

UNA FORMULA PARA CALCULAR LA *r* DE PEARSON

El cálculo de la *r* de Pearson a partir de los puntajes *z* ayuda a relacionar el tema de la correlación con nuestro anterior estudio de los puntajes estándar y la curva normal. Sin embargo, la fórmula de los puntajes *z* para la *r* de Pearson requiere cálculos largos y demorados. Afortunadamente existe una fórmula alternativa para la *r* de Pearson que trabaja directamente con puntajes crudos, eliminando con ello la necesidad de obtener puntajes *z* productos para las variables *X* y *Y*. De acuerdo con la fórmula para calcular la *r* de Pearson,

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

donde:

r = el coeficiente de correlación de Pearson
N = el número total de pares de puntajes X y Y
X = puntaje crudo en la variable X
Y = puntaje crudo en la variable Y

Para ilustrar el uso de la fórmula para calcular la *r* de Pearson volvamos a los datos de la Tabla 11.1 respecto a la relación entre el número de años de estudio que completó el padre (X) y el número de años que completó su hijo (Y). Para aplicar la fórmula de la *r* de Pearson debemos obtener primero X, Y, XY, X^2 y Y^2, como sigue:

X	X^2	Y	Y^2	XY
12	144	12	144	144
10	100	8	64	80
6	36	6	36	36
16	256	11	121	176
8	64	10	100	80
9	81	8	64	72
12	144	11	121	132
$\Sigma X = 73$	$\Sigma X^2 = 825$	$\Sigma Y = 66$	$\Sigma Y^2 = 650$	$\Sigma XY = 720$

$$r = \frac{7(720) - (73)(66)}{\sqrt{[7(825) - (73)^2][7(650) - (66)^2]}}$$

$$= \frac{5040 - 4818}{\sqrt{(5775 - 5329)(4550 - 4356)}}$$

$$= \frac{222}{\sqrt{(446)(194)}}$$

$$= \frac{222}{\sqrt{86524}}$$

$$= \frac{222}{294{,}15}$$

$$= +0{,}75$$

Comprobando la significancia de la *r* de Pearson

El coeficiente de correlación de Pearson nos da una medida exacta de la fuerza y la dirección de la correlación en la *muestra* que se está estudiando. Si hemos tomado una muestra aleatoria de una población específica, es posible que aún busquemos determinar si la asociación obtenida entre X y Y existe en la *población* y no se debe solamente al error de muestreo.

Para comprobar la significancia de una medida de correlación, usualmente planteamos la hipótesis nula de que no existe correlación en la población. Con respecto al coeficiente de correlación de Pearson, la hipótesis nula afirma que

$$r = 0$$

en tanto que la hipótesis de investigación establece que

$$r \neq 0$$

Como sucedió en capítulos anteriores, comprobamos la hipótesis nula seleccionando un nivel de confianza tal como 0,05 o 0,01 y calculando una prueba de significancia apropiada. Para comprobar la significancia de la *r* de Pearson podemos calcular una razón *t* con los grados de libertad iguales a $N - 2$ (N es igual al número de pares de puntajes). Con este fin, la razón *t* se puede calcular por la fórmula,

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

donde

t = la razón *t* para comprobar la significancia estadística de la *r* de Pearson
N = el número de pares de puntajes X y Y
r = el coeficiente de correlación de Pearson obtenido

Volviendo al ejemplo anterior, podemos comprobar la significancia de un coeficiente de correlación igual a +0,754 entre el nivel educativo del entrevistado y el de su padre.

$$t = \frac{0{,}754\sqrt{5}}{\sqrt{1 - (0{,}754)^2}}$$
$$= \frac{0{,}754(2{,}236)}{\sqrt{1 - 0{,}569}}$$
$$= \frac{1{,}69}{\sqrt{0{,}431}}$$
$$= \frac{1{,}69}{0{,}656}$$
$$= 2{,}58$$

Al consultar la Tabla C, al final del texto, encontramos que una razón *t* significativa debe ser igual o mayor que 2,57 al nivel de confianza 0,05 con 5 grados de libertad. Ya que nuestra razón *t* calculada ($t = 2{,}58$) es mayor que el valor de la tabla requerido, podemos rechazar la hipótesis nula de que $r = 0$ y aceptar la hipótesis de investigación de que $r \neq 0$. Los niveles educativos del entrevistado y de su padre están realmente asociados en la población.

Un método simplificado para comprobar la significancia de *r*

Afortunadamente, el proceso que se ilustró anteriormente para comprobar la significancia de la *r* de Pearson ha sido simplificado, de manera que es innecesario calcular realmente una razón *t*. En lugar de esto vamos a la Tabla F de la parte final del texto, donde encontramos una lista de valores significativos de la *r* de Pearson para los niveles de confianza de 0,05 y 0,01 con el número de grados de libertad de 1 a 90. Comparando directamente nuestro valor calculado de *r* con el valor correspondiente en la tabla, se produce el mismo resultado que si hubiéramos calculado realmente una razón *t*. Si el coeficiente de correlación de Pearson calculado es menor que el valor correspondiente en la tabla, debemos aceptar la hipótesis nula de que $r = 0$; si, por otra parte, el *r* calculado es igual o mayor que el valor de la tabla, rechazamos la hipótesis nula y aceptamos la hipótesis de investigación de que existe una correlación en la población.

Volvamos, con fines ilustrativos, sobre nuestro ejemplo anterior en el cual se comprueba un coeficiente de correlación igual a +0,754 por medio de una razón *t* que se encontró estadísticamente significativa. Mirando la Tabla F, al final del texto, encontramos ahora que el valor de *r* debe ser de por lo menos 0,754 para rechazar la hipótesis nula al nivel de confianza de 0,05 con 5 grados de libertad. Por lo tanto, este método simplificado nos lleva a la misma conclusión que el procedimiento más largo del cálculo de la razón *t*.

La correlación: una ilustración

Para ilustrar el procedimiento paso a paso para obtener un coeficiente de correlación

de Pearson (r), examinemos la relación entre los años de estudio completados (X) y los prejuicios (Y) tal como se encontró en la siguiente muestra de diez entrevistados:

Entrevistado	*Años de estudio (X)*	*Prejuicios (Y)*[a]
A	10	1
B	3	7
C	12	2
D	11	3
E	6	5
F	8	4
G	14	1
H	9	2
I	10	3
J	2	10

[a] Los datos más altos sobre la medida de los prejuicios (de 1 a 10) indican mayores prejuicios.

Para encontrar la r de Pearson seguimos los siguientes pasos:

PASO 1: Encontrar los valores de (1) ΣX, (2) ΣX^2, (3) ΣY, (4) ΣX^2, y (5) ΣXY

Entrevistado	X	X^2	Y	Y^2	XY
A	10	100	1	1	10
B	3	9	7	49	21
C	12	144	2	4	24
D	11	121	3	9	33
E	6	36	5	25	30
F	8	64	4	16	32
G	14	196	1	1	14
H	9	81	2	4	18
I	10	100	3	9	30
J	2	4	10	100	20
	$\Sigma X = 85$	$\Sigma X^2 = 855$	$\Sigma Y = 38$	$\Sigma Y^2 = 218$	$\Sigma XY = 232$
	(1)	(2)	(3)	(4)	(5)

PASO 2: Sustituir los valores del paso 1 en la fórmula para el coeficiente de correlación de Pearson

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

$$= \frac{10(232) - (85)(38)}{\sqrt{[10(855) - (85)^2][10(218) - (38)^2]}}$$

$$= \frac{2320 - 3230}{\sqrt{(8550 - 7225)(2180 - 1444)}}$$

$$= \frac{-910}{\sqrt{(1325)(736)}}$$

$$= \frac{-910}{\sqrt{975200}}$$

$$= \frac{-910}{987{,}52}$$

$$= -0{,}92$$

Nuestro resultado indica una correlación negativa bastante fuerte entre la educación y los prejuicios.

PASO 3: Hallar los grados de libertad

$$\begin{aligned} gl &= N - 2 \\ &= 10 - 2 \\ &= 8 \end{aligned}$$

PASO 4: Comparar la *r* de Pearson obtenida con el valor correspondiente de la *r* de Pearson en la Tabla F

$$\begin{aligned} r \text{ obtenida} &= -0{,}92 \\ r \text{ de la tabla} &= 0{,}63 \\ gl &= 8 \\ P &= 0{,}05 \end{aligned}$$

Como se indica más arriba, para rechazar la hipótesis nula de que $r = 0$ al nivel de confianza de 0,05 con 8 grados de libertad, nuestro valor calculado para la *r* de Pearson debe ser de por lo menos 0,63. Ya que nuestra *r* obtenida es igual a –0,92, rechazamos la hipótesis nula y aceptamos la hipótesis de investigación. Esto es, nuestro resultado sugiere que hay una correlación entre la educación y los prejuicios que está presente en la población de la cual se extrajo nuestra muestra.

Requisitos para el uso del coeficiente de correlación de Pearson

Con el fin de emplear correctamente el coeficiente de correlación de Pearson, como medida de asociación entre las variables *X* y *Y*, se deben tomar en cuenta los siguientes requisitos:

1. Una relación lineal en línea recta: la *r* de Pearson es útil solamente para detectar una correlación lineal en línea recta entre *X* y *Y*.
2. Los datos de intervalo: ambas variables, *X* y *Y*, deben medirse al nivel por intervalos de manera que se pueda asignar puntajes a los entrevistados.

3. El muestreo aleatorio: los miembros de la muestra deben haberse extraído aleatoriamente de una población específica. De esta manera no puede aplicarse una prueba de significancia.
4. Las características normalmente distribuidas: la prueba de la significación de la *r* de Pearson requiere que tanto la variable *X* como la *Y* estén normalmente distribuidas en la población. En muestras pequeñas, el no llenar el requisito de características normalmente distribuidas puede menoscabar seriamente la validez de la *r* de Pearson. No obstante, este requisito es secundario cuando la magnitud de la muestra es igual o mayor que 30 casos.

ANALISIS DE REGRESION

Establecer una correlación entre dos variables puede ser útil para predecir los valores de una variable (*Y*) conociendo los valores de otra variable (*X*). La técnica que se emplea para hacer tal predicción se conoce como *análisis de regresión.*

Hemos visto anteriormente en este capítulo que la fuerza de una correlación entre *X* y *Y* aumenta a medida que los puntos del diagrama de dispersión se estrechan formando una línea recta imaginaria. Podemos ahora identificar esa línea como una *línea de regresión,* línea recta que se dibuja a través del diagrama de dispersión, la cual representa la mayor "conveniencia" posible para hacer predicciones de *X* a *Y.*

Predicción de *Y* a partir de *X*

Imaginemos un estudio que trata de la correlación entre el número de años de estudio completados (*X*) y el ingreso anual (*Y*) en el que obtenemos una correlación positiva perfecta ($r = +1{,}00$) y los siguientes resultados para una muestra de seis entrevistados:

Entrevistado	*Años de estudio (X)*	*Ingreso (Y)*
A	18	$30 000
B	6	10 000
C	9	15 000
D	15	25 000
E	12	20 000
F	3	5 000

Como muestra la Figura 11.5, podemos marcar los puntajes anteriores y dibujar una línea recta a través de ellos, una línea de regresión que conecta los puntajes de cada entrevistado de la muestra. Una línea de regresión de este tipo permite la siguiente predicción: un individuo con 18 años de estudio ganará $ 30 000; un individuo con 3 años de estudio ganará $ 5 000 y así sucesivamente.

Como se señaló anteriormente, en la investigación social son pocas las correla-

ciones perfectas, ya sea +1,00 o −1,00. Esto es importante ya que por regla general las predicciones se vuelven más exactas a medida que aumenta el tamaño de una correlación. Para las correlaciones que son menos que perfectas, podemos construir aún una predicción o línea de regresión que se "ajuste" mejor a la dirección de los puntos en un diagrama de dispersión. Esto es cierto incluso aunque todos los puntos nunca estén sobre esa línea y nuestras predicciones sean menos que exactas. La línea de regresión para esa correlación que es menos que perfecta se presenta en la Figura 11.6.

La ecuación de regresión

La línea de regresión puede describirse mediante la fórmula

$$Y' = r\left(\frac{s_Y}{s_X}\right)X - r\left(\frac{s_Y}{s_X}\right)\overline{X} + \overline{Y}$$

donde

Y' = el valor calculado para Y (*Nota:* Es sólo una predicción y puede variar de Y.)
r = el coeficiente de correlación de Pearson para la relación entre las variables X y Y
s_Y = desviación estándar muestral de la distribución de la variable Y
s_X = desviación estándar muestral de la distribución de la variable X
X = un valor dado de X
$\overline{X}$ = media muestral de la distribución de la variable X
$\overline{Y}$ = media muestral de la distribución de la variable Y

Para ilustrar el uso de la fórmula de regresión para predecir los valores de Y, supongamos que hemos obtenido un coeficiente de correlación igual a +0,85 entre los años de estudio (X) y el ingreso anual (Y).

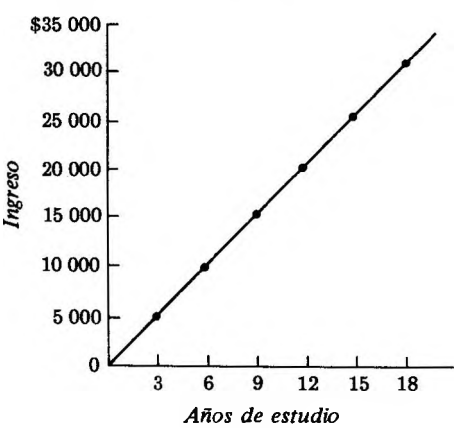


FIGURA 11.5 Una línea de regresión para la relación entre los años de estudio completados (X) y el ingreso anual (Y) (r =+1,00)

Dados los datos

$$\begin{aligned} r &= +0{,}85 \\ s_Y &= 0{,}50 \\ s_X &= 0{,}40 \\ \overline{X} &= 10 \text{ años} \\ \overline{Y} &= \$5000 \end{aligned}$$

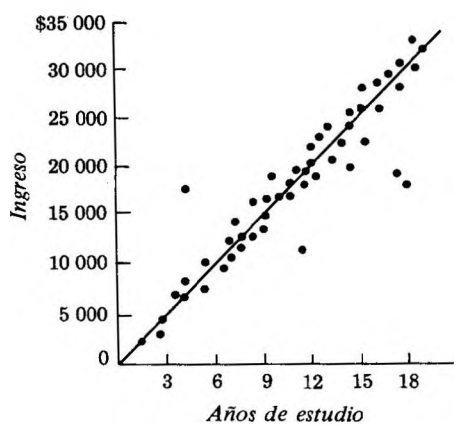
ahora podemos calcular la ecuación de regresión como sigue:

$$\begin{aligned} Y' &= 0{,}85\left(\frac{0{,}5}{0{,}4}\right) X - 0{,}85\left(\frac{0{,}5}{0{,}4}\right) 10 + 5000 \\ &= 1{,}06X - 1{,}06(10) + 5000 \\ &= 1{,}06X - 10{,}6 + 5000 \\ &= 1{,}06X + 4989{,}4 \end{aligned}$$

Para predecir el valor de Y por cada X, simplemente "sustituimos" los valores de X. Por ejemplo: ¿cuál es el ingreso anual calculado para un individuo que ha terminado 12 años de estudio? Sustituyendo en la ecuación de regresión,

$$\begin{aligned} Y' &= 1{,}06(12) + 4989{,}4 \\ &= 12{,}72 + 4989{,}4 \\ &= 5002{,}12 \end{aligned}$$

FIGURA 11.6 Una línea de regresión para la relación entre los años de estudio completados (X) y el ingreso anual (Y) ($r < +1{,}00$)



Por lo tanto, predecimos que el ingreso anual de alguien que tiene 12 años de estudio es de $ 5 002,12.

Del mismo modo, podemos predecir que un individuo que completa 6 años de estudio gana $ 4 995,76, o

$$Y' = 1{,}06(6) + 4989{,}4$$
$$= 6{,}36 + 4989{,}4$$
$$= \$4995{,}76$$

El análisis de regresión: una ilustración

El análisis de regresión se puede ilustrar más volviendo a examinar la relación entre el nivel educativo logrado por los padres (X) y el de sus hijos (Y). Como se anotó anteriormente en este capítulo, esta relación produjo un coeficiente de correlación de Pearson igual a 0,75 en una muestra de siete entrevistados:

	Educación	
Entrevistado	*Padres* (X)	*Entrevistados* (Y)
A	12	12
B	10	8
C	6	6
D	16	11
E	8	10
F	9	8
G	12	11

Podemos predecir los valores de Y (educación del hijo) del conocimiento de los valores de X (educación del padre) mediante los pasos siguientes:

PASO 1: Encontrar el coeficiente de correlación de Pearson

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$
$$= \frac{7(720) - (73)(66)}{\sqrt{[7(825) - (73)^2][7(650) - (66)^2]}}$$
$$= \frac{5040 - 4818}{\sqrt{(5775 - 5329)(4550 - 4356)}}$$
$$= \frac{222}{\sqrt{86524}}$$
$$= \frac{222}{294{,}15}$$
$$= +0{,}754$$

PASO 2: Obtener la media muestral para X y Y

$$\overline{X} = \frac{\Sigma X}{N} \qquad \overline{Y} = \frac{\Sigma Y}{N}$$

$$= \frac{73}{7} \qquad = \frac{66}{7}$$
$$= 10{,}43 \qquad = 9{,}43$$

PASO 3: Obtener la desviación estándar muestral para X y Y

$$s_X = \sqrt{\frac{\Sigma X^2}{N} - \overline{X}^2} \qquad s_Y = \sqrt{\frac{\Sigma Y^2}{N} - \overline{Y}^2}$$
$$= \sqrt{\tfrac{825}{7} - (10{,}43)^2} \qquad = \sqrt{\tfrac{650}{7} - (9{,}43)^2}$$
$$= \sqrt{117{,}86 - 108{,}79} \qquad = \sqrt{92{,}86 - 88{,}93}$$
$$= \sqrt{9{,}07} \qquad = \sqrt{3{,}93}$$
$$= 3{,}01 \qquad = 1{,}98$$

PASO 4: Sustituir los valores de los pasos 1, 2 y 3 en la ecuación de regresión

$$Y' = r\left(\frac{s_Y}{s_X}\right) X - r\left(\frac{s_Y}{s_X}\right) \overline{X} + \overline{Y}$$
$$= 0{,}75\left(\frac{1{,}98}{3{,}01}\right) X - 0{,}75\left(\frac{1{,}98}{3{,}01}\right) 10{,}43 + 9{,}43$$
$$= 0{,}75(0{,}66)X - 0{,}75(0{,}66)10{,}43 + 9{,}43$$
$$= 0{,}50X - 5{,}22 + 9{,}43$$
$$= 0{,}50X + 4{,}21$$

PASO 5: Determinar el valor de Y' para los valores de X

[*Ejemplos*]

1. Para un entrevistado cuyo padre completó 16 años de estudio:

$$Y' = 0{,}50X + 4{,}21$$
$$= 0{,}50(16) + 4{,}21$$
$$= 8{,}0 + 4{,}21$$
$$= 12{,}21$$

2. Para un entrevistado cuyo padre completó 6 años de estudio:

$$Y' = 0{,}50X + 4{,}21$$
$$= 0{,}50(6) + 4{,}21$$
$$= 3{,}0 + 4{,}21$$
$$= 7{,}21$$

Conclusión: Podemos predecir que los entrevistados cuyos padres han completado 16 años de estudio habrán completado 12,21 años de educación; los entrevistados

cuyos padres han completado 6 años de estudio habrán completado 7,21 años de educación.

COEFICIENTE DE CORRELACION PARA LOS DATOS ORDINALES.

Hasta este punto hemos presentado la *r* de Pearson un coeficiente de correlación para aplicarse a los datos que se pueden marcar en el nivel de medición por intervalos. Vamos ahora al problema de encontrar el grado de asociación para los datos ordinales: datos que han sido colocados por rangos u ordenados en relación a la presencia de una característica dada.

Para tomar un ejemplo de la investigación social, considérese la relación entre el estatus socioeconómico y la cantidad de tiempo empleado en mirar televisión. Imaginemos que una muestra de ocho entrevistados pudiera colocarse por rangos como sigue:

Entrevistado	*Estatus socioeconómico (X)* *Rango*		*Tiempo empleado en ver TV (Y)*	*Rango*
Miguel	1	más alto	2	mayor
Araceli	2	estatus socio-	1	tiempo
Juan	3	económico	3	viendo TV
Norma	4		5	
María	5		4	
Tomás	6		8	
Rafael	7		6	
Alejandra	8		7	

Como se muestra aquí, Miguel ocupó el primer rango con respecto al estatus socioeconómico, pero el segundo en relación con la cantidad de tiempo empleado en mirar televisión; la posición de Araceli fue segunda con respecto al estatus socioeconómico y primera en términos del tiempo empleado en mirar televisión, y así sucesivamente.

Para determinar el grado de asociación entre el estatus socioeconómico y la cantidad de tiempo empleado en ver televisión, aplicamos el *coeficiente de correlación por rangos ordenados* (r_s) de Spearman. Por fórmula.

$$r_s = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}$$

donde:

r_s = el coeficiente de correlación por rangos ordenados
D = la diferencia de rangos entre las variables X y Y
N = el número total de casos

Exponemos el presente ejemplo tal como se muestra en la Tabla 11.2.

TABLA 11.2 La relación entre el status socio-económico y el tiempo empleado en ver televisión

Entrevistado	*Estatus socio-económico X*	*Tiempo empleado en ver TV Y*	*D*	D^2
1	1	2	−1	1
2	2	1	1	1
3	3	3	0	0
4	4	5	−1	1
5	5	4	1	1
6	6	8	−2	4
7	7	6	1	1
8	8	7	1	1
				$\Sigma D^2 = 10$

Aplicando el coeficiente de correlación por rangos ordenados a los datos de la Tabla 11.2

$$r_s = 1 - \frac{6(10)}{8(64 - 1)}$$

$$= 1 - \frac{60}{8(63)}$$

$$= 1 - \frac{60}{504}$$

$$= 1 - 0{,}12$$

$$= + 0{,}88$$

Por lo tanto, encontramos una fuerte correlación positiva (r_s = + 0,88) entre el estatus socioeconómico y el tiempo empleado en ver televisión: los entrevistados con un alto estatus socioeconómico tienden a ver bastante televisión; los entrevistados con bajo estatus socioeconómico tienden a pasar poco tiempo viendo televisión.

Como tratar los rangos empatados

En la práctica real no es siempre posible colocar a nuestros entrevistados por rangos u ordenados evitando los empates en todas y cada una de las posiciones. Podríamos encontrar, por ejemplo, que dos o más entrevistados pasan exactamente la misma cantidad de tiempo frente al televisor, que el rendimiento académico de dos o más estudiantes es indistinguible, o que varios entrevistados tienen el mismo puntaje de coeficiente intelectual.

Para ilustrar el procedimiento de obtención de un coeficiente de correlación por rangos ordenados, en el caso de un empate entre ellos, digamos que estamos interesados en determinar el grado de asociación entre las categorías en un grupo que se gradúa y el coeficiente intelectual (C.I.). Supóngase también que podemos

colocar por rangos una muestra de 10 bachilleres, que están por graduarse, con respecto a su posición en la clase y que podemos obtener sus puntajes de C.I. como sigue:

Entrevistado	*Posición en la clase X*	*C.I. Y*
Jaime	10 ← (último)	110
Juan	9	90
Araceli	8	104
Norma	7	100
Carlos	6	110
Rosa María	5	110
Alejandra	4	132
Paco	3	115
Ricardo	2	140
Aldo	1 ← (primero)	140

Antes de seguir con el procedimiento estándar para obtener un coeficiente de correlación por rangos ordenados, coloquemos primero, por rangos, los puntajes de C.I. de nuestros futuros bachilleres:

Entrevistado	*C.I.*	*Rango C.I.*	
Jaime	110	7	
Juan	90	10	
Araceli	104	8	las posiciones 5, 6
Norma	100	9	y 7 están empatadas
Carlos	110	6	
Rosa María	110	5	
Alejandra	132	3	
Paco	115	4	
Ricardo	140	2	las posiciones 1 y
Aldo	140	1	2 están empatadas

Como se muestra aquí, Ricardo y Aldo recibieron los puntajes de C.I. más altos, y, por lo tanto, están empatados para el primero y segundo puestos. Igualmente, Rosa María, Carlos y Jaime lograron un puntaje de C.I. de 110 que los deja empatados en los puestos quinto, sexto y séptimo.

Para determinar la posición exacta en el caso de un empate, debemos *sumar los rangos empatados y dividir entre el número de empates.* Por lo tanto, la posición de un C.I. de 140, que se ha categorizado como 1 y 2, constituiría el rango "promedio".

$$\frac{1+2}{2} = 1{,}5$$

Del mismo modo, encontramos que la posición de un puntaje de C.I. de 110 es

$$\frac{5 + 6 + 7}{3} = 6{,}0$$

Habiendo encontrado la posición por rango de cada puntaje de C.I. podemos proceder a exponer este problema tal como se muestra en la Tabla 11.3.

Tabla 11.3 la relación entre la posición en la clase y el C.I.

Entrevistado	*Posición en la clase* (*X*)	*C.I.* (*Y*)	$X - Y = D$	D^2
1	10	6	4,0	16,00
2	9	10	−1,0	1,00
3	8	8	0	0
4	7	9	−2,0	4,00
5	6	6	0	0
6	5	6	−1,0	1,00
7	4	3	1,0	1,00
8	3	4	−1,0	1,00
9	2	1,5	0,5	0,25
10	1	1,5	−0,5	0,25
				$\Sigma D^2 = 24{,}50$

Obtenemos el coeficiente de correlación por rangos ordenados para el problema de la Tabla 11.3 como sigue:

$$r_s = 1 - \frac{6(24{,}50)}{10(100 - 1)}$$

$$= 1 - \frac{147}{990}$$

$$= 1 - 0{,}15$$

$$= +0{,}85$$

El coeficiente por rangos ordenados resultante indica una correlación *positiva* bastante fuerte entre la posición en clase y el C.I. o sea que los estudiantes con puntajes de C.I. altos tendieron a ocupar un *alto* rango en su clase; los estudiantes con puntajes de C.I. *bajos* tendieron a lograr *bajos* rangos en el grupo.

Prueba de significancia del coeficiente de correlación por rangos ordenados

¿Cómo hacemos para comprobar la significancia de un coeficiente por rangos ordenados? Por ejemplo: ¿Cómo podemos determinar a la correlación obtenida de +0,85 entre la posición en la clase y el C.I. puede generalizarse a una población mayor? Para comprobar la significancia de un r_s calculando simplemente vamos al final del texto, a la Tabla G, donde encontramos los valores significativos del coeficiente de correlación por rangos ordenados para los niveles de confianza de 0,05 y 0,01. Nótese que nos referimos directamente el número de pares de puntajes

(N) más que a un número de grados de libertad en particular. En el presente caso $N = 10$ y un r_s significativo debe ser igual o mayor que 0,648. Por lo tanto, rechazamos la hipótesis nula de que $r_s = 0$ y aceptamos la hipótesis de investigación de que la posición en la clase y el C.I. en realidad están relacionados en la población de la cual se extrajo nuestra muestra.

Correlación por rangos ordenados: una ilustración

Podemos resumir el procedimiento paso a paso para obtener el coeficiente de correlación por rangos ordenados en relación entre el grado de participación en las asociaciones voluntarias y el número de amigos cercanos. Esta relación se indica en la siguiente muestra de cinco entrevistados:

Entrevistado	*Participación en asociaciones voluntarias (X)*	*Rango*	*Número de amigos (Y)*
A	1	← mayor participación	6
B	2		4
C	3		6
D	4		2
E	5	← menor participación	2

Para determinar el grado de asociación entre la participación en las asociaciones voluntarias y el número de amigos, llevamos a cabo los siguientes pasos.

PASO 1: Colocar por rangos a los entrevistados sobre las variables X y Y. Como antes se mostró, colocamos por rangos a los entrevistados en relación a X, participación en asociaciones voluntarias, asignando el rango de 1 al entrevistado que participa más y el rango de 5 al entrevistado que participa menos.

También colocamos por rangos a los entrevistados en términos de Y, número de amigos. En el presente ejemplo tenemos casos de rangos empatados como se muestra a continuación:

Número de amigos (Y)	*Rango*	
6 ←	1 ←	Empatados en primero y segundo
4	3	
6 ←	2 ←	
2 ←	4 ←	Empatados en cuarto y quinto
2 ←	5 ←	

Para transformar los rangos empatados, tomamos un "promedio" de las posiciones empatadas:

Para las posiciones primera y segunda: $\frac{1 + 2}{2} = 1{,}5$

Para las posiciones cuarta y quinta: $\frac{4 + 5}{2} = 4{,}5$

Por lo tanto,

X	Y
1	1,5
2	3,0
3	1,5
4	4,5
5	4,5

PASO 2: Buscar ΣD^2. Debemos encontrar la diferencia entre los rangos X y Y (D), elevar al cuadrado cada diferencia (D^2) y sumar estos cuadrados (ΣD^2):

X	Y	D	D^2
1	1,5	−0,5	0,25
2	3,0	−1,0	1,00
3	1,5	1,5	2,25
4	4,5	−0,5	0,25
5	4,5	0,5	0,25
			$\Sigma D^2 = 4{,}00$

PASO 3: Sustituir el resultado del paso 2 en la fórmula para el coeficiente de correlación por rangos ordenados

$$r_s = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}$$

$$= 1 - \frac{6(4)}{5(24)}$$

$$= 1 - \frac{24}{120}$$

$$= 1 - 0{,}20$$

$$= +0{,}80$$

PASO 4: Comparar el coeficiente de correlación por rangos ordenados obtenido con el valor correspondiente de r_s en la Tabla G

r_s obtenido = 0,80
r_s de la tabla = 1,00
$N = 5$
$P = 0{,}05$

Al consultar la Tabla G al final del libro encontramos que un coeficiente de correlación de 1,00 (correlación perfecta) es necesario para rechazar la hipótesis nula al nivel de confianza de 0,05 con un tamaño muestral de 5. Por lo tanto, aunque hemos descubierto una fuerte correlación positiva entre la participación en asociaciones voluntarias y el número de amigos, aún debemos aceptar la hipótesis nula de que $r_s = 0$. Nuestro resultado no puede generalizarse a la población de la que extrajimos nuestra muestra.

Requisitos para el uso del coeficiente de correlación por rangos ordenados

El coeficiente de correlación por rangos ordenados deberá emplearse cuando se puedan cumplir las siguientes condiciones:

1. Una correlación lineal: el coeficiente por rangos ordenados detecta relaciones lineales entre X y Y.
2. Los datos ordinales: las variables X y Y deben ordenarse o colocarse por rangos.
3. El muestreo aleatorio: los miembros de la muestra deben haber sido extraídos aleatoriamente de una población mayor.

LA GAMMA DE GOODMAN Y KRUSKAL

La correlación puede mirarse en términos del grado hasta el cual se pueden predecir o adivinar los valores de una variable conociendo los valores de otra. Esto se puede ver muy directamente en la *gamma* (G) de Goodman y Kruskal, una alternativa para el coeficiente de correlación por rangos ordenados que prefieren muchos investigadores sociales para medir el grado de asociación entre variables de nivel ordinal.

La fórmula básica para *gamma* es

$$G = \frac{\Sigma f_c - \Sigma f_i}{\Sigma f_c + \Sigma f_i}$$

donde

f_c = la frecuencia de coincidencias
f_i = la frecuencia de las inversiones

Las coincidencias y las inversiones se pueden entender como expresiones de la dirección de la correlación entre las variables X y Y. Una coincidencia perfecta indica una correlación positiva perfecta (+ 1,00): todos los individuos que se están estudiando se han colocado por rangos exactamente en el mismo orden sobre ambas variables. Como se muestra a continuación, un individuo que logra un primer rango sobre X también lo logra sobre Y; un individuo que tiene un segundo rango sobre X también lo tiene sobre Y; y así sucesivamente.

Individuos	*Rango*	*Sobre*
	X	*Y*
A	1	1
B	2	2
C	3	3
D	4	4
E	5	5
F	6	6

Por contraste, la inversión perfecta indica una correlación negativa perfecta (−1,00), de manera que los individuos en estudio se colocan por rangos en un orden exactamente inverso sobre dos variables. Así, un individuo que logra un primer rango sobre *X* obtiene el último rango sobre *Y*; un individuo que tiene un segundo rango sobre *X* logra el penúltimo sobre *Y*, y así sucesivamente.

	Rango	*Sobre*
Individuos	*X*	*Y*
A	1	6
B	2	5
C	3	4
D	4	3
E	5	2
F	6	1

Cuando ocurre perfecta coincidencia o inversión se hace posible predecir con total exactitud el rango de un individuo sobre una variable, conociendo el rango que ocupa sobre la otra variable. En el caso de la coincidencia perfecta, por ejemplo, sabemos que una persona que obtiene el tercer rango sobre *X* también lo hace sobre *Y*. Sin embargo, ya que la correlación perfecta rara vez ocurre en la práctica de la investigación social, nuestra habilidad para hacer predicciones correctas acerca de una variable, basándonos en el conocimiento de otra, debe depender de la *cantidad* de coincidencia o inversión en el orden de los rangos de los individuos sobre las dos variables.

El coeficiente gamma: una ilustración

Para ilustrar el uso de *gamma,* digamos que estuviéramos estudiando la magnitud de la población negra en las áreas metropolitanas de los Estados Unidos en relación con su nivel de discriminación laboral. Tal estudio podría desarrollarse, por ejemplo, analizando los datos de población e ingreso disponibles en la Oficina de censos de los Estados Unidos.

Supóngase que pudiéramos ordenar por rangos las seis áreas metropolitanas más grandes de los Estados Unidos con respecto tanto a la magnitud de su población negra (X) y su nivel de discriminación (Y) como sigue:

Area metropolitana	*Magnitud de la población negra (X)*	*Nivel de discriminación laboral (Y)*
A	6	4
B	1	2
C	2	3
D	5	5
E	4	6
F	3	1

Así, vemos que el área metropolitana A tenía el número más pequeño de negros y era la cuarta más alta respecto a la discriminación: el área metropolitana B tenía la población negra más grande y fue segunda respecto a la discriminación, y así sucesivamente.

PASO 1: Reordenar los datos de manera que la variable X quede perfectamente ordenada de mayor a menor. Para determinar el grado de asociación entre el tamaño de la población negra y la discriminación laboral, colocamos primero los datos en una tabla en la que la variable X (en este caso el tamaño de la población negra) haya sido perfectamente ordenada de primero (1) a último (6) y la variable Y (en este caso el nivel de discriminación) se haya dejado desordenada. La frecuencia de coincidencias e inversiones en la columna desordenada (variable Y) indica cuánto difiere, esta columna de rangos, de una colocación por rangos perfectamente ordenada, ya sea positiva (1, 2, 3, 4, 5, 6) o negativa (6, 5, 4, 3, 2, 1):

Area metropolitana	*Tamaño de la población negra (X)*	*Nivel de discriminación laboral (Y)*
B	1	2
C	2	3
F	3	1
E	4	6
D	5	5
A	6	4

PASO 2: Obtener la frecuencia de las coincidencias. Para obtener la frecuencia de las coincidencias (f_c) empezamos con el rango más alto en la columna Y (área metropolitana B). Para cada rango contamos *el número de rangos que caen sobre él en la tabla y que son menores en valor numérico.* El número de rangos que ocurren por encima del rango más alto es siempre cero (puesto que no hay ningún rango por

encima de la cifra más alta en la tabla). Como resultado, escribimos un cero en la columna de las coincidencias para el área metropolitana B. Pasando al segundo rango de la columna *Y* (área metropolitana C) contamos el número de rangos que caen sobre él y que son menores en valor numérico. Vemos que solamente el rango de 2 cae por encima de eso para el área metropolitana C. Luego, como este rango es menor que 3, añadimos un 1 en la columna de las coincidencias. Pasando al siguiente rango de la lista (área metropolitana F) encontramos un rango de 1. Como los rangos sobre él (3 y 2) son mayores que 1, anotamos un cero en la columna de las coincidencias. Bajando una vez más por la columna *Y* al área metropolitana E, contamos el número de rangos sobre él y que son menores de 6. Como los tres rangos arriba mencionados (1, 3, 2) son menores, colocamos un 3 en la columna de coincidencias. Seguimos hacia los rangos restantes de la columna *Y* y repetimos el procedimiento de contar y poner coincidencias.

Area metropolitana	*Tamaño de la población negra (X)*	*Nivel de discriminación (Y) laboral*	*Coincidencias*
B	1	2	0
C	2	3	1
F	3	1	0
E	4	6	3
D	5	5	3
A	6	4	3

PASO 3: Obtener la frecuencia de las inversiones. Para encontrar la frecuencia de inversiones, comenzamos de nuevo con la anotación más alta en la columa *Y* (área metropolitana B). Sin embargo, esta vez contamos para cada rango *el número de rangos que caen sobre él y que son mayores en valor numérico.* Comenzando con el rango más alto, vemos nuevamente que no existen rangos sobre él y añadimos un cero en la columna de inversiones. Continuando con el segundo rango de la lista en la columna *Y* (área metropolitana C), contamos el número de rangos que caen sobre 3 y que son mayores en valor. Sólo el rango de 2 cae sobre eso para el área metropolitana C. Ya que este rango es menor, no mayor, que 3, agregamos un cero en la columna de inversiones. Bajando al siguiente rango en la lista (área metropoli-

Area metropolitana	*Tamaño de la población negra (X)*	*Nivel de discriminación laboral (Y)*	*Inversiones*
B	1	2	0
C	2	3	0
F	3	1	2
E	4	6	0
D	5	5	1
A	6	4	2

tana F), encontramos un rango de 1. Ya que los dos rangos sobre él (3 y 2) son mayores que 1, añadimos un 2 en la columna de inversiones. Bajando una vez más, encontramos un rango de 6 para el área metropolitana E. Como ninguno de los rangos sobre él (1, 3, 2) es mayor que 6, colocamos un cero en la columna de inversiones. Continuamos entonces con los rangos restantes y repetimos el procedimiento de contar o agregar inversiones.

PASO 4: Obtener Σf_c y Σf_i. Una vez que se han contado todas las coincidencias e inversiones, sumamos las coincidencias (Σf_c) y las inversiones (Σf_i) como se muestra a continuación:

	Coincidencias	*Inversiones*
B	0	0
C	1	0
F	0	2
E	3	0
D	3	1
A	3	2
	$\Sigma f_c = 10$	$\Sigma f_i = 5$

PASO 5: "Sustituir" Σf_c y Σf_i en la fórmula para gamma

$$G = \frac{\Sigma f_c - \Sigma f_i}{\Sigma f_c + \Sigma f_i}$$

$$= \frac{10 - 5}{10 + 5}$$

$$= \frac{5}{15}$$

$$= +0{,}33$$

Un coeficiente *gamma* igual a +0,33 indica la presencia de una correlación positiva débil. Esta es una correlación basada en la predominancia de coincidencias: hay un 33 por ciento de mayor coincidencia que de inversión entre el tamaño de la población negra y la discriminación laboral.

Como manejar los rangos empatados

Como vimos en relación con el coeficiente de correlación por rangos ordenados, no siempre es posible evitar empates en los rangos al nivel ordinal de medición. En efecto, los investigadores sociales trabajan frecuentemente con medidas ordinales brutas que producen un sinnúmero de rangos empatados. Cuando ocurre un número muy grande de empates, los procedimientos de cálculo simples de gamma la convierten en una medida de asociación especialmente útil. Para los rangos empatados se

emplea la fórmula básica para gamma, pero las frecuencias de las coincidencias y las inversiones se calculan de manera algo distinta.

Ilustremos el procedimiento para obtener un coeficiente con rangos empatados. Supongamos que un investigador quiera examinar la relación entre la clase social y la afiliación a determinada asociación voluntaria y obtenga los siguientes datos de un estudio con cuestionarios de 80 residentes de una ciudad: entre 29 entrevistados de la clase alta, 15 eran de la "alta", 10 eran de la "media" y 4 eran de la "baja" respecto a la afiliación a asociaciones voluntarias; entre 25 entrevistados de la clase media, 8 eran de la "alta", 10 eran de la "media" y 7 eran de la "baja" respecto a la afiliación mencionada; y entre 26 entrevistados de la clase baja, 7 eran de la "alta", 8 eran de la "media" y 11 eran de la "baja" respecto a la afiliación a tales asociaciones voluntarias. Nótese que en cada posición ocurren rangos empatados. Por ejemplo, hubo 29 entrevistados que empataron en el rango de clase social alta, el rango más alto sobre la variable *X*.

PASO 1: Reordenar los datos en forma de tabla de frecuencia:

Afiliación a las asociaciones voluntarias (Y)	*Clase Social (X)*		
	Alta	*Media*	*Baja*
Alta	15	8	7
Media	10	10	8
Baja	4	7	11
	29	25	26
		$N = 80$	

Nótese que la tabla anterior es una tabla de frecuencia de 3 × 3 que contiene 9 casillas (3 filas × 3 columnas = 9). Para asegurar que el signo del coeficiente gamma está representado con exactitud como positivo o negativo, la variable *X* de las columnas debe ordenarse siempre en orden decreciente de izquierda a derecha. En la tabla, por ejemplo, la clase social disminuye –alta, media, baja– de la columna izquierda a la de la derecha. Igualmente, la variable *Y* en los renglones debe disminuir de arriba hacia abajo. En la tabla anterior, la afiliación a las asociaciones voluntarias disminuye –alta, media, baja– de los renglones de arriba hacia los de abajo.

PASO 2: Obtener Σf_c. Para encontrar Σf_c se comienza con la casilla ($f = 15$) de la esquina superior izquierda. Luego se multiplica este número por la suma de todos los números que caigan *por debajo y a la derecha de él.* Leyendo de izquierda a derecha vemos que todas las frecuencias que están por debajo *y* a la derecha de 15 son 10, 8, 7 y 11. Ahora repita este procedimiento para todas las frecuencias que tienen casillas por debajo y a la derecha de ellas. Trabajando de izquierda a derecha en la tabla:

Clase alta/afiliación
alta — $15(10 + 8 + 7 + 11) = 15(36) = 540$
Clase media/afiliación
alta — $8(8 + 11) = 8(19) = 152$
Clase alta/afiliación
media — $10(7 + 11) = 10(18) = 180$
Clase media/afiliación
media — $10(11) = 110$

(Nótese que ninguna de las otras frecuencias de casilla de la tabla —7 en el renglón de arriba, 8 en el siguiente y 4, 7 y 11 en el de abajo— tienen casillas por debajo *y* a la derecha)

Σf_c es la suma de los productos obtenidos arriba. Por lo tanto,

$$\Sigma f_c = 540 + 152 + 180 + 110$$
$$= 982$$

PASO 3: Obtener Σf_i. Para obtener Σf_i se invierte el procedimiento para encontrar coincidencias y se comienza en la esquina superior *derecha* de la tabla. Esta vez, cada número se multiplica por la suma de todos los números que caen *por debajo y a la izquierda de él.* Leyendo de derecha a izquierda, vemos que las frecuencias por debajo *y* a la izquierda de 7 son 10, 10, 7 y 4. Al igual que en el paso anterior, se repite este procedimiento para todas las frecuencias que tienen casillas por debajo y a la derecha de ellas.

Trabajando de derecha a izquierda,

Clase baja/afiliación
alta — $7(10 + 10 + 7 + 4) = 7(31) = 217$
Clase media/afiliación
alta — $8(10 + 4) = 8(14) = 112$
Clase baja/afiliación
media — $8(7 + 4) = 8(11) = 88$
Clase media/afiliación
media — $10(4) = 40$

(Nótese que ninguna de las otras frecuencias de casilla de la tabla —15 en el renglón de arriba, 10 en el de en medio, 11, 7 y 4 en el de abajo— tienen casillas por debajo y a la izquierda.)

Σf_i es la suma de los productos antes calculados. Por lo tanto,

$$\Sigma f_i = 217 + 112 + 88 + 40$$
$$= 457$$

PASO 4: "Sustituir" los resultados de los pasos 2 y 3 en la fórmula para gamma

$$G = \frac{\Sigma f_a - \Sigma f_i}{\Sigma f_a + \Sigma f_i}$$
$$= \frac{992 - 457}{992 + 457}$$
$$= \frac{535}{1449}$$
$$= +0{,}37$$

Un coeficiente gamma de +0,37 indica una correlación positiva moderadamente débil entre la clase social y la afiliación a las asociaciones voluntarias. Nuestro resultado sugiere una correlación basada en una predominancia de coincidencias: existe un 37 por ciento de mayor coincidencia que de inversión entre la clase social y la afiliación a las asociaciones voluntarias. (Nótese en cambio, que un coeficiente gamma de –0,37 nos habría indicado una correlación *negativa* moderadamente débil basada en una predominancia de *inversiones*.)

Prueba de la significancia

Para comprobar la hipótesis nula de que X y Y no están asociadas en la población, convertimos nuestra G calculada a un puntaje z mediante la fórmula siguiente:

$$z = G\sqrt{\frac{\Sigma f_a - \Sigma f_i}{N(1 - G^2)}}$$

donde

G = el coeficiente gamma calculado
f_c = la frecuencia de coincidencias
f_i = la frecuencia de inversiones

En la ilustración anterior encontramos que $G = +0{,}37$ para la correlación entre la clase social y la afiliación a las asociaciones voluntarias. Para comprobar la significancia de nuestro resultado, reemplazamos en la fórmula:

$$z = (0{,}37)\sqrt{\frac{992 - 457}{80(1 - 0{,}37^2)}}$$
$$= (0{,}37)\sqrt{\frac{535}{80(0{,}86)}}$$
$$= (0{,}37)\sqrt{\frac{535}{68{,}80}}$$
$$= (0{,}37)\sqrt{7{,}78}$$
$$= (0{,}37)(2{,}79)$$
$$= 1{,}03$$

Consultando la Tabla B al final del libro, vemos que z debe ser igual o mayor que 1,96 para rechazar la hipótesis nula al nivel de confianza de 0,05. Ya que nuestra z calculada (z = 1,03) es menor que el valor requerido por la tabla, debemos aceptar la hipótesis nula de que $G = 0$ y rechazar la hipótesis de investigación de que $G = 0$. Nuestra correlación obtenida no puede generalizarse a la población de la que extrajimos nuestra muestra.

Requisitos para el uso de gamma

Deben tomarse en cuenta los siguientes factores para poder emplear gamma como medida de asociación:

1. Una correlación lineal: gamma detecta relaciones lineales entre X y Y.
2. Los datos ordinales: tanto X como Y deben estar colocadas por rangos u ordenadas.
3. El muestreo aleatorio: para comprobar la hipótesis nula ($G = 0$), los miembros de la muestra deben haberse tomado sobre una base aleatoria de una población específica.

COEFICIENTE DE CORRELACION PARA DATOS NOMINALES ORGANIZADO EN UNA TABLA 2 × 2

En el capitulo anterior se nos presentó una prueba de significancia para los datos de frecuencia que se conoce como chi cuadrada. Por una simple extensión de la prueba de chi cuadrada, podemos determinar ahora el grado de asociación entre variables al nivel nominal de medición.

Miremos nuevamente la hipótesis nula de que:

la proporción de fumadores de mariguana entre los estudiantes de Bachillerato orientados a estudios universitarios es igual que la proporción de fumadores de mariguana que no piensan asistir a la universidad.

En el Capítulo 10 se comprobó esta hipótesis nula en una muestra de 21 estudiantes que desean entrar a la universidad y una muestra de 15 estudiantes que no tenían planes de asistir a ella. Se determinó que 15 de 21 estudiantes iban a la universidad, pero sólo 5 de 15 estudiantes que no pensaban ir a la universidad, eran fumadores de mariguana (ver Capítulo 10). Así, tenemos el problema 2 × 2 en la Tabla 11.4.

Esta relación entre la orientación a estudios universitarios y el uso de la mariguana se comprobó aplicando la fórmula 2 × 2 para calcular chi cuadrada como sigue:

$$\chi^2 = \frac{36[(15)(10) - (5)(6)]^2}{(15 + 5)(6 + 10)(15 + 6)(5 + 10)}$$

TABLA 11.4 Uso de la mariguana entre estudiantes con y sin orientación hacia la universidad: datos de la Tabla 10.3

	Fumadores	*No fumadores*	
Orientación hacia la universidad	15	6	21
Sin orientación hacia la universidad	5	10	15
	20	16	$N = 36$

$$= \frac{36(150 - 30)^2}{(20)(16)(21)(15)}$$
$$= 5{,}14$$

Habiendo calculado un valor de chi cuadrada de 5,14, podemos obtener el *coeficiente phi* (ϕ), que es una medida del grado de asociación para las tablas 2 X 2 Por fórmula,

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

donde

ϕ = el coeficiente phi
χ^2 = el valor chi cuadrada calculado
N = el número total de casos

Aplicando la fórmula anterior al problema presente

$$\phi = \sqrt{\frac{5{,}14}{36}}$$
$$= \sqrt{0{,}14}$$
$$= 0{,}37$$

Nuestro coeficiente phi obtenido de 0,37 indica la presencia de una correlación moderada entre la orientación a los estudios universitarios y el uso de la mariguana.

Prueba de la significancia de phi

Afortunadamente, el coeficiente phi puede comprobarse fácilmente por medio de la chi cuadrada, cuyo valor ya se ha determinado, y la Tabla E al final del libro:

$$\chi^2 \text{ obtenido} = 5{,}14$$
$$\chi^2 \text{ de la tabla} = 3{,}84$$
$$\text{gl} = 1$$
$$P = 0{,}05$$

Dado que nuestro valor de chi cuadrada calculado de 5,14 es mayor que el valor requerido por la tabla, rechazamos la hipótesis nula de que $\phi = 0$ y aceptamos la hipótesis de investigación de que la orientación política y el uso de la mariguana están asociados en la población.

Requisitos para el uso del coeficiente phi

A fin de emplear el coeficiente phi como medida de asociación entre las variables *X* y *Y*, debemos tomar en cuenta los siguientes requisitos:

1. Los datos nominales: sólo se requieren datos de frecuencia.
2. Una tabla 2 × 2: los datos deben poder colocarse en forma de tabla 2 × 2 (2 filas por 2 columnas). Es inadecuado aplicarle el coeficiente phi a tablas mayores que 2 × 2, en las cuales se están comparando varios grupos o categorías.
3. El muestreo aleatorio: para poder comprobar la significancia del coeficiente phi, los miembros de la muestra deben haberse extraído, sobre una base aleatoria, de una población mayor.

COEFICIENTES DE CORRELACION PARA DATOS NOMINALES MAYORES QUE TABLAS 2 × 2

Hasta aquí hemos estudiado el coeficiente de correlación para datos nominales colocados en una tabla 2 × 2. Como vimos en el Capítulo 10, hay ocasiones en que tenemos datos nominales pero estamos comparando varios grupos o categorías. Para ilustrar, estudiemos nuevamente la hipótesis de que

> *la frecuencia relativa de los métodos no rígidos, moderados y autoritarios de crianza de los niños es igual para los liberales, los moderados y los conservadores.*

En el Capítulo 10 se comprobó esta hipótesis con los datos de la tabla 3 × 3, Tabla 11.5.

La relación entre el método de crianza de los niños y la orientación política se comprobó aplicando la fórmula para chi cuadrada como sigue:

$$\chi^2 = \frac{(7 - 10{,}79)^2}{10{,}79} + \frac{(10 - 10{,}07)^2}{10{,}07} + \frac{(15 - 11{,}14)^2}{11{,}14}$$

TABLA 11.5 Crianza de los niños según la orientación política: datos de la Tabla 10.4

	Conservador	*Moderado*	*Liberal*	
No rígido	7	9	14	30
Moderado	10	10	8	28
Autoritario	15	11	5	31
	32	30	27	$N = 89$

$$+ \frac{(9 - 10{,}11)^2}{10{,}11} + \frac{(10 - 9{,}44)^2}{9{,}44} + \frac{(11 - 10{,}45)^2}{10{,}45}$$
$$+ \frac{(14 - 9{,}10)^2}{9{,}10} + \frac{(8 - 8{,}49)^2}{8{,}49} + \frac{(5 - 9{,}40)^2}{9{,}40}$$
$$= 7{,}58$$

En el presente contexto, buscamos determinar la correlación o grado de asociación entre la orientación política (X) y el método de crianza de los niños (Y). Esto puede hacerse en una tabla mayor que 2 × 2 por una simple extensión de la prueba de chi cuadrada, a la cual nos referimos como el *coeficiente de contingencia* (C). El valor de C puede encontrarse por la fórmula

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}$$

donde

χ^2 = el valor calculado de chi cuadrada
N = el número total de casos
C = el coeficiente de contingencia

Al verificar el grado de asociación entre la orientación política y el método de crianza de los niños,

$$C = \sqrt{\frac{7{,}58}{89 + 7{,}58}}$$
$$= \sqrt{\frac{7{,}58}{96{,}58}}$$
$$= \sqrt{0{,}08}$$
$$= 0{,}28$$

Nuestro coeficiente de contingencia obtenido de 0,28 indica que la correlación entre la orientación política y la crianza de los niños puede considerarse bastante débil. La orientación política y el método de crianza de los niños están relacionados, pero se pueden encontrar muchas excepciones.

Prueba de significancia del coeficiente de contingencia

Tal como en el caso del coeficiente phi, la significancia estadística del coeficiente de contingencia se puede determinar fácilmente de la magnitud del valor de chi cuadrada obtenido. En el presente ejemplo, encontramos que la relación entre la orientación política y la crianza de los niños no es significativa y se limita a los miembros de nuestras muestras. Esto es cierto ya que el valor calculado de chi cuadrada, 7,58, es menor que el valor requerido por la tabla:

$$\chi^2 \text{ obtenido} = 7{,}58$$
$$\chi^2 \text{ de la tabla} = 9{,}49$$
$$\text{gl} = 4$$
$$P = 0{,}05$$

Requisitos para el uso del coeficiente de contingencia

Para aplicar el coeficiente de contingencia adecuadamente, debemos estar conscientes de los siguientes requisitos:

1. Los datos nominales: sólo se requieren datos de frecuencia. Estos datos pueden colocarse en forma de tabla 2 × 2 o más.
2. El muestreo aleatorio: a fin de comprobar la significancia del coeficiente de contingencia, todos los miembros de la muestra deben haber sido tomados aleatoriamente de una población mayor.

Una alternativa al coeficiente de contingencia

A pesar de su gran popularidad entre los investigadores sociales, el coeficiente de contingencia tiene una importante desventaja: el número de renglones y columnas en una tabla de chi cuadrada influirá en el tamaño máximo que C pueda alcanzar. Esto es, el valor del coeficiente de contingencia no siempre variará entre 0 y 1,0 (aunque nunca excederá de 1,0). Bajo ciertas condiciones el máximo valor de C puede ser 0,94; otras veces el valor máximo de C será 0,89, y así sucesivamente.

Para evitar esta desventaja de C podríamos decidir emplear otro coeficiente de correlación que exprese el grado de asociación entre las variables de nivel nominal en

una tabla mayor que 2 × 2. Este coeficiente, que se conoce como la *V* de Cramér no depende del tamaño de la tabla χ^2 y tiene los mismos requisitos que el coeficiente de contingencia. Por fórmula,

$$V = \sqrt{\frac{\chi^2}{N(k-1)}}$$

donde

V = la V de Cramér,
N = el número total de casos
k = el número de renglones *o* columnas, cualquiera que sea menor (si el número de renglones es igual al número de columnas como en el caso de una tabla 3 × 3, 4 × 4, o 5 × 5, se puede usar cualquiera de los números para k).

Volviendo a la relación entre la orientación política y la crianza de los niños como se ve en la Tabla 11.5 (una tabla 3 × 3),

$$\begin{aligned} V &= \sqrt{\frac{7{,}58}{89(3-1)}} \\ &= \sqrt{\frac{7{,}58}{89(2)}} \\ &= \sqrt{\frac{7{,}58}{178}} \\ &= \sqrt{0{,}04} \\ &= 0{,}20 \end{aligned}$$

Resultado: Encontramos un coeticiente de correlación V de Cramér igual a 0,20 que indica una relación débil entre la orientación política y las prácticas de crianza de los niños.

RESUMEN

En este capítulo se nos han presentado los coeficientes de correlación que expresan numéricamente el grado de asociación entre las variables X y Y. Con ayuda del coeficiente de correlación de Pearson (r), podemos determinar tanto la fuerza como la dirección de la relación entre las variables que se han medido al nivel por intervalos. Podemos usar también la r de Pearson para predecir los valores de una variable (Y) a partir del conocimiento de los valores de otra variable (X)

Hay varias alternativas no paramétricas para la r de Pearson. Para determinar la correlación entre las variables al nivel ordinal de medición, podemos aplicar el coeficiente de correlación por rangos ordenados de Spearman (r_s). Para utilizar esta medida de correlación, ambas variables, X y Y, deben estar colocadas u ordenadas

por rangos. Cuando ocurre un gran número de empates entre los rangos, el coeficiente gamma de Kruskal y Goodman (G) es una alternativa más efectiva que el coeficiente de correlación por orden de los rangos.

Por una simple extensión de la prueba de significancia chi cuadrada, podemos determinar el grado de asociación entre las variables al nivel nominal de medición. Para un problema 2 × 2 empleamos el coeficiente phi (ϕ); para un problema mayor que este usamos ya sea el coeficiente de contingencia o la V de Cramér.

PROBLEMAS

1. Se interrogó a seis estudiantes respecto de (X) su actitud hacia los judíos y sus actitudes hacia los portorriqueños (Y). Calcular un coeficiente de correlación Pearson para estos datos y determinar si la correlación es significativa.

Estudiante,	X	Y
A	1	2
B	6	5
C	4	3
D	3	3
E	2	1
F	7	4

2. Calcular un coeficiente de correlación de Pearson para los siguientes conjuntos de puntajes e indicar si la correlación es significativa.

X	Y
2	5
1	4
5	3
4	1

3. Calcular un coeficiente de correlación de Pearson para el siguiente conjunto de puntajes e indicar si la correlación es significativa.

X	Y
3	8
4	9
1	5
6	10
2	4

4. Calcular un coeficiente de correlación de Pearson para el siguiente conjunto de puntajes e indicar si la correlación es significativa.

X	Y
2	1
5	5
1	2
6	8
4	4

5. Calcular un coeficiente de correlación de Pearson para el siguiente conjunto de puntajes e indicar si la correlación es significativa.

X	Y
10	2
8	2
6	4
3	9
1	10
4	6
5	5

6. Empleando los datos del problema 1, calcular una ecuación de regresión para predecir el valor de Y (actitud hacia los portorriqueños) para los siguientes valores de X (actitud hacia los judíos): (a) $X = 5$, (b) $X = 2$, (c) $X = 9$.
7. Empleando los datos del problema 5, calcular una ecuación de regresión para predecir el valor de Y para los siguientes valores de X: (a) $X = 10$; (b) $X = 2$.
8. Cinco estudiantes fueron colocados por rangos en términos del tiempo que tardaban en terminar un examen (1 = el primero en terminar, 2 = el segundo en terminar, y así sucesivamente) y el instructor dio las calificaciones de los exámenes. Probar la hipótesis nula de la no relación entre (X), la calificación, y (Y), el periodo de tiempo necesario para terminar el examen (esto es, calcular un coeficiente de correlación por rangos ordenados e indicar si es significativo).

X	Y
53	1
91	2
70	3
85	4
91	5

9. Los ocho individuos siguientes han sido colocados por rangos sobre X y se les ha dado puntajes sobre Y. Para estos datos, calcular un coeficiente de correlación por rangos ordenados e indicar si existe una relación significativa entre X y Y.

X	Y
1	32
2	28
3	45
4	60
5	45
6	60
7	53
8	55

10. Los siete individuos siguientes se han colocado por rangos sobre X y Y. Calcular un coeficiente de correlación por rangos ordenados para estos datos e indicar si existe una relación significativa entre X y Y.

X	Y
1	7
3	6
2	5
4	3
5	4
7	2
6	1

11. Los cinco individuos siguientes se han colocado por rango de 1 a 5 sobre X y Y. Calcular un coeficiente de correlación por rangos ordenados para estos datos e indicar si existe una relación significativa entre X y Y.

X	Y
1	4
3	2
2	5
4	3
5	1

12. Los cinco individuos siguientes se han colocado por rangos de 1 a 5 sobre X y Y. Calcular un coeficiente gamma para estos datos e indicar si existe una relación significativa entre X y Y.

X	Y
2	3
1	2
3	1
5	5
4	4

13. 96 estudiantes fueron colocados por rangos de mayor a menor con respecto a (X), consumo de bebidas alcohólicas, y (Y), uso diario de la mariguana. Calcular un coeficiente gamma para estos datos a fin de determinar el grado de asociación entre el consumo de alcohol y el uso de la mariguana e indicar si existe una relación significativa entre X y Y.

	Consumo de alcohol		
Uso de mariguana	*Alto* f	*Medio* f	*Bajo* f
Alto	5	7	20
Medio	10	8	15
Bajo	15	6	10
		$N = 96$	

14. En el problema 2 del Capítulo 10, $\chi^2 = 8{,}29$ para la relación entre la asistencia a clases y las calificaciones de un examen final de estadística. Dada la información de que $N = 58$, calcular un coeficiente phi para determinar el grado de asociación entre estas variables.
15. Dado un problema 2×2 en el que $N = 138$ y $\chi^2 = 4{,}02$, calcular un coeficiente phi para determinar el grado de asociación entre las variables X y Y.
16. Dado un problema 2×2 en el que $N = 150$ y $\chi^2 = 3{,}90$, calcular un coeficiente phi para determinar el grado de asociación entre las variables X y Y.
17. Para determinar el grado de asociación entre X y Y para un problema 4×3 en el que $N = 100$ y $\chi^2 = 8{,}05$, calcular (a) un coeficiente de contingencia y (b) una V de Cramér.
18. En el problema 5 del Capítulo 10 se determinó que $N = 118$ y $\chi^2 = 17{,}75$. Determinar el grado de asociación entre X y Y para este problema 4×2 (a) calculando un coeficiente de contingencia (b) por la V de Cramér.
19. Para determinar el grado de asociación entre X y Y para un problema 3×3 en el que $N = 138$ y $\chi^2 = 10{,}04$, calcular (a) un coeficiente de contingencia y (b) la V de Cramér.

12

Aplicación de métodos estadísticos a problemas de investigación

La Parte III del texto contiene varias técnicas estadísticas que se pueden aplicar a los diferentes problemas de la investigación social. Los Capítulos 8, 9 y 10 presentaron las diversas técnicas utilizadas para determinar si las diferencias muestrales obtenidas son estadísticamente significativas o sólo un simple producto del error de muestreo. Las técnicas del Capítulo 11 tienen por objeto determinar el grado de asociación, la correlación entre dos variables.

Como se ha hecho notar, a través de todo el texto, cada técnica estadística tiene un conjunto de hipótesis para su correcta aplicación. En la selección de las técnicas, cualquier investigador deberá tener en cuenta varios factores, tales como:

1. si el investigador busca contrastar diferencias estadísticamente significativas, el grado de asociación, o ambos;
2. si el investigador ha alcanzado el nivel de medición nominal, ordinal o por intervalos de las variables en estudio;
3. si las variables que se están estudiando están o no distribuidas normalmente en la población de donde fueron extraídas; y
4. si el investigador está estudiando muestras independientes o la misma muestra medida más de una vez.

El presente capítulo proporciona una serie de situaciones hipotéticas de investigación en las que se especifican los criterios anteriores. Se pide al estudiante que escoja la técnica estadística más apropiada para cada situación de investigación de entre las siguientes pruebas que se vieron en la Parte III del texto:

1. la razón *t*
2. el análisis de varianza

3. la chi cuadrada
4. la prueba de la mediana
5. el análisis de varianza en una dirección de Kruskal-Wallis
6. el análisis de varianza en dos direcciones de Friedman
7. la *r* de Pearson
8. el orden de los rangos de Spearman
9. gamma de Goodman y Kruskal
10. phi
11. el coeficiente de contingencia
12. la *V* de Cramér

La Tabla 12.1 (p. 244) sitúa cada técnica estadística con respecto a algunas de las suposiciones importantes que se deben tener en cuenta para su correcta aplicación. Mirando las columnas de la tabla nos encontramos frente a la primera decisión importante relacionada con la selección de una técnica estadística: ¿Deseamos determinar si existe o no una relación? Las pruebas de significancia estudiadas en los Capítulos 8, 9 y 10 tienen por objeto determinar si una diferencia muestral obtenida refleja una diferencia poblacional verdadera. O acaso ¿buscamos establecer la fuerza de la relación entre dos variables? Esta es una cuestión de correlación a la que nos podemos dirigir por medio de las técnicas estadísticas presentadas en el Capítulo 11. Los subtítulos de las columnas de la Tabla 12.1 indican que un investigador que decide emplear una prueba de significancia en lugar de una técnica de correlación debe saber si está estudiando muestras independientes o la misma muestra medida más de una vez.

Los renglones de la Tabla 12.1 dirigen nuestra atención hacia el nivel al que están medidas nuestras variables. Si hemos logrado el nivel de medición por intervalos bien podríamos pensar en el empleo de una técnica paramétrica como *t*, *F* o *r*. Sin embargo, ya sea que hayamos llegado al nivel de medición nominal o al ordinal, la elección se limitará a varias alternativas no paramétricas.

Al final del capítulo se pueden encontrar las soluciones a las siguientes situaciones de investigación.

SITUACIONES DE INVESTIGACION

Situación de investigación 1

Un investigador realizó un experimento para determinar el efecto de la edad de un conferencista sobre la preferencia de los estudiantes para escuchar sus conferencias. En una situación normal, dentro del salón de clases, se dijo a 20 estudiantes que la administración quería saber acerca de sus preferencias respecto a una próxima serie de conferencistas visitantes. Específicamente, se les pidió evaluar a un profesor que "podría venir de visita a la universidad". A todos los estudiantes se les describió del mismo modo el profesor excepto porque: a la mitad de los alumnos se le dijo que el profesor tenía 65 años de edad; a la otra mitad se le dijo que el profesor tenía sólo 25. Se pidió entonces a todos los estudiantes que indicaran su disposición

para asistir a la conferencia de dicho profesor (los datos más altos indican una mayor disposición). Se obtuvieron los siguientes resultados:

X_1 *(Puntajes de estudiantes a quienes se dijo que el profesor tenía 25 años)*	X_2 *(Puntajes de estudiantes a quienes se dijo que el profesor tenía 65 años)*
65	78
38	42
52	77
71	50
69	65
72	70
55	55
78	51
56	33
80	59

¿Qué procedimiento estadístico se podría aplicar para determinar si existe una diferencia significativa entre estos grupos de estudiantes con respecto a su disposición para asistir a la conferencia?

Situación de investigación 2

Un investigador llevó a cabo un experimento para determinar el efecto de la edad de un conferencista sobre la preferencia de los estudiantes para escuchar sus conferencias. En una situación normal dentro del salón de clase, se dijo a 30 estudiantes que la administración deseaba conocer sus preferencias en relación con una futura serie de conferencistas visitantes. Concretamente se les pidió que evaluaran a un profesor que "podría venir de visita a la universidad". El profesor fue descrito a todos los estudiantes de la misma manera, sólo que a un tercio de los alumnos se les dijo que el profesor tenía 75 años de edad; a un tercio se le dijo que tenía 50; y a un tercio se le dijo que tenía sólo 25. Luego se pidió a todos los estudiantes que indicaran su disposición para asistir a la conferencia del profesor. Se obtuvieron los siguientes resultados:

X_1 *(Puntajes de estudiantes a quienes se dijo que el profesor tenía 25 años)*	X_2 *(Puntajes de estudiantes a quienes se dijo que el profesor tenía 50 años)*	X_3 *(Puntajes de estudiantes a quienes se dijo que el profesor tenía 75 años)*
65	63	67
38	42	42

TABLA 12.1 Criterios para escoger una técnica estadística apropiada

	Pruebas de significancia (Capítulos 8, 9, 10)		*Correlación (Capítulo 11*
Nivel de medición	*Muestras independientes*	*La misma muestra medida dos veces*	
Nominal	Chi cuadrada (prueba no paramétrica para comparar dos o más muestras)		Coeficiente phi (tabla 2 X 2 no paramétrico) Contingencia y *V* de Cramér (para tablas mayores de 2 X 2 no paramétricas)
Ordinal	Prueba de la mediana no paramétrica para comparar dos muestras) Análisis de varianza en una dirección de Kruskal-Wallis (no paramétrico para comparar tres o más muestras)	Análisis de varianza en dos direcciones de Friedman (no paramétrico para comparar la misma muestra medida por lo menos dos veces)	Orden de rango de Spearman (no paramétrico) Gamma de Goodman y Kruskal (no paramétrica para tratar un gran número de rangos empatados)
Intervalo	Razón *t* (paramétrica para comparar dos muestras) Análisis de varianza (paramétrico para comparar tres o más muestras)	Razón *t* (paramétrica para comparar la misma muestra medida dos veces)	*r* de Pearson (paramétrica)

X_1 *(Puntajes de estudiantes a quienes se dijo que el profesor tenía 25 años)*	X_2 *(Puntajes de estudiantes a quienes se dijo que el profesor tenía 50 años)*	X_3 *(Puntajes de estudiantes a quienes se dijo que el profesor tenía 75 años)*
52	60	77
71	55	32
69	43	52
72	36	34
55	69	45
78	57	38
56	67	39
80	79	46

¿Qué procedimiento estadístico se podría aplicar para determinar si existe una diferencia significativa entre estos grupos de estudiantes con respecto a su disposición para asistir a la conferencia?

Situación de investigación 3

Para investigar la relación entre la ortografía y la habilidad para la lectura, un investigador aplicó exámenes de ortografía y de lectura a un grupo de 20 estudiantes seleccionados aleatoriamente de una gran población de estudiantes no graduados. Se obtuvieron los siguientes resultados (los puntajes más altos indican una mayor habilidad):

Estudiante	X *(Puntaje de ortografía)*	Y *(Puntaje de lectura)*
A	52	56
B	90	81
C	63	75
D	81	72
E	93	50
F	51	45
G	48	39
H	99	87
I	85	59
J	57	56
K	60	69
L	77	78
M	96	69
N	62	57
O	28	35
P	43	47
Q	88	73
R	72	76
S	75	63
T	69	79

¿Qué procedimiento estadístico se podría aplicar para determinar el grado de asociación entre la ortografía y la habilidad para la lectura?

Situación de investigación 4

Para averiguar la validez de un determinado examen de lectura, los investigadores lo aplicaron a una muestra de 20 estudiantes cuya habilidad para leer había sido previamente colocada por rangos por su profesor. El puntaje del examen y el rango que el profesor dio para cada estudiante se enumeran a continuación:

Estudiante	*X (Puntaje de lectura)*	*Y (Rango del profesor)*
A	28	18
B	50	17
C	92	1
D	85	6
E	76	5
F	69	10
G	42	11
H	53	12
I	80	3
J	91	2
K	73	4
L	74	9
M	14	20
N	29	19
O	86	7
P	73	8
Q	39	16
R	80	13
S	91	15
T	72	14

¿Qué procedimiento estadístico se podría aplicar para determinar el grado de asociación entre los puntajes de lectura y la categorización del profesor?

Situación de investigación 5

Para estudiar las diferencias regionales relacionadas con el espíritu servicial hacia los desconocidos, un investigador dejó caer 400 llaves (todas las cuales habían sido marcadas y señaladas con una dirección de remitente) en los alrededores de los buzones de las regiones norte, sur, este y oeste de una ciudad. El número de llaves devueltas por región (como un indicador del espíritu servicial) se indica a continuación:

	Región			
	Norte	*Sur*	*Este*	*Oeste*
	f	*f*	*f*	*f*
Devueltas	55	69	82	61
No devueltas	45	31	18	39
	100	100	100	100

¿ Qué procedimiento estadístico se podría aplicar para determinar si estas diferencias regionales son estadísticamente significativas?

Situación de investigación 6

Para examinar la relación entre el autoritarismo y los prejuicios, un investigador administró medidas de autoritarismo (la escala F) y prejuicio (una lista de confrontación de los adjetivos negativos generalmente asignados a los norteamericanos negros) a una muestra nacional de 950 norteamericanos adultos. Se obtuvieron los siguientes resultados: de 500 entrevistados autoritarios, 350 estaban "prejuiciados" y 150 eran "tolerantes". De 450 entrevistados no autoritarios, 125 estaban "prejuiciados" y 325 eran "tolerantes".

¿Qué procedimiento estadístico se podría aplicar para estudiar el grado de asociación entre el autoritarismo y el prejuicio?

Situación de investigación 7

Para investigar la relación entre el año escolar y el promedio de calificaciones, los investigadores examinaron los antecedentes académicos de 186 estudiantes universitarios seleccionados sobre una base aleatoria de la población no graduada de cierta universidad. Los investigadores obtuvieron los siguientes resultados:

	Año escolar				
Promedio de calificaciones	*1o.* *f*	*2o.* *f*	*3o.* *f*	*4o.* *f*	
MB	6	5	7	10	
B	10	16	19	18	
S	23	20	15	7	
NA	15	7	6	2	
	54	48	47	37	(N = 186)

¿Qué procedimiento estadístico se podría aplicar para determinar el grado de asociación entre el promedio de calificaciones y el año escolar de los alumnos?

Situación de investigación 8

Para investigar la influencia de la frustración sobre los prejuicios, se pidió a 10 sujetos que asignaran adjetivos negativos como perezoso, sucio e inmoral, para describir a los miembros de un grupo minoritario (una medida de prejuicio). Todos los sujetos describieron al grupo minoritario tanto antes como después de que habían tomado una serie de exámenes largos y difíciles (la situación frustrante). Se obtuvieron los siguientes resultados (los puntajes más altos representan un mayor prejuicio):

Sujeto	X_1 *(Puntajes de prejuicio antes de tomar los exámenes frustrantes)*	X_2 *(Puntajes de prejuicio después de tomar los exámenes frustrantes)*
A	22	26
B	39	45
C	25	24
D	40	43
E	36	36
F	27	29
G	44	47
H	31	30
I	52	52
J	48	59

¿Qué procedimiento estadístico se podría aplicar para determinar si existe una diferencia estadísticamente significativa en los prejuicios antes y después de la administración de los exámenes frustrantes?

Situación de investigación 9

Para investigar la relación entre el estatus ocupacional real de un entrevistado y su clase social subjetiva (o sea, su propia identificación de clase social), se pidió a 677 individuos que indicaran su ocupación y la clase social a la que pertenecían. De 190 entrevistados con ocupaciones de estatus superior (profesional-técnico-gerencial), 56 se identificaron como miembros de la clase alta, 122 de la clase media, y 12 de la clase baja; de 221 entrevistados con ocupaciones de estatus medio (vendedores-oficinistas-trabajadores calificados), 42 se identificaron como miembros de la clase alta, 163 de la clase media, y 16 de la clase baja; de 266 entrevistados con ocupaciones de estatus bajo (trabajadores de mano de obra semi calificada y no calificada), 15 se identificaron como miembros de la clase alta, 202 de la clase media y 49 de la clase baja.

¿Qué procedimiento estadístico se podría aplicar para determinar el grado de asociación entre el estatus ocupacional y la clase social subjetiva?

Situación de investigación 10

Para investigar la influencia de la especialización universitaria en el sueldo inicial de los graduados universitarios, los investigadores entrevistaron a un grupo de estudiantes recién graduados, especializados en ingeniería, ciencias sociales o administración de empresas, en relación con sus primeros empleos. Los resultados obtenidos para estos 21 entrevistados son los siguientes:

Salarios iniciales

Ingeniería	*Ciencias sociales*	*Administración de empresas*
$ 10 500	$ 7 000	$ 7 500
12 300	9 500	9 000
14 000	10 000	8 000
9 500	11 000	9 300
9 000	8 500	10 500
8 500	7 500	10 000
7 500	7 000	7 000

¿Qué procedimiento estadístico se podría aplicar para determinar si existe una diferencia significativa entre estos grupos de entrevistados con respecto a sus salarios iniciales?

Situación de investigación 11

Para investigar la influencia de la especialización universitaria en el salario inicial de los graduados universitarios, los investigadores entrevistaron a un grupo de estudiantes recién graduados, especializados en ciencias sociales o en administración, en relación con sus primeros empleos. Los resultados obtenidos para estos 16 entrevistados son los siguientes:

Salarios iniciales

Ciencias sociales	*Administración*
$ 7 000	$ 7 500
9 500	9 000
10 000	8 000
11 000	9 300
8 500	10 500
7 500	10 000
7 000	7 000
	8 000
	9 300

¿Qué procedimiento estadístico se podría aplicar para determinar si existe una diferencia significativa entre los especialistas en ciencias sociales y los especialistas en administración con respecto a sus salarios iniciales?

Situación de investigación 12

Un investigador llevó a cabo un experimento para determinar el efecto de la edad de un conferencista sobre la disposición estudiantil para escuchar sus conferencias. En una situación normal, dentro del salón de clases, se dijo a 130 estudiantes que la administración deseaba conocer sus preferencias respecto a una próxima serie de conferencistas visitantes. Específicamente, se les pidió evaluar a un profesor que "podría venir de visita a la universidad". El profesor fue descrito igualmente para todos, a no ser porque: a la mitad de los estudiantes se le dijo que el profesor tenía 65 años de edad y a la otra mitad se le dijo que el profesor tenía sólo 25. Más tarde se pidió a todos los estudiantes que indicaran su disposición para asistir a la conferencia del profesor y se obtuvieron los siguientes resultados: de los estudiantes a quienes se dijo que el profesor tenía 65 años, 22 manifestaron su disposición para asistir a las conferencias y 43 expresaron su renuencia; de los estudiantes a quienes se dijo que el profesor tenía 25 años, 38 manifestaron su disposición de asistir a las conferencias y 27 expresaron su renuencia.

¿Qué procedimiento estadístico se podría aplicar para determinar si existe una diferencia significativa entre estos grupos de estudiantes con respecto a su disposición para asistir a la conferencia del profesor?

SOLUCION A LAS INVESTIGACIONES

Solución a la situación de investigación 1

(Razón t o prueba de la mediana)

La situación de investigación 1 representa una comparación entre los puntajes de dos muestras independientes de estudiantes. La razón *t* (Capítulo 8) se emplea con el fin de hacer comparaciones entre dos medias cuando se han obtenido datos por intervalos. La prueba de la mediana (Capítulo 10) es una alternativa no paramétrica que se puede aplicar cuando sospechemos que los puntajes no están distribuidos normalmente en la población o que no se ha logrado el nivel de medición por intervalos.

Solución a la situación de investigación 2

(Análisis de varianza o *análisis de varianza en una dirección de Kruskal-Wallis)*

La situación de investigación 2 representa una comparación de los puntajes de tres muestras independientes de estudiantes. La razón *F* (análisis de varianza, Capítulo 9) se emplea para hacer comparaciones entre tres o más medias independientes cuando se han obtenido datos por intervalos. El análisis de varianza en una dirección de Kruskal-Wallis (Capítulo 10) puede aplicarse como una alternativa no paramétrica cuando tenemos motivos para sospechar que los puntajes no están distribuidos

normalmente en la población o cuando no se ha alcanzado el nivel de medición por intervalos.

Solución a la situación de investigación 3
(La r de Pearson)

La situación de investigación 3 es un problema de correlación puesto que pide el grado de asociación entre X (habilidad en ortografía) y Y (habilidad para la lectura). La r de Pearson (Capítulo 11) puede emplearse para detectar una correlación lineal entre las variables X y Y cuando ambas han sido medidas al nivel por intervalos. Si X (habilidad en ortografía) y Y (habilidad en lectura) no están distribuidas normalmente en la población, habrá que pensar en la aplicación de una alternativa no paramétrica tal como el coeficiente de correlación por rangos ordenados de Spearman (Capítulo 11).

Solución a la situación de investigación 4
(Rangos ordenados de Spearman)

La situación de investigación 4 es un problema de correlación que pregunta por el grado de asociación entre X (puntajes de lectura) y Y (evaluación del profesor respecto a la habilidad para la lectura). El coeficiente de correlación por rangos ordenados de Spearman (Capítulo 11) puede emplearse para detectar una relación lineal entre las variables X y Y, cuando ambas variables han sido ordenadas o colocadas por rangos. La r de Pearson no se puede emplear pues requiere el nivel de medición por intervalos para X y Y. En el presente caso, los puntajes de lectura (X) deben ser colocados por rangos 1 a 20 antes de aplicar el coeficiente por rangos ordenados.

Solución a la situación de investigación 5
(Chi cuadrada)

La situación de investigación 5 representa una comparación entre las frecuencias (llaves devueltas contra llaves no devueltas) encontradas en cuatro grupos (norte, sur, este y oeste). La prueba de significancia chi cuadrada (Capítulo 10) se utiliza para hacer comparaciones entre dos o más muestras. Sólo se requieren los datos nominales. Los presentes resultados se pueden colocar en forma de tabla 2 × 4, representando 2 renglones y 4 columnas. Nótese que el grado de asociación entre la tasa de devolución (X) y la región (Y) se puede medir con el coeficiente de contingencia (C) o la V de Cramér (Capítulo 11).

Solución a la situación de investigación 6
(Coeficiente phi)

La situación de investigación 6 es un problema de correlación que pregunta por el grado de asociación entre X (autoritarismo) y Y (prejuicio). El coeficiente phi (Capí-

tulo 11) es una medida de asociación que puede emplearse cuando los datos de frecuencia o nominales se pueden colocar en forma de tabla 2 × 2 (2 renglones y 2 columnas). En el presente problema, dicha tabla tomaría la forma siguiente:

Nivel de prejuicio	*Nivel de autoritarismo*		
	Autoritario	*No autoritario*	
Prejuiciado	350	120	
Tolerante	150	325	$N = 950$

Solución a la situación de investigación 7
(Gamma de Goodman y Kruskal)

La situación de investigación 7 es un problema de correlación que pregunta por el grado de asociación entre X (promedio de calificaciones) y Y (año escolar). El coeficiente gamma de Goodman y Kruskal (Capítulo 11) se emplea para detectar una relación lineal entre X y Y cuando ambas variables se han colocado por rangos y ha ocurrido un gran número de empates. En el presente problema, el promedio de calificaciones se ha colocado por rangos desde MB hasta NA y el año escolar se ha colocado por rangos de 1o. a 4o. Ambas medidas ordinales crudas han generado numerosos rangos empatados (por ejemplo, 54 estudiantes estaban en su primer año escolar; 48 el segundo, y así sucesivamente). El coeficiente de contingencia (C) o la V de Cramér (Capítulo 11) representa una alternativa en relación con gamma, la cual supone únicamente datos de nivel nominal.

Solución a la situación de investigación 8
(Razón t o análisis de varianza en dos direcciones por rangos)

La situación de investigación 8 representa una comparación antes-después de una sola muestra medida en dos puntos diferentes en el tiempo. La razón t (Capítulo 8) puede emplearse para comparar dos medias de una sola muestra ordenada en un diseño de panel antes-después. El análisis de varianza en dos direcciones de Friedman (Capítulo 10) es una alternativa no paramétrica que se puede aplicar a la situación antes-después cuando tenemos motivos para sospechar que los puntajes no están distribuidos normalmente en la población o cuando no hemos alcanzado el nivel de medición por intervalos.

Solución a la situación de investigación 9
(Gamma de Goodman y Kruskal)

La situación de investigación 9 es un problema de correlación que pregunta por el grado de asociación entre X (estatus ocupacional) y Y (clase social subjetiva).

El coeficiente gamma (Capítulo 11) es especialmente apropiado para el problema de detectar una relación lineal entre X y Y, cuando ambas variables pueden colocarse por rangos y ha ocurrido un gran número de empates. En la presente situación, el estatus ocupacional y la clase social subjetiva se han ordenado de "alta" a "media" y a "baja", generando un número muy grande de rangos empatados (por ejemplo, 221 entrevistados tenían ocupaciones de estatus medio). Para obtener el coeficiente gamma, se deben reordenar los datos en forma de tabla de frecuencia como sigue:

	Estatus ocupacional (X)		
Clase social subjetiva (Y)	*Alto* f	*Medio* f	*Bajo* f
Alta	56	42	15
Media	122	163	202
Baja	12	16	49
	190	221	266

El coeficiente de contingencia (C) y la V de Cramér son alternativas para gamma que suponen sólo datos nominales.

Solución a la situación de investigación 10
(Análisis de varianza o análisis de varianza en una dirección de Kruskal-Wallis)

La situación de investigación 10 representa una comparación de los puntajes de tres muestras independientes de entrevistados. La razón F (Capítulo 9) se utiliza para hacer comparaciones entre tres o más medias independientes cuando se han obtenido datos por intervalos. El análisis de varianza en una dirección de Kruskal-Wallis (Capítulo 10) es una alternativa no paramétrica que puede emplearse cuando sospechamos que los puntajes pueden no estar distribuidos normalmente en la población o cuando no se ha logrado el nivel de medición por intervalos.

Solución a la situación de investigación 11
(Razón t o prueba de la mediana)

La situación de investigación 11 representa una comparación entre los puntajes de dos muestras independientes de entrevistados. La razón t (Capítulo 8) se emplea para comparar dos medias cuando se han obtenido datos por intervalos. La prueba de la mediana (Capítulo 10) es una alternativa no paramétrica que puede aplicarse cuando no podemos suponer que los puntajes están distribuidos normalmente en la población o cuando no se ha alcanzado el nivel de medición por intervalos.

Solución a la situación de investigación 12
(Chi cuadrada)

La situación de investigación 12 representa una comparación de las frecuencias (disposición contra renuencia) en dos grupos de estudiantes (aquéllos a quienes se dijo que el profesor tenía 65 años contra aquéllos a quienes se dijo que tenía 25). La prueba de significancia chi cuadrada (Capítulo 10) se usa para hacer comparaciones entre dos o más muestras cuando se han obtenido datos nominales o de frecuencia. Los presentes resultados pueden colocarse en forma de la siguiente tabla 2 × 2, que representen 2 renglones y 2 columnas:

	Condición experimental		
Disposición para asistir	*Estudiantes a quienes se dijo que el profesor tenía 65 años* *f*	*Estudiantes a quienes se dijo que el profesor tenía 25 años* *f*	
Dispuesto	22	38	
Renuente	43	27	$N = 130$

Apéndices

Apéndice A
Una revisión de algunos aspectos fundamentales de matemáticas

Para los alumnos de estadística que necesitan repasar algunos de los fundamentos del álgebra y la aritmética, este apéndice incluye los problemas del trabajo con decimales, números negativos y raíces cuadradas. Otros problemas de las matemáticas se han estudiado en las partes apropiadas a través del texto. Por ejemplo, el Capítulo 1 identifica, define y compara tres niveles de medición; el Capítulo 2 estudia porcentajes, proporciones, razones y tasas; y el Capítulo 4 explica la sumatoria ($\leqslant$).

TRABAJANDO CON DECIMALES

Al sumar y restar decimales hay que asegurarse de colocar las comas decimales de los números directamente unas debajo de las otras. Por ejemplo, para sumar 3210,76, 2,541 y 98,3,

$$\begin{array}{r} 3210{,}76 \\ 2{,}541 \\ 98{,}3 \\ \hline 3311{,}601 \end{array}$$

Para restar 34,1 de 876,62,

$$\begin{array}{r} 876{,}62 \\ -34{,}1 \\ \hline 842{,}52 \end{array}$$

Al multiplicar decimales hay que asegurarse de que la respuesta contiene el mismo número de lugares decimales de su multiplicando y su multiplicador combinados. Por ejemplo,

Multiplicando →	63,41	2,6	0,000$	0,5
Multiplicador →	×0,05	× 1,4	×0,03	×0,5
Producto →	3,1705	3,64	0,000009	0,25

Antes de dividir conviene eliminar siempre los decimales del divisor, corriendo el punto decimal hacia la derecha tantos lugares como sea necesario para convertir al divisor en un número entero. Debe hacerse el correspondiente cambio del mismo número de lugares para los decimales del dividendo (esto es, si se corren dos lugares decimales en el divisor, entonces habrá que mover dos lugares en el dividendo). Este procedimiento indicará el número de lugares decimales de su respuesta.

$\frac{2,44}{0,02} = 122$ divisor → 0,02)2,44 ← dividendo, cociente 122,

$\frac{2,2}{0,4} = 2,2$ 0,4)0,88 — 2,2

$\frac{10,10}{0,10} = 1,01$ 10)10,10 — 1,01

$\frac{1010}{0,10} = 10100$ 0,10)1010,00 — 10100,0

Las operaciones aritméticas producen frecuentemente respuestas en forma decimal; por ejemplo, 2,034, 24,7, 86,001, y así sucesivamente. La pregunta que surge es sobre cuántos lugares decimales habremos de tener en nuestras respuestas. Una regla simple es la de llevar toda operación a tres lugares decimales más y redondear en dos lugares decimales más que los que se encontraron en el conjunto original de números.

Para ilustrar, si los datos se derivan de un conjunto original de números enteros (por ejemplo, 12, 9, 49 o 15), relizaríamos operaciones a tres lugares decimales (a milésimos) y expresaríamos nuestra respuesta en la centena más cercana. Por ejemplo,

3,889 = 3,89
1,224 = 1,22
7,761 = 7,76

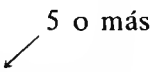
Generalmente se redondea al lugar decimal más cercano como sigue: se elimina el último dígito si es menor que 5 (en los ejemplos siguientes, el último dígito es el que indica los milésimos):

menor que 5

26,234 = 26,23

14,891 = 14,89
1,012 = 1,01

Hay que sumar un uno al dígito anterior si el último de ellos es igual a cinco o mayor (en los ejemplos siguientes el dígito precedente es el de las centenas):



26,236 = 26,24
14,899 = 14,90
1,015 = 1,02

Los siguientes se han redondeado al número entero más próximo:

3,1 = 3
3,5 = 4
4,5 = 5
4,8 = 5

Los siguientes se han redondeado a la decena más próxima:

3,11 = 3,1
3,55 = 3,6
4,45 = 4,5
4,17 = 4,2

Los siguientes se han redondeado a la centena más próxima:

3,328 = 3,33
4,823 = 4,82
3,065 = 3,07
3,055 = 3,06

EMPLEANDO LOS NUMEROS NEGATIVOS

Al sumar una serie de números negativos conviene asegurarse de dar un signo negativo a la suma. Por ejemplo,

−20	−3
−12	−9
−6	−4
−38	−16

Para sumar una serie que contenga números negativos y positivos se agrupan primero todos los negativos y los positivos por separado; se suma cada grupo y se restan sus sumas (la diferencia toma el signo del número mayor). Por ejemplo,

−6	+4	−6	+6
+4	+2	−1	−10
+2	+6	−3	−4
−1		−10	
−3			
−4			

Para restar un número negativo primero se le debe dar un signo positivo y luego seguir el procedimiento para sumar. La diferencia toma el signo del número mayor. Por ejemplo,

$$\begin{array}{r} 24 \\ -(-6) \\ \hline 30 \end{array}$$

−6 toma un signo positivo y, por lo tanto, se suma con el 24. Como el valor mayor es un número positivo (24), la diferencia (30) es un valor positivo.

$$\begin{array}{r} -6 \\ -(-24) \\ \hline 18 \end{array}$$

−24 toma un signo positivo y, por lo tanto se resta. Como el valor mayor es un número positivo (recuerde que se ha cambiado el signo a −24), la diferencia (18) es un valor positivo.

$$\begin{array}{r} -24 \\ -(-6) \\ \hline -18 \end{array}$$

−6 toma un signo positivo y, por lo tanto, se resta. Como el valor mayor es un número negativo (−24), la diferencia (−18) es valor negativo.

Al multiplicar (o dividir) dos números que tienen el mismo signo, hay que asignar siempre un signo positivo a su producto (o cociente). Por ejemplo,

$$(+8) \times (+5) = +40 \qquad \begin{array}{r} +8 \\ +5\overline{)+40} \end{array} \qquad \begin{array}{r} +8 \\ -5\overline{)-40} \end{array}$$
$$(-8) \times (-5) = +40$$

En el caso de dos números de signo diferente, hay que asignar un signo negativo (o cociente). Por ejemplo,

$$(-8) \times (+5) = -40 \qquad \begin{array}{r} -8 \\ -5\overline{)+40} \end{array}$$

COMO BUSCAR RAICES CUADRADAS CON LA TABLA A

Con la ayuda de la Tabla A, al final del libro, se puede encontrar fácilmente la raíz cuadrada ($\sqrt{n}$) de cualquier número entero (n) desde 1 hasta 1000.

Para encontrar la raíz cuadrada de números decimales, así como de números sobre 1000, puede ser útil comenzar con la columna de los cuadrados (n^2) de la Tabla A. La raíz cuadrada de cualquier número multiplicador por sí mismo debe ser igual a ese número. Como resultado, n, en la Tabla A, es en realidad la raíz cuadrada de n^2.

Para aprovechar plenamente la columna n^2 a fin de encontrar raíces cuadradas, debemos determinar cuántos dígitos preceden a la coma decimal en cualquier valor de raíz cuadrada. Una regla simple es aparear los dígitos que están antes de la coma decimal en una cifra. El número de pares equivale al número de dígitos que deben incluirse en la raíz cuadrada de la cifra. Por ejemplo,

$\sqrt{\underline{55}\underline{55},} = \underline{7}4,53$ (2 pares = 2 dígitos)
$\sqrt{\underline{55},55} = \ \underline{7},45$ (1 par = 1 dígito)

Cuando una cifra contiene un número impar de dígitos, el dígito non que precede a la coma decimal agrega otro dígito a la raíz cuadrada del número, como si se tratara de un par completo. Por ejemplo:

$\sqrt{\underline{55}5,5} = \underline{2}3,57$ (1 par + 1 dígito non = 2 dígitos)
$\sqrt{\underline{5},555} = \ \underline{2},36$ (1 dígito non = 1 dígito)

Para encontrar la raíz cuadrada de cualquier número menor que 1 se puede seguir este procedimiento:

1. Redondear a la centena más próxima

$\sqrt{0,328} = \sqrt{0,33}$
$\sqrt{0,823} = \sqrt{0,82}$
$\sqrt{0,0651} = \sqrt{0,07}$
$\sqrt{0,035} = \sqrt{0,04}$

2. Localizar la raíz cuadrada del número entero correspondiente en la Tabla A (Para encontrar el número entero simplemente se elimina la coma decimal)

$\sqrt{33} = 5,74$
$\sqrt{82} = 9,06$
$\sqrt{7} = 2,65$
$\sqrt{4} = 2$

3. Correr la coma decimal un lugar hacia la izquierda y redondear

$\sqrt{0,33} = 0,57$
$\sqrt{0,82} = 0,91$
$\sqrt{0,07} = 0,27$
$\sqrt{0,04} = 0,2$

Apéndice B

Tablas

TABLA A Cuadros, Raíces cuadradas e inversos de los números del 1 al 1 000

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
1	1	1.0000	1.000000	1.0000
2	4	1.4142	.500000	.7071
3	9	1.7321	.333333	.5774
4	16	2.0000	.250000	.5000
5	25	2.2361	.200000	.4472
6	36	2.4495	.166667	.4082
7	49	2.6458	.142857	.3780
8	64	2.8284	.125000	.3536
9	81	3.0000	.111111	.3333
10	100	3.1623	.100000	.3162
11	121	3.3166	.090909	.3015
12	144	3.4641	.083333	.2887
13	169	3.6056	.076923	.2774
14	196	3.7417	.071429	.2673
15	225	3.8730	.066667	.2582
16	256	4.0000	.062500	.2500
17	289	4.1231	.058824	.2425
18	324	4.2426	.055556	.2357
19	361	4.3589	.052632	.2294
20	400	4.4721	.050000	.2236
21	441	4.5826	.047619	.2182
22	484	4.6904	.045455	.2132
23	529	4.7958	.043478	.2085
24	576	4.8990	.041667	.2041
25	625	5.0000	.040000	.2000
26	676	5.0990	.038462	.1961
27	729	5.1962	.037037	.1925
28	784	5.2915	.035714	.1890
29	841	5.3852	.034483	.1857
30	900	5.4772	.033333	.1826
31	961	5.5678	.032258	.1796
32	1024	5.6569	.031250	.1768
33	1089	5.7446	.030303	.1741
34	1156	5.8310	.029412	.1715
35	1225	5.9161	.028571	.1690

* NOTA: Recuérdese que las Tablas son copias fieles del original en inglés, por lo tanto no se ha sustituido el punto que divide las fracciones de los enteros, por la coma decimal.

TABLA A
(Continuación)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
36	1296	6.0000	.027778	.1667
37	1369	6.0828	.027027	.1644
38	1444	6.1644	.026316	.1622
39	1521	6.2450	.025641	.1601
40	1600	6.3246	.025000	.1581
41	1681	6.4031	.024390	.1562
42	1764	6.4807	.023810	.1543
43	1849	6.5574	.023256	.1525
44	1936	6.6332	.022727	.1508
45	2025	6.7082	.022222	.1491
46	2116	6.7823	.021739	.1474
47	2209	6.8557	.021277	.1459
48	2304	6.9282	.020833	.1443
49	2401	7.0000	.020408	.1429
50	2500	7.0711	.020000	.1414
51	2601	7.1414	.019608	.1400
52	2704	7.2111	.019231	.1387
53	2809	7.2801	.018868	.1374
54	2916	7.3485	.018519	.1361
55	3025	7.4162	.018182	.1348
56	3136	7.4833	.017857	.1336
57	3249	7.5498	.017544	.1325
58	3364	7.6158	.017241	.1313
59	3481	7.6811	.016949	.1302
60	3600	7.7460	.016667	.1291
61	3721	7.8102	.016393	.1280
62	3844	7.8740	.016129	.1270
63	3969	7.9373	.015873	.1260
64	4096	8.0000	.015625	.1250
65	4225	8.0623	.015385	.1240
66	4356	8.1240	.015152	.1231
67	4489	8.1854	.014925	.1222
68	4624	8.2462	.014706	.1213
69	4761	8.3066	.014493	.1204
70	4900	8.3666	.014286	.1195
71	5041	8.4261	.014085	.1187
72	5184	8.4853	.013889	.1179
73	5329	8.5440	.013699	.1170
74	5476	8.6023	.013514	.1162
75	5625	8.6603	.013333	.1155
76	5776	8.7178	.013158	.1147
77	5929	8.7750	.012987	.1140
78	6084	8.8318	.012821	.1132
79	6241	8.8882	.012658	.1125
80	6400	8.9443	.012500	.1118
81	6561	9.0000	.012346	.1111
82	6724	9.0554	.012195	.1104
83	6889	9.1104	.012048	.1098
84	7056	9.1652	.011905	.1091
85	7225	9.2195	.011765	.1085

TABLA A
(Continuación)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
86	7396	9.2736	.011628	.1078
87	7569	9.3274	.011494	.1072
88	7744	9.3808	.011364	.1066
89	7921	9.4340	.011236	.1060
90	8100	9.4868	.011111	.1054
91	8281	9.5394	.010989	.1048
92	8464	9.5917	.010870	.1043
93	8649	9.6437	.010753	.1037
94	8836	9.6954	.010638	.1031
95	9025	9.7468	.010526	.1026
96	9216	9.7980	.010417	.1021
97	9409	9.8489	.010309	.1015
98	9604	9.8995	.010204	.1010
99	9801	9.9499	.010101	.1005
100	10000	10.0000	.010000	.1000
101	10201	10.0499	.009901	.0995
102	10404	10.0995	.009804	.0990
103	10609	10.1489	.009709	.0985
104	10816	10.1980	.009615	.0981
105	11025	10.2470	.009524	.0976
106	11236	10.2956	.009434	.0971
107	11449	10.3441	.009346	.0967
108	11664	10.3923	.009259	.0962
109	11881	10.4403	.009174	.0958
110	12100	10.4881	.009091	.0953
111	12321	10.5357	.009009	.0949
112	12544	10.5830	.008929	.0945
113	12769	10.6301	.008850	.0941
114	12996	10.6771	.008772	.0937
115	13225	10.7238	.008696	.0933
116	13456	10.7703	.008621	.0928
117	13689	10.8167	.008547	.0925
118	13924	10.8628	.008475	.0921
119	14161	10.9087	.008403	.0917
120	14400	10.9545	.008333	.0913
121	14641	11.0000	.008264	.0909
122	14884	11.0454	.008197	.0905
123	15129	11.0905	.008130	.0902
124	15376	11.1355	.008065	.0898
125	15625	11.1803	.008000	.0894
126	15876	11.2250	.007937	.0891
127	16129	11.2694	.007874	.0887
128	16384	11.3137	.007813	.0884
129	16641	11.3578	.007752	.0880
130	16900	11.4018	.007692	.0877
131	17161	11.4455	.007634	.0874
132	17424	11.4891	.007576	.0870
133	17689	11.5326	.007519	.0867
134	17956	11.5758	.007463	.0864
135	18225	11.6190	.007407	.0861

TABLA A
(Continuación)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
136	18496	11.6619	.007353	.0857
137	18769	11.7047	.007299	.0854
138	19044	11.7473	.007246	.0851
139	19321	11.7898	.007194	.0848
140	19600	11.8322	.007143	.0845
141	19881	11.8743	.007092	.0842
142	20164	11.9164	.007042	.0839
143	20449	11.9583	.006993	.0836
144	20736	12.0000	.006944	.0833
145	21025	12.0416	.006897	.0830
146	21316	12.0830	.006849	.0828
147	21609	12.1244	.006803	.0825
148	21904	12.1655	.006757	.0822
149	22201	12.2066	.006711	.0819
150	22500	12.2474	.006667	.0816
151	22801	12.2882	.006623	.0814
152	23104	12.3288	.006579	.0811
153	23409	12.3693	.006536	.0808
154	23716	12.4097	.006494	.0806
155	24025	12.4499	.006452	.0803
156	24336	12.4900	.006410	.0801
157	24649	12.5300	.006369	.0798
158	24964	12.5698	.006329	.0796
159	25281	12.6095	.006289	.0793
160	25600	12.6491	.006250	.0791
161	25921	12.6886	.006211	.0788
162	26244	12.7279	.006173	.0786
163	26569	12.7671	.006135	.0783
164	26896	12.8062	.006098	.0781
165	27225	12.8452	.006061	.0778
166	27556	12.8841	.006024	.0776
167	27889	12.9228	.005988	.0774
168	28224	12.9615	.005952	.0772
169	28561	13.0000	.005917	.0769
170	28900	13.0384	.005882	.0767
171	29241	13.0767	.005848	.0765
172	29584	13.1149	.005814	.0762
173	29929	13.1529	.005780	.0760
174	30276	13.1909	.005747	.0758
175	30625	13.2288	.005714	.0756
176	30976	13.2665	.005682	.0754
177	31329	13.3041	.005650	.0752
178	31684	13.3417	.005618	.0750
179	32041	13.3791	.005587	.0747
180	32400	13.4164	.005556	.0745
181	32761	13.4536	.005525	.0743
182	33124	13.4907	.005495	.0741
183	33489	13.5277	.005464	.0739
184	33856	13.5647	.005435	.0737
185	34225	13.6015	.005405	.0735

TABLA A
(continuación)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
186	34596	13.6382	.005376	.0733
187	34969	13.6748	.005348	.0731
188	35344	13.7113	.005319	.0729
189	35721	13.7477	.005291	.0727
190	36100	13.7840	.005263	.0725
191	36481	13.8203	.005236	.0724
192	36864	13.8564	.005208	.0722
193	37249	13.8924	.005181	.0720
194	37636	13.9284	.005155	.0718
195	38025	13.9642	.005128	.0716
196	38416	14.0000	.005102	.0714
197	38809	14.0357	.005076	.0712
198	39204	14.0712	.005051	.0711
199	39601	14.1067	.005025	.0709
200	40000	14.1421	.005000	.0707
201	40401	14.1774	.004975	.0705
202	40804	14.2127	.004950	.0704
203	41209	14.2478	.004926	.0702
204	41616	14.2829	.004902	.0700
205	42025	14.3178	.004878	.0698
206	42436	14.3527	.004854	.0697
207	42849	14.3875	.004831	.0695
208	43264	14.4222	.004808	.0693
209	43681	14.4568	.004785	.0692
210	44100	14.4914	.004762	.0690
211	44521	14.5258	.004739	.0688
212	44944	14.5602	.004717	.0687
213	45369	14.5945	.004695	.0685
214	45796	14.6287	.004673	.0684
215	46225	14.6629	.004651	.0682
216	46656	14.6969	.004630	.0680
217	47089	14.7309	.004608	.0679
218	47524	14.7648	.004587	.0677
219	47961	14.7986	.004566	.0676
220	48400	14.8324	.004545	.0674
221	48841	14.8661	.004525	.0673
222	49284	14.8997	.004505	.0671
223	49729	14.9332	.004484	.0670
224	50176	14.9666	.004464	.0668
225	50625	15.0000	.004444	.0667
226	51076	15.0333	.004425	.0665
227	51529	15.0665	.004405	.0664
228	51984	15.0997	.004386	.0662
229	52441	15.1327	.004367	.0661
230	52900	15.1658	.004348	.0659
231	53361	15.1987	.004329	.0658
232	53824	15.2315	.004310	.0657
233	54289	15.2643	.004292	.0655
234	54756	15.2971	.004274	.0654
235	55225	15.3297	.004255	.0652

TABLA A
(Continuación)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
236	55696	15.3623	.004237	.0651
237	56169	15.3948	.004219	.0650
238	56644	15.4272	.004202	.0648
239	57121	15.4596	.004184	.0647
240	57600	15.4919	.004167	.0645
241	58081	15.5242	.004149	.0644
242	58564	15.5563	.004132	.0643
243	59049	15.5885	.004115	.0642
244	59536	15.6205	.004098	.0640
245	60025	15.6525	.004082	.0639
246	60516	15.6844	.004065	.0638
247	61009	15.7162	.004049	.0636
248	61504	15.7480	.004032	.0635
249	62001	15.7797	.004016	.0634
250	62500	15.8114	.004000	.0632
251	63001	15.8430	.003984	.0631
252	63504	15.8745	.003968	.0630
253	64009	15.9060	.003953	.0629
254	64516	15.9374	.003937	.0627
255	65025	15.9687	.003922	.0626
256	65536	16.0000	.003906	.0625
257	66049	16.0312	.003891	.0624
258	66564	16.0624	.003876	.0623
259	67081	16.0935	.003861	.0621
260	67600	16.1245	.003846	.0620
261	68121	16.1555	.003831	.0619
262	68644	16.1864	.003817	.0618
263	69169	16.2173	.003802	.0617
264	69696	16.2481	.003788	.0615
265	70225	16.2788	.003774	.0614
266	70756	16.3095	.003759	.0613
267	71289	16.3401	.003745	.0612
268	71824	16.3707	.003731	.0611
269	72361	16.4012	.003717	.0610
270	72900	16.4317	.003704	.0609
271	73441	16.4621	.003690	.0607
272	73984	16.4924	.003676	.0606
273	74529	16.5227	.003663	.0605
274	75076	16.5529	.003650	.0604
275	75625	16.5831	.003636	.0603
276	76176	16.6132	.003623	.0602
277	76729	16.6433	.003610	.0601
278	77284	16.6733	.003597	.0600
279	77841	16.7033	.003584	.0599
280	78400	16.7332	.003571	.0598
281	78961	16.7631	.003559	.0597
282	79524	16.7929	.003546	.0595
283	80089	16.8226	.003534	.0594
284	80656	16.8523	.003521	.0593
285	81225	16.8819	.003509	.0592

TABLA A
(Continuación)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
286	81796	16.9115	.003497	.0591
287	82369	16.9411	.003484	.0590
288	82944	16.9706	.003472	.0589
289	83521	17.0000	.003460	.0588
290	84100	17.0294	.003448	.0587
291	84681	17.0587	.003436	.0586
292	85264	17.0880	.003425	.0585
293	85849	17.1172	.003413	.0584
294	86436	17.1464	.003401	.0583
295	87025	17.1756	.003390	.0582
296	87616	17.2047	.003378	.0581
297	88209	17.2337	.003367	.0580
298	88804	17.2627	.003356	.0579
299	89401	17.2916	.003344	.0578
300	90000	17.3205	.003333	.0577
301	90601	17.3494	.003322	.0576
302	91204	17.3781	.003311	.0575
303	91809	17.4069	.003300	.0574
304	92416	17.4356	.003289	.0574
305	93025	17.4642	.003279	.0573
306	93636	17.4929	.003268	.0572
307	94249	17.5214	.003257	.0571
308	94864	17.5499	.003247	.0570
309	95481	17.5784	.003236	.0569
310	96100	17.6068	.003226	.0568
311	96721	17.6352	.003215	.0567
312	97344	17.6635	.003205	.0566
313	97969	17.6918	.003195	.0565
314	98596	17.7200	.003185	.0564
315	99225	17.7482	.003175	.0563
316	99856	17.7764	.003165	.0563
317	100489	17.8045	.003155	.0562
318	101124	17.8326	.003145	.0561
319	101761	17.8606	.003135	.0560
320	102400	17.8885	.003125	.0559
321	103041	17.9165	.003115	.0558
322	103684	17.9444	.003106	.0557
323	104329	17.9722	.003096	.0556
324	104976	18.0000	.003086	.0556
325	105625	18.0278	.003077	.0555
326	106276	18.0555	.003067	.0554
327	106929	18.0831	.003058	.0553
328	107584	18.1108	.003049	.0552
329	108241	18.1384	.003040	.0551
330	108900	18.1659	.003030	.0550
331	109561	18.1934	.003021	.0550
332	110224	18.2209	.003012	.0549
333	110889	18.2483	.003003	.0548
334	111556	18.2757	.002994	.0547
335	112225	18.3030	.002985	.0546

TABLA A
(Continuación)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
336	112896	18.3303	.002976	.0546
337	113569	18.3576	.002967	.0545
338	114244	18.3848	.002959	.0544
339	114921	18.4120	.002950	.0543
340	115600	18.4391	.002941	.0542
341	116281	18.4662	.002933	.0542
342	116964	18.4932	.002924	.0541
343	117649	18.5203	.002915	.0540
344	118336	18.5472	.002907	.0539
345	119025	18.5742	.002899	.0538
346	119716	18.6011	.002890	.0538
347	120409	18.6279	.002882	.0537
348	121104	18.6548	.002874	.0536
349	121801	18.6815	.002865	.0535
350	122500	18.7083	.002857	.0535
351	123201	18.7350	.002849	.0534
352	123904	18.7617	.002841	.0533
353	124609	18.7883	.002833	.0532
354	125316	18.8149	.002825	.0531
355	126025	18.8414	.002817	.0531
356	126736	18.8680	.002809	.0530
357	127449	18.8944	.002801	.0529
358	128164	18.9209	.002793	.0529
359	128881	18.9473	.002786	.0528
360	129600	18.9737	.002778	.0527
361	130321	19.0000	.002770	.0526
362	131044	19.0263	.002762	.0526
363	131769	19.0526	.002755	.0525
364	132496	19.0788	.002747	.0524
365	133225	19.1050	.002740	.0523
366	133956	19.1311	.002732	.0523
367	134689	19.1572	.002725	.0522
368	135424	19.1833	.002717	.0521
369	136161	19.2094	.002710	.0521
370	136900	19.2354	.002703	.0520
371	137641	19.2614	.002695	.0519
372	138384	19.2873	.002688	.0518
373	139129	19.3132	.002681	.0518
374	139876	19.3391	.002674	.0517
375	140625	19.3649	.002667	.0516
376	141376	19.3907	.002660	.0516
377	142129	19.4165	.002653	.0515
378	142884	19.4422	.002646	.0514
379	143641	19.4679	.002639	.0514
380	144400	19.4936	.002632	.0513
381	145161	19.5192	.002625	.0512
382	145924	19.5448	.002618	.0512
383	146689	19.5704	.002611	.0511
384	147456	19.5959	.002604	.0510
385	148225	19.6214	.002597	.0510

TABLA A
(Continuación)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
386	148996	19.6469	.002591	.0509
387	149769	19.6723	.002584	.0508
388	150544	19.6977	.002577	.0508
389	151321	19.7231	.002571	.0507
390	152100	19.7484	.002564	.0506
391	152881	19.7737	.002558	.0506
392	153664	19.7990	.002551	.0505
393	154449	19.8242	.002545	.0504
394	155236	19.8494	.002538	.0504
395	156025	19.8746	.002532	.0503
396	156816	19.8997	.002525	.0503
397	157609	19.9249	.002519	.0502
398	158404	19.9499	.002513	.0501
399	159201	19.9750	.002506	.0501
400	160000	20.0000	.002500	.0500
401	160801	20.0250	.002494	.0499
402	161604	20.0499	.002488	.0499
403	162409	20.0749	.002481	.0498
404	163216	20.0998	.002475	.0498
405	164025	20.1246	.002469	.0497
406	164836	20.1494	.002463	.0496
407	165649	20.1742	.002457	.0496
408	166464	20.1990	.002451	.0495
409	167281	20.2237	.002445	.0494
410	168100	20.2485	.002439	.0494
411	168921	20.2731	.002433	.0493
412	169744	20.2978	.002427	.0493
413	170569	20.3224	.002421	.0492
414	171396	20.3470	.002415	.0491
415	172225	20.3715	.002410	.0491
416	173056	20.3961	.002404	.0490
417	173889	20.4206	.002398	.0490
418	174724	20.4450	.002392	.0489
419	175561	20.4695	.002387	.0489
420	176400	20.4939	.002381	.0488
421	177241	20.5183	.002375	.0487
422	178084	20.5426	.002370	.0487
423	178929	20.5670	.002364	.0486
424	179776	20.5913	.002358	.0486
425	180625	20.6155	.002353	.0485
426	181476	20.6398	.002347	.0485
427	182329	20.6640	.002342	.0484
428	183184	20.6882	.002336	.0483
429	184041	20.7123	.002331	.0483
430	184900	20.7364	.002326	.0482
431	185761	20.7605	.002320	.0482
432	186624	20.7846	.002315	.0481
433	187489	20.8087	.002309	.0481
434	188356	20.8327	.002304	.0480
435	189225	20.8567	.002299	.0479

TABLA A
(Continuación)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
436	190096	20.8806	.002294	.0479
437	190969	20.9045	.002288	.0478
438	191844	20.9284	.002283	.0478
439	192721	20.9523	.002278	.0477
440	193600	20.9762	.002273	.0477
441	194481	21.0000	.002268	.0476
442	195364	21.0238	.002262	.0476
443	196249	21.0476	.002257	.0475
444	197136	21.0713	.002252	.0475
445	198025	21.0950	.002247	.0474
446	198916	21.1187	.002242	.0474
447	199809	21.1424	.002237	.0473
448	200704	21.1660	.002232	.0472
449	201601	21.1896	.002227	.0472
450	202500	21.2132	.002222	.0471
451	203401	21.2368	.002217	.0471
452	204304	21.2603	.002212	.0470
453	205209	21.2838	.002208	.0470
454	206116	21.3073	.002203	.0469
455	207025	21.3307	.002198	.0469
456	207936	21.3542	.002193	.0468
457	208849	21.3776	.022188	.0468
458	209764	21.4009	.002183	.0467
459	210681	21.4243	.002179	.0467
460	211600	21.4476	.002174	.0466
461	212521	21.4709	.002169	.0466
462	213444	21.4942	.002165	.0465
463	214369	21.5174	.002160	.0465
464	215296	21.5407	.002155	.0464
465	216225	21.5639	.002151	.0464
466	217156	21.5870	.002146	.0463
467	218089	21.6102	.002141	.0463
468	219024	21.6333	.002137	.0462
469	219961	21.6564	.002132	.0462
470	220900	21.6795	.002128	.0461
471	221841	21.7025	.002123	.0461
472	222784	21.7256	.002119	.0460
473	223729	21.7486	.002114	.0460
474	224676	21.7715	.002110	.0459
475	225625	21.7945	.002105	.0459
476	226576	21.8174	.002101	.0458
477	227529	21.8403	.002096	.0458
478	228484	21.8632	.002092	.0457
479	229441	21.8861	.002088	.0457
480	230400	21.9089	.002083	.0456
481	231361	21.9317	.002079	.0456
482	232324	21.9545	.002075	.0455
483	233289	21.9773	.002070	.0455
484	234256	22.0000	.002066	.0455
485	235225	22.0227	.002062	.0454

TABLA A
(Continuación)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
486	236196	22.0454	.002058	.0454
487	237169	22.0681	.002053	.0453
488	238144	22.0907	.002049	.0453
489	239121	22.1133	.002045	.0452
490	240100	22.1359	.002041	.0452
491	241081	22.1585	.002037	.0451
492	242064	22.1811	.002033	.0451
493	243049	22.2036	.002028	.0450
494	244036	22.2261	.002024	.0450
495	245025	22.2486	.002020	.0449
496	246016	22.2711	.002016	.0448
497	247009	22.2935	.002012	.0449
498	248004	22.3159	.002008	.0449
499	249001	22.3383	.002004	.0448
500	250000	22.3607	.002000	.0447
501	251001	22.3830	.001996	.0447
502	252004	22.4054	.001992	.0446
503	253009	22.4277	.001988	.0446
504	254016	22.4499	.001984	.0445
505	255025	22.4722	.001980	.0445
506	256036	22.4944	.001976	.0445
507	257049	22.5167	.001972	.0444
508	258064	22.5389	.001969	.0444
509	259081	22.5610	.001965	.0443
510	260100	22.5832	.001961	.0443
511	261121	22.6053	.001957	.0442
512	262144	22.6274	.001953	.0442
513	263169	22.6495	.001949	.0442
514	264196	22.6716	.001946	.0441
515	265225	22.6936	.001942	.0441
516	266256	22.7156	.001938	.0440
517	267289	22.7376	.001934	.0440
518	268324	22.7596	.001931	.0439
519	269361	22.7816	.001927	.0439
520	270400	22.8035	.001923	.0439
521	271441	22.8254	.001919	.0438
522	272484	22.8473	.001916	.0438
523	273529	22.8692	.001912	.0437
524	274576	22.8910	.001908	.0437
525	275625	22.9129	.001905	.0436
526	276676	22.9347	.001901	.0436
527	277729	22.9565	.001898	.0436
528	278784	22.9783	.001894	.0435
529	279841	23.0000	.001890	.0435
530	280900	23.0217	.001887	.0434
531	281961	23.0434	.001883	.0434
532	283024	23.0651	.001880	.0434
533	284089	23.0868	.001876	.0433
534	285156	23.1084	.001873	.0433
535	286225	23.1301	.001869	.0432

TABLA A
(Continuación)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
536	287296	23.1517	.001866	.0432
537	288369	23.1733	.001862	.0432
538	289444	23.1948	.001859	.0431
539	290521	23.2164	.001855	.0431
540	291600	23.2379	.001852	.0430
541	292681	23.2594	.001848	.0430
542	293764	23.2809	.001845	.0430
543	294849	23.3024	.001842	.0429
544	295936	23.3238	.001838	.0429
545	297025	23.3452	.001835	.0428
546	298116	23.3666	.001832	.0428
547	299209	23.3880	.001828	.0428
548	300304	23.4094	.001825	.0427
549	301401	23.4307	.001821	.0427
550	302500	23.4521	.001818	.0426
551	303601	23.4734	.001815	.0426
552	304704	23.4947	.001812	.0426
553	305809	23.5160	.001808	.0425
554	306916	23.5372	.001805	.0425
555	308025	23.5584	.001802	.0424
556	309136	23.5797	.001799	.0424
557	310249	23.6008	.001795	.0424
558	311364	23.6220	.001792	.0423
559	312481	23.6432	.001789	.0423
560	313600	23.6643	.001786	.0423
561	314721	23.6854	.001783	.0422
562	315844	23.7065	.001779	.0422
563	316969	23.7276	.001776	.0421
564	318096	23.7487	.001773	.0421
565	319225	23.7697	.001770	.0421
566	320356	23.7908	.001767	.0420
567	321489	23.8118	.001764	.0420
568	322624	23.8328	.001761	.0420
569	323761	23.8537	.001757	.0419
570	324900	23.8747	.001754	.0419
571	326041	23.8956	.001751	.0418
572	327184	23.9165	.001748	.0418
573	328329	23.9374	.001745	.0418
574	329476	23.9583	.001742	.0417
575	330625	23.9792	.001739	.0417
576	331776	24.0000	.001736	.0417
577	332929	24.0208	.001733	.0416
578	334084	24.0416	.001730	.0416
579	335241	24.0624	.001727	.0416
580	336400	24.0832	.001724	.0415
581	337561	24.1039	.001721	.0415
582	338724	24.1247	.001718	.0415
583	339889	24.1454	.001715	.0414
584	341056	24.1661	.001712	.0414
585	342225	24.1868	.001709	.0413

TABLA A
(Continuación)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
586	343396	24.2074	.001706	.0413
587	344569	24.2281	.001704	.0413
588	345744	24.2487	.001701	.0412
589	346921	24.2693	.001698	.0412
590	348100	24.2899	.001695	.0412
591	349281	24.3105	.001692	.0411
592	350464	24.3311	.001689	.0411
593	351649	24.3516	.001686	.0411
594	352836	24.3721	.001684	.0410
595	354025	24.3926	.001681	.0410
596	355216	24.4131	.001678	.0410
597	356409	24.4336	.001675	.0409
598	357604	24.4540	.001672	.0409
599	358801	24.4745	.001669	.0409
600	360000	24.4949	.001667	.0408
601	361201	24.5153	.001664	.0408
602	362404	24.5357	.001661	.0408
603	363609	24.5561	.001658	.0407
604	364816	24.5764	.001656	.0407
605	366025	24.5967	.001653	.0407
606	367236	24.6171	.001650	.0406
607	368449	24.6374	.001647	.0406
608	369664	24.6577	.001645	.0406
609	370881	24.6779	.001642	.0405
610	372100	24.6982	.001639	.0405
611	373321	24.7184	.001637	.0405
612	374544	24.7386	.001634	.0404
613	375769	24.7588	.001631	.0404
614	376996	24.7790	.001629	.0404
615	378225	24.7992	.001626	.0403
616	379456	24.8193	.001623	.0403
617	380689	24.8395	.001621	.0403
618	381924	24.8596	.001618	.0402
619	383161	24.8797	.001616	.0402
620	384400	24.8998	.001613	.0402
621	385641	24.9199	.001610	.0401
622	386884	24.9399	.001608	.0401
623	388129	24.9600	.001605	.0401
624	389376	24.9800	.001603	.0400
625	390625	25.0000	.001600	.0400
626	391876	25.0200	.001597	.0400
627	393129	25.0400	.001595	.0399
628	394384	25.0599	.001592	.0399
629	395641	25.0799	.001590	.0399
630	396900	25.0998	.001587	.0398
631	398161	25.1197	.001585	.0398
632	399424	25.1396	.001582	.0398
633	400689	25.1595	.001580	.0397
634	401956	25.1794	.001577	.0397
635	403225	25.1992	.001575	.0397

TABLA A
(Continuación)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
636	404496	25.2190	.001572	.0397
637	405769	25.2389	.001570	.0396
638	407044	25.2587	.001567	.0396
639	408321	25.2784	.001565	.0396
640	409600	25.2982	.001563	.0395
641	410881	25.3180	.001560	.0395
642	412164	25.3377	.001558	.0395
643	413449	25.3574	.001555	.0394
644	414736	25.3772	.001553	.0394
645	416025	25.3969	.001550	.0394
646	417316	25.4165	.001548	.0393
647	418609	25.4362	.001546	.0393
648	419904	25.4558	.001543	.0393
649	421201	25.4755	.001541	.0393
650	422500	25.4951	.001538	.0392
651	423801	25.5147	.001536	.0392
652	425104	25.5343	.001534	.0392
653	426409	25.5539	.001531	.0391
654	427716	25.5734	.001529	.0391
655	429025	25.5930	.001527	.0391
656	430336	25.6125	.001524	.0390
657	431649	25.6320	.001522	.0390
658	432964	25.6515	.001520	.0390
659	434281	25.6710	.001517	.0390
660	435600	25.6905	.001515	.0389
661	436921	25.7099	.001513	.0389
662	438244	25.7294	.001511	.0389
663	439569	25.7488	.001508	.0388
664	440896	25.7682	.001506	.0388
665	442225	25.7876	.001504	.0388
666	443556	25.8070	.001502	.0387
667	444889	25.8263	.001499	.0387
668	446224	25.8457	.001497	.0387
669	447561	25.8650	.001495	.0387
670	448900	25.8844	.001493	.0386
671	450241	25.9037	.001490	.0386
672	451584	25.9230	.001488	.0386
673	452929	25.9422	.001486	.0385
674	454276	25.9615	.001484	.0385
675	455625	25.9808	.001481	.0385
676	456976	26.0000	.001479	.0385
677	458329	26.0192	.001477	.0384
678	459684	26.0384	.001475	.0384
679	461041	26.0576	.001473	.0384
680	462400	26.0768	.001471	.0383
681	463761	26.0960	.001468	.0383
682	465124	26.1151	.001466	.0383
683	466489	26.1343	.001464	.0383
684	467856	26.1534	.001462	.0382
685	469225	26.1725	.001460	.0382

TABLA A
(Continuación)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
686	470596	26.1916	.001458	.0382
687	471969	26.2107	.001456	.0382
688	473344	26.2298	.001453	.0381
689	474721	26.2488	.001451	.0381
690	476100	26.2679	.001449	.0381
691	477481	26.2869	.001447	.0380
692	478864	26.3059	.001445	.0380
693	480249	26.3249	.001443	.0380
694	481636	26.3439	.001441	.0380
695	483025	26.3629	.001439	.0379
696	484416	26.3818	.001437	.0379
697	485809	26.4008	.001435	.0379
698	487204	26.4197	.001433	.0379
699	488601	26.4386	.001431	.0378
700	490000	26.4575	.001429	.0378
701	491401	26.4764	.001427	.0378
702	492804	26.4953	.001425	.0377
703	494209	26.5141	.001422	.0377
704	495616	26.5330	.001420	.0377
705	497025	26.5518	.001418	.0377
706	498436	26.5707	.001416	.0376
707	499849	26.5895	.001414	.0376
708	501264	26.6083	.001412	.0376
709	502681	26.6271	.001410	.0376
710	504100	26.6458	.001408	.0375
711	505521	26.6646	.001406	.0375
712	506944	26.6833	.001404	.0375
713	508369	26.7021	.001403	.0375
714	509796	26.7208	.001401	.0374
715	511225	26.7395	.001399	.0374
716	512656	26.7582	.001397	.0374
717	514089	26.7769	.001395	.0373
718	515524	26.7955	.001393	.0373
719	516961	26.8142	.001391	.0373
720	518400	26.8328	.001389	.0373
721	519841	26.8514	.001387	.0372
722	521284	26.8701	.001385	.0372
723	522729	26.8887	.001383	.0372
724	524176	26.9072	.001381	.0372
725	525625	26.9258	.001379	.0371
726	527076	26.9444	.001377	.0371
727	528529	26.9629	.001376	.0371
728	529984	26.9815	.001374	.0371
729	531441	27.0000	.001372	.0370
730	532900	27.0185	.001370	.0370
731	534361	27.0370	.001368	.0370
732	535824	27.0555	.001366	.0370
733	537289	27.0740	.001364	.0369
734	538756	27.0924	.001362	.0369
735	540225	27.1109	.001361	.0369

TABLA A
(Continuación)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
736	541696	27.1293	.001359	.0369
737	543169	27.1477	.001357	.0368
738	544644	27.1662	.001355	.0368
739	546121	27.1846	.001353	.0368
740	547600	27.2029	.001351	.0368
741	549081	27.2213	.001350	.0367
742	550564	27.2397	.001348	.0367
743	552049	27.2580	.001346	.0367
744	553536	27.2764	.001344	.0367
745	555025	27.2947	.001342	.0366
746	556516	27.3130	.001340	.0366
747	558009	27.3313	.001339	.0366
748	559504	27.3496	.001337	.0366
749	561001	27.3679	.001335	.0365
750	562500	27.3861	.001333	.0365
751	564001	27.4044	.001332	.0365
752	565504	27.4226	.001330	.0365
753	567009	27.4408	.001328	.0364
754	568516	27.4591	.001326	.0364
755	570025	27.4773	.001325	.0364
756	571536	27.4955	.001323	.0364
757	573049	27.5136	.001321	.0363
758	574564	27.5318	.001319	.0363
759	576081	27.5500	.001318	.0363
760	577600	27.5681	.001316	.0363
761	579121	27.5862	.001314	.0363
762	580644	27.6043	.001312	.0362
763	582169	27.6225	.001311	.0362
764	583696	27.6405	.001309	.0362
765	585225	27.6586	.001307	.0362
766	586756	27.6767	.001305	.0361
767	588289	27.6948	.001304	.0361
768	589824	27.7128	.001302	.0361
769	591361	27.7308	.001300	.0361
770	592900	27.7489	.001299	.0360
771	594441	27.7669	.001297	.0360
772	595984	27.7849	.001295	.0360
773	597529	27.8029	.001294	.0360
774	599076	27.8209	.001292	.0359
775	600625	27.8388	.001290	.0359
776	602176	27.8568	.001289	.0359
777	603729	27.8747	.001287	.0359
778	605284	27.8927	.001285	.0359
779	606841	27.9106	.001284	.0358
780	608400	27.9285	.001282	.0358
781	609961	27.9464	.001280	.0358
782	611524	27.9643	.001279	.0358
783	613089	27.9821	.001277	.0357
784	614656	28.0000	.001276	.0357
785	616225	28.0179	.001274	.0357

TABLA A
(Continuación)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
786	617796	28.0357	.001272	.0357
787	619369	28.0535	.001271	.0356
788	620944	28.0713	.001269	.0356
789	622521	28.0891	.001267	.0356
790	624100	28.1069	.001266	.0356
791	625681	28.1247	.001264	.0356
792	627264	28.1425	.001263	.0355
793	628849	28.1603	.001261	.0355
794	630436	28.1780	.001259	.0355
795	632025	28.1957	.001258	.0355
796	633616	28.2135	.001256	.0354
797	635209	28.2312	.001255	.0354
798	636804	28.2489	.001253	.0354
799	638401	28.2666	.001252	.0354
800	640000	28.2843	.001250	.0354
801	641601	28.3019	.001248	.0353
802	643204	28.3196	.001247	.0353
803	644809	28.3373	.001245	.0353
804	646416	28.3549	.001244	.0353
805	648025	28.3725	.001242	.0352
806	649636	28.3901	.001241	.0352
807	651249	28.4077	.001239	.0352
808	652864	28.4253	.001238	.0352
809	654481	28.4429	.001236	.0352
810	656100	28.4605	.001235	.0351
811	657721	28.4781	.001233	.0351
812	659344	28.4956	.001232	.0351
813	660969	28.5132	.001230	.0351
814	662596	28.5307	.001229	.0351
815	664225	28.5482	.001227	.0350
816	665856	28.5657	.001225	.0350
817	667489	28.5832	.001224	.0350
818	669124	28.6007	.001222	.0350
819	670761	28.6182	.001221	.0349
820	672400	28.6356	.001220	.0349
821	674041	28.6531	.001218	.0349
822	675684	28.6705	.001217	.0349
823	677329	28.6880	.001215	.0349
824	678976	28.7054	.001214	.0348
825	680625	28.7228	.001212	.0348
826	682276	28.7402	.001211	.0348
827	683929	28.7576	.001209	.0348
828	685584	28.7750	.001208	.0348
829	687241	28.7924	.001206	.0347
830	688900	28.8097	.001205	.0347
831	690561	28.8271	.001203	.0347
832	692224	28.8444	.001202	.0347
833	693889	28.8617	.001200	.0346
834	695556	28.8791	.001199	.0346
835	697225	28.8964	.001198	.0346

TABLA A
(Continuación)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
836	698896	28.9137	.001196	.0346
837	700569	28.9310	.001195	.0346
838	702244	28.9482	.001193	.0345
839	703921	28.9655	.001192	.0345
840	705600	28.9828	.001190	.0345
841	707281	29.0000	.001189	.0345
842	708964	29.0172	.001188	.0345
843	710649	29.0345	.001186	.0344
844	712336	29.0517	.001185	.0344
845	714025	29.0689	.001183	.0344
846	715716	29.0861	.001182	.0344
847	717409	29.1033	.001181	.0344
848	719104	29.1204	.001179	.0343
849	720801	29.1376	.001178	.0343
850	722500	29.1548	.001176	.0343
851	724201	29.1719	.001175	.0343
852	725904	29.1890	.001174	.0343
853	727609	29.2062	.001172	.0342
854	729316	29.2233	.001171	.0342
855	731025	29.2404	.001170	.0342
856	732736	29.2575	.001168	.0342
857	734449	29.2746	.001167	.0342
858	736164	29.2916	.001166	.0341
859	737881	29.3087	.001164	.0341
860	739600	29.3258	.001163	.0341
861	741321	29.3428	.001161	.0341
862	743044	29.3598	.001160	.0341
863	744769	29.3769	.001159	.0340
864	746496	29.3939	.001157	.0340
865	748225	29.4109	.001156	.0340
866	749956	29.4279	.001155	.0340
867	751689	29.4449	.001153	.0340
868	753424	29.4618	.001152	.0339
869	755161	29.4788	.001151	.0339
870	756900	29.4958	.001149	.0339
871	758641	29.5127	.001148	.0339
872	760384	29.5296	.001147	.0339
873	762129	29.5466	.001145	.0338
874	763876	29.5635	.001144	.0338
875	765625	29.5804	.001143	.0338
876	767376	29.5973	.001142	.0338
877	769129	29.6142	.001140	.0338
878	770884	29.6311	.001139	.0337
879	772641	29.6479	.001138	.0337
880	774400	29.6648	.001136	.0337
881	776161	29.6816	.001135	.0337
882	777924	29.6985	.001134	.0337
883	779689	29.7153	.001133	.0337
884	781456	29.7321	.001131	.0336
885	783225	29.7489	.001130	.0336

TABLA A
(Continuación)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
886	784996	29.7658	.001129	.0336
887	786769	29.7825	.001127	.0336
888	788544	29.7993	.001126	.0336
889	790321	29.8161	.001125	.0335
890	792100	29.8329	.001124	.0335
891	793881	29.8496	.001122	.0335
892	795664	29.8664	.001121	.0335
893	797449	29.8831	.001120	.0335
894	799236	29.8998	.001119	.0334
895	801025	29.9166	.001117	.0334
896	802816	29.9333	.001116	.0334
897	804609	29.9500	.001115	.0334
898	806404	29.9666	.001114	.0334
899	808201	29.9833	.001112	.0334
900	810000	30.0000	.001111	.0333
901	811801	30.0167	.001110	.0333
902	813604	30.0333	.001109	.0333
903	815409	30.0500	.001107	.0333
904	817216	30.0666	.001106	.0333
905	819025	30.0832	.001105	.0332
906	820836	30.0998	.001104	.0332
907	822649	30.1164	.001103	.0332
908	824464	30.1330	.001101	.0332
909	826281	30.1496	.001100	.0332
910	828100	30.1662	.001099	.0331
911	829921	30.1828	.001098	.0331
912	831744	30.1993	.001096	.0331
913	833569	30.2159	.001095	.0331
914	835396	30.2324	.001094	.0331
915	837225	30.2490	.001093	.0331
916	839056	30.2655	.001092	.0330
917	840889	30.2820	.001091	.0330
918	842724	30.2985	.001089	.0330
919	844561	30.3150	.001088	.0330
920	846400	30.3315	.001087	.0330
921	848241	30.3480	.001086	.0330
922	850084	30.3645	.001085	.0329
923	851929	30.3809	.001083	.0329
924	853776	30.3974	.001082	.0329
925	855625	30.4138	.001081	.0329
926	857476	30.4302	.001080	.0329
927	859329	30.4467	.001079	.0328
928	861184	30.4631	.001078	.0328
929	863041	30.4795	.001076	.0328
930	864900	30.4959	.001075	.0328
931	866761	30.5123	.001074	.0328
932	868624	30.5287	.001073	.0328
933	870489	30.5450	.001072	.0327
934	872356	30.5614	.001071	.0327
935	874225	30.5778	.001070	.0327

TABLA A
(Continuación)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
936	876096	30.5941	.001068	.0327
937	877969	30.6105	.001067	.0327
938	879844	30.6268	.001066	.0327
939	881721	30.6431	.001065	.0326
940	883600	30.6594	.001064	.0326
941	885481	30.6757	.001063	.0326
942	887364	30.6920	.001062	.0326
943	889249	30.7083	.001060	.0326
944	891136	30.7246	.001059	.0325
945	893025	30.7409	.001058	.0325
946	894916	30.7571	.001057	.0325
947	896809	30.7734	.001056	.0325
948	898704	30.7896	.001055	.0325
949	900601	30.8058	.001054	.0325
950	902500	30.8221	.001053	.0324
951	904401	30.8383	.001052	.0324
952	906304	30.8545	.001050	.0324
953	908209	30.8707	.001049	.0324
954	910116	30.8869	.001048	.0324
955	912025	30.9031	.001047	.0324
956	913936	30.9192	.001046	.0323
957	915849	30.9354	.001045	.0323
958	917764	30.9516	.001044	.0323
959	919681	30.9677	.001043	.0323
960	921600	30.9839	.001042	.0323
961	923521	31.0000	.001041	.0323
962	925444	31.0161	.001040	.0322
963	927369	31.0322	.001038	.0322
964	929296	31.0483	.001037	.0322
965	931225	31.0644	.001036	.0322
966	933156	31.0805	.001035	.0322
967	935089	31.0966	.001034	.0322
968	937024	31.1127	.001033	.0321
969	938961	31.1288	.001032	.0321
970	940900	31.1448	.001031	.0321
971	942841	31.1609	.001030	.0321
972	944784	31.1769	.001029	.0321
973	946729	31.1929	.001028	.0321
974	948676	31.2090	.001027	.0320
975	950625	31.2250	.001026	.0320
976	952576	31.2410	.001025	.0320
977	954529	31.2570	.001024	.0320
978	956484	31.2730	.001022	.0320
979	958441	31.2890	.001021	.0320
980	960400	31.3050	.001020	.0319
981	962361	31.3209	.001019	.0319
982	964324	31.3369	.001018	.0319
983	966289	31.3528	.001017	.0319
984	968256	31.3688	.001016	.0319
985	970225	31.3847	.001015	.0319

TABLA A
(Continuación)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
986	972196	31.4006	.001014	.0318
987	974169	31.4166	.001013	.0318
988	976144	31.4325	.001012	.0318
989	978121	31.4484	.001011	.0318
990	980100	31.4643	.001010	.0318
991	982081	31.4802	.001009	.0318
992	984064	31.4960	.001008	.0318
993	986049	31.5119	.001007	.0317
994	988036	31.5278	.001006	.0317
995	990025	31.5436	.001005	.0317
996	992016	31.5595	.001004	.0317
997	994009	31.5753	.001003	.0317
998	996004	31.5911	.001002	.0317
999	998001	31.6070	.001001	.0316
1000	1000000	31.6228	.001000	.0316

TABLA B Porcentaje del área bajo la curva normal entre $\bar{X}$ y z

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	00.00	00.40	00.80	01.20	01.60	01.99	02.39	02.79	03.19	03.59
0.1	03.98	04.38	04.78	05.17	05.57	05.96	06.36	06.75	07.14	07.53
0.2	07.93	08.32	08.71	09.10	09.48	09.87	10.26	10.64	11.03	11.41
0.3	11.79	12.17	12.55	12.93	13.31	13.68	14.06	14.43	14.80	15.17
0.4	15.54	15.91	16.28	16.64	17.00	17.36	17.72	18.08	18.44	18.79
0.5	19.15	19.50	19.85	20.19	20.54	20.88	21.23	21.57	21.90	22.24
0.6	22.57	22.91	23.24	23.57	23.89	24.22	24.54	24.86	25.17	25.49
0.7	25.80	26.11	26.42	26.73	27.04	27.34	27.64	27.94	28.23	28.52
0.8	28.81	29.10	29.39	29.67	29.95	30.23	30.51	30.78	31.06	31.33
0.9	31.59	31.86	32.12	32.38	32.64	32.90	33.15	33.40	33.65	33.89
1.0	34.13	34.38	34.61	34.85	35.08	35.31	35.54	35.77	35.99	36.21
1.1	36.43	36.65	36.86	37.08	37.29	37.49	37.70	37.90	38.10	38.30
1.2	38.49	38.69	38.88	39.07	39.25	39.44	39.62	39.80	39.97	40.15
1.3	40.32	40.49	40.66	40.82	40.99	41.15	41.31	41.47	41.62	41.77
1.4	41.92	42.07	42.22	42.36	42.51	42.65	42.79	42.92	43.06	43.19
1.5	43.32	43.45	43.57	43.70	43.83	43.94	44.06	44.18	44.29	44.41
1.6	44.52	44.63	44.74	44.84	44.95	45.05	45.15	45.25	45.35	45.45
1.7	45.54	45.64	45.73	45.82	45.91	45.99	46.08	46.16	46.25	46.33
1.8	46.41	46.49	46.56	46.64	46.71	46.78	46.86	46.93	46.99	47.06
1.9	47.13	47.19	47.26	47.32	47.38	47.44	47.50	47.56	47.61	47.67
2.0	47.72	47.78	47.83	47.88	47.93	47.98	48.03	48.08	48.12	48.17
2.1	48.21	48.26	48.30	48.34	48.38	48.42	48.46	48.50	48.54	48.57
2.2	48.61	48.64	48.68	48.71	48.75	48.78	48.81	48.84	48.87	48.90
2.3	48.93	48.96	48.98	49.01	49.04	49.06	49.09	49.11	49.13	49.16
2.4	49.18	49.20	49.22	49.25	49.27	49.29	49.31	49.32	49.34	49.36
2.5	49.38	49.40	49.41	49.43	49.45	49.46	49.48	49.49	49.51	49.52
2.6	49.53	49.55	49.56	49.57	49.59	49.60	49.61	49.62	49.63	49.64
2.7	49.65	49.66	49.67	49.68	49.69	49.70	49.71	49.72	49.73	49.74
2.8	49.74	49.75	49.76	49.77	49.77	49.78	49.79	49.79	49.80	49.81
2.9	49.81	49.82	49.82	49.83	49.84	49.84	49.85	49.85	49.86	49.86
3.0	49.87									
4.0	49.997									

FUENTE: Karl Pearson, *Tables for Statisticians and Biometricians*, Cambridge University Press, Londres, pp. 98-101, con autorización de Biometrika Trustees.

TABLA C Valores de *t* a los niveles de confianza de 0,05 y 0,01

gl	.05	.01
1	12.706	63.657
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
6	2.447	3.707
7	2.365	3.499
8	2.306	3.355
9	2.262	3.250
10	2.228	3.169
11	2.201	3.106
12	2.179	3.055
13	2.160	3.012
14	2.145	2.977
15	2.131	2.947
16	2.120	2.921
17	2.110	2.898
18	2.101	2.878
19	2.093	2.861
20	2.086	2.845
21	2.080	2.831
22	2.074	2.819
23	2.069	2.807
24	2.064	2.797
25	2.060	2.787
26	2.056	2.779
27	2.052	2.771
28	2.048	2.763
29	2.045	2.756
30	2.042	2.750
40	2.021	2.704
60	2.000	2.660
120	1.980	2.617
∞	1.960	2.576

FUENTE: Ronald A. Fisher y Frank Yates, *Statistical Tables for Biological, Agricultural, and Medical Research*, 4a. ed., Oliver & Boyd, Edimburgo. Tabla III, con autorización de los autores y el editor.

TABLA D Valores de *F* al Nivel de Confianza de 0,05 y 0,01

(gl para el numerador) P = .05

gl (gl para el denominador)	1	2	3	4	5	6	8	12
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83
∞	3.84	2.99	2.60	2.37	2.21	2.09	1.94	1.75

FUENTE: Fisher y F. Yates, *Statistical Tables for Biological, Agricultural, and Medical Research*, 4a. ed., Oliver & Boyd, Edimburgo, Tabla V, con autorización de los autores y el editor.

TABLA D
(Continuación)

(gl para el numerador) P = .01

gl	1	2	3	4	5	6	8	12
1	4052	4999	5403	5625	5764	5859	5981	6106
2	98.49	99.01	99.17	99.25	99.30	99.33	99.36	99.42
3	34.12	30.81	29.46	28.71	28.24	27.91	27.49	27.05
4	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.37
5	16.26	13.27	12.06	11.39	10.97	10.67	10.27	9.89
6	13.74	10.92	9.78	9.15	8.75	8.47	8.10	7.72
7	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.47
8	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67
9	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.11
10	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.71
11	9.65	7.20	6.22	5.67	5.32	5.07	4.74	4.40
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.16
13	9.07	6.70	5.74	5.20	4.86	4.62	4.30	3.96
14	8.86	6.51	5.56	5.03	4.69	4.46	4.14	3.80
15	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.67
16	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.55
17	8.40	6.11	5.18	4.67	4.34	4.10	3.79	3.45
18	8.28	6.01	5.09	4.58	4.25	4.01	3.71	3.37
19	8.18	5.93	5.01	4.50	4.17	3.94	3.63	3.30
20	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.23
21	8.02	5.78	4.87	4.37	4.04	3.81	3.51	3.17
22	7.94	5.72	4.82	4.31	3.99	3.76	3.45	3.12
23	7.88	5.66	4.76	4.26	3.94	3.71	3.41	3.07
24	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.03
25	7.77	5.57	4.68	4.18	3.86	3.63	3.32	2.99
26	7.72	5.53	4.64	4.14	3.82	3.59	3.29	2.96
27	7.68	5.49	4.60	4.11	3.78	3.56	3.26	2.93
28	7.64	5.45	4.57	4.07	3.75	3.53	3.23	2.90
29	7.60	5.42	4.54	4.04	3.73	3.50	3.20	2.87
30	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84
40	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.66
60	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.50
120	6.85	4.79	3.95	3.48	3.17	2.96	2.66	2.34
∞	6.64	4.60	3.78	3.32	3.02	2.80	2.51	2.18

(gl para el denominador)

TABLA E Valores de Chi Cuadrada a los Niveles de Confianza de 0,05 y 0,01

gl	.05	.01
1	3.841	6.635
2	5.991	9.210
3	7.815	11.345
4	9.488	13.277
5	11.070	15.086
6	12.592	16.812
7	14.067	18.475
8	15.507	20.090
9	16.919	21.666
10	18.307	23.209
11	19.675	24.725
12	21.026	26.217
13	22.362	27.688
14	23.685	29.141
15	24.996	30.578
16	26.296	32.000
17	27.587	33.409
18	28.869	34.805
19	30.144	36.191
20	31.410	37.566
21	32.671	38.932
22	33.924	40.289
23	35.172	41.638
24	36.415	42.980
25	37.652	44.314
26	38.885	45.642
27	40.113	46.963
28	41.337	48.278
29	42.557	49.588
30	43.773	50.892

FUENTE: Fisher y F. Yates, *Statistical Tables for Biological, Agricultural, and Medical Research*, 4a. ed., Oliver & Boyd, Edimburgo, Tabla IV, con autorización de los autores y el editor.

TABLA F Valores de r a los Niveles de Confianza de 0,05 y 0,01

gl	.05	.01
1	.99692	.999877
2	.95000	.990000
3	.8783	.95873
4	.8114	.91720
5	.7545	.8745
6	.7067	.8343
7	.6664	.7977
8	.6319	.7646
9	.6021	.7348
10	.5760	.7079
11	.5529	.6835
12	.5324	.6614
13	.5139	.6411
14	.4973	.6226
15	.4821	.6055
16	.4683	.5897
17	.4555	.5751
18	.4438	.5614
19	.4329	.5487
20	.4227	.5368
25	.3809	.4869
30	.3494	.4487
35	.3246	.4182
40	.3044	.3932
45	.2875	.3721
50	.2732	.3541
60	.2500	.3248
70	.2319	.3017
80	.2172	.2830
90	.2050	.2673

FUENTE: Fisher y F. Yates, *Statistical Tables for Biological, Agricultural, and Medical Research*, 4a. ed., Oliver & Boyd, Edimburgo, Tabla IV, con autorización de los autores y el editor.

TABLA G Valores de r_s a los Niveles de confianza de 0,05 y 0,01

N	.05	.01
5	1.000	—
6	.886	1.000
7	.786	.929
8	.738	.881
9	.683	.833
10	.648	.794
12	.591	.777
14	.544	.714
16	.506	.665
18	.475	.625
20	.450	.591
22	.428	.562
24	.409	.537
26	.392	.515
28	.377	.496
30	.364	.478

FUENTE: E. G. Olds, *The Annals of Mathematical Statistics*, "Distribution of the Sum of Squares of Rank Differences for Small Numbers of Individuals," 1938, vol. 9 y "The 5 Percent Significance Levels for Sums of Squares of Rank Differences and a Correction," 1949, vol. 20, por autorización del Instituto de Estadísticas Matemáticas.

TABLA H Números Aleatorios

Renglón	Número de columna																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	9	8	9	6	9	9	0	9	6	3	2	3	3	8	6	8	4	4	2
2	3	5	6	1	7	4	1	3	2	6	8	6	0	4	7	5	2	0	3
3	4	0	6	1	6	9	6	1	5	9	5	4	5	4	8	6	7	4	0
4	6	5	6	3	1	6	8	6	7	2	0	7	2	3	2	1	5	0	9
5	2	4	9	7	9	1	0	3	9	6	7	4	1	5	4	9	6	9	8
6	7	6	1	2	7	5	6	9	4	8	4	2	8	5	2	4	1	8	0
7	8	2	1	3	4	7	4	6	3	0	7	5	0	9	2	9	0	6	1
8	6	9	5	6	5	6	0	9	0	7	7	1	4	1	8	3	1	9	3
9	7	2	1	9	9	8	0	1	6	1	6	2	3	6	9	5	5	8	4
10	2	9	0	7	3	0	8	9	6	3	3	8	5	5	6	5	2	0	9
11	9	3	5	4	5	7	4	0	3	0	1	0	4	3	3	9	5	3	2
12	9	7	5	7	9	4	8	6	8	7	6	1	6	8	2	5	5	5	3
13	4	1	7	8	6	8	1	0	5	8	8	6	1	6	8	2	9	0	4
14	5	0	8	3	3	4	5	4	4	2	5	3	0	4	9	6	1	2	3
15	3	5	0	2	9	4	1	0	0	3	9	0	5	8	6	0	9	9	6
16	0	3	8	2	3	5	1	0	1	0	6	8	5	2	4	8	0	3	8
17	1	7	2	9	1	2	7	8	4	7	0	3	3	1	5	8	2	7	3
18	5	0	5	7	9	5	8	7	8	9	3	5	3	4	4	6	1	1	3
19	7	7	3	3	5	3	6	1	3	2	8	5	4	1	4	8	3	9	0
20	1	0	9	1	3	8	2	5	3	0	3	8	0	9	3	3	0	4	5
21	1	3	8	5	1	8	5	9	4	1	9	3	9	3	6	5	9	8	4
22	8	6	4	7	8	7	5	9	4	1	9	3	9	3	6	5	9	8	4
23	0	6	9	6	5	1	0	3	2	6	7	7	4	9	6	0	3	4	0
24	7	6	7	4	7	0	8	3	8	7	3	2	5	1	2	4	2	9	7
25	3	2	3	8	1	3	1	8	7	4	5	9	0	0	2	4	1	2	1
26	9	2	1	6	4	2	3	8	7	6	2	6	2	6	4	8	1	0	1
27	3	7	4	2	2	8	1	7	8	0	6	0	0	0	3	2	2	9	7
28	0	7	8	0	8	5	1	5	2	6	5	8	7	5	3	0	5	9	6
29	7	4	2	3	3	2	6	0	0	6	5	2	2	3	6	3	9	0	4
30	1	8	2	7	5	9	5	3	6	5	2	9	9	1	1	7	3	4	3
31	4	3	1	8	7	0	6	0	8	6	5	0	1	0	4	0	6	1	5
32	8	5	8	0	6	1	4	1	2	0	4	4	1	4	7	6	3	5	1
33	4	5	8	5	0	4	5	8	3	9	2	8	7	8	9	0	8	4	3
34	5	0	2	5	4	9	2	2	1	1	0	0	5	4	8	7	6	4	0
35	0	8	1	7	0	6	3	3	4	7	6	2	6	8	9	3	4	1	4
36	2	5	9	3	4	6	0	7	5	2	0	0	9	6	0	8	2	2	5
37	2	1	3	1	3	7	8	9	8	4	9	3	8	0	2	2	1	8	1
38	3	8	8	6	8	5	1	3	3	4	6	7	2	6	3	4	8	6	7
39	0	9	9	8	5	9	8	4	4	2	2	1	1	0	1	7	6	1	3
40	2	2	3	5	3	9	7	4	4	2	1	4	0	5	8	2	3	0	8

TABLA H
(Continuación)

Número de columna																					Renglón
20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	
0	9	7	1	1	9	1	2	7	3	5	1	8	4	0	4	1	0	6	0	3	1
8	3	7	7	9	1	4	9	9	5	9	2	0	1	6	1	2	6	6	7	0	2
2	5	6	3	7	8	3	3	8	4	3	9	3	9	0	0	9	8	3	5	2	3
4	7	0	8	6	6	5	9	6	2	7	3	5	9	0	1	8	0	9	6	9	4
0	9	8	7	3	5	6	8	8	1	2	0	2	3	2	6	4	3	1	9	7	5
5	1	8	8	4	7	0	1	7	6	8	2	1	6	3	2	1	8	1	8	3	6
1	3	7	8	6	9	5	4	1	7	3	8	7	1	5	6	5	6	4	3	6	7
5	9	0	1	5	2	8	6	5	5	7	8	1	8	7	1	2	4	0	4	1	8
2	2	5	5	2	1	8	6	9	8	9	8	0	5	8	9	9	4	1	3	4	9
1	3	4	2	8	5	0	7	9	8	4	3	5	8	0	9	4	6	6	0	5	10
2	6	8	6	6	4	7	1	5	1	6	4	6	7	6	0	8	7	3	5	2	11
8	6	0	1	4	2	9	8	6	8	0	7	6	5	1	9	1	3	7	0	3	12
9	5	7	0	9	8	7	6	9	0	6	5	4	0	3	6	5	6	3	5	0	13
2	2	3	4	7	8	0	2	0	8	0	3	4	9	2	5	7	7	8	6	4	14
2	4	6	1	0	5	0	6	1	4	9	4	7	3	9	1	7	6	4	5	8	15
6	3	4	8	1	6	9	5	6	2	0	4	6	1	6	8	1	9	9	1	1	16
9	0	5	1	3	6	1	9	5	4	1	2	5	4	2	9	5	6	2	4	0	17
3	6	7	0	3	5	3	7	4	1	7	5	4	8	3	7	4	8	5	7	2	18
4	3	6	6	3	6	3	0	0	9	4	2	2	5	1	8	9	5	1	9	7	19
1	0	6	9	0	2	7	3	9	8	4	0	6	9	8	2	3	2	8	0	4	20
9	1	3	5	7	9	6	2	4	3	4	6	4	9	1	3	1	7	5	2	2	21
6	4	2	2	2	1	4	5	2	2	8	3	2	1	2	6	6	0	1	8	9	22
7	2	6	9	0	7	5	3	2	5	6	2	7	6	3	8	1	4	1	5	1	23
8	2	8	2	4	4	4	2	9	1	9	8	3	4	4	1	0	4	6	9	6	24
7	3	1	4	3	0	4	7	1	3	7	4	8	6	7	3	2	6	6	2	0	25
0	6	4	5	8	3	1	4	8	1	8	3	1	6	4	3	0	2	8	7	3	26
4	2	2	8	3	2	1	9	3	0	1	7	5	9	0	9	1	2	5	8	2	27
2	9	8	7	2	0	6	4	0	2	7	1	3	1	6	8	7	0	9	2	5	28
0	8	0	5	6	8	2	4	3	6	1	3	5	2	3	5	9	8	6	2	1	29
0	1	7	6	1	5	7	9	0	3	5	3	4	2	4	8	5	6	4	0	6	30
5	1	9	8	5	2	4	5	1	7	5	3	2	4	6	7	9	9	6	7	2	31
0	3	6	6	3	7	8	6	9	7	2	8	9	0	7	2	9	4	0	8	6	32
5	0	0	0	2	0	8	9	0	1	0	6	2	0	4	6	9	6	5	4	9	33
1	9	4	4	2	6	4	2	4	1	0	2	7	9	6	8	7	5	6	9	3	34
0	0	5	3	8	3	2	7	5	0	4	7	6	4	6	3	0	4	7	5	3	35
6	2	6	2	0	6	0	1	4	8	9	6	5	9	7	3	6	7	6	5	4	36
6	3	9	0	3	5	0	9	1	2	0	5	9	7	3	2	5	9	3	0	2	37
9	7	3	3	5	4	0	6	4	9	4	7	9	1	4	3	9	7	7	1	8	38
1	9	6	2	9	4	2	9	7	0	3	8	9	5	7	0	6	9	7	2	5	39
5	9	4	5	8	6	2	3	0	6	2	9	8	6	3	0	4	1	0	7	6	40

FUENTE: N.M. Downie y R.W. Heath, *Basic Statistical Methcds*, 3a. ed., Harper & Row, Nueva York, 1970. Reeditado con autorización de Harper & Row.

TABLA I Puntos de porcentaje del rango student

μC gl	α	*k = Número de medias*									
		2	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
5	.05	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99	7.17
	.01	5.70	6.98	7.80	8.42	8.91	9.32	9.67	9.97	10.24	10.48
6	.05	3.46	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49	6.65
	.01	5.24	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.10	9.30
7	.05	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16	6.30
	.01	4.95	5.92	6.54	7.01	7.37	7.68	7.94	8.17	8.37	8.55
8	.05	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.05
	.01	4.75	5.64	6.20	6.62	6.96	7.24	7.47	7.68	7.86	8.03
9	.05	3.20	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74	5.87
	.01	4.60	5.43	5.96	6.35	6.66	6.91	7.13	7.33	7.49	7.65
10	.05	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60	5.72
	.01	4.48	5.27	5.77	6.14	6.43	6.67	6.87	7.05	7.21	7.36
11	.05	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49	5.61
	.01	4.39	5.15	5.62	5.97	6.25	6.48	6.67	6.84	6.99	7.13
12	.05	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.39	5.51
	.01	4.32	5.05	5.50	5.84	6.10	6.32	6.51	6.67	6.81	6.94
13	.05	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32	5.43
	.01	4.26	4.96	5.40	5.73	5.98	6.19	6.37	6.53	6.67	6.79
14	.05	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.36
	.01	4.21	4.89	5.32	5.63	5.88	6.08	6.26	6.41	6.54	6.66
15	.05	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20	5.31
	.01	4.17	4.84	5.25	5.56	5.80	5.99	6.16	6.31	6.44	6.55
16	.05	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15	5.26
	.01	4.13	4.79	5.19	5.49	5.72	5.92	6.08	6.22	6.35	6.46
17	.05	2.98	3.63	4.02	4.30	4.52	4.70	4.86	4.99	5.11	5.21
	.01	4.10	4.74	5.14	5.43	5.66	5.85	6.01	6.15	6.27	6.38
18	.05	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07	5.17
	.01	4.07	4.70	5.09	5.38	5.60	5.79	5.94	6.08	6.20	6.31
19	.05	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04	5.14
	.01	4.05	4.67	5.05	5.33	5.55	5.73	5.89	6.02	6.14	6.25
20	.05	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01	5.11
	.01	4.02	4.64	5.02	5.29	5.51	5.69	5.84	5.97	6.09	6.19
24	.05	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92	5.01
	.01	3.96	4.55	4.91	5.17	5.37	5.54	5.69	5.81	5.92	6.02
30	.05	2.89	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82	4.92
	.01	3.89	4.45	4.80	5.05	5.24	5.40	5.54	5.65	5.76	5.85
40	.05	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73	4.82
	.01	3.82	4.37	4.70	4.93	5.11	5.26	5.39	5.50	5.60	5.69
60	.05	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65	4.73
	.01	3.76	4.28	4.59	4.82	4.99	5.13	5.25	5.36	5.45	5.53
120	.05	2.80	3.36	3.68	3.92	4.10	4.24	4.36	4.47	4.56	4.64
	.01	3.70	4.20	4.50	4.71	4.87	5.01	5.12	5.21	5.30	5.37
∞	.05	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47	4.55
	.01	3.64	4.12	4.40	4.60	4.76	4.88	4.99	5.08	5.16	5.23

FUENTE: E.S. Pearson y H.O. Hartley, *Biometrika Tables for Statisticians,* vol. 1, 3a. ed., Cambridge Press, Nueva York, 1966, con autorización de Biometrika Trustees.

Apéndice C

Lista de fórmulas

FORMULA	PAGINA
$P = \frac{f}{N}$	17
$\% = (100)\frac{f}{N}$	17
Razón $= \frac{f_1}{f_2}$	18
Razón de sexo $= (100)\frac{f \text{ hombres}}{f \text{ mujeres}}$	19
Tasa de nacimientos $= (1000)\frac{f \text{ casos reales}}{f \text{ casos potenciales}}$	20
Tasa de cambio $= (100)\frac{\text{tiempo } 2f - \text{tiempo } 1f}{\text{tiempo } 1f}$	20
Punto medio $= \frac{\text{puntaje más bajo} + \text{puntaje más alto}}{2}$	23
$c\% = (100)\frac{fa}{N}$	25

$$\text{Rango percentil} = \begin{array}{l}\text{c\% por debajo del}\\ \text{límite inferior del}\\ \text{intervalo crítico}\end{array} + \left[\frac{\text{puntaje} - \begin{array}{l}\text{límite inferior}\\ \text{del intervalo}\\ \text{crítico}\end{array}}{\text{magnitud del intervalo crítico}} \left(\begin{array}{l}\%\ \text{en}\\ \text{el}\\ \text{inter-}\\ \text{valo}\\ \text{crítico}\end{array}\right)\right] \qquad 26$$

$$\text{Posición de la mediana} = \frac{N+1}{2} \qquad 40$$

$$\overline{X} = \frac{\Sigma X}{N} \qquad 42$$

$$x = X - \overline{X} \qquad 43$$

$$\overline{X} = \frac{\Sigma fX}{N} \qquad 44$$

$$\text{Mediana} = \begin{array}{l}\text{límite inferior}\\ \text{del intervalo}\\ \text{de la mediana}\end{array} + \left(\frac{\frac{N}{2} - \begin{array}{l}fa\ \text{por debajo}\\ \text{del límite de in-}\\ \text{ferior del in-}\\ \text{tervalo de la}\\ \text{mediana}\end{array}}{f\ \text{en el intervalo de la mediana}}\right) \text{magnitud del intervalo} \qquad 50$$

$$\text{DM} = \frac{\Sigma|x|}{N} \qquad 57$$

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} \qquad 59$$

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - \overline{X}^2} \qquad 61$$

$$\sigma = \sqrt{\frac{\Sigma fX^2}{N} - \overline{X}^2} \qquad 62$$

$$z = \frac{X - \overline{X}}{\sigma} \qquad 84$$

$$X = z\sigma + \overline{X} \qquad 85$$

$$\text{Probabilidad} = \frac{\text{número de veces que puede ocurrir el suceso}}{\text{número total de sucesos}} \qquad 85$$

$$z = \frac{\overline{X} - M}{\sigma_{\overline{X}}} \qquad 105$$

$$\sigma_{\overline{X}} = \frac{s}{\sqrt{N - 1}} \qquad 106$$

$$\text{Intervalo de confianza del 95\%} = \overline{X} \pm (1{,}96)\,\sigma_{\overline{X}} \qquad 109$$

$$\text{Intervalo de confianza del 99\%} = \overline{X} \pm (2{,}58)\,\sigma_{\overline{X}} \qquad 111$$

$$\sigma_P = \sqrt{\frac{P(1 - P)}{N}} \qquad 114$$

$$\text{Intervalo de confianza del 95\%} = P \pm (1{,}96)\,\sigma_P \qquad 115$$

$$z = \frac{(\overline{X}_1 - \overline{X}_2) - 0}{\sigma_{\text{dif}}} \qquad 128$$

$$\sigma_{\text{dif}} = \sqrt{\sigma_{\overline{X}_1}^2 + \sigma_{\overline{X}_2}^2} \qquad 132$$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sigma_{\text{dif}}} \qquad 136$$

$$\sigma_{\text{dif}} = \sqrt{\left(\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)} \qquad 140$$

$$s = \sqrt{\frac{\Sigma D^2}{N} - (\overline{X}_1 - \overline{X}_2)^2} \qquad 144$$

$$\text{SC}_{\text{dentro}} = \Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 + \Sigma X_4^2 \qquad 153$$

$$\text{SC}_{\text{ent}} = \Sigma(\overline{X} - \overline{X}_{\text{total}})^2 N \qquad 154$$

$$\text{SC}_{\text{total}} = \text{SC}_{\text{ent}}\ \text{SC}_{\text{dentro}} \qquad 155$$

$$\text{SC}_{\text{total}} = \Sigma(X - \overline{X}_{\text{total}})^2 \qquad 155$$

$$\text{SC}_{\text{total}} = \Sigma X^2{}_{\text{total}} \frac{-(\Sigma X_{\text{total}})^2}{N_{\text{total}}} \qquad 156$$

$$\text{SC}_{\text{ent}} = \left[\sum \frac{(\Sigma X)^2}{N}\right] - \frac{(\Sigma x_{\text{total}})^2}{N_{\text{total}}} \qquad 157$$

$$\text{SC}_{\text{dentro}} = \sum\left[(\Sigma X^2) - \frac{(\Sigma X)^2}{N}\right] \qquad 157$$

$$\mu C_{ent} = \frac{SC_{ent}}{gl_{ent}} \qquad 158$$

$$SC_{dentro} = \frac{\mu C_{dentro}}{gl_{dentro}} \qquad 158$$

$$F = \frac{\mu C_{ent}}{\mu C_{dentro}} \qquad 160$$

$$DSH = q\alpha \sqrt{\frac{\mu C_{dentro}}{n}} \qquad 165$$

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e} \qquad 171$$

$$\chi^2 = \frac{N(AD - BC)^2}{(A + B)(C + D)(A + C)(B + D)} \qquad 178$$

$$\chi^2 = \sum \frac{(|f_o - f_e| - 0{,}50)^2}{f_e} \qquad 179$$

$$\chi^2 = \frac{N(|AD - BC| - N/2)^2}{(A + B)(C + D)(A + C)(B + D)} \qquad 180$$

$$\chi_r^2 = \frac{12}{N_k(k + 1)} \Sigma (\Sigma R_i)^2 - 3N(k + 1) \qquad 189$$

$$H = \frac{12}{N(N + 1)} \Sigma \left[\frac{(\Sigma R_i)^2}{n} \right] - 3(N + 1) \qquad 192$$

$$r = \frac{\Sigma (z_X z_y)}{N} \qquad 204$$

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}} \qquad 207$$

$$t = \frac{r\sqrt{N - 2}}{\sqrt{1 - r^2}} \qquad 208$$

$$Y' = r \left(\frac{s_y}{s_x} \right) X - r \left(\frac{s_y}{s_x} \right) \overline{X} + \overline{Y} \qquad 213$$

$$r_s = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)} \qquad 217$$

$$G = \frac{\Sigma f_a - \Sigma f_i}{\Sigma f_a + \Sigma f_i} \qquad 223$$

$$z = G\sqrt{\frac{\Sigma f_a - \Sigma f_i}{\mathrm{N}(1 - G^2)}} \qquad 230$$

$$\phi = \sqrt{\frac{\chi^2}{N}} \qquad 232$$

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}} \qquad 234$$

$$V = \sqrt{\frac{\chi^2}{N(k - 1)}} \qquad 236$$

Respuestas a los problemas seleccionados

Capítulo 2

1. (a) 51%, (b) 27%, (c) $P = 0{,}51$, (d) $P = 0{,}27$
2. (a) 71%, (b) 74%, (c) $P = 0{,}71$, (d) $P = 0{,}74$
3. $\frac{4}{24} = \frac{1}{6}$
4. 156,25
5. $\frac{15}{20} = \frac{3}{4}$
6. Hay 85,71 nacimientos vivos por cada 1000 mujeres en edad de concebir.
7. 66,67%
8.

Intervalo de clase	*f*
10–12	11
7–9	16
4–6	9
1–3	4
	$N = 40$

a. 3

b. 9,5– 12,5
6,5– 9,5
3,5– 6,5
0,5– 3,5

c. 11
8
5
2

d. *fa*
40
29
13
4

e. $c\%$
100
72,5
32,5
10,0

9. (a) 59,38, (b) 12.59
10. (a) 84,82, (b) 29,64

Capítulo 4

1. (a) 9, (b) 6, (c) 5,71
2. (a) 9 y 1, (b) 5, (c) 5,13
3. (a) 5, (b) 5, (c) 32,71
4. (a) 1, (b) 2,5, (c) 3
5. (a) 10, (b) 10, (c) 9,63
6. (a) 3 y 6, (b) 4, (c) 4,1
7. (a) 8, (b) 8, (c) 7,67
8. (a) 6, (b) 4,5, (c) 4,17
9. (a) 4, (b) 5, (c) 6
10. (a) 12, (b) 7, (c) 7,86
11. (a) 0, (b) + 12,5, (c) −5,5, (d) + 0,5
12. (a) + 1,0, (b) − 0,5, (c) +3,3, (d) 0
13. (a) −12, (b) 7,5, (c) 0. (d) −4,5
14. (a) 4, (b) 4, (c) 4,13
15. (a) 3, (b) 3, (c) 3,19
16. (a) 6, (b) 6, (c) 6,26
17. (a) 12, (b) 12,3, (c) 12,79
18. (a) 84,5, (b) 82,4, (c) 80,39
19. (a) 12, (b) 11,76, (c) 12

Capítulo 5

1. (a) 6, (b) 1,92, (c) 2,15
2. (a) Clase A = 5, Clase B = 3, (b) Clase A = 1,67, Clase B = 0,83, (c) Clase A = 1,89, Clase B = 0,96
3. (a) 4, (b) 1,28, (c) 1,50
4. 2,70
5. 1,6
6. 1,19
7. 1,54
8. 1,40
9. (a) 49, (b) 10,51, (c) 12,46
10. (a) 14, (b) 2,47, (c) 3,25
11. (a) 19, (b) 3,71, (c) 4,66

Capítulo 6

1. (a) 68,26%, (b) 95,44%, (c) 99,74%
2. (a) + 0,38, (b) − 1,15, (c) − 1,69, (d) + 2,08, (e) 0, (f) 0,77, (g) + 2,69
3. (a) −0,75, (b) +0,18, (c) +0,96, (d) −1,96, (e) + 1,61, (f) + 0,36, (g) −0,54
4. (a) 5,37%, (b) $P = 0{,}05$, (c) 7,14%, (d) $P = 0{,}07$, (e) $P = 0{,}43$, (f) $P = 0{,}86$ (g) $P = 0{,}18$
5. (a) 0,38%, (b) P es menor que 0,01, (c) 40,82%, (d) $P = 0{,}41$ (e) 25,14%, (f) $P = 0{,}25$

Capítulo 7

1. 0,27
2. (a) 2,40 $\longleftrightarrow$ 3,46, (b) 2,23 $\longleftrightarrow$ 3,63

3. 0,35
4. (a) 5,10 ⟷ 6,48, (b) 4,89 ⟷ 6,69
5. 0,39
6. (a) 4,24 ⟷ 5,76, (b) 3,99 ⟷ 6,01
7. (a) 0,07, (b) 0,43 ⟷ 0,71
8. (a) 0,04, (b) 0,24 ⟷ 0,40
9. (a) 0,03, (b) 0,19 ⟷ 0,31

Capítulo 8

1. $z = 2{,}50, P = 0{,}01$, rechazar la hipótesis nula a 0,05
2. $t = 1{,}47$, gl = 6, aceptar la hipótesis nula a 0,05
3. $t = 1{,}84$, gl = 12, aceptar la hipótesis nula a 0,05
4. $t = 2{,}03$, gl = 16, aceptar la hipótesis nula a 0,05
5. $t = 4{,}31$, gl = 8, rechazar la hipótesis nula a 0,05
6. $t = 0{,}67$, gl = 8, aceptar la hipótesis nula a 0,05
7. $t = 3{,}90$, gl = 13, rechazar la hipótesis nula a 0,05
8. $t = 4{,}32$, gl = 10, rechazar la hipótesis nula a 0,05
9. $t = 2{,}51$, gl = 10, rechazar la hipótesis nula a 0,05
10. $t = 3{,}12$, gl = 5, rechazar la hipótesis nula a 0,05
11. $t = 3{,}85$, gl = 3, rechazar la hipótesis nula a 0,05
12. $t = 6{,}0$, gl = 4, rechazar la hipótesis nula a 0,05

Capítulo 9

1. $F = 2{,}71$, gl = $_{12}^{3}$, aceptar la hipótesis nula a 0,05
2. $F = 46{,}33$, gl = $\frac{2}{9}$, rechazar la hipótesis nula a 0,05
3. $F = 6{,}99$, gl = $\frac{2}{13}$, rechazar la hipótesis nula a 0,05
4. $F = 4{,}23$, gl = $\frac{2}{12}$, rechazar la hipótesis nula a 0,05
5. DSH = 2,11. Por lo tanto sólo $\bar{X}_1 - \bar{X}_3$ es estadísticamente significativo
6. $F = 8{,}16$, gl = $\frac{3}{20}$, rechazar la hipótesis nula a 0,05
7. DSH = 1,98. Por lo tanto, $\bar{X}_1 - \bar{X}_2$, $\bar{X}_1 - \bar{X}_3$, y $\bar{X}_1 - \bar{X}_4$ son estadísticamente significativos

Capítulo 10

1. $\chi^2 = 1{,}36$, gl = 1, aceptar la hipótesis nula a 0,05
2. $\chi^2 = 8{,}29$, gl = 1, rechazar la hipótesis nula a 0,05
3. $\chi^2 = 2{,}17$, gl = 1, aceptar la hipótesis nula a 0,05
4. $\chi^2 = 1{,}50$, gl = 1, aceptar la hipótesis nula a 0,05
5. $\chi^2 = 1{,}78$, gl = 1, aceptar la hipótesis nula a 0,05
6. $\chi^2 = 17{,}77$, gl = 4, rechazar la hipótesis nula a 0,05
7. $\chi^2 = 17{,}75$, gl = 3, rechazar la hipótesis nula a 0,05
8. $\chi^2 = 2{,}24$, gl = 2, aceptar la hipótesis nula a 0,05
9. Mdn = 5, $\chi^2 = 2{,}07$, gl = 1, aceptar la hipótesis nula a 0,05
10. Mdn = 6, $\chi^2 = 19{,}57$, gl = 1, rechazar la hipótesis nula a 0,05
11. $\chi_r^2 = 1{,}96$, gl = 1, aceptar la hipótesis nula a 0,05
12. $\chi_r^2 = 10{,}20$, gl = 2, rechazar la hipótesis nula a 0,05
13. $H = 1{,}97$, gl = 2, aceptar la hipótesis nula a 0,05
14. $H = 10{,}64$, gl = 2, rechazar la hipótesis nula a 0,05

Capítulo 11

1. $r = +0{,}85$, gl = 4, significativo a 0,05
2. $r = -0{,}64$, gl = 2, no significativo a 0,05
3. $r = +0{,}76$, gl = 3, no significativo a 0,05

4. $r = +0{,}93$, gl = 3, significativo a 0,05.
5. $r = -0{,}91$, gl = 5, significativo a 0,05
6. $Y' = 0{,}52X + 1{,}01$; (a) $Y' = 3{,}61$, (b) $Y' = 2{,}05$, (c) $Y' = 5{,}69$
7. $Y' = -0{,}90X + 10{,}19$; (a) $Y' = 1{,}19$, (b) $Y' = 8{,}39$
8. $r_s = -0{,}53$, $N = 5$, no significativo a 0,05
9. $r_s = -0{,}65$, $N = 8$, no significativo a 0,05
10. $r_s = -0{,}89$, $N = 7$, significativo a 0,05
11. $r_s = -0{,}80$, $N = 5$, no significativo a 0,05.
12. $G = +0{,}60$, $z = 0{,}82$, no significativo a 0,05
13. $G = -0{,}39$, $z = 1{,}15$, no significativo a 0,05
14. $\phi = 0{,}37$
15. $\phi = 0{,}17$
16. $\phi = 0{,}17$
17. (a) $C = 0{,}26$, (b) $V = 0{,}20$
18. (a) $C = 0{,}36$, (b) $V = 0{,}39$
19. (a) $C = 0{,}27$, (b) $V = 0{,}20$

Referencias

Anderson, Theodore R. y Morris Zelditch, Jr., *A Basic Course in Statistics*, Holt, Rinehart y Winston, Nueva York, 1968.

Blalock, Hubert. M., *Social Statistics,* McGraw-Hill, Nueva York, 1960.

Campbell, Stephen K., *Flaws and Fallacies in Statistical Thinking*, Prentice-Hall, Englewood Cliffs, N.J., 1974.

Champion, Dean J., *Basic Statistics for Social Research*, Chandler, San Francisco, 1970.

Chase, Clinton I., *Elementary Statistical Procedures*, McGraw-Hill, Nueva York, 1967.

Cohen, Lillian, *Statistical Methods for Social Scientists,* Prentice-Hall, Englewood Cliffs, N.J., 1954.

Courts, Frederick A., *Psychological Statistics,* The Dorsey Press, Homewood, Ill., 1966.

Dixon, Wilfrid J. y Frank J. Massey, *Introduction to Statistical Analysis*, McGraw-Hill, Nueva York, 1957.

Dornbusch, Sanford M. y Calvin F. Schmid, *A primer of Social Statistics*, McGraw-Hill, Nueva York, 1955.

Downey, Kenneth J., *Elementary Social Statistics*, Random House, Nueva York, 1975.

Downie, Norville M. y R. W. Heat, *Basic Statistical Methods*, Harper & Row, Nueva York, 1974.

Edwards, A. L., *Experimental Design in Psychological Research*, Holt, Rinehart y Winston, Nueva York, 1960.

Edwards, Allen L., *Statistical Methods for the Behavioral Sciences,* Holt, Rinehart y Winston, Nueva York, 1967.

Ferguson, George A., *Statistical Analysis in Psychology and Education*, McGraw-Hill, Nueva York, 1966.

Freeman, Linton C., *Elementary Applied Statistics*, Wiley, Nueva York, 1965.

Freund, John E., *Modern Elementary Statistics*, Prentice-Hall, Englewood Cliffs, N.J., 1960.

Fried, Robert, *Introduction to Statistics*, Oxford University, 1969.

Guilford, Jay P., *Fundamental Statistics in Psychology and Education*, McGraw-Hill, Nueva York, 1956.

Hagood, Margaret J. y Daniel O. Price, *Statistics for Sociologists*, Holt Rinehart y Winston, Nueva York, 1952.

Hammond, Kenneth R. y James E. Householder, *Introduction to the Statistical Method*, Knopf, Nueva York, 1963.

Huff, Darrell, *How to Lie With Statistics*, Wiley, Nueva York, 1966.

Loether, Herman J. y Donald G. McTavish, *Inferential Statistics for Sociologists*, Allen y Bacon, Boston, 1974.

McNemar, Quinn, *Psychological Statistics*, Wiley, Nueva York, 1962.

Meyers, Lawrence S. y Neal E. Grossen, *Behavioral Research*, Freeman, San Francisco, 1974.

Mueller, John H., Karl F. Schuessler, y Herbert L. Costner, *Statistical Reasoning in Sociology*, Houghton Mifflin, Boston, 1970.

Palumbo, Dennis J., *Statistics in Political and Behavioral Science*, Appleton, Nueva York, 1969.

Popham, W. James y Kenneth A. Sirotnik, *Educational Statistics*, Harper & Row, Nueva York, 1973.

Runyon, Richard P. y Audrey Haber, *Fundamentals of Behavioral Statistics*, Addison-Wesley, Reading, Mass., 1971.

Siegel, Sidney, *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill, Nueva York, 1956.

Spence, Janet T., Benthon J. Underwood, Carl P. Duncan y John W. Cotton, *Elementary Statistics*, Appleton, Nueva York, 1968.

Walker, Helén Mary y Joseph Lev, *Elementary Statistical Methods*, Holt, Rinehart y Winston, Nueva York, 1958.

Wallis, Wilson A. y Harry Roberts, *The Nature of Statistics*, Free Press, Nueva York, 1965.

Welkowitz, Joan, Robert B. Ewen y Jacob Cohen, *Introductory Statistics for the Behavioral Sciences*, Academic, Nueva York, 1971.

Williams, Frederick, *Reasoning with Statistics*, Holt, Rinehart y Winston, Nueva York, 1968.

Winer, B. J., *Statistical Principles in Experimental Desingn*, McGraw-Hill, Nueva York, 1962.

Indice

Análisis de varianza, 151-168
  comparación múltiple de medias, 164-166
  lógica, 152-153
  media cuadrática, 158-159
  razón *F*, 160
  requisitos, 166
  suma de los cuadrados, 153
Análisis de varianza en una dirección de Kruskal-Wallis, 192-194
Análisis de varianza en dos direcciones de Friedman, 189-192
Aplicación de la estadística, 243-254

Coeficiente de contingencia, 234
Coeficiente de correlación de Pearson
  fórmula, 207-209
  grados de libertad, 211
  requisitos, 211
  significancia, 210-212
Coeficiente de correlación por rangos ordenados
  fórmula, 217
  rangos empatados, 218
  requisitos, 222
  significancia, 220
Coeficiente phi, 232
Corrección de Yates, 180
Correlación, 200
  coeficiente, 203
  coeficiente de contingencia, 235
  coeficiente phi, 232
  curvilínea, 202
  dirección, 201
  fuerza de, 200
  rangos ordenados, *r* de Pearson, 207
  V de Cramér, 236
Cuartiles, 29-30
Curtosis, 37
Curva normal, 75
  área, 78-80
  características, 76
  y el mundo real, 76-77
Chi cuadrada, 170
  cálculo, 173-175
  comparación de varios grupos, 181-185
  frecuencias esperadas, 174
  fórmula de cálculo, 178
  grados de libertad, 173
  pequeñas frecuencias esperadas, 179
  como prueba de significancia, 170
  requisitos, 185-186

Deciles, 29
Decimales, 257-259
Desviación
  cálculo, 42-43
  definida, 42
Desviación estándar,
  cálculo, 59-61
  comparada con otras medidas de variabilidad, 66
  definida, 59-60
  fórmula para datos crudos, 61-62
  para distribuciones de frecuencia agrupada 68-69

para distribuciones de frecuencia simple, 62-63
significado, 64-66
Desviación media,
cálculo, 57-59
comparada con otras medidas de variabilidad, 66
definida, 56
para distribuciones de frecuencia agrupadas, 67-68

Diagrama de dispersión, 204
Distribuciones acumuladas, 24-26
Distribución de frecuencia acumulada, 24-26
agrupada, 22-24
datos nominales, 15
datos ordinales y por intervalos, 20-21
forma, 37
sesgada, 37
simétrica, 37
Distribución muestral de diferencias, 123-129
características, 124
comprobación de hipótesis, 126
Distribución muestral de medias, 100-101
características, 101-102
como curva normal, 103-104
DFS de Tukey, 164-166

Error, alfa y beta, 132
Error de muestreo, 99
Error estándar de la diferencia, 132-133
Error estándar de la media, 106-107
Error estándar de la proporción, 113
Estadística, funciones, 7-12
Estadística no paramétrica, 171-172

Gamma de Goodman y Kruskal,
fórmula, 223
rangos empatados, 227
requisitos, 231
significancia, 230-231
Grados de libertad, 137-138
Chi cuadrada, 172
r de Pearson, 211
razón *F*, 159
razón *t*, 137
Gráficas de barras, 34-35
construcción de, 36
Gráficas de sectores, 33

Hipótesis
de investigación, 123-124
nula, 122-123
prueba, 2
Hipótesis de investigación, 122-123
Hipótesis nula, 121-122
Histograma, 33-35

Intervalo de clase, 21-23
definido, 22
límites, 22-24
número de, 24
puntos medios, 22
tamaño, 22
Intervalo de confianza,
cálculo, 107
definido, 107
95%, 108
99%, 111
proporciones, 113
Investigación social, 3-4

Línea de regresión, 214
ecuación de la, 215-217

Media,
cálculo, 44
comparada con otras medidas de tendencia central, 44-48
definida, 42
para distribuciones de frecuencia agrupada, 49-50
para distribuciones de frecuencia simple, 44
Media cuadrática, 158-159
Mediana
cálculo, 41
comparada con otras medidas de tendencia central, 44-48
definida, 40
para distribuciones de frecuencia agrupada, 48-49
para distribuciones de frecuencia simple, 41-42
Métodos de muestreo, 94
Moda,
comparada con otras medidas de tendencia central, 44-48
definida, 39
en distribuciones bimodales, 40

para distribuciones de frecuencia agrupada, 48
Muestras,
aleatoria, 95-98
definida, 93
no aleatoria, 94
Muestra aleatoria, 96-99
Muestras no aleatorias, 94

Nivel de confianza, 130-131
Nivel de medición, 4-7
nominal, 4
ordinal, 6
por intervalos, 6
Nivel de medición por intervalos, 6
Nivel de significancia.
Ver Nivel de confianza
Nivel nominal de medición, 4
Nivel ordinal de medición, 5
Números negativos, 258-259

Polígono de frecuencia, 35-36
construcción de, 36
Porcentaje
cálculo, 17
definido, 17
Potencia, 169-170
Probabilidad, 85-92
Proporción,
cálculo, 56
definida, 56
Prueba de la mediana, 186-188
requisitos, 188
Puntaje estándar.
Ver Puntaje *Z*
Puntaje *Z*, 83-84
cálculo, 85
definido, 84
para la diferencia entre medias, 129-130
requisitos, 145-146

Raíces cuadradas, 259-260
Rango,
cálculo, 56
comparada con otras medidas de variabilidad, 66
definida, 56
Rango percentil, 26-30
Razón,
cálculo, 19
definida, 19
Razón o cociente *F*, 159
fórmula, 160
grados de libertad, 159
requisitos, 166
Razón *t*, 137-138
grados de libertad, 137-138
la misma muestra medida dos veces, 143-145
muestras de tamaño distinto, 140-143
muestras de igual tamaño, 138-140
requisitos, 145-146

Sesgo, 37-38
Sumas de cuadrados,
cálculo, 155-156
definida, 152
dentro de los grupos, 153
entre grupos, 153-154
total, 155

Tasa,
cálculo, 20
definida, 19
Tasa de cambio,
cálculo, 20
definida, 20
Tendencia central, 39
comparación de medidas, 44
media, 41
mediana, 40
moda, 39

Variabilidad, 55
comparación de medidas, 66
desviación estándar, 59-66
desviación media, 56-58
rango, 56
V de Cramér, 236