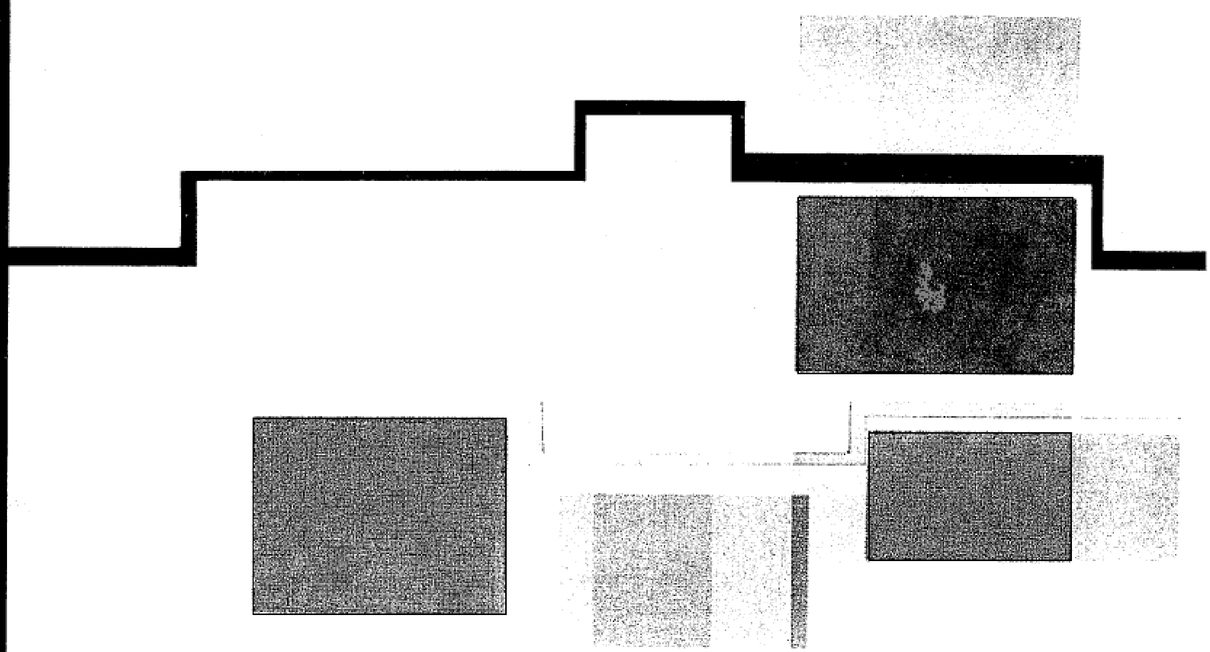
Estadística

Richard C. Weimer



CECSA

ESTADÍSTICA

SEGUNDA EDICIÓN EN INGLÉS
(PRIMERA EDICIÓN EN ESPAÑOL)

Richard C. Weimer
Universidad Estatal Frostburg

Revisión técnica:

Dr. Piotr Marian Wisniewski
Profesor de Planta del Depto. de Matemáticas
División de Ingeniería y Arquitectura
Instituto Tecnológico de Monterrey, Campus Estado de México

QUINTA REIMPRESIÓN
MÉXICO, 2003

COMPAÑÍA EDITORIAL CONTINENTAL

Para establecer comunicación
con nosotros puede hacerlo por:

correo:
Renacimiento 180, Col. San Juan
Tlihuaca, Azcapotzalco,
02400, México, D.F.

fax pedidos:
(015) 561 4063 • 561 5231

e-mail:
info@patriacultural.com.mx

home page:
http://www.patriacultural.com.mx

Título original:
STATISTICS, Second Edition
ISBN 0-697-12146-1

Traducción autorizada por:

Traducción:
M. en C. Ana Irene Ramírez Galarza
Facultad de Ciencias, UNAM

Revisión Técnica:
M. en C. René Valverde Ventura
ITESM Campus Estado de México

Estadística

bajo el sello de Compañía Editorial Continental
Renacimiento 180, Colonia San Juan Tlihuaca,
Delegación Azcapotzalco, C.P. 02400, México, D.F.

Miembro de la Cámara Nacional de la Industria Editorial
Registro núm. 43

ISBN 968-26-1261-6

Impreso en México
Printed in Mexico

Primera edición: 1996
Cuarta reimpresión: 2002
Quinta reimpresión: 2003

Este libro está dedicado con amor a mi esposa,
Marlene H. Weimer

Contenido

Prólogo

El libro *Estadística* se desarrolló durante un periodo de ocho años a partir de un conjunto de notas de clase, para trabajar en el salón, de un curso introductorio sobre probabilidad y estadística impartido a estudiantes avanzados de finanzas, ciencias sociales, ciencias biológicas y ciencias naturales. Se pretende que este libro sea una introducción a la probabilidad y a las técnicas modernas aplicadas en estadística, para aquellos estudiantes que han tomado al menos un curso de álgebra elemental y desean una buena comprensión de los conceptos y principios básicos de probabilidad y estadística; la lógica y la teoría de la estadística aplicada están desarrolladas en forma clara y natural.

Mi propósito principal al escribir este texto fue lograr una exposición cuidadosa y legible que ayudara a los estudiantes a entender las técnicas de estadística elemental; en consecuencia, se insiste en la comprensión e interpretación de los temas más que en la aplicación de una serie de "recetas de cocina".

Enfoque

Se motiva a los alumnos a estudiar información reciente e ideas nuevas, mediante una exposición clara con aplicaciones prácticas y ejemplos. El enfoque de las aplicaciones prácticas se utiliza a lo largo de la *Estadística* para transmitir las ideas esenciales de la probabilidad y la estadística modernas. Cada tema está motivado usando una aplicación práctica seguida de una generalización o regla, luego se resuelve al menos una aplicación para ampliar la explicación y justificar el tratamiento; por ejemplo, muchos textos introductorios presentan la media, la mediana y la moda como medidas de tendencia central y generalmente, no se hace intento alguno por justificar la necesidad de estudiar estos conceptos ni de por qué las tres medidas deben ser investigadas; con *Estadística*, se introduce al estudiante en la tendencia central estableciendo asimismo la utilidad de su estudio, en seguida tenemos un escenario que usa cuatro tipos distintos de datos los cuales sirven para motivar la razón de estudiar cuatro medidas de tendencia central, entonces cada medida se comenta y se desarrolla.

A lo largo del texto el rigor matemático se ha sacrificado en aras de un completo entendimiento: es preferible entender los conceptos y principios estadísticos, que memorizar una lista de fórmulas y términos sin ser capaz de aplicarlos o de comprender sus implicaciones. El lenguaje y las notaciones son de uso común y preciso.

Acerca del libro

El libro está organizado en tres unidades: Unidad I, Estadística descriptiva; Unidad II, Probabilidad básica; y Unidad III, Estadística inferencial. Las primeras dos unidades sirven como trampolín para la estadística inferencial, y la probabilidad se presenta como una forma de salvar el espacio entre la estadística descriptiva y la inferencial; se usa siempre un acercamiento en espiral para presentar y organizar el material. Por ejemplo, regresión y correlación se introducen en el capítulo 4 como conceptos descriptivos, y se usan para caracterizar a las poblaciones estadísticas, estas ideas se estudian de nuevo en el capítulo 14 y se extienden hacia el desarrollo de la estadística inferencial y de la construcción de modelos; el concepto de suma de cuadrados se presenta primero en el capítulo 3, y se usa en todo el texto para simplificar ideas y fórmulas que involucran varianza; las distribuciones muestrales se exponen en el capítulo 8 y después se usan en el desarrollo de técnicas inferenciales; la importancia del papel de las distribuciones muestrales al hacer inferencias, está completamente demostrada y enfatizada como un concepto central necesario para el desarrollo de los procedimientos de prueba de hipótesis y estimación; las distribuciones t se introducen en el capítulo 8 y se usan para construir intervalos de confianza para la media en el capítulo 9. De esta manera el estudiante siente cómodo el aprender estadística, ya que él o ella han visto antes muchos de los conceptos señalados.

La estadística descriptiva y la probabilidad se exponen de forma más balanceada que en muchos textos introductorios, lo que proporciona al maestro una gran flexibilidad al escoger el contenido del curso. Por ejemplo, el material sobre datos agrupados de la sección 3.3 puede no ser expuesto por muchos maestros, pero otros considerarán que este material es esencial porque sus estudiantes deberán tratar con reportes publicados por el gobierno. Ya que una comprensión adecuada de la estadística inferencial requiere de un cierto nivel de asimilación de la teoría de probabilidad elemental, se deja al maestro la elección del nivel de exposición en probabilidad. Las reglas del conteo y de la probabilidad se presentan en secciones separadas del capítulo 5 para minimizar la confusión entre los estudiantes.

Este libro difiere de muchos textos introductorios de estadística en el desarrollo claro de las distribuciones muestrales del capítulo 8, ahí se introduce primero al estudiante en el muestreo, el error muestral, las distribuciones muestrales y el teorema central del límite; también en este capítulo se tratan las distribuciones t, ya que se ha demostrado la importancia de hacerlo así para que los estudiantes se sientan más a gusto al usarlas para obtener inferencias en los últimos capítulos, habiéndolas entendido de antemano. En muchos tratados, las distribuciones t y las estimaciones sobre medias poblacionales se trabajan simultáneamente.

Estadística utiliza ejemplos y aplicaciones prácticas para motivar y explicar conceptos y principios estadísticos. Los **ejemplos** se usan para explicar conceptos y mostrar cómo se realizan los procedimientos estadísticos y las **aplicaciones**, para explicar las soluciones de problemas estadísticos. Algunos autores usan los términos "ejemplos" y "aplicaciones" indistintamente, nosotros preferimos distinguirlos. Los estudiantes aprenden un concepto tratando con ejemplos y no ejemplos del mismo, luego se usan las aplicaciones prácticas para resolver algún problema propuesto anteriormente o para responder una pregunta previa. Las aplicaciones prácticas son semejantes al tipo de ejercicios que el estudiante encontrará en los ya denominados **Aplicaciones adicionales** y **Un paso más allá** al final de varias secciones; esta diferencia entre ejemplos y aplicaciones permitirá a los estudiantes aprender estadística más fácilmente porque si no entiende un concepto, puede revisar ejemplos relevantes o si tiene dudas al resolver un ejercicio, puede recurrir a las aplicaciones prácticas apropiadas.

Los capítulos del libro y las secciones de los mismos, están organizados con una secuencia lógica. Así, los primeros 11 conforman el contenido esencial de un curso básico; hay secciones y temas en esos 11 capítulos que pueden omitirse sin pérdida de continuidad. Sin embargo, no se ha hecho intento alguno de clasificar algún material como opcional, ya que lo que una persona juzga esencial, para otra puede ser opcional y viceversa.

Cambios en esta edición

La filosofía básica del texto y su organización han permanecido sin cambios, pero el contenido se ha extendido sustancialmente para un curso de una o dos partes en probabilidad aplicada y estadística. Esta edición pone también un mayor énfasis en el uso de la computadora en todos los capítulos. A continuación se describe parte del nuevo contenido y de los cambios en la presentación.

- La presentación al principio de cada capítulo incluye un panorama del mismo y un motivador introductorio.
- Hay tres niveles de ejercicios en lugar de dos, y muchas secciones concluyen con gran conjunto de ejercicios dividido en tres partes: Habilidades básicas, Aplicaciones adicionales y Un paso más allá. Los ejercicios de **Habilidades básicas** ofrecen entrenamiento adicional y práctica con aplicaciones en computadora que son esenciales para una mejor comprensión; los de **Aplicaciones adicionales** son parecidos a los ya vistos en el texto; representan aplicaciones prácticas de una gran variedad de temas y obligan a desarrollar las habilidades que se requieren para clasificar problemas por distribución, así como a traducir problemas coloquiales a símbolos matemáticos y a fórmulas para resolverlos; los ejercicios de **Un paso más allá** son de naturaleza más difícil, desarrollan las ideas planteadas en el texto, piden al estudiante que proporcione pruebas simples de ciertos hechos importantes y, a menudo, conducen a nuevas ideas; el nivel de este tipo de ejercicios ya se considera apropiado para una sección superior de estadística introductoria. En total hay más de 2,000 ejercicios en el libro.
- Ejercicios paralelos. En general, en los ejercicios de habilidades básicas y aplicaciones adicionales, cada cuestionamiento con número par es semejante a otro con número impar; las respuestas de todos los ejercicios impares de Habilidades básicas y Aplicaciones adicionales se incluyen en la sección **Respuestas a ejercicios impares** de los apéndices, donde se encuentran también respuestas a los ejercicios de repaso impares y las preguntas sobre el dominio de los capítulos.
- La presentación del final de un capítulo incluye **Aplicaciones en computación** y **Experimentos con datos reales.** Las aplicaciones en computadora utilizan típicamente grandes conjuntos de datos, cálculos repetitivos y cálculos largos o simulaciones. Asimismo, los experimentos con datos reales que se incluyen en los capítulos del 8 al 15 sirven para proporcionar experiencias que manejen datos reales; en el apéndice C incluimos una base de datos consistente en información selecta sobre la salud de 720 individuos.
- Capítulo 1 Introducción

  Sección 1.5 Se ha añadido "El papel de la computadora en la estadística" para destacar el desempeño de las computadoras en esta materia.

- Capítulo 3 Estadística descriptiva-análisis de datos univariados

  Sección 3.2 Se ha extendido el tema "Medidas de dispersión" o "variabilidad" para incluir comentarios sobre el rango intercuartílico. Este concepto es prerrequisito para entender los comentarios sobre gráficas de caja que aparecen en la sección 3.4.

  Las medidas de tendencia central y de dispersión para datos agrupados, se tratan ahora en una sección separada, (la sección 3.3, Tendencia central y dispersión para datos contenidos en tablas de frecuencia agrupada). Este cambio da más flexibilidad al instructor al diseñar su curso. Los maestros que utilicen computadoras en sus cursos tal vez prefieran no cubrir este material.

  Sección 3.4 Valores estándar y aberrantes (antes, sección 3.3) contiene ahora material sobre gráficas de caja y valores aberrantes, que se ha añadido para quienes deseen incluir más material sobre análisis exploratorio de datos.

- Capítulo 4 Análisis de datos bivariados

  Se ha omitido el material de la sección 4.1, Ecuaciones lineales y sus gráficas, respecto a la primera edición; ahora contiene material de covarianza con el propósito de facilitar la comprensión por parte del estudiante del concepto de correlación, presentado en la sección 4.2.

- Capítulo 5 Introducción a la probabilidad elemental

  Las notaciones $\cup$ y $\cap$ se usan en lugar de "o" e "y", respectivamente, ya que provocan menos confusión en el enunciado de resultados y permiten al estudiante aprovechar el material de teoría de conjuntos aprendido en secundaria.

  Se ha añadido material sobre permutaciones y combinaciones en la sección 5.3 para permitir un conteo más sofisticado.

- Capítulo 6 Distribuciones discretas

  Se ha cambiado el nombre del capítulo, Distribuciones binomiales, por el de Distribuciones discretas, y permite mayor flexibilidad en la enseñanza porque ahora cuenta con tres nuevas secciones.

- Capítulo 7 Distribuciones continuas

  El nombre del capítulo varió de Distribuciones normales a Distribuciones continuas. Para permitir más flexibilidad en la enseñanza se han agregado dos secciones, la 7.1 y la 7.4.

- Capítulo 11 Inferencias concernientes a dos parámetros

  El material sobre muestras pequeñas se ha separado en dos, en las secciones 11.5 y 11.6, para dar mayor claridad y disminuir la confusión en los estudiantes.

- Capítulo 13 Análisis de varianza

  Se modificó el nombre al capítulo, de Análisis de varianza de un solo factor a Análisis de varianza, a fin de resaltar que el contenido del capítulo fue ampliado. Tiene también ahora dos nuevas secciones (13.4 y 13.5), con lo que el maestro diseñará su curso con libertad de uso de computadoras para un mayor desarrollo de ANOVA.

  En la sección 13.3 se usa el procedimiento de Bonferroni en lugar del de Sheffé. Este cambio permite al estudiante utilizar los resultados previamente obtenidos en distribuciones t y propicia una mejor comprensión.

- Capítulo 14 Análisis de regresión lineal

  La sección 14.4 sobre regresión lineal múltiple es nueva y permitirá a los maestros el uso de computadoras a fin de investigar mejor el modelo lineal.

  Se ha modificado la sección 14.3 de la primera edición moviendo el material sobre prueba de Spearman al capítulo 15; en cambio el maestro podrá trabajar con técnicas no paramétricas en un solo capítulo.

- Capítulo 15 Pruebas no paramétricas para muestras grandes

  Una nueva sección, la 15.2, incluye el estudio de la prueba de Wilcoxon para rangos con signo, ya que en muchos casos, esta prueba es más contundente que la prueba del signo presentada en la sección 15.1.

  En el apartado 15.4 tratamos la prueba de Friedman como el análogo no paramétrico del ANOVA doble cuando no se presenta la interacción.

El texto está diseñado para usarse en cursos con o sin computadoras, y proporciona una introducción específica para el uso de éstas que puede ser opcional; para aquellas clases que deseen auxiliarse con las computadoras, *Estadística* proporciona algunas de las siguientes ventajas:

- No hay necesidad de estudiar las fórmulas para hacer cálculos porque éstas se manejan más eficientemente vía la computadora.
- Pueden usarse valores simulados de variables aleatorias para distribuciones empíricas.
- Se facilita la comprensión de las distribuciones muestrales y del teorema central del límite.
- También es sencilla la comprensión del error muestral y de las propiedades de la probabilidad.
- Puede dejar de enfatizarse sobre el uso de fórmulas de ANOVA y de regresión lineal múltiple, y emplear el tiempo ahorrado en interpretar los resultados.
- El estudiante es capaz de manejar una mayor cantidad de datos reales.

Debido a su gran preponderancia en colegios y universidades, se ha escogido a MINITAB como el paquete estadístico que ilustre el uso de la computadora en probabilidad y estadística aplicadas. Las pantallas de computadora de MINITAB se usan en muchas de nuestras secciones para ilustrar los comandos de MINITAB y las respuestas correspondientes; además, sirven al menos para los siguientes cinco propósitos:

1. Ilustrar la facilidad del uso de MINITAB.
2. Visualizar los comandos que necesita manejar el usuario a fin de alcanzar los resultados estadísticos deseados.
3. Mostrar el formato y la notación empleados por la computadora en sus respuestas.
4. Conocer el alcance de la potencia del cálculo estadístico obtenido con MINITAB y otros paquetes estadísticos semejantes.
5. Enseñar el uso de MINITAB para llevar a cabo algunas tareas estadísticas comunes.

Organización

El libro está organizado para permitir una gran flexibilidad al maestro. Quienes deseen pasar rápidamente por los capítulos 1 y 2, sólo necesitan enseñar el concepto de histograma del capítulo 2; si usted no planea un estudio profundo de probabilidad, sólo necesitará impartir las secciones 5.6 y 5.7 del capítulo 5 como una base para secciones posteriores, ya que la sección 5.6 es fundamental para el capítulo 12 y la 5.7 para las secciones 6.3, 8.1 y 8.2; si pretende omitir distribuciones binomiales y los conceptos relacionados con ellas, sáltese las secciones 6.1, 6.3, 7.3, 8.5, 9.3, parte de la sección 10.4, la 11.3 y la 12.1. Después de cubrir hasta el capítulo 11, los restantes pueden estudiarse en cualquier orden.

A menudo los estudiantes encuentran confuso estudiar distribuciones normales, la aproximación normal a la binomial y el teorema central del límite en el mismo capítulo, por ello se ha tenido mucho cuidado de secuenciar estos tres temas; asimismo, el teorema central del límite está tratado de forma que destaque su importancia en el desarrollo de métodos estadísticos para hacer inferencias en el caso de muestras grandes. La distribución muestral de la suma de la muestra se describe también como aproximadamente normal y relacionada con distribuciones binomiales que se aproximan vía distribuciones normales.

En el capítulo 9 se presenta la estimación de parámetros únicos siguiendo un desarrollo lógico, y en el 10 se desarrollan los procedimientos de prueba de hipótesis para pruebas de una sola muestra; estos dos procedimientos los exponemos juntos en el capítulo 11 para mostrar a los estudiantes cómo hacer comparaciones entre dos parámetros usando dos poblaciones, y para lograr resultados consistentes cuando se usan pruebas no direccionales.

El concepto de error muestral introducido en el capítulo 8, sirve como unificador para explicar la variación natural de los valores de los estadísticos muestrales cuando se muestrea. Por ejemplo, en el capítulo 10, si no hay evidencia que sugiera la falsedad de la hipótesis nula, la diferencia entre el valor del estadístico de prueba y el hipotético valor poblacional se atribuye al error muestral.

A menudo se pide a los estudiantes que clasifiquen aplicaciones de acuerdo con la distribución muestral y con las hipótesis necesarias; las pruebas no paramétricas del capítulo 15 se presentan como una alternativa cuando los supuestos son débiles o se sabe que serán violados y tanto la interpretación de los resultados como la identificación de los riesgos asociados se resaltan en los capítulos del 8 al 15.

Características principales

Este texto difiere de otros que tratan el mismo tema, al presentar la estadística como una forma de trabajar con la eventualidad y la variabilidad, mostrando los verdaderos papeles jugados por la variabilidad, por el error muestral y por la distribución muestral, en el desarrollo de la inferencia estadística. Los estudiantes que entiendan intuitivamente los principios subyacentes y la manera en que se relacionan, estarán mejor equipados para el futuro que aquellos que sólo han estado en condiciones de sustituir números en la fórmula apropiada. Otras características distintivas serían:

1. El texto motiva y explica la lógica en que se basan los principios de la estadística inferencial, impulsando a los lectores a practicar esta lógica; enseña a los estudiantes no sólo la forma de usar las técnicas estadísticas, sino también a entender cómo trabajan éstas, para lo cual el contenido didáctico del libro se ha ordenado de manera apropiada, unificada.

2. Los histogramas, elementos fundamentales de la estadística, se usan siempre para indicar propiedades básicas de distribuciones teóricas.
3. Siempre que es posible se resaltan las relaciones importantes que conciernen a pruebas de hipótesis; por ejemplo, el análisis de varianza se presenta como una extensión de la prueba de t con dos muestras; se comenta la relación entre t^2 y F; incluimos la prueba de ji-cuadrada para probar proporciones de más de dos poblaciones y demostramos que es una extensión de la prueba de z para proporciones de dos muestras; también se examina la relación entre z^2 y χ^2.
4. Las proporciones para poblaciones dicotómicas y binomiales se introducen primero asociándolas con medias para una población consistente de unos y ceros. Después se obtienen las propiedades de la distribución muestral de la proporción muestral, usando las propiedades de la distribución muestral de la media muestral.
5. Se pone énfasis en las condiciones indispensables para aplicar las distintas pruebas de hipótesis; la prueba de F para probar la hipótesis de homogeneidad de varianza se estudia antes que la prueba de t de dos muestras, posibilitando así probar esta hipótesis antes de aprender a usar la prueba de t de dos muestras.
6. El ANOVA simple se introduce y promueve en relación con el concepto de rango. Se da el procedimiento de Bonferroni para probar diferencias pareadas, y el estadístico de Hays omega-cuadrada se usa para medir la potencia de la relación obtenida de una prueba de F significativa. El ANOVA con dos criterios se introduce primero usando el diseño de bloques aleatorizados y a continuación se comentan los diseños de interacción y factoriales; insistimos mucho en la utilización del enfoque computacional en el ANOVA de datos en diseños de dos factores, cuando el estudiante ya entiende los conceptos básicos utilizados.
7. Para regresión lineal múltiple se toma el enfoque computacional, que es comprensible y completo; con la ayuda de la computadora se estudia el modelo lineal con más de una variable independiente, como una extensión del modelo básico de una variable, y los resultados principales se demuestran y explican usando los que proporciona MINITAB. Una vez más se pone énfasis en comprender el uso y las implicaciones de la regresión lineal múltiple, así como en la capacidad de interpretar los resultados de la computadora, mas no en el uso de fórmulas para obtener los resultados.
8. El sistema pedagógico aplicado guiará y auxiliará a los estudiantes. Para ello cada capítulo comienza con una lista introductoria de objetivos y termina con un resumen conciso. Además, cada capítulo tiene los siguientes elementos:
   - El **motivador** que abre el capítulo, enfatiza la necesidad de comentar cada contenido. Estos motivadores constan de breves resúmenes sobre investigación actualizada, aplicaciones reales de lo escrito en la materia y problemas interesantes. Sirven para responder en parte la pregunta "¿por qué debo estudiar el material de este capítulo?"
   - El **panorama** del capítulo proporciona un espectro para estudiarlo.
   - Los **ejemplos** y **aplicaciones** se usan libremente para explicar y justificar nuevos conceptos y procedimientos. Las aplicaciones proporcio-

nadas son realistas porque su fundamento no es de naturaleza técnica. Como están obtenidas de una amplia variedad de áreas en ciencias naturales, biológicas y sociales, pero no incluyen una gran cantidad de información previa o de lenguaje técnico, permiten resolver adecuadamente los ejercicios a estudiantes de distintas disciplinas.

- Los ejercicios del final de muchas secciones le dan al estudiante la oportunidad de practicar lo que aprendió en la sección.
- Al final de cada capítulo hay una lista de notaciones importantes.
- También se encuentra al final de cada capítulo, una lista de hechos y fórmulas más preponderantes.
- Se incluyen asimismo ejercicios de revisión que no siguen un orden particular de temas.
- Al terminar los capítulos del 2 al 15, hay ejercicios opcionales de aplicación en computadora y éstos piden al estudiante que efectúe alguna simulación o que analice un conjunto de datos relativamente grande.
- Un examen de conocimientos que permitirá al estudiante evaluar la comprensión acumulada en cada capítulo, desde el 2 hasta el 15, lo hallará al final de cada uno de ellos.

Agradecimientos

Agradezco la colaboración de mucha gente que me ayudó a preparar este libro, y en gran medida, las sugerencias y las críticas constructivas de cientos de estudiantes de la Frostburg State University, quienes estudiaron *Estadística* desde la primera edición y en los primeros borradores del texto durante los pasados siete años. Los estudiantes Peter Tirrell, Stewart Crall, Donna Pope y William Byers me ayudaron a calcular y a verificar las respuestas a los problemas en la primera edición. Aprecio también la colaboración y asesoría proporcionadas por mis colegas, los doctores Kil Lee, Lance Revennaugh y Kurtis Lemmert. Un gran número de observaciones y sugerencias para la primera edición se deben a los siguientes revisores: James Baker, Jefferson Community College; James Baldwin, Nassau Community College; Pat Cerrito, University of South Florida; James Daly, California Polytechnic State University; Ken Eberhard, Chabot College; Antanas Gilvydis, Malcom X College; Raymond Guzmán, Pasadena City College; Gary Itzkowitz, Glassboro State College; Keith Nelson, Beloit College; Jim Ridenhour, Austin Peay State University; Larry Ringler, Texas A&M University; Ann Thomas, University of Northern Colorado; William Tomhave, Concordia College; y John Van Druff, Fort Steilacoom Community College.

También deseo agradecer a los revisores de la segunda edición: Derek K. Chang, California State University, Los Ángeles; William H. Beyer, The University of Akron; Nancy J. Carter, California State University, Chico; Kenneth R. Eberhard, Chabot College; Theodore S. Erickson, Wheeling Jesuit College; H. Joseph Heffelfinger, Anne Arundel Community College, doctora Ángela Hernández, University of Montevallo; R. Bruce Lind, University of Puget Sound; Cameron Neal, hijo, Temple Junior College; Steve Patch, University of North Carolina at Ashville; Jim Ridenhour, Austin Peay State University; Joseph F. Stokes, Western Kentucky University; Deborah A. Vrooman, Coastal Carolina College of the University of South Carolina; June Miller White, St. Petersburg Junior College y Ellen T. Wood, Stephen F. Austin State University.

Aprecio especialmente la experimentada asesoría ofrecida por la profesora Susan L. Reiland y por la señora Patricia Steele. La profesora Reiland leyó el manuscrito completo para la primera edición, antes de que se le preparara la prueba final, y fungió como editora de la copia para la primera edición del manuscrito. La señora Steele fungió como editora de la segunda edición y su profesionalismo, dedicación y meticuloso cuidado y tratamiento de los detalles merecen un agradecimiento especial.

Mi agradecimiento también para el equipo editorial y de producción de Wm. C. Brown Editores, especialmente a Jan Scotchmer por su amplia asesoría y estímulo a lo largo del proyecto.

Finalmente, también estoy en deuda con el doctor Larry Brant del Gerontology Research Center y del National Institute of Aging Baltimore, Maryland, por permitirme usar una base de datos consistente de 720 registros relacionados con la salud y que viene incluida en el apéndice C; la cual es utilizada a lo largo del texto. Esos datos fueron obtenidos como parte del estudio longitudinal de Baltimore sobre Aging (BLSA) y representan sólo una fracción de la información recabada durante el estudio.

Richard C. Weimer
Enero de 1992

UNIDAD UNO

Estadística descriptiva

1 Introducción

DESCRIPCIÓN

1.1 ¿Por qué estudiar estadística?
1.2 El lenguaje de la estadística
1.3 Estadística descriptiva e inferencial
1.4 Inferencias y deducciones
1.5 El papel de la computadora en la estadística

OBJETIVOS DEL CAPÍTULO

En este capítulo estudiaremos:

- *Los dos significados del término estadística.*
- *La diferencia entre una población y una muestra.*
- *La diferencia entre un estadístico y un parámetro.*
- *La diferencia entre estadística descriptiva y estadística inferencial*
- *Cómo se relacionan la probabilidad y la estadística con la inducción y la deducción.*
- *Por qué es importante estudiar estadística.*
- *El papel de la computadora en la estadística.*

MOTIVADOR 1

La razón principal de que una empresa sea abandonada por sus clientes es un mal servicio. Los directores y los profesionales del desarrollo organizativo tienen el reto de ayudar a sus empresas a mejorar el servicio al cliente, satisfacer al mismo y mantener el lugar de ventas. Según un artículo de Stum y Church[1], la investigación lo demanda; ellos citan los siguientes hechos obtenidos mediante técnicas estadísticas y que ilustran las distintas formas de usar la información estadística para el manejo de la toma de decisiones.

- De acuerdo con un estudio de Forum, el motivo principal de que un cliente acuda a un competidor es un mal servicio.
- La American Management Association asegura que el 60% de ventas nuevas deben provenir de clientes antiguos, que muestran lealtad de recompra.
- El consultor R. L. Desatnick señala que, por ejemplo, en la industria automotriz un cliente leal representa un ingreso de 140,000 dólares a lo largo de su vida.
- La Consumer Affairs Office advierte que siete de cada diez personas pueden suspender una relación con un proveedor basándose en el tipo de trato recibido durante el primer contacto.
- La ATT reporta que el número de 800 líneas telefónicas usadas a menudo por compañías que desean proporcionar información o asistencia al cliente, crece anualmente un 25 por ciento.
- El Technical Assistance Research Project (TARP) afirma que una compañía nunca tendrá conocimiento del 90% de sus clientes insatisfechos, aunque estas personas molestas les cuenten a otras 10 sus experiencias negativas, pero que cuando los clientes insatisfechos los emplazan legalmente, su lealtad aumentará significativamente si sus demandas se resuelven en forma satisfactoria.

Panorama del capítulo

Las personas vemos la estadística desde perspectivas distintas, suele vérsele como algo relacionado con porcentajes, promedios, cuentas y gráficas; para algunos, la estadística es un área de estudio consistente en reglas y métodos para tratar información; para otros, la estadística es una forma de actuar y de pensar con respecto a los sucesos mundanos que ocurren irregularmente y que están gobernados por ciertas leyes de incertidumbre. Este capítulo introduce las ideas básicas y el lenguaje de la estadística.

SECCIÓN 1.1 *¿Por qué estudiar estadística?*

Existen cuando menos cuatro buenas razones para estudiar estadística, al hacerlo seremos capaces de:

1. Aprender las reglas y métodos para tratar información estadística.
2. Evaluar y cuantificar la importancia de los resultados estadísticos que veamos publicados.
3. Conocer los aspectos del pensamiento estadístico como un componente esencial de una educación humanística.
4. Entender mejor el mundo real de nuestro entorno.

Quizá una de las razones más importantes para estudiar estadística en este nivel, sea que nos permite tomar críticamente la información estadística proporcionada por los medios de comunicación, por ejemplo, consideremos las afirmaciones siguientes hechas en algunos anuncios en estos medios. Usaremos X en lugar del nombre de una marca comercial:

1. La llanta marca X frena un 35% más rápido. (¿Más rápido que qué?)
2. En un periodo de cuatro años, el rendimiento de la gasolina para el coche X aumentó en 50%. (¿50% de qué?)
3. El jabón marca X es 99.44% puro. (¿Puro en qué? ¿El jabón?)
4. Noventa por ciento de todos los coches de la marca X vendidos en los últimos 10 años están todavía en circulación. (Con esta afirmación, supondremos erróneamente que los coches se vendieron en aproximadamente el mismo número todos los años, pero muchos de los autos en circulación fueron adquiridos durante los últimos tres o cuatro años.)
5. El calmante marca X contiene el doble de calmante. (¿Significa eso que calma más eficazmente el dolor que cualquier otro calmante?)
6. Cuatro de cada cinco dentistas interrogados declararon preferir la pasta dentífrica X. (¿Cuántos dentistas fueron interrogados? ¿Cómo fueron escogidos?)
7. Ninguna aspirina calma mejor el dolor que la de marca X. (Esta afirmación no dice que la marca X sea mejor que cualquier otra, sólo dice que la marca X es tan buena como cualquiera otra.)
8. Las mujeres que usan la marca X reportaron un 80% de alivio durante las primeras horas. (¿Puede medirse el alivio en términos de porcentaje? ¿Qué significa esto?)
9. El calmante marca X es recomendado por mucha de la gente más conocedora. (¿Más conocedora sobre qué? ¿Sobre cualquier tema?)

Como consumidores de información estadística y usuarios potenciales de técnicas estadísticas, necesitamos entender las ideas básicas y las herramientas de esta disciplina. Muchos de nosotros estamos influidos diariamente por algún aspecto de la estadística manejada en la información que obtenemos de la radio y de la televisión, o de periódicos y revistas. Por ejemplo, podemos leer u oír que:

1. Algunos estudios sugieren que alrededor del 50% de todos los ahogamientos de adolescentes y adultos están asociados con el uso del alcohol.
2. Las familias con sólo uno de los padres constituyen actualmente el 26% de todas las familias estadounidenses con niños menores de 18 años contra sólo el 13% en 1970.
3. Siete de cada diez estadounidenses no tienen facultad de decisión.
4. El predominio de la diabetes en personas con sobrepeso es casi el triple que en personas sin sobrepeso.
5. Más de 3,000 compañías aseguradoras pagan arriba de 8,800 millones de dólares anuales por reclamaciones.
6. Hay un 50% de probabilidades de que el perdedor nunca más vuelva a competir.
7. Los niños que cepillan sus dientes con la pasta dentífrica marca X tienen un 35% menos caries.
8. El importe neto medio de las jubilaciones recientes para los beneficiarios de la Seguridad Social de 1981 a 1982 estuvo entre los 64,700 y los 68,300 dólares para matrimonios y entre los 17,000 y los 30,000 dólares para los solteros.
9. En 1960 se estimó que sólo el 1% de los estudiantes del último año de bachillerato había probado la mariguana, mientras que en 1980 se estimó que el 60% lo había hecho.
10. Algunos estudios sugieren que el sentimiento de desamparo está correlacionado con una marcada disminución de las células que combaten enfermedades en varios sistemas inmunológicos.

Estos ejemplos indican que la información estadística se usa por una gran cantidad de razones. Entre ellas encontramos:

- Informar al público, como en los ejemplos anteriores.
- Proporcionar comparaciones, ejemplos 2, 4, 8 y 9.
- Explicar acciones que se han efectuado, ejemplos 1, 4 y 10.
- Influir en decisiones que han de tomarse, ejemplos 1 y 7.
- Justificar un reclamo o afirmación, ejemplos 1, 7 y 10.
- Predecir futuros resultados, ejemplo 6.
- Estimar cantidades desconocidas, ejemplos 1 y 9.
- Establecer una relación o asociación entre dos factores, ejemplos 1, 4 y 10.

Como somos consumidores de información estadística, podemos usar la estadística para estudiar y entender mejor muchos sucesos cambiantes que contribuirán a nuestra comprensión del mundo. Estudiar estadística nos permitirá dar una interpretación razonable a cada uno de los ejemplos anteriores; así, la cantidad 35% del ejemplo 7 puede interpretarse libremente porque no conoce-

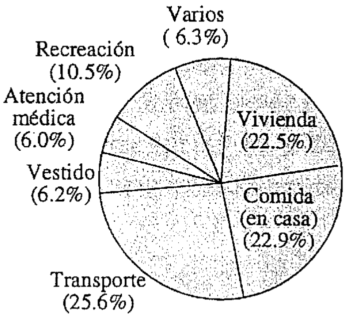


FIGURA 1.1

Cómo se gastan 142,700 dólares para mentener a un niño hasta los 17 años de edad

mos la base de la comparación; puede ser difícil si no imposible, encontrar una pasta de dientes que permita tener 35% menos caries que cualquier otra pasta de dientes cuando se prueba bajo condiciones similares en grupos de niños semejantes e independientes, pero sería extremadamente simple encontrar un niño que use pasta de la marca X y que tiene 35% menos caries que algún otro niño, por ejemplo uno que no se los lava. Como consumidores cotidianos de información estadística, debemos estar conscientes de los usos y abusos en el manejo de dicha información. En este curso aprenderemos cómo puede obtenerse cada uno de los números mencionados en los ejemplos de esta sección.

Como consumidores de información estadística encontramos con frecuencia gráficas que la proporcionan. Por ejemplo, los padres han gastado aproximadamente 142,700 dólares en un hijo cuando éste llega a los 17 años. La gráfica de pastel de la figura 1.1 indica cómo se ha distribuido ese dinero[2]; podemos ver de un vistazo que del presupuesto total, los gastos debidos a comida, vivienda y transporte tienen montos más o menos iguales y los hechos en diversiones, atención médica y varios, también tienen montos casi iguales.

Como usuarios potenciales de los métodos y técnicas estadísticos, necesitamos estar familiarizados con el quehacer de la investigación estadística básica, con la descripción de los resultados de nuestra búsqueda científica, con la toma de decisiones basadas en ésta y con la estimación de cantidades desconocidas. Los ejemplos 1.1, 1.2 y 1.3 ilustrarán cómo puede usarse la estadística en general, y los ejemplos 1.4, 1.5 y 1.6 mostrarán aplicaciones de la estadística que responden a cuestiones de interés general.

EJEMPLO 1.1

En comercial de televisión se afirma que una marca de un producto es superior a todas las otras marcas: si la afirmación se basa en una encuesta científica se estará usando para educar a los televidentes; pero si dudamos de la afirmación, en un intento para desmentirla podemos recopilar datos relevantes sobre todas las marcas del producto en cuestión, analizar los resultados usando procedimientos estadísticos apropiados y tomar una decisión respecto a la afirmación del anuncio. Con frecuencia las afirmaciones de los anuncios se basan en información insuficiente o en análisis defectuosos de la misma.

EJEMPLO 1.2

Suponga que queremos determinar quién es el mejor maestro del Excel College. ¿Cómo debemos proceder para hacerlo? Podemos preguntar a los estudiantes de Excel quién es el mejor maestro, analizar los resultados y llegar a una conclusión. ¿Debemos preguntar a cada estudiante?, ¿cómo debe conducirse la encuesta?, ¿cómo se analizará la información? y ¿de qué manera se determinará quién es el mejor maestro? Uno de los propósitos centrales de la estadística es dar respuestas a éstas y a otras preguntas.

EJEMPLO 1.3

Una compañía de seguros de vida está pensando ofrecer primas reducidas a los asegurados que se enrolen en un programa de ejercicios. Para ayudar a la compañía de seguros a tomar la decisión, se recolectará y analizará información sobre mortalidad: ¿qué tipo de programa de ejercicios permitirá calificar a un asegurado para obtener una prima reducida?, ¿de cuánto debe ser la reducción? y ¿qué factores de riesgo deben descalificar a un asegurado enrolado en un programa de ejercicios para obtener una reducción en la prima? Una persona con una sólida formación estadística es quien podrá asesorar a la compañía de seguros para evaluar los méritos del nuevo programa.

EJEMPLO 1.4

¿Qué papel juega la dieta en las afecciones cardiacas de coronaria? Durante casi dos generaciones se ha debatido el papel de la dieta en las afecciones cardiacas de coronaria. La teoría dieta-corazón afirma que la reducción del colesterol en la sangre mediante la dieta disminuye el riesgo de contraer afecciones cardiacas de coronaria; para probar parcialmente la relación entre la reducción del colesterol en la sangre y las afecciones cardiacas de coronaria se emprendió un estudio que utiliza el fármaco colesteramina, medicamento reductor del colesterol. El estudio utilizó a 3,800 hombres de edad intermedia; todos tenían niveles de colesterol en la sangre de al menos 265 miligramos por decilitro de sangre, que los colocaban dentro del 5% de adultos estadounidenses con más altos niveles de colesterol, y se encontró que todos estaban libres de cualquier síntoma de afecciones cardiacas de la coronaria al empezar el estudio. Mil novecientos hombres fueron asignados aleatoriamente a cada uno de los dos grupos: un grupo con tratamiento y un grupo de control. Los participantes en el grupo con tratamiento recibieron dosis diarias de colestiramina y una dieta para reducir el colesterol, durante 7.4 años en promedio, mientras que los participantes en el grupo de control no recibieron colestiramina, sino un placebo indistinguible de ésta. El estudio concluyó que el grupo con fármaco tuvo menos ataques al corazón (155 personas contra 187 del grupo de control) y menos muertes por ataques al corazón (30 personas contra 38). Se juzgó que la diferencia entre los dos grupos es estadísticamente significativa; la probabilidad no se debe sólo a factores de suerte, los hallazgos apoyan la creencia de que la reducción de colesterol en la sangre usando colestiramina en hombres de edad intermedia con niveles de colesterol superiores a 265 miligramos, es eficaz para reducir las afecciones cardiacas de coronaria.[3]

EJEMPLO 1.5

¿Es buena la nueva Coca Cola? A principios de 1985, la compañía Coca Cola anunció que estaba cambiando en secreto la fórmula para fabricar esa bebida, una fórmula que había usado desde 1886; luego de que la nueva Coca Cola fue lanzada al mercado, los miembros de *Consumer Reports* intentaron responder preguntas como éstas: ¿a qué sabe realmente la nueva Coca Cola?, ¿es mejor que la antigua? y ¿cómo se compara con la Pepsi Cola? El equipo de investigación realizó tres pruebas de sabor a ciegas con 95 de sus miembros y 532 copas de plástico. Los resultados del estudio no mostraron diferencia en los gustos entre Pepsi Cola y la nueva Coca Cola; ambos productos fueron preferidos sobre la vieja Coca Cola por un margen de 2 a 1. Se encontró que las tres fórmulas consistían en cerca de 99% de agua carbonatada y azúcar, cada una con entre 6.14% y 6.22% de fructosa y entre 4.54% y 4.73% de dextrosa, azúcar de maíz.[4] Tradición y otros factores humanos diversos pueden afectar las preferencias de los consumidores; aunque el experimento parece indicar que la nueva Coca Cola es superior en sabor a la antigua, la Coca Cola clásica está desplazando a la nueva en muchas regiones de Estados Unidos.

EJEMPLO 1.6

¿Daña el humo del tabaco a los no fumadores? Se sabe desde hace mucho que a los fumadores les hace daño fumar. Para determinar si el humo del tabaco es dañino para los no fumadores, en la Universidad de California en San Diego se realizó un estudio basado en pruebas sobre funciones pulmonares; estas pruebas fueron realizadas en 200 no fumadores de mediana edad, cuyo entorno estaba relativamente libre de humo de tabaco, y en otro grupo de 200 no fumadores de mediana edad que habían estado expuestos rutinariamente al humo del tabaco durante 20 años o más. Ambos grupos se compararon con fumadores que no aspiran el humo, fumadores ligeros, fumadores moderados y fumadores empedernidos; los investigadores concluyeron que los dos grupos de no fumadores no diferían significativamente en los resultados de pruebas pulmonares que medían el daño de la capacidad vital y la razón de expiración inicial, sin embargo, reportaron una diferencia estadísticamente significativa entre los dos grupos en la cantidad de pequeñas vías pulmonares dañadas; los no fumadores expuestos con pasividad al humo en su trabajo tuvieron puntajes que se juzgaron indicativos de enfermedades pulmonares, pero sus puntajes fueron similares a los de los fumadores ligeros, uno a diez cigarros por día, y a los de los fumadores que no aspiran el humo. El estudio sugiere que la exposición crónica al humo del tabaco en el trabajo es dañina para los no fumadores y que reduce en mucho las funciones de las vías pulmonares pequeñas.[5]

GRUPO DE EJERCICIOS 1.1

Para cada una de las afirmaciones siguientes: (a) establezca la conclusión a la que, en su opinión, llega un lector de la afirmación, y (b) liste las preguntas que considera deben hacerse sobre la afirmación para evitar llegar a una conclusión falsa.

1. Entre los 35 y los 65 años de edad, siete de cada diez trabajadores sufrirán una lesión que durará tres meses o más.
2. A la edad de 32 años, una lesión que dure tres meses o más es seis veces más probable que la muerte.
3. El personal de nuestro autoservicio tiene 100 años de experiencia.
4. En más de un millón de millas de prueba, nuestros automóviles han tenido un promedio de reparaciones menor al 1 por ciento.
5. Noventa y ocho por ciento de los médicos prescriben el calmante encontrado en la marca X.
6. El alimento contenido en nuestra marca ayuda a reducir los niveles de colesterol de la sangre en los adultos.
7. Los productos alimenticios bajos en grasas ayudan a reducir los ataques al corazón.
8. Los alimentos con alto contenido de fibra disminuyen las posibilidades de tener cáncer de colon.

SECCIÓN 1.2 El lenguaje de la estadística

Como todas las ciencias, la estadística tiene su lenguaje propio. Comencemos examinando el término **estadística** que tiene dos significados.

> La estadística, en singular, es la ciencia de recolectar, organizar, analizar e interpretar información; las estadísticas, en plural, son números obtenidos de un conjunto o colección de informaciones.

Como ciencia, la estadística se encarga de describir los resultados de una investigación científica, de tomar decisiones basadas en dicha investigación y de estimar cantidades desconocidas. Las características numéricas usadas como estimaciones sirven como ejemplo de un **estadístico.**

EJEMPLO 1.7

Los investigadores calculan que toda la familia de computadoras personales de la marca IBM, controla alrededor del 40% de las microcomputadoras vendidas en Estados Unidos. El número 40% es un ejemplo de un estadístico.

Una diferencia básica en estadística es la que existe entre una población y una muestra.

Una **población** es el total de la información o de los objetos de interés para un estadístico en una investigación particular.

Una **muestra** es cualquier subconjunto de una población.

EJEMPLO 1.8

La colección de promedios por grado de los estudiantes (PPG) en un plantel de bachillerato local puede servir como población estadística, y cualquier subcolección, digamos los PPG de los estudiantes en una clase de matemáticas 101, puede servir como muestra de esa población.

EJEMPLO 1.9

Para ejemplo del Excel College dado en la sección 1.1, la población consiste en las respuestas de todo el estudiantado a la pregunta: "¿quién es el mejor maestro?" Como sería extremadamente difícil y llevaría mucho tiempo preguntar a cada estudiante, en vez de ello podemos preguntar a un subconjunto representativo del plantel; este subconjunto representativo de la población constituye una muestra, la información de la muestra puede usarse para **estimar** quién es el mejor maestro del Excel College.

APLICACIÓN 1.1

Un fabricante de calentadores de petróleo quiere determinar si los consumidores están satisfechos con la hechura de sus aparatos; con ese propósito localiza a 5,000 de sus 200,000 clientes y les pregunta: "¿está satisfecho con la hechura del calentador que compró?" Identificar la población y la muestra para este caso.

Solución: La población es la colección hipotética de respuestas de los 200,000 clientes; no hemos preguntado a toda la población pero esperamos aprender algo mediante la muestra. La muestra la constituyen las 5,000 respuestas dadas por los clientes interrogados. ■

EJEMPLO 1.10

Sin embargo, una población estadística no necesita ser real. Por ejemplo, si un investigador está interesado en los posibles precios de venta de automóviles de 1995, la información deseada no existe, pero aun cuando no esté disponible, los precios de venta de automóviles de varios años pasados, junto con la información relativa al índice inflacionario, pueden usarse para predecir los precios finales de los automóviles en 1995.

Un valor usado en estadística puede constituir un estadístico o un **parámetro,** depende de la extensión de la información. Examinaremos las definiciones siguientes en el ejemplo 1.11 y después mostraremos sus usos en las aplicaciones 1.2 y 1.3.

> Un estadístico es cualquier característica numérica de una muestra.
>
> Un parámetro es cualquier característica numérica de una población.

EJEMPLO 1.11

En un estudio realizado en 1989 por el Food Marketing Institute sobre modas de compras en supermercados, una muestra de respuestas de compradores reveló que el promedio de consumo familiar de alimentos era de 74 dólares; el valor de 74 dólares es un ejemplo de un estadístico; el estudio reveló también que por cada minuto adicional de permanencia en la tienda sobre el promedio de 80 a 90 minutos por semana, hay un gasto adicional de 1.89 dólares; las cifras 80, 90 y 1.89 dólares son también ejemplos de estadísticos. El estudio se realizó para obtener información sobre la población de todos los supermercados, sobre el promedio de consumo de alimentos en todos los consumidores y el promedio de tiempo semanal dedicado a las compras por todos ellos, además del monto adicional gastado por cada minuto extra dedicado a las compras sobre el promedio por consumidor; todos son ejemplos de parámetros desconocidos. Si un estudio de todos los clientes de supermercados revela que el promedio familiar es de 2.3 viajes al establecimiento cada semana y las respuestas de todos los clientes de supermercados comprenden a toda la población, entonces el valor 2.3 es un ejemplo de un parámetro.

APLICACIÓN 1.2

Para estimar la población de estudiantes que fuman cigarrillos en un cierto colegio, un administrador tomó una muestra de 200 estudiantes y determinó la proporción de estudiantes en la muestra que fuman cigarrillos. Identifique el parámetro y el estadístico.

Solución: El parámetro es la proporción de todos los estudiantes en el colegio que el administrador determinó que fuman cigarrillos, mientras que el estadístico es la proporción de estudiantes en la muestra de 200 que sí fuman cigarrillos. ■

APLICACIÓN 1.3

Una propina es la cantidad de dinero que sobre el total del consumo se otorga por un servicio satisfactorio. A los asistentes a 1,500 centros nocturnos se les dio un cuestionario confidencial preguntándoles cuánta propina habían dejado; los cálculos posteriores demostraron que la propina promedio fue de alrededor de 15% sobre el total del consumo. ¿Es parámetro o estadístico 15%? Explique su respuesta.

Solución: Si sólo están en estudio los 1,500 establecimientos, entonces la información sobre las propinas de esos establecimientos constituye la población y 15% del consumo es el parámetro; sin embargo, si el dato de las propinas de los 1,500 establecimientos forma una muestra de una población mayor de datos de propinas, entonces 15% del consumo es un estadístico. ■

GRUPO DE EJERCICIOS 1.2

1. En un intento por reducir el número de accidentes en carretera, el estado de Maryland ha llevado a cabo una campaña que se propone reducir el número de corredores y de conductores que sufren los efectos del alcohol. Un investigador que está interesado en determinar hasta qué punto el alcohol es un factor que contribuye a las muertes en carretera en el estado de Maryland, obtuvo información del mes de junio de cinco de las 22 patrullas de caminos del estado.
   a) ¿Cuál es la población de interés para el investigador?
   b) Describa la muestra.
   c) ¿Cómo puede usar el investigador la información muestral para estimar hasta qué punto el alcohol es un factor que contribuye a las muertes en carretera dentro del estado de Maryland?
2. El problema médico del síndrome de inmunodeficiencia adquirida (SIDA), ha creado altos niveles de ansiedad y preocupación entre el público. Los puntajes revelan que 71% de todos los casos de SIDA en Estados Unidos se han dado entre hombres homosexuales o bisexuales, y que alrededor del 18% han ocurrido entre los usuarios de drogas intravenosas; mucha gente se pregunta sobre la posibilidad de contraer el SIDA a través de una transfusión sanguínea. Aunque la sangre para transfusión se analiza respecto al SIDA, un investigador médico quiere estudiar los registros clínicos de 50 hospitales localizados en ciudades de todo Estados Unidos para determinar la cantidad de casos de SIDA que se ha comprobado se deben a transfusión sanguínea.
   a) ¿Cuál es la población de interés para el investigador?
   b) Describa la muestra.
3. Un doctor afirmó recientemente que una cucharada diaria de aceite de hígado de bacalao puede curar la artritis y un investigador está interesado en probar la afirmación. Se usan dos grupos, cada uno con 50% de pacientes artríticos, y sólo a los pacientes de uno de los grupos se les administra una cucharada diaria de aceite de hígado de bacalao durante un año, después del cual todos los sujetos de ambos grupos se examinarán respecto a los síntomas de la artritis.
   a) ¿Cuáles son las dos poblaciones de interés?
   b) Describa las dos muestras.
4. Desde 1971 hasta principios de 1985, la National Highway Traffic Safety Association (NHTSA) ha atribuido al menos 207 muertes a vehículos fabricados por la Ford Motor Company que inesperadamente se echan para atrás y sobre la gente; también se han reportado 4,597 demandas a resultas de movimientos inesperados de reversa en los vehículos Ford; hacia junio de 1980, la NHTSA había recibido más de 23,000 reportes de coches de la Ford que tenían fallas en el engranaje o en el arranque. El gobierno, en lugar de emitir una orden respecto a los vehículos afectados, negoció un acuerdo con la compañía Ford en un intento de prevenir futuras demandas o muertes. La empresa convino en mandar avisos junto con calcomanías preventivas para colocarse en el tablero de los vehículos, a los propietarios de los vehículos afectados, unos 23 millones de coches y camiones; a mediados de 1981, el Center for Auto Safety verificó 700 vehículos marca Ford en cuatro ciudades para asegurarse de que las calcomanías preventivas habían sido pegadas en los tableros, pero sólo 7% de los carros tenían entonces adherida dicha calcomanía.[6]
   a) Identifique la población de interés para el Center for Auto Safety en este ejemplo.
   b) Describa la muestra.
   c) ¿Diría usted que la campaña de la calcomanía tuvo éxito en reducir movimientos de reversa inesperados en los vehículos Ford? Explique.
   d) Identifique una población de interés para la NHTSA.
5. Se realizó un estudio de seis meses para determinar si el estrés y el estado de ánimo están ligados a la presencia de ciertas células del sistema inmunológico; en él participaron 36 personas con distintos niveles de estrés y dicho estudio requería examinar muestras de sangre, tomadas a intervalos regulares, respecto de cambios en el número de células auxiliares y de tipo T que regulan las funciones inmunológicas. Los resultados revelaron que el aumento en los niveles de estrés parecía estar directamente relacionado con una disminución de células tipo T y ataques de herpes.[7]
   a) Identifique a la población de interés.
   b) Describa la muestra.

6. Un censo completo del plantel de estudiantes de una universidad reveló que el número de estudiantes con 50 años de edad o mayores era de 515. ¿Este número 515 es un estadístico o un parámetro?
7. Se hizo una encuesta telefónica a 100 familias de una comunidad a fin de detectar ciudadanos interesados en pagar mayores impuestos para mejorar la calidad de la educación pública. La encuesta reveló que un 37% sí pagaría mayores impuestos para lograr tal fin. ¿37% es un estadístico o un parámetro?

SECCIÓN 1.3 Estadística descriptiva e inferencial

Los procedimientos y análisis que aparecen en estadística caen en dos categorías generales, descriptiva e inferencial, dependiendo del propósito del estudio.

> La **estadística descriptiva** comprende aquellos métodos usados para organizar y describir la información recabada.

Estos métodos se usan para analizar la información y desplegarla en forma gráfica tal, que permita interpretaciones con significado. Los métodos de la estadística descriptiva nos ayudan a describir el mundo en torno nuestro. Usamos estadística descriptiva cuando recolectamos información: como la producción promedio de trigo por acre en una cierta región agrícola, el número de personas con distintos niveles de ingresos, o el promedio de puntos obtenidos por un equipo de futbol americano durante el primer cuarto de juego. Esperamos saber cómo son las cosas mediante la estadística descriptiva.

EJEMPLO 1.12

Las situaciones siguientes utilizan estadística descriptiva.

1. Un jugador de boliche quiere conocer su promedio de anotaciones en los pasados 12 juegos.
2. Una mujer dedicada a la política desea saber el porcentaje exacto de votos que obtuvo en la última elección.
3. María quiere describir la variación que hay en las cinco calificaciones de exámenes que comprenden la primera cuarta parte de su curso de cálculo.
4. Al señor Smith le interesa determinar el promedio semanal total de sus gastos en comestibles durante los últimos tres meses.

Por otro lado, la estadística inferencial involucra teoría de probabilidad.

> La **estadística inferencial** comprende aquellos métodos y técnicas usados para hacer generalizaciones, predicciones o estimaciones sobre poblaciones a partir de una muestra.

La habilidad para hacer generalizaciones sobre la población a partir de una muestra es un aspecto importante en estadística. Rara vez tenemos la información completa que necesitamos para llegar a la verdad absoluta sobre algún evento total. Las decisiones e inferencias se basan en información limitada e incompleta; los métodos de la estadística inferencial y el conocimiento obtenido al usarlos, nos permiten utilizar información disponible limitada para entender y tratar con las incertidumbres de este mundo cambiante y azaroso. Por ejemplo, podemos predecir el trigo que se producirá el año entrante si nos basamos en las producciones de los años próximos pasados; estimar el crecimiento del ingreso promedio de un periodo de cinco años con base en el conocimiento del promedio de ingresos en el pasado y de otros estadísticos descriptivos; también podríamos tratar de predecir el total de puntos alcanzados durante la temporada por un equipo particular de futbol, si conocemos los ya obtenidos en los primeros siete juegos; con estadística inferencial establecemos cómo serán las cosas *probablemente* o a veces sólo cómo *pueden ser*. Usando métodos de probabilidad, intentaremos medir el grado de incertidumbre asociado con una inferencia.

EJEMPLO 1.13

Las situaciones siguientes, que son paralelas a las situaciones descriptivas dadas en el ejemplo 1.12, requieren estadística inferencial.

1. Un jugador de boliche quiere estimar la oportunidad que tiene de ganar un torneo próximo con base en su promedio de la temporada actual y en los promedios de sus futuros contrincantes.
2. Con base en una encuesta de opinión, a un político le gustaría calcular la oportunidad de reelegirse en las próximas elecciones.
3. Con apoyo en la variación de sus calificaciones de exámenes en la primera cuarta parte del curso de cálculo, María desea predecir la que tendrá en las calificaciones de exámenes de la segunda cuarta parte del curso de cálculo.
4. El señor Smith desea calcular el monto semanal promedio que gastará en comestibles el año próximo, tomando como base sus facturas de comestibles del último año.

GRUPO DE EJERCICIOS 1.3

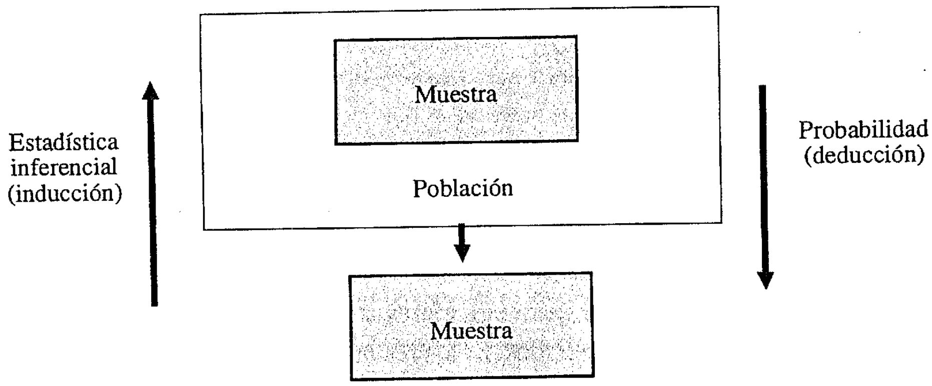
1. El señor Jackson, candidato a alcalde de un pueblo pequeño, quiere determinar si debe hacer una campaña más fuerte contra su oponente; para ello entrevistará a 500 de los 1,500 votantes registrados. Si los resultados indican que tiene 25% más votos que su oponente, no intensificará sus esfuerzos de campaña contra su rival.
   a) Identifique la población.
   b) ¿Cuál es la muestra?
   c) Señale un estadístico.
   d) Ubique un parámetro.
   e) ¿Qué haría el señor Jackson si tuviera el 65% de los votos de la muestra?
2. Dé un ejemplo no mencionado en el texto de cada uno de los siguientes conceptos:
   a) población
   b) muestra
   c) estadístico
   d) parámetro

3. Un agente independiente de mercado realizó un estudio de precios de alimentos en cuatro de diez establecimientos expendedores de comestibles en una ciudad pequeña. Los precios que siguen corresponden a bolsas de azúcar de cinco libras: 1.25 dólares, 1.18 dólares, 1.20 dólares y 1.30 dólares; el agente hizo las cuatro afirmaciones que anotamos abajo. ¿Cuáles se obtuvieron usando estadística inferencial y cuáles con estadística descriptiva? Explique sus respuestas.
   a) El precio más alto cobrado en el pueblo es 1.30 dólares.
   b) Dos tiendas cobran más de 1.20 dólares por una bolsa de 5 libras de azúcar.
   c) La cuarta parte de las tiendas cobran más de 0.25 dólares por una libra de azúcar.
   d) Los precios en todos los mercados para una bolsa de 5 libras de azúcar varían entre 1.18 y 1.30 dólares.
4. Un anuncio comercial afirma: "cuatro de cada cinco médicos recomiendan el preparado A". ¿Cree usted que esta conclusión proviene de una muestra o de una población? Explique.
5. Se estableció que el costo promedio de los textos escolares en un colegio pequeño durante el último semestre fue de 135 dólares, con base en una inscripción de 1,200 estudiantes. Como un trabajo de clase en el colegio, un grupo de estadística encuestó a 25 estudiantes para determinar el promedio del costo de un libro de texto en el último semestre y se concluyó que fue de 152.25 dólares. Identifique:
   a) la población.
   b) la muestra.
   c) los parámetros.
   d) dos estadísticos y
   e) ¿qué podría concluir el grupo de estadística si el costo promedio de un libro para la muestra de 25 estudiantes fuera de 400 dólares?
6. Clasifique la naturaleza de cada una de las afirmaciones siguientes como inferencial o descriptiva; también diga cuáles son las hipótesis en que basa su respuesta.
   a) Una familia de cinco o más miembros tiene un gasto semanal promedio en compras de comestibles de 109 dólares.
   b) El 66% de todas las compras de comestibles no son planeadas.
   c) En 1978 hubo un total de 11,767 artículos comestibles comprados.
   d) Cada año sale del mercado un 80% de artículos comestibles recientes.
   e) El número de artículos comestibles diferentes existentes en un establecimiento común en 1989 era de 26,430.

SECCIÓN 1.4 *Inferencias y deducciones*

El estudio de la estadística requiere tanto inducción como deducción (véase la figura 1.2).

FIGURA 1.2
Inducción *versus* deducción



La **inducción** consiste en razonar desde los ejemplos específicos al caso general.

La **deducción** consiste en razonar desde el caso general hasta los ejemplos específicos.

Cuando hacemos una generalización sobre un parámetro poblacional basándonos en la información derivada de una muestra, estamos usando inducción.

EJEMPLO 1.14

Si después de probar un cierto número de uvas de un platón llegamos a la generalización de que todas las uvas contenidas ahí están agrias, estamos usando un razonamiento inductivo; la generalización de que todas las uvas del platón están agrias es un ejemplo de inferencia.

Una **inferencia** es una generalización obtenida mediante inducción.

En estadística inferencial, las inferencias se hacen típicamente sobre un parámetro poblacional empleando sólo una muestra específica, en lugar de muchas muestras como uno esperaría por el uso de la inducción. Cuando se hace esto, debe tenerse gran cuidado de asegurarse que la muestra sea verdaderamente representativa de la aplicación.

Al adjudicar propiedades de una población a las muestras, estamos deduciendo; las deducciones requerirán probabilidad, el estudio de la incertidumbre. Estudiaremos probabilidad en los capítulos del 5 al 7 y estadística inferencial del 8 al 15.

EJEMPLO 1.15

(Deducción y probabilidad). Supongamos que 1,000 automóviles son de fabricación reciente, que 5% de ellos tienen un componente direccional defectuoso y que un comerciante local tiene una muestra de diez de esos coches. Como una aplicación de la probabilidad, podemos intentar determinar la posibilidad de que al menos dos de esos diez automóviles tengan componentes defectuosos en la dirección. Como la muestra es un subconjunto de la población, podríamos esperar que 5% represente la posibilidad de que un cierto coche de la muestra esté defectuoso. En el capítulo 5 aprenderemos a determinar la probabilidad de que al menos dos de las diez unidades tengan componentes defectuosos.

EJEMPLO 1.16

(Estadística inferencial). Suponga que 1,000 autos son de fabricación reciente y que no se sabe cuántos tienen defectos en la dirección. Para estimar el porcentaje de autos con sistema direccional defectuoso en esta población, inspeccionaremos una muestra de diez unidades. Si se encuentra que dos de ellas tienen sistema direccional defectuoso, podremos inferir usando la inducción que 20%, es decir 200 de los 1,000 automóviles, tienen fallas en el sistema direccional; la proporción de sistemas direccionales defectuosos en la muestra es un ejemplo de un estadístico; su valor es 0.20. El porcentaje de automóviles, en esta población, que tienen el sistema direccional defectuoso es un ejemplo de un parámetro. En el capítulo 9 aprenderemos a usar la estadística para estimar parámetros desconocidos.

La **confiabilidad** de una inferencia es un aspecto fundamental de la estadística inferencial. Una inferencia es *confiable* si se puede depender de ella con una cierta seguridad, ya que no puede describirse con exactitud una característica de una población si la inferencia no es confiable. La teoría de la probabilidad debe usarse al determinar la confiabilidad de una inferencia. Estudiaremos este tema en los capítulos 9 y 10, y la probabilidad en los capítulos 5 al 7.

SECCIÓN 1.5 El papel de la computadora en la estadística

Con la introducción de las microcomputadoras, el trabajo pesado de cálculos asociado con un gran número de datos y con análisis complicados, ha sido relegado a las computadoras. Como las manipulaciones tediosas de los datos se hacen con la computadora, el usuario puede concentrarse en el análisis de los resultados. Hay muchos programas computacionales amigables disponibles en el mercado que permiten a los estudiantes y a los especialistas realizar los cálculos estadísticos tediosos con poca o ninguna dificultad. Algunos de los programas más usuales incluyen MINITAB, SPSSx, SAS, BMDP y SYSTAT. Todos permiten al usuario comunicarse con el sistema de la computadora mediante comandos sencillos.

En este texto hemos escogido usar MINITAB para ilustrar las aplicaciones estadísticas que utilizan computadora. Este programa, desarrollado originalmente en la Pennsylvania State University como una herramienta para enseñar estadística, hoy en día se usa ampliamente tanto en la enseñanza como en la investigación en todo Estados Unidos; puede instalarse en unidades centrales de procesamiento, así como en mini y microcomputadoras, ya que ofrece una gran capacidad de cálculo tanto para el estudiante como para el investigador de la estadística.

MINITAB es un programa interactivo operado mediante órdenes; una vez cargado en la computadora, el usuario se comunica con el sistema y da órdenes que son ejecutadas de inmediato. Es muy fácil de usar. La aparición del símbolo MTB > en la pantalla del monitor le informa al usuario que el sistema está listo para aceptar una orden de un dispositivo de entrada, como el teclado. Después de que los datos se han introducido en el programa, el usuario da una orden oprimiendo la tecla de entrada y el sistema proporciona inmediatamente el valor deseado.

EJEMPLO 1.17

La pantalla 1.1 muestra las órdenes de MINITAB y las correspondientes respuestas usadas para determinar la media (promedio) de las cifras 34, 68, 39, 21 y 42. Después del símbolo del sistema MYB >, el usuario escribe SET C1 y presiona la tecla de entrada (la computadora no se entera de la orden sino hasta que la tecla de entrada es oprimida). Esta acción informa a MINITAB que debe crear una columna, identificada en la memoria como C1, que contendrá los datos. El sistema responderá con el símbolo DATA >; aquí el usuario escribe las cifras: 34, 68, 39, 21 y 42, y oprime la tecla de entrada para registrarlos en la memoria del sistema, donde se usan espacios en vez de comas para separar los números; la computadora responde otra vez con el símbolo DATA > y como no habrá más información, el usuario escribe END para entrar en el sistema que responde ahora con el símbolo del programa MTB >, para informar que está esperando otra orden del usuario, quien escribirá la orden MEAN C1 y oprimirá en seguida la tecla de entrada para pedir el promedio de los números en la columna etiquetada con C1; la computadora responde inmediatamente con MEAN = 40.800 y otro símbolo MTB >.

Note que en este ejemplo las instrucciones SET C1 y MEAN C1 son las **órdenes** dadas por el usuario. Al final de cada orden el usuario debe oprimir la tecla de entrada, *enter* o *return* para que ingresen las órdenes en el sistema de la computadora; este modo de comunicación con el sistema de la computadora es la razón de que a MINITAB se le llame **sistema operado mediante órdenes**, en oposición a un sistema operado por menú, donde la selección en el menú da lugar a una acción particular de la computadora.

Pantalla 1.1

```
MTB >SET C1
DATA > 34 68 39 21 42
DATA > END
MTB > MEAN C1
  MEAN = 40.800
MTB >
```

Las computadoras son herramientas muy eficaces cuando se necesita procesar una gran cantidad de datos, realizar alguna tarea en forma repetitiva o cuando los resultados deben analizarse rápida y cuidadosamente. Los problemas que se encuentran en este libro utilizarán conjuntos de datos relativamente pequeños; pero aun así algunos de los cálculos pueden resultar tediosos en una calculadora; es deseable que se entiendan los cálculos manuales hechos en el texto y realizar cada una de sus etapas en forma sucesiva al resolver muchos de los ejercicios. Una vez hecho esto, usted comprenderá los usos y las limitaciones de cada procedimiento. También será capaz de entender e interpretar los resultados que ellos proporcionan; si usted tiene acceso a una computadora, puede usar un paquete estadístico, como MINITAB, para efectuar procedimientos similares en el futuro.

Muchos de los conjuntos de datos, en las aplicaciones prácticas de este texto, se usan para mostrar en pantalla el uso de MINITAB y cumplir al menos cinco propósitos:

1. Ilustrar la sencillez del uso de MINITAB.
2. Visualizar las órdenes que debe proporcionar el usuario para lograr los resultados estadísticos deseados.
3. Conocer el formato y la notación usados en las respuestas de la computadora.
4. Saber la magnitud de la potencia de cálculo estadístico disponible con MINITAB y con otros paquetes estadísticos.
5. Enseñar el uso de MINITAB en el desarrollo de algunas tareas estadísticas.

REPASO DEL CAPÍTULO

■ TÉRMINOS IMPORTANTES ■

Los términos siguientes, pertenecientes al capítulo, se han mezclado para proporcionar una práctica más eficaz. Para cada uno dé una definición con sus propias palabras; después verifique sus respuestas con las proporcionadas en el texto.

orden	estimar	sistema operado por órdenes
deducción	estadística descriptiva	inducción
inferencia	estadística inferencial	población
muestra	estadística	estadística
MINITAB	confiabilidad	parámetro

2 Estadística descriptiva: organización de datos

DESCRIPCIÓN

OBJETIVOS DEL CAPÍTULO

En este capítulo estudiaremos:

- *Datos.*
- *Dos tipos generales de datos.*
- *Datos clasificados según el tipo de escala de medición usado.*
- *Cómo organizar y resumir los datos mediante tablas.*
- *Cómo mostrar los datos mediante distintos tipos de gráficas.*

MOTIVADOR 2

Se realizó un estudio dividido en dos partes con dos muestras a fin de medir el sentir de los usuarios de servicios.[8] Una muestra consistió en más de 1,300 usuarios de una vasta región que incluía Estados Unidos, Canadá y Gran Bretaña; la otra muestra constó de cerca de 900 personas prestadoras de servicios representantes de nueve organizaciones diferentes. La primera parte de los cuestionarios del estudio se construyó sobre 17 aspectos del servicio al cliente y los entrevistados se refirieron a cada aspecto en dos niveles: importancia y pericia. Los niveles de importancia preguntaban a los entrevistados: "¿Qué tan importante considera usted que es esta dimensión para un servicio eficiente al cliente?" Las respuestas posibles iban desde 5, extremadamente importantes, hasta uno, no importante. La encuesta sobre pericia preguntaba a los entrevistados: "¿Qué tan bien cree usted que el personal de prestación de servicios al cliente aprovecha este aspecto cuando interactúa?" En donde las respuestas iban desde 5, siempre lo aprovechan bien, hasta uno, nunca lo aprovechan. La segunda parte del estudio se diseñó para medir el impacto del servicio al cliente sobre la decisión de los consumidores de repetir el negocio; una pregunta era: "¿Cuánto influye un buen servicio en su decisión de volver a tratar con esta organización?" En la escala de cinco puntos, las respuestas iban desde 5, tiene un gran efecto, hasta uno, poco o ningún efecto. Una segunda pregunta a los entrevistados era: "¿Qué tan seguido comenta usted con otras personas si recibió un servicio a clientes excelente o malo?" Las categorías de respuestas fueron "nunca", "ocasionalmente" y "con regularidad".

Los resultados indicaron que el sentir de los clientes sobre la calidad del servicio difiere del sentir de los prestadores de servicios. La tabla 2.1 lista los puntajes medios del sentir de los clientes y del personal de servicios al cliente para 5 de los 17 aspectos del servicio al cliente; las figuras 2.1 y 2.2 muestran el impacto del servicio al cliente en los negocios. En este capítulo, conoceremos distintas clases de datos y cómo organizarlos y presentarlos usando tablas y también gráficas, como aquí.

TABLA 2.1
Puntajes de clientes y personal de prestación de servicios

Aspecto	Muestra de clientes: Importancia	Muestra de clientes: Provecho	Muestra de prestadores de servicios: Importancia	Muestra de prestadores de servicios: Provecho
Comunicación	4.05	2.95	4.55	3.64
Sensibilidad del cliente	3.92	2.67	4.38	3.56
Capacidad de decisión	3.84	2.74	4.34	3.53
Conocimiento del trabajo	4.10	2.96	4.54	3.56
Motivación para servir a los clientes	3.97	2.73	4.27	3.32

FIGURA 2.1
Efectos de un buen servicio

	Gran efecto	Efecto moderado	Poco o ningún efecto
Muestra de contacto	97%	2%	1%
personal con clientes	83%	13%	4%

FIGURA 2.2
Comentario verbal sobre el servicio

	Excelente servicio	Servicio malo
Muestra de contacto	38%	75%
personal con clientes	57%	65%

Panorama del capítulo

El aspecto fundamental de la estadística es la información que contiene; sin información que recabar, organizar, analizar e interpretar, no habría razón para usar o estudiar estadística; a la información usada en estadística se le llama **datos.** Para que sea útil dicha información en la toma de decisiones, debe organizarse y mostrarse apropiadamente. El tipo de datos indicará los métodos a usar en su análisis. Comenzamos este capítulo con un estudio de los distintos tipos de datos.

SECCIÓN 2.1 *Datos: los bloques de construcción de la estadística*

Cabe distinguir entre el término *"datos"* y *"dato"*. *Dato* es una porción de información. *Datos* es sinónimo de *muestra*. Los datos pueden clasificarse en dos categorías generales, cuantitativos y cualitativos.

> Los **datos cuantitativos** se refieren a información numérica, como cuánto o cuántos, y se miden en una escala numérica.

EJEMPLO 2.1

Ejemplos de datos cuantitativos son el peso dado en kilos, la edad en años, la longitud en centímetros, el volumen en metros cúbicos, el precio en pesos.

> Los **datos cualitativos** representan categoría o atributos que pueden clasificarse según un criterio o cualidad.

EJEMPLO 2.2

Ejemplos de datos cualitativos son el sexo: hombre, mujer; el color: rojo, verde, azul; la religión: católica, protestante, judía; el tipo de sangre: A, B, AB, O; la marca favorita de coche: Ford, Chevrolet; o una marca de computadora: IBM, Kaypro, Zenith, Compaq.

Los datos consistentes en números se pueden clasificar en términos cuantitativos o cualitativos, dependiendo de cómo se usen. Si se usan como una etiqueta para propósitos de identificación, son cualitativos; en otro caso, son cuantitativos (véase el ejemplo 2.3); sin embargo, algunas mediciones pueden hacerse mediante escalas cuantitativas o cualitativas, como en el ejemplo 2.4.

EJEMPLO 2.3

Si un número de serie de un radio se usa para identificar el número de radios fabricados hasta ese momento, será una medida cuantitativa, pero si se usa sólo para propósitos de identificación, es un elemento de información cualitativa.

EJEMPLO 2.4

Si la estatura de un individuo se mide en pies y pulgadas, entonces la información es cuantitativa; pero si se mide como bajo, medio o alto, es cualitativa. Además, la estatura puede medirse usando datos cuantitativos, pies y pulgadas, pero representarse por datos cualitativos, bajo, medio o alto.

Los datos cuantitativos pueden clasificarse como discretos o continuos.

> Los datos obtenidos de un proceso de conteo son **datos discretos.**

> Los datos obtenidos de un proceso de medición, donde la característica que se mide puede tomar cualquier valor numérico en un intervalo, son **datos continuos.**

EJEMPLO 2.5

Ejemplos de datos discretos son el número de niños en una familia, la cantidad de coches en un estacionamiento, el salario de un individuo, el conjunto de personas en una fila; el número de pulsaciones del corazón por minuto y la presión sanguínea, medida con instrumento digital, también son ejemplos de datos discretos; sin embargo, la velocidad de un coche en millas por hora, no da lugar a datos discretos porque puede llegar a ser cualquier cifra, desde 0 millas por hora hasta la velocidad máxima del coche.

EJEMPLO 2.6

Los datos continuos no se pueden contar. El peso en kilogramos, la estatura en metros, el tiempo en minutos y la distancia en kilómetros, son ejemplos de datos continuos; la presión barométrica y el tiempo que tarda usted en llegar a la escuela son ejemplos de datos continuos, pero el número de personas en una playa un fin de semana concurrido no sería continuo porque es una cantidad que sí se puede contar.

Cualquier proceso de medición que proporcione datos continuos está limitado por la precisión del instrumento de medición utilizado. Por ejemplo, si un instrumento es preciso hasta los décimos de pulgada y se usa para medir la altura de un individuo, entonces hay sólo un número finito de medidas posibles que pueden obtenerse y las estaturas así medidas se redondearán hasta décimos de pulgada. Una medida de este tipo representa una aproximación a la medida real. Las medidas reales son teóricas y representan datos continuos, mientras que las medidas aproximadas son datos discretos porque hay sólo un número finito de formas de medir algo con un instrumento de precisión dado. En realidad, todas las medidas físicas son discretas; la restricción de una precisión limitada se aplica sólo a los instrumentos de medición, no a los datos: éstos son de naturaleza continua y se redondea su valor de acuerdo con la precisión de los instrumentos usados para obtenerlos.

Nuestro propósito principal al analizar datos es efectuar una interpretación que tenga sentido. Como regla general, la cantidad de información contenida en los datos depende de su naturaleza; las dicotomías cuantitativo-cualitativo y discreto-continuo no siempre son adecuadas para la clasificación de datos según la cantidad de información que contienen; los datos también se pueden clasificar según la escala de medición o el procedimiento que los generó.

Considere el dígito 4 en las siguientes situaciones:

a) El número de la camiseta de futbol de Juan es el 4.

b) Juan está en el 4° grado.

c) Juan registró la temperatura como 4° Celsius.

d) Juan cultivó un pepino que midió 4 pulgadas de largo.

Estas situaciones representan cuatro niveles distintos de información, resultantes del uso de escalas diferentes de medición. La medida en la situación del inciso *a*, por ejemplo, se usa sólo para identificar o clasificar a Juan como el jugador de futbol número 4; el 4° grado en el caso de la situación del inciso *b*, también es una clasificación, pero da más información porque nos da el nivel del grado, más avanzado que el 3° grado y menos que el 5°, aunque qué tanto más o menos es algo que no podemos medir.

En la situación del inciso *c*, de nuevo vemos niveles de comparación, pues 4 indica que la temperatura es más alta que una temperatura de 2° Celsius y más baja que una temperatura de 7° Celsius. Es más, una temperatura de 4° Celsius es 1.5° más alta que una de 2.5°, porque la diferencia entre 4° y 2.5° es 1.5°. Sin embargo, una temperatura de 4° Celsius no es el doble de caliente que una temperatura de 2° Celsius.

Finalmente, en el inciso *d*, la medida 4 identifica al pepino como miembro de una clase de pepinos que miden 4 pulgadas de largo; sabemos también que este pepino es más largo que uno de 3 pulgadas de longitud, que excede de 1 pulgada a uno de 3 y que es el doble de largo de un pepino de 2 pulgadas de longitud.

Las situaciones vistas de *a* a *d* son representativas de cuatro tipo de escalas de medición que discutiremos con detalle porque el tipo de escala de medida usada determina la cantidad de información contenida en cualquier dato proporcionado.

Cuatro tipos de escalas de medición usados en estadística

1. Nominal
2. Ordinal
3. De intervalo
4. De razón

Escala nominal

Existen escalas nominales tanto para los datos cuantitativos como para los cualitativos. Una *escala nominal para datos numéricos* asigna números a las categorías para distinguirlas como en el ejemplo 2.7. Una *escala nominal para datos cualitativos*, como en el ejemplo 2.8, es un agrupamiento no ordenado de los datos en categorías discretas, donde cada dato puede incluirse solamente en uno de los grupos. Las escalas nominales se usan principalmente con propósitos de identificación o de clasificación.

EJEMPLO 2.7

Entre los datos numéricos que son nominales se incluyen los números en las camisetas deportivas, los números de código de las zonas postales, los números telefónicos y los puntajes de futbol americano, 6 puntos por un *touchdown*, 1 punto por la patada extra, 2 puntos por una escapada extra y 3 puntos por un gol de campo.

EJEMPLO 2.8

Los **datos nominales** que son cualitativos incluyen el género, la raza, el tipo de sangre y la religión.

Escala ordinal

Los datos medidos en una escala nominal ordenada de alguna manera se denominan **datos ordinales**. Una escala ordinal coloca las medidas en categorías, cada una de las cuales indica un nivel distinto respecto a un atributo que se está midiendo.

EJEMPLO 2.9

La lista de datos ordinales comprende:

1. Clasificaciones por letra: A, B, C, D y F; estos grados indican categorías de perfeccionamiento, así como los niveles alcanzados.
2. Rangos académicos: instructor, asistente de profesor, profesor asociado y profesor, donde un profesor tiene mayor rango académico que un instructor.
3. La numeración de las casas en las calles: calle Norte 421, calle Norte 423 y así sucesivamente. La casa correspondiente al domicilio calle Norte 423 se localiza entre las casas localizadas en calle Norte 421 y calle Norte 425.
4. La evaluación de un maestreo: pobre, razonable, buena y superior.
5. Los grados de la escuela: primero, segundo, tercero, etcétera.

No es posible determinar la diferencia o distancia entre los valores medidos en una escala ordinal. Aun cuando solemos codificar la letra del grado A como 4, B como 3, C como 2, D como 1 y F como 0, no diríamos, por ejemplo, que una A es el doble de buena que una C o que un estudiante con A sabe el doble de un estudiante con C; todo lo que podemos decir es que la calificación A es mejor o de un grado superior a la C, ya que una escala ordinal no admite unidad de distancia.

Escala de intervalo

Los datos medidos en una escala ordinal para los cuales pueden calcularse las distancias entre valores, se llaman **datos de intervalo.** La distancia entre dos valores es importante y los datos de intervalo son cuantitativos por necesidad; una escala de intervalo no siempre tiene un punto cero, un punto que indique la ausencia de lo que se quiere medir.

EJEMPLO 2.10

Las listas de datos de intervalo comprenden:

1. **Puntajes en las pruebas de inteligencia:** un puntaje de inteligencia de 110 es cinco puntos superior a uno de 105 (datos ordinales). En este caso, no sólo podemos decir que un puntaje de 110 es superior a uno de 105, sino que también podemos decir que es cinco puntos más alto; pero no podemos decir que una persona con un puntaje de inteligencia de 180 es doblemente lista que una persona que tiene uno de 90, y una determinada diferencia entre dos puntajes de inteligencia no siempre tiene el mismo significado: por ejemplo, las diferencias entre 100 y 90 y entre 150 y 140, pueden tener interpretaciones distintas aunque ambas sean iguales a 10. Aunque una persona con 140 es más inteligente de acuerdo con la prueba de inteligencia que una persona con 100, no podemos decir que quien tiene un puntaje de 150 es tanto más inteligente que una persona con 140, o que lo es una persona con un cociente de inteligencia de 100 respecto a una persona con uno de 90.
2. **Temperaturas Celsius.** Una temperatura de 80° es 40° más caliente que una temperatura de 40°, pero no es correcto decir que 80° es el doble de caliente que 40°. Nótese también que una temperatura de 0°no representa la ausencia total de calor. El punto cero en la escala de temperatura Celsius fue escogido arbitrariamente como el punto de congelamiento e indica que está presente algo de calor. (Teóricamente, –273°C representa el mínimo absoluto de temperatura, la temperatura en la que las moléculas de una sustancia se mueven a una velocidad casi de cero).
3. **Fechas**. Ronald Reagan fue investido como el 40º presidente de Estados Unidos en 1981, 192 años después de George Washington (1789). Podemos especificar la distancia entre estos dos sucesos ordenados, 192 años, pero si existiera el año cero, no representaría la ausencia de tiempo.

Escala de razón

Los datos medidos en una escala de intervalo con un punto cero que significa "ninguno", se llaman **datos de razón.** Con datos medidos en una escala de razón, podemos determinar cuántas veces es mayor una medida que otra. Como el punto cero de la escala de temperatura Celsius no representa la ausencia completa de calor, la escala Celsius no es una escala de razón; por otra parte, la escala Kelvin de temperatura, donde 0 K corresponde a –273° C, es un ejemplo de una escala de razón de temperatura.

EJEMPLO 2.11

Las escalas de razón incluyen escalas usadas comúnmente para medir unidades como pies, libras, dólares y centímetros; los resultados de contar objetos también son datos de razón; diez manzanas es el doble que cinco manzanas. Con una escala de razón, una persona que pesa 200 libras siempre pesará el doble que una persona de 100 libras, aunque se use otra escala de razón, como onzas, gramos o kilos.

APLICACIÓN 2.1

Suponga que se hace una encuesta a un grupo de maestros con respecto a su religión y que 15 son protestantes, 21 católicos y 7 judíos. ¿Qué tipo de datos son éstos?

Solución: La respuesta de cada profesor es protestante, católico o judío, y estas respuestas constituyen datos nominales de categorías o cualitativos; por otro lado, los números 15, 21 y 7 resultan de contar los datos cuantitativos. Las cifras obtenidas al realizar operaciones con datos, como la suma, no deben confundirse con la colección de datos. ■

Es importante ser capaz de clasificar datos de acuerdo con la escala de medida usada. Al realizar una inferencia sobre una población de interés, las técnicas usadas dependen del tipo de escala medida. Por ejemplo, si se trabaja con una muestra de datos ordinales, debe utilizarse una técnica estadística que use datos ordinales; al clasificar los datos según el tipo de escala de medida usada, el investigador puede identificar la mejor estadística para analizar los datos.

GRUPO DE EJERCICIOS 2.1

Habilidades básicas

1. Clasifique los datos siguientes en cuantitativos y cualitativos:
   a) Estaturas en pulgadas de cinco jugadores de basquetbol.
   b) Peso en onzas de doce pollitos.
   c) Clasificación étnica de 20 empleados.
   d) Números telefónicos de amigos.

2. Clasifique también como cuantitativos o cualitativos:
   a) Calificaciones numéricas PPG de los miembros de la clase elemental.
   b) Calificaciones con letra de 15 estudiantes del grupo 209 de filosofía.
   c) Número de dulces en un paquete de 70 gramos.
   d) Fechas de cumpleaños de los miembros de su familia.

3. Clasifique los datos siguientes como discretos o continuos:
   a) El número de defectos en cada unidad de un lote de 50 coches nuevos.
   b) Puntajes de matemáticas en la prueba de aptitud académica de 30 alumnos del último año de preparatoria.
   c) Distancia en yardas recorrida por un mediocampista en cada juego durante la última temporada.
   d) Peso perdido en libras por 20 personas debido a una dieta.

4. Clasifique los datos siguientes como discretos o continuos:
   a) El número de carreras anotadas en cada juego por los Piratas en la temporada de 1900.
   b) Los sueldos ganados en el último mes por 50 directores de institutos.
   c) Las temperaturas promedio diarias de los últimos 30 días.
   d) El número de granos de arena en cada una de 100 playas.

5. La tabla siguiente contiene la distribución de vehículos registrados en Excel College.

Clase	Tipo de vehículo	Cifra registrada
1	Coche	150
2	Camioneta	25
3	Motocicleta	15
4	Bicicleta	10

   a) Clasifique los datos de cada una de las tres columnas como cuantitativos.
   b) Identifique los datos de la tercera columna como discretos o continuos.
   c) Determine los datos de cada una de las tres columnas como nominales, ordinales, de intervalo o de razón.

6. El Memorial Hospital registra la información siguiente de cada uno de los pacientes:
   número de seguridad social
   fecha del último ingreso
   fecha de nacimiento
   compañía de seguros
   patrón
   dirección particular
   teléfono particular

   a) Diga si la información anterior es cuantitativa o cualitativa.
   b) Clasifique los datos de cada categoría como discretos, continuos o como ninguno de los dos.
   c) Ordene la información como nominal, ordinal, de intervalo o de razón.

7. La figura de abajo muestra una escala numérica para medir la efectividad de la enseñanza.

1	3	5
Necesita mejorarse	Efectiva y competente	Verdaderamente extraordinaria

   a) Identifique el tipo de escala de medición.
   b) Suponga que 30 estudiantes usan esta escala para evaluar a su maestro de estadística. ¿Será más fácil interpretar esos resultados que los que se obtendrían si los 30 estudiantes evaluaran a su maestro mediante una opinión escrita de respuesta libre? Explique.

8. Los estudiantes de una universidad se clasifican como de primer año, de segundo año, de penúltimo año y de último año. ¿Qué tipo de escala de medición es ésta?

9. Dé un ejemplo distinto de los mencionados en la sección de una escala ordinal para datos cuantitativos.

10. ¿Toda información numérica proporciona datos cuantitativos? ¿Por qué?

11. ¿Toda información no numérica nos ofrece datos cualitativos? Explique.

SECCIÓN 2.2 *Organización de datos mediante tablas*

El objetivo de la organización de datos es acomodar un conjunto de datos en forma útil para revelar sus características esenciales y simplificar ciertos análisis. Los datos que no están organizados se denominan **datos no agrupados.** Una manera de acomodarlos es construir un arreglo ordenado: esto es, acomodando los datos de abajo hacia arriba o al revés; si el número de datos es grande, el arreglo puede ser difícil de manejar o de comprender; por eso a menudo se usan tablas como una aproximación general a la organización de **datos no agrupados**. En esta sección estudiaremos varios tipos de tablas usadas para organizar datos; en la seccion 2.3 discutiremos medios gráficos para mostrar datos no agrupados organizados en forma tabular. El tipo de datos nominal, ordinal, de intervalo o de razón determinará la forma en que se coloquen.

La **frecuencia** de una medida o de una categoría, es el número de veces que aparecen en una colección de datos. El uso de frecuencias es más conveniente para datos cualitativos o discretos; el símbolo f se usa para denotar la frecuencia de una medida. La muestra de datos siguiente representa el número de tiros libres fallados por un equipo de basquetbol durante los últimos siete juegos:

$$7\ 2\ 8\ 4\ 2\ 7\ 2$$

El número 7 aparece con una frecuencia de $f = 2$, 2 aparece con una frecuencia de $f = 3$, 8 y 4 aparecen con una frecuencia de $f = 1$.

Existen dos tipos generales de tablas para reportar datos usando frecuencias, éstas son: **tablas de frecuencias no agrupadas** y **tablas de frecuencias agrupadas.** Ambas tablas se mencionan como **tablas de frecuencia** y estudiaremos primero las tablas de frecuencias no agrupadas.

Tablas de frecuencias no agrupadas

Los datos sobre tiros libres citados anteriormente pueden resumirse como lo muestra la tabla 2.2, donde x denota las medidas y f, la frecuencia de cada medida; la tabla 2.2 es un ejemplo de una tabla de frecuencias no agrupadas para datos discretos.

TABLA 2.2

Tabla de frecuencias de datos sobre tiros libres

x	f
2	3
4	1
7	2
8	1

APLICACIÓN 2.2

Construya una tabla de frecuencias para los datos siguientes, correspondientes al número de faltas a clases durante el periodo de otoño de 1988 para estudiantes inscritos en la materia Estadística 101.

9	8	7	8	4	3
2	1	0	5	3	2
1	1	7	3	2	8
7	6	6	4	3	2
2	0	9	4	6	9
6	9	4	3	5	7
3	2	1	4	4	2

Solución: Como paso intermedio usaremos **marcas de cuenta** para ayudar a determinar la frecuencia f de cada observación, donde x representa el número de faltas.

Número de faltas (x)	Cuenta	Frecuencia (f)
0	II	2
1	IIII	4
2	~~IIII~~ II	7
3	~~IIII~~ I	6
4	~~IIII~~ I	6
5	II	2
6	IIII	4
7	IIII	4
8	III	3
9	IIII	4
		42

En correspondencia con cada observación, hacemos una marca (I) en la columna de marcas al lado del valor observado; cuando se han hecho todas las marcas se cuentan las de cada medida x para determinar la frecuencia. Note que la suma de todas las frecuencias de una tabla de frecuencias es igual al número de datos de la colección. En este caso, la suma de las frecuencias (42) representa las 42 clases para las cuales se registraron las faltas. ■

APLICACIÓN 2.3

Cinco miembros, Jones, Smith, Baker, Brown y Thomas, de la junta directiva de una pequeña universidad, fueron nominados para presidirla y los datos siguientes muestran el resultado de la elección; construya una tabla de frecuencias para ellos.

Jones	Jones	Smith	Smith	Jones	Thomas
Smith	Baker	Baker	Jones	Thomas	Jones
Smith	Smith	Smith	Brown	Brown	Jones
Smith	Thomas	Smith	Brown	Brown	Brown

Solución: La tabla de frecuencias es como sigue:

Miembro de la junta	Frecuencia (f)
Baker	2
Brown	5
Jones	6
Smith	8
Thomas	3

■

Tablas de frecuencias agrupadas

Las tablas de frecuencias como la tabla 2.2 se denominan apropiadamente **tablas de frecuencias no agrupadas** porque cada medida tiene la frecuencia correspondiente. Una **tabla de frecuencias agrupadas,** en contraste, presenta las frecuencias de acuerdo con grupos o clases de medidas. Las tablas de frecuencias agrupadas se usan comúnmente para resumir grandes cantidades de datos continuos que contienen relativamente pocas repeticiones; tales resúmenes facilitan ciertos cálculos estadísticos y presentaciones gráficas cuando no se usa la computadora: para usar una tabla de frecuencias agrupadas a fin de resumir los datos, éstos deben medirse al menos con una escala de intervalo, y para cantidades grandes de datos que no se midan con al menos una escala de intervalo, debe usarse una tabla de frecuencias no agrupadas.

Supongamos que el Memorial Hospital quiere saber si su servicio en la sala de emergencias es adecuado. Para empezar el estudio, el gerente del departamento correspondiente registra el número de personas que ocupan la sala de emergencias cada día durante un periodo de 12 días, con los resultados siguientes:

Día	1	2	3	4	5	6	7	8	9	10	11	12
Núm. de pacientes	7	43	8	22	13	28	36	18	23	21	15	52

Para simplificar los datos, el gerente construye seis agrupamientos o clases: la primera clase representa de 1 a 10 pacientes; la segunda, de 11 a 20; la tercera, de 21 a 30 y así sucesivamente. A partir de esta clasificación, prepara una tabla de frecuencias agrupadas (tabla 2.3) para mostrar qué tan a menudo, a lo largo de los doce días, cae en cada grupo el número de pacientes.

TABLA 2.3

Tabla de frecuencias agrupadas para los datos de la sala de emergencias

Clase	Frecuencia (f)
1-10	2
11-20	3
21-30	4
31-40	1
41-50	1
51-60	1

Las clases de frecuencias agrupadas poseen lo que se llama **límites de clase.** Para la clase 1-10, a 1 se le llama *límite inferior de clase*, y a 10, *límite superior de clase*. Existen dos medidas que caen entre 1 y 10, inclusive; tres medidas que caen entre 11 y 20, inclusive; cuatro medidas que caen entre 21 y 30, inclusive; una medida que cae entre 31 y 40, inclusive y así sucesivamente. La distancia entre cualquiera de dos límites superiores consecutivos o entre cualquiera de dos límites inferiores consecutivos es llamada **amplitud de clase.** La amplitud de cada clase en la tabla 2.3 es 10. La distancia entre el límite superior de la primera clase y el límite superior de la segunda clase es $20 - 10 = 10$. Cada clase en una tabla de frecuencia tiene límites de clase teóricos llamados **fronteras de clase;** al límite superior teórico se le llama *frontera superior* y al límite inferior teórico de clase se le llama *frontera inferior*. La frontera inferior para la primera clase es 0.5 y la frontera superior para esa misma clase es 10.5. Para esta tabla de frecuencias, la frontera superior de cada clase se encuentra sumando 0.5 al límite superior, y la frontera inferior de cada clase se encuentra restando 0.5 del límite inferior de cada clase.

Note que cuando se examina una tabla de frecuencias agrupadas sin los datos no agrupados, esto es, antes del procesamiento estadístico, no conocemos las medidas individuales; por ejemplo, en la tabla 2.3 vemos que dos medidas caen en la clase 1 a 10, pero no sabemos cuáles son éstas, lo cual no sería el caso para una tabla de frecuencias no agrupadas donde se conocen todas las medidas.

Cualquier tabla de frecuencias agrupadas debería poseer las tres características siguientes:

1. Uniformidad: cada clase debería tener la misma amplitud.
2. Unicidad: dos clases no se traslapan.
3. Completez: cada uno de los datos debe pertenecer a alguna clase.

Las fronteras de clase y las amplitudes de clase de una tabla de frecuencias agrupadas se determinan considerando la **unidad** o precisión de la medida. Para las clases de la tabla 2.3, la precisión de la medida es el número entero más cercano, ya que estamos contando individuos, así que la unidad de medida es 1.

La frontera inferior de clase de un intervalo se localiza media unidad abajo del límite, y la frontera superior de clase de un intervalo se localiza media unidad arriba del límite.

Para la primera clase de la tabla 2.3, la frontera inferior de clase es $[1 - 0.5(1)] = 0.5$ y la frontera superior es $[10 + 0.5(1)] = 10.5$. Ninguno de los datos cae en la frontera de un intervalo, por lo tanto, las medidas 0.5 y 10.5 no pueden caer en la primera clase, pero cualquiera de las medidas *entre* 0.5 y 10.5 sí. Desde luego, 0.5 y 10.5 no son medidas posibles, así que las fronteras de clase sólo tienen significado matemático.

La amplitud w de cualquier clase de una tabla de frecuencias agrupadas puede encontrarse restando la frontera inferior de la clase de su frontera superior.

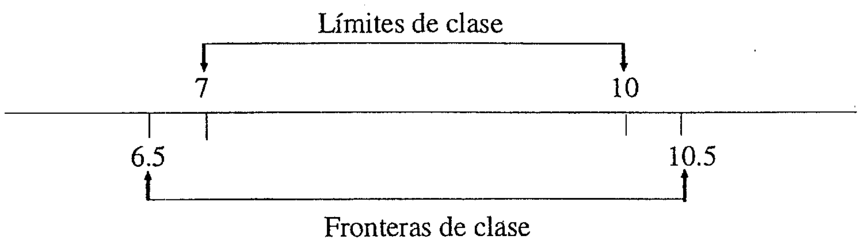
Entonces, para la primera clase en la tabla 2.3, $w = 10.5 - 0.5 = 10$. Tome en cuenta los ejemplos 2.12 y 2.13.

EJEMPLO 2.12

La siguiente es una tabla de frecuencias agrupadas para el número de semillas en 21 naranjas.

Clase	Frecuencia (f)
3-6	5
7-10	6
11-14	7
15-18	3

La precisión de medida para las clases es 1 porque los datos de la tabla son números enteros. Para la clase 7-10, si sumamos $(0.5)(1) = 0.5$ al límite superior de clase 10, obtendremos la frontera superior de clase 10.5. Para conocer la frontera inferior restamos 0.5 del límite inferior de clase y nos resulta $7 - 0.5 = 6.5$ (véase el ejemplo 2.13). El ancho de la clase 7-10 se encuentra entonces restando la frontera inferior de clase de la frontera superior, es decir, $w = 10.5 - 6.5 = 4$.



EJEMPLO 2.13

La siguiente es una tabla de frecuencias agrupadas para el peso en libras de 18 recién nacidos.

Clase	Frecuencia (f)
3.0-4.4	1
4.5-5.9	1
6.0-7.4	7
7.5-8.9	8
9.0-10.4	1

La precisión de la medida de las clases es 0.1 libras. Para la clase 7.5 - 8.9, al restar la mitad de una unidad del límite inferior de clase se obtiene $7.5 - (0.5)(0.1) = 7.5 - 0.05 = 7.45$, la frontera inferior de clase. La frontera superior de clase se encuentra sumando media unidad al límite superior de clase, obteniéndose $8.9 + (0.5)(0.1) = 8.95$; note que ningún peso corresponde a alguna frontera porque la precisión de la medida es el décimo de libra más cercano.

En cualquier tabla de frecuencias agrupadas, la amplitud de clase puede encontrarse simplemente realizando el procedimiento siguiente:

> **Determinación de la amplitud de clase**
>
> Réstense dos límites superiores de clase consecutivos o inferiores de clase consecutivos, o dos fronteras inferiores consecutivas, o dos fronteras inferiores consecutivas, o réstese la frontera inferior de una clase de la frontera superior de dicha clase.

Para los datos de la sala de emergencia proporcionados originalmente en la tabla 2.3, podemos calcular la amplitud de clase como se indica en la tabla 2.4.

TABLA 2.4

Cálculo de la amplitud de clase para la tabla 2.3

	Clase		Frecuencia (f)
	1-10	$w = 20 - 10$ $= 10$	2
	11-20		3
$w = 31 - 21$ $= 10$	21-30		4
	31-40		1
	41-50		1
	51-60		1

Sin embargo, note que la amplitud de clase no se encuentra restando el límite inferior de clase del límite superior.

Elección de clases para tablas de frecuencias agrupadas

Si se quiere construir una tabla de frecuencias agrupadas para una cierta colección de datos, es necesario responder tres preguntas relativas a las clases.

1. ¿Cuántas clases deben usarse?
2. ¿Cuál debe ser la amplitud de clase?
3. ¿En qué valor debe empezar la primera clase?

Escoger el número de clases requiere varias consideraciones. Si todos los datos se agrupan en un número pequeño de clases, las características de los datos originales se ocultan y puede perderse información relevante; por otro lado, demasiadas clases dan demasiados detalles y se pierde el propósito del agrupamiento, que es condensar los datos de manera significativa y fácil de interpretar. Además, demasiadas clases pueden dar lugar a que muchas clases queden vacías quitándole sentido al agrupamiento de los datos.

El número de clases, denotado por c, depende de la situación y del total de los datos obtenidos. Como no hay un acuerdo general entre los estadísticos acerca del número de clases que deben usarse y dado que la elección es arbitraria, en este texto usaremos de 5 a 15 clases, inclusive.

Número de clases para una tabla de frecuencias agrupadas:
Entre 5 y 15 clases (inclusive).

Una sugerencia útil para el número de clases está dado por la **regla de Sturges,** que establece como número de clases necesario, aproximadamente,

Regla de Sturges

$$c = 3.3\,(\log n) + 1$$

donde n es el número de medidas y $\log n$ es el logaritmo de n en base 10

El valor de c es común redondearlo al entero más cercano.

EJEMPLO 2.14

Si el número de medidas proporcionadas es $n = 25$, la regla de Sturges sugiere usar seis clases, porque

$$\begin{aligned} c &= 3.3(\log n) + 1 \\ &= 3.3(\log 25) + 1 \\ &= 3.3(1.3979) + 1 \approx 6 \end{aligned}$$

donde $\approx$ significa aproximadamente igual que.

Algunos investigadores piensan que, en muchas situaciones, la regla de Sturges da un valor de c que permite la construcción de una tabla de frecuencias agrupadas que da una imagen realista de los datos no agrupados. Una vez establecido el número de intervalos de clase que se usarán, la amplitud de clase se encuentra usando el **rango** R, que es la diferencia entre la medida mayor U y la medida menor L en la muestra:

Rango

$$R = U - L$$

Como c clases deben cubrir el rango, dividimos éste entre el número de clases para encontrar la amplitud de clase w:

Amplitud de clase

$$w = \frac{R}{c}$$

Como la medida menor debe caer en la primera clase, el límite inferior de la primera clase debe estar en, o un poco antes de, la medida menor L. Así que podemos establecer un acuerdo general sobre las clases de nuestras tablas de frecuencias agrupadas, empezando siempre la primera clase con la medida menor; esto nos será especialmente útil cuando verifiquemos nuestras respuestas. En la práctica, es común que la primera clase empiece en un número que permita expresar las clases de intervalos convenientes, pero hay ocasiones en que se justifica una excepción a la regla (véase el ejercicio 33 al final de esta sección).

Cuando la primera clase comienza con la menor de las medidas, el valor mínimo que puede tomar w depende de la unidad de medida. El valor mínimo para la amplitud de clase w se determina redondeando el cociente R/c al siguiente valor entero.

> El valor de w se toma como el mínimo entero mayor que R/c

APLICACIÓN 2.4

El profesor Smith puso un examen final consistente en 100 preguntas a su grupo de Introducción a la contabilidad. Los datos siguientes representan el número de respuestas correctas en cada examen; construya una tabla de frecuencias agrupadas con cinco clases que ayude al profesor Smith a analizar los resultados.

17	15	78	21	10	32	7	65	18	87
4	22	34	42	9	9	82	79	98	4
44	64	62	77	2	81	45	37	83	44
77	13	41	16	17	13	82	37	5	54
7	67	88	41	61	22	92	16	67	85

Solución:

Paso 1. Primero determinamos el rango R. Como la medida mayor es U = 98 y la mínima $L = 2$, el rango es

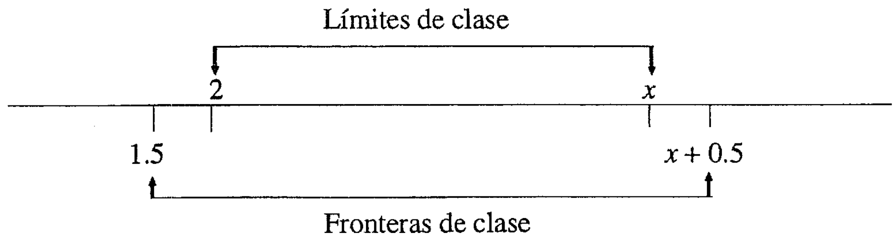
$$\begin{aligned} R &= U - L \\ &= 98 - 2 = 96 \end{aligned}$$

Paso 2. A continuación determinamos w, la amplitud de cada clase. Vea que el número de clases está dado por $c = 5$.

$$w = \frac{R}{c} = \frac{96}{5} = 19.2$$

Como la unidad de precisión para los puntajes de exámenes es 1, escogemos el mínimo entero mayor que 19.2 como el valor de la amplitud. Para nuestra aplicación, el mínimo entero mayor que 19.2 es 20; en consecuencia, $w = 20$.

Paso 3. Comenzamos con $L = 2$ y construimos la primera clase con un ancho de 20. Supongamos que la primera clase se extiende de 2 a x, donde x representa la incógnita límite superior de clase (véase el diagrama adjunto).



Como la unidad de medida es 1 y 0.5(1) = 0.5, la frontera superior de clase puede representarse como $x + 0.5$. La amplitud de la primera clase se encuentra restando la frontera inferior de clase de la frontera superior de clase, por lo tanto,

$$w = (x + 0.5) - 1.5$$

Como $w = 20$, tenemos

$$20 = x - 1$$

Al resolver esta ecuación encontramos que el límite superior de clase es $x = 21$. En consecuencia, la primera clase resulta $2 - 21$.

Paso 4. Para obtener cada una de las clases siguientes a esta primera, sumamos $w = 20$ a los límites inferior y superior de la clase precedente. Así,

2-21	
22-41	(Note: $41 = 21 + 20$)
42-61	(Note: $42 = 22 + 20$)
62-81	
82-101	(Note: $101 = 81 + 20$)

Paso 5. Para determinar la frecuencia de cada clase usamos una columna de marcas de cuenta. Si uno de los datos cae en una clase, anotamos una marca (I) en la columna correspondiente a esa clase. La tabla 2.5 contiene nuestra tabla de frecuencias agrupadas para los 50 puntajes del examen.

TABLA 2.5
Tabla de frecuencias agrupadas de los puntajes del examen final

Clase	Cuenta	Frecuencia (f)
2-21	卌 卌 卌 III	18
22-41	卌 III	8
42-61	卌 I	6
62-81	卌 卌	10
82-101	卌 III	8
		50

■

APLICACIÓN 2.5

Los datos adjuntos representan el número de clientes que visitan una tienda en un periodo de 22 días. Use seis clases y construya una tabla de frecuencias agrupadas para los datos.

28	42	52	50	29	31	34	45	48	38	28
33	33	49	32	37	41	43	46	49	34	49

Solución:

Paso 1. La medida mayor es $U = 52$, y la menor es $L = 28$. Determinamos el rango:

$$R = U - L$$
$$= 52 - 28 = 24$$

Paso 2. Determinamos el ancho de cada clase:

$$w = \frac{R}{c} = \frac{24}{6} = 4$$

Debemos usar un ancho de $w = 5$ en este caso; si no, los seis intervalos pueden no contener todos los datos. Con $w = 4$, no hay una clase que contenga al valor mayor de los datos o el valor menor. Como hemos convenido en comenzar el primer intervalo con el valor menor de los datos, nos arriesgamos a que el valor mayor de los datos no pertenezca al último intervalo. Para ver por qué es éste el caso, supongamos que el ancho es $w = 4$. La primera clase es 28-31, y las seis clases serán:

Clase
28-31
32-35
36-39
40-43
44-47
48-51

Advierta que el valor $U = 52$ no pertenece a ninguna clase; para remediar la situación debemos escoger el mínimo entero mayor que 4, que es 5. Por lo tanto, el ancho de la clase debe ser $w = 5$.

Paso 3. La tabla 2.6 es la tabla completa de frecuencias agrupadas.

TABLA 2.6
Tabla de frecuencias agrupadas para los datos de la aplicación 2.5

Clase	Cuenta	f
28-32	~~IIII~~	5
33-37	~~IIII~~	5
38-42	III	3
43-47	III	3
48-52	~~IIII~~ I	6
53-57		0

En este caso la última clase está vacía. Para remediar esta situación podemos empezar la primera clase en un valor menor, digamos 26. La tabla de frecuencias sería entonces:

Clase	f
26-30	3
31-35	6
36-40	2
41-45	4
46-50	6
51-55	1

■

Marca de clase

El punto medio de cada clase se denomina **marca de clase** y se denota por X. Cuando los datos se condensan en una tabla de frecuencias agrupadas se pierde información y no sabemos el valor exacto de las medidas que caen en cada clase; por eso lo mejor que podemos hacer es permitir que cada una de las medidas de una clase dada esté representada por la marca de esa clase; al usar marcas de clase en lugar de los datos sin agrupar, los cálculos se facilitan aunque se pierde precisión. Para una clase dada, la marca de clase se encuentra usando la fórmula

$$X = \frac{l_1 + l_2}{2}$$

donde l_1 es el límite inferior de clase y l_2 es el límite superior.

EJEMPLO 2.15

Para la primera clase de la tabla 2.5, la marca de clase es

$$X = \frac{l_1 + l_2}{2}$$
$$= \frac{2 + 21}{2} = 11.5$$

Para la segunda clase, la marca de clase es

$$X = \frac{l_1 + l_2}{2}$$
$$= \frac{22 + 41}{2} = 31.5$$

Advierta que la marca de clase para la clase 2 también se puede encontrar sumando $w = 20$ a la marca de clase para la clase 1 ($11.5 + 20 = 31.5$). En general, cada marca de clase para las clases siguientes a la primera puede encontrarse sumando $w = 20$ a la marca de clase precedente, por lo tanto, resulta que las tres marcas de clase restantes son 51.5, 71.5 y 91.5. La tabla 2.7 muestra una tabla de frecuencias agrupadas que contiene las marcas de clase.

TABLA 2.7

Marcas de clase para la tabla de frecuencias agrupadas de la aplicación 2.4 o de la tabla 2.5

Número de clase	Clase	Cuenta	f	Marca de *la clase X*
1	2-21	𝍸 𝍸 𝍸 III	18	11.5
2	22-41	𝍸 III	8	31.5
3	42-61	𝍸 I	6	51.5
4	62-81	𝍸 𝍸	10	71.5
5	82-101	𝍸 III	8	91.5

APLICACIÓN 2.6

El conjunto de datos siguiente representa los totales de efectivo (en dólares) gastados en un cierto fin de semana por 25 estudiantes graduados. Construya una tabla de frecuencias agrupadas que contenga cinco clases.

	39.78	28.30	28.31	17.95	44.47	
	46.65	31.47	33.45	29.17	48.39	
$U \longrightarrow$	82.71	43.63	41.17	47.32	52.16	
	25.94	50.32	35.25	35.70	17.89	$\longleftarrow L$
	60.20	48.14	22.78	38.22	23.25	

Solución:

Paso 1. Calcule el rango R. Como $U = 82.71$ y $L = 17.89$, el rango es

$$R = U - L$$
$$= 82.71 - 17.89 = 64.82$$

Paso 2. Calcule el ancho de clase w. Como $c = 5$, tenemos

$$\frac{R}{c} = \frac{64.82}{5} = 12.96$$

El mínimo entero mayor que 12.96 es 13. En consecuencia, el ancho de clase es $w = 13$.

Paso 3. Comenzamos con $L = 17.89$ y construimos una clase con ancho $w = 13$. La unidad es 0.01 y $(0.5)(0.01) = 0.005$. Representemos con x la frontera superior de la primera clase. Entonces, el ancho se obtiene restando la frontera inferior de clase de la frontera superior.

$$\begin{aligned} w &= x + 0.005 - 17.885 \\ 13 &= x + 0.005 - 17.885 \\ 13 &= x - 17.88 \\ x &= 30.88 \end{aligned}$$

En consecuencia, la primera clase es 17.89-30.88.

Paso 4. Para obtener las clases restantes, sumamos 13 a los límites de clase precedentes.

Clase	
17.89 - 30.88	Primera clase
13 13	sume $w = 13$ en ambos lados de la primera clase
30.89 - 43.88	Segunda clase
13 13	sume $w = 13$ en ambos lados de la segunda clase
43.89-56.88	Tercera clase

Las dos clases restantes se encuentran de manera análoga. Son:

56.89 - 69.88
69.89 - 82.88

Paso 5. Las frecuencias para las cinco clases se encuentran usando marcas de cuenta como se ve en la tabla 2.8.

Paso 6. Las marcas de cada clase se encuentran usando la fórmula para el punto medio dada previamente. La marca de clase para la primera es

$$\begin{aligned} X &= \frac{l_1 + l_2}{2} \\ &= \frac{17.89 + 30.88}{2} = 24.385 \end{aligned}$$

Cada marca de clase sucesiva se encuentra sumando $w = 13$ a la marca anterior. La tabla 2.8 de frecuencias agrupadas de los datos, muestra también las marcas de clase.

TABLA 2.8

Tabla de frecuencias agrupadas de los datos de la aplicación 2.6

Número de clase	Clase	Cuenta	f	X
1	17.89-30.88	~~IIII~~ III	8	24.385
2	30.89-43.88	~~IIII~~ III	8	37.385
3	43.89-56.88	~~IIII~~ II	7	50.385
4	56.89-69.88	I	1	63.385
5	69.89-82.88	I	1	76.385
			25	

■

Tablas de frecuencia relativas

A veces es útil expresar cada valor o clase de una tabla de frecuencias como una fracción o porcentaje del total de las medidas. La **frecuencia relativa** de una medida o clase se encuentra dividiendo la frecuencia f de dicha medida entre el total n de medidas; entonces a la tabla se le llama tabla de frecuencia relativa (véase el ejemplo 2.16). Una **tabla de frecuencia relativa** tiene varias ventajas sobre otra tabla de frecuencias cualquiera; una ventaja importante es, como se ve en el ejemplo 2.17, que podemos hacer comparaciones llenas de significado entre conjuntos similares de datos que tengan las mismas clases pero frecuencias totales distintas.

EJEMPLO 2.16

La tabla de frecuencias relativas para la clase 1 de la tabla 2.8 es $n = 8/25 = 0.32$. La tabla 2.9 nos enseña una tabla de frecuencias relativas para los datos de la aplicación 2.6. Note que la suma de la columna de frecuencias relativas es 1.00.

TABLA 2.9

Tabla de frecuencia relativa

Clase	Frecuencia relativa
17.89-30.88	0.32
30.89-43.88	0.32
43.89-56.88	0.28
56.89-69.88	0.04
69.89-82.88	0.04
	1.00

EJEMPLO 2.17

Considere la tabla 2.10, que exhibe los salarios iniciales de matemáticos recién graduados en dos universidades estatales, A y B.

TABLA 2.10

Tabla de frecuencias para los salarios iniciales en dos universidades

Universidad A		Universidad B	
Clase salarial	f	Clase salarial	f
$10,000-12,999	0	$10,000-12,999	1
13,000-15,999	2	13,000-15,999	1
16,000-18,999	7	16,000-18,999	2
19,000-21,999	6	19,000-21,999	2
22,000-24,999	3	22,000-24,999	3
25,000-27,999	2	25,000-27,999	1

Al examinar las dos partes de la tabla vemos que en cada universidad hay tres matemáticos recién graduados con salarios iniciales de entre 22,000 y 24,999 dólares. Pero si comparamos las frecuencias relativas, vemos que la universidad A tiene un 3/20 = 15% de sus matemáticos ganando entre 22,000 y 24,999 dólares, mientras que la universidad B tiene a un 3/10 = 30% ganando entre 22,000 y 24,999 dólares (véase la tabla 2.11).

TABLA 2.11

Frecuencias relativas de salarios iniciales en dos universidades

Universidad A		Universidad B	
Clase salarial	Relativa f	Clase salarial	Relativa f
$10,000-12,999	0/20 = 0	$10,000-12,999	1/10 = 0.10
13,000-15,999	2/20 = 0.10	13,000-15,999	1/10 = 0.10
16,000-18,999	7/20 = 0.35	16,000-18,999	2/10 = 0.20
19,000-21,999	6/20 = 0.30	19,000-21,999	2/10 = 0.20
22,000-24,999	3/20 = 0.15	22,000-24,999	3/10 = 0.30
25,000-27,999	2/20 = 0.10	25,000-27,999	1/10 = 0.10

En la tabla de frecuencias relativas, la suma de la columna de frecuencias relativas, sin error de redondeo, es siempre 1, lo cual no es sorprendente pues una frecuencia relativa equivale a un porcentaje; para convertir una frecuencia relativa en un porcentaje, la multiplicamos por 100%. Encontraremos de utilidad las tablas de frecuencia relativa cuando estudiemos probabilidad en el capítulo 5, donde la suma de las probabilidades estudiadas será siempre igual a 1.

Tablas de frecuencia acumulada

La **frecuencia acumulada** de cualquier medida, o clase, es la suma de las frecuencias de esa misma medida o clase, y de las frecuencias de todas las demás de menor valor. En muchas ocasiones estaremos interesados en el número de observaciones menores o iguales que algún valor dado. La tabla 2.12 ilustra una tabla de frecuencia acumulada para los datos de la aplicación 2.4.

EJEMPLO 2.18

En el caso de que la frecuencia acumulada sea de interés, se incluyen los siguientes ejemplos:

1. Un ingeniero especialista en control de calidad desearía conocer el número de días en que un proceso de producción originó, cuando mucho 100 artículos defectuosos.
2. Un maestro puede estar interesado en conocer el número de estudiantes que obtuvieron un puntaje menor o igual al 70% en un examen.
3. A un entrenador de basquetbol le interesaría saber el número de juegos en que los contrarios anotaron a lo más 60 puntos.

TABLA 2.12

Tabla de frecuencia relativa acumulada para los datos de la aplicación 2.4

Tabla de frecuencia		Tabla de frecuencia acumulada	
Clase	f	Clase	Frecuencia acumulada
2-21	18	2-21	18
22-24	8	22-24	26 = (18 + 8)
42-61	6	42-61	32 = (26 + 6)
62-81	10	62-81	42 = (32 + 10)
82-101	8	82-101	50 = (42 + 8)
	50		

Tablas de frecuencia relativa acumulada

También pueden construirse tablas de frecuencia acumulada para tablas que contienen frecuencias relativas o porcentajes. Cuando se hace esto, la tabla se denomina **tabla de frecuencia relativa acumulada**. Una tabla de esta naturaleza para los datos de la aplicación 2.4 se muestra en la tabla 2.13; se obtuvo de la tabla 2.12 calculando frecuencias relativas acumuladas para la frecuencia acumulada.

TABLA 2.13

Tabla de frecuencia relativa acumulada para los datos de la aplicación 2.4

Clase	Frecuencia relativa acumulada
2-21	18/50 = 0.36
22-41	26/50 = 0.52
42-61	32/50 = 0.64
62-81	42/50 = 0.84
82-101	50/50 = 1.00

Las frecuencias relativas acumuladas tienen muchos usos, uno es la calificación de pruebas escolares de aplicación generalizada, como la Prueba de Aptitud Académica (PAA) y muchos otros exámenes de ingreso; los puntajes de esas pruebas suelen darse como **percentiles.** Un *puntaje percentil* nos dice que parte de la población examinada quedó *abajo del puntaje dado.*

EJEMPLO 2.19

Si se dice que 590 es el nonagésimo percentil en la parte de matemáticas de la PAA; eso significa que 90% de los puntajes en la porción de matemáticas de esa prueba estuvieron abajo de 590.

EJEMPLO 2.20

La tabla 2.14 registra las estaturas en pulgadas de 200 alumnos de nuevo ingreso del sexo masculino, en una preparatoria.

TABLA 2.14

Estaturas en pulgadas de alumnos de nuevo ingreso

Estaturas	f	Frecuencias relativas	Frecuencia relativa acumulada
59.5-62.5	2	0.01	0.01
62.5-65.5	12	0.06	0.07
65.5-68.5	24	0.12	0.19
68.5-71.5	46	0.23	0.42
71.5-74.5	62	0.31	0.73
74.5-77.5	36	0.18	0.91
77.5-80.5	16	0.08	0.99
80.5-83.5	2	0.01	1.00

Esta tabla, que utiliza fronteras de clase y frecuencias relativas acumuladas, se puede usar para determinar percentiles. Las conclusiones siguientes son manifiestas en la observación de la tabla anterior:

1. Una estatura de 74.5 pulgadas es el percentil septuagésimo.
2. El percentil número cincuenta está entre 71.5 y 74.5 pulgadas.
3. El percentil 19 es 68.5 pulgadas.
4. El septuagésimo quinto percentil se ubica entre 74.5 y 77.5 pulgadas.

APLICACIÓN 2.7

Supongamos que el percentil número setenta de peso de los hombres adultos es 175 libras y que el 85 es 195 libras. ¿Qué porcentaje de hombres tienen pesos mayores que 175 libras y menores que 195?

Solución: Por definición, 70% de los hombres adultos pesan menos de 175 libras y 85% pesan menos de 195; por lo tanto: 0.85 – 0.70 = 0.15 = 15% de los hombres adultos tienen pesos comprendidos entre 175 y 195 libras. ▪

Tablas bivariadas

Si tenemos datos resultantes de medir dos aspectos distintos de los miembros de una población, entonces los llamamos **datos bivariados.** Se usan dos variables para representar los dos aspectos de cada miembro; un miembro puede ser un objeto, persona o fuente. Supongamos que queremos investigar la altura y el peso de todos los jugadores de basquetbol de las preparatorias en Allegany County, Maryland: cada miembro es un jugador de basquetbol y asociado a él tenemos dos medidas, altura en pulgadas y peso en libras; si asignamos a la variable x la representación de la altura de un jugador, y a la variable y el peso del mismo jugador, entonces la pareja ordenada (x, y) representará la altura y el peso, respectivamente, de un miembro. Hasta ahora, habíamos estado midiendo sólo un aspecto de cada miembro de una población, por eso usamos una única variable para representar las medidas; cuando se usa sólo una variable para representar los datos obtenidos de los miembros de una población, los nombramos **datos univariados.**

EJEMPLO 2.21

(Datos univariados). Si los datos son los pesos en libras de un grupo de 30 estudiantes de estadística, entonces un miembro es un estudiante y el aspecto medido es el peso del mismo.

EJEMPLO 2.22

(Datos bivariados). Supongamos que estamos interesados en el promedio diario de precipitación pluvial y temperatura ambiente habidos en Athens, Georgia, durante los diez años pasados. La población consiste de los diez años pasados, donde un miembro de la población es cada año, y en este caso, un año es una fuente para dos piezas de la información que vendrían a ser el promedio diario de la precipitación pluvial y el promedio diario de la temperatura ambiente. Cada año da lugar a dos medidas.

Una **tabla de frecuencia bivariada** es un arreglo de datos clasificados en dos categorías; la información usada para construir las tablas de frecuencia bivariada se obtienen generalmente de contar frecuencias. Cada categoría se identifica con un símbolo llamado variable, cada variable representa datos de una categoría; las categorías pueden ser números discretos, intervalos numéricos o valores cualitativos como género, color de cabello o religión.

EJEMPLO 2.23

Vamos a suponer que la información se obtuvo de una muestra de votantes a los que se preguntó su filosofía política y su filiación partidista; a cada uno se le pidió identificar su filosofía política como: liberal, conservadora u otra, y su filiación partidista como demócrata, republicana u otra; las dos variables de clasificación son filosofía política y filiación partidista. La variable filosofía política tiene tres categorías o niveles de clasificación: liberal, conservadora u otra; la segunda variable tiene también tres categorías o niveles: demócrata, republicana, otra; los datos están tabulados en la tabla 2.15.

TABLA 2.15

Tabla de frecuencia bivariada

	Filosofía política			
Filiación partidista	Liberal	Conservadora	Otra	Total
Demócrata	78	65	37	180
Republicana	84	79	7	170
Otra	38	46	16	100
Total	200	190	60	450

La información que sigue, entre otras, puede leerse fácilmente de la tabla:

1. Hubo 78 votantes que dijeron ser liberales demócratas.
2. 79 personas manifestaron ser conservadores republicanos.
3. Se entrevistó a 450 individuos.

4. Hubo 170 republicanos entrevistados.
5. 60 votantes clasificaron su filosofía política como otra.

APLICACIÓN 2.8

En la tabla siguiente están anotadas las calificaciones en estadística y el sexo de 32 estudiantes universitarios. Construya una tabla de frecuencia para los datos bivariados.

Estudiante	Calificación	Sexo	Estudiante	Calificación	Sexo
1	B	M	17	C	F
2	C	F	18	E	F
3	C	F	19	C	M
4	C	M	20	B	F
5	B	F	21	D	M
6	B	F	22	E	M
7	A	M	23	B	M
8	C	M	24	B	M
9	D	F	25	C	M
10	C	M	26	C	F
11	B	F	27	D	M
12	A	F	28	B	F
13	C	M	29	D	F
14	D	F	30	A	M
15	D	F	31	E	M
16	A	F	32	A	F

Solución:

Paso 1. Usamos marcas de cuenta para determinar los totales para cada una de las diez combinaciones sexo-calificación.

<table>
<tr><td></td><td colspan="5">Calificación</td></tr>
<tr><td>Sexo</td><td>A</td><td>B</td><td>C</td><td>D</td><td>E</td></tr>
<tr><td>H</td><td>||</td><td>|||</td><td>卌|</td><td>||</td><td>||</td></tr>
<tr><td>M</td><td>|||</td><td>卌</td><td>||||</td><td>||||</td><td>|</td></tr>
</table>

Paso 2. A continuación encontramos los totales para los dos renglones, cinco columnas y diez combinaciones sexo/calificación.

	Calificación					
Sexo	A	B	C	D	E	Total
H	2	3	6	2	2	15
M	3	5	4	4	1	17
Total	5	8	10	6	3	32

■

APLICACIÓN 2.9

Se hizo un estudio entre los miembros del colegio de académicos para considerar sus actitudes hacia el contrato colectivo de trabajo, celebrado entre la administración y el sindicato del personal académico; los resultados se resumen en la tabla siguiente.

	Actitud hacia el contrato colectivo			
Cargo	A favor	En contra	Abstención	Total
Profesor	45	8	2	55
Prof. Asociado	31	16	3	50
Prof. Asistente	42	19	4	65
Instructor	12	4	14	30
Total	130	47	23	200

Use la tabla para responder cada una de estas preguntas:

a) ¿Qué porcentaje de los académicos está en contra del contrato colectivo?

b) ¿Qué porcentaje corresponde a profesores asociados?

c) ¿Qué tanto por ciento de los profesores están a favor del contrato colectivo?

d) ¿Instructores que están a favor de dicho contrato, en porcentaje?

e) ¿Quiénes se oponen al contrato? ¿Qué tanto por ciento corresponde a profesores?

f) ¿Qué porcentaje del total de los académicos representan los profesores, asociados o de rango superior, que están a favor del contrato?

Solución:

a) $47/200 = 23.5\%$.

b) $50/20 = 25\%$.

c) $45/55 = 81.8\%$.

d) $12/30 = 40\%$.

e) Cuarenta y siete académicos se oponen al contrato colectivo de trabajo y, de ellos, 8 son profesores. En consecuencia, los profesores constituyen el $8/47 = 17.02\%$ de quienes se oponen al contrato colectivo.

f) $(42 + 31 + 45)/200 = 59\%$. ■

GRUPO DE EJERCICIOS 2.2

Habilidades básicas

1. Determine la amplitud de clase w para cada uno de los conjuntos siguientes:
   a) $L = 17$, $U = 81$, $c = 8$
   b) $L = 14.5$, $U = 102.3$, $c = 7$
   c) $L = 23$, $U = 204$, $c = 11$
   d) $L = 23.65$, $U = 67.24$, $c = 10$
   e) $L = 13.6$, $U = 73.6$, $c = 12$

2. Determine el ancho de clase w para cada uno de estos conjuntos de condiciones:
   a) $L = 27$, $U = 87$, $c = 7$
   b) $L = 24.3$, $U = 112.5$, $c = 9$
   c) $L = 39$, $U = 130$, $c = 13$
   d) $L = 13.64$, $U = 75.24$, $c = 10$
   e) $L = 15.2$, $U = 75.2$, $c = 12$

3. Calcule los límites superiores para el primer intervalo de clase en cada una de las siguientes condiciones:
   a) $L = 17$, $U = 81$, $c = 8$, unidad de medida = 1
   b) $L = 14.5$, $U = 102.3$, $c = 7$, unidad de medida = 0.1
   c) $L = 23$, $U = 204$, $c = 11$, unidad de medida = 1
   d) $L = 23.65$, $U = 67.24$, $c = 10$, unidad de medida = 0.01
   e) $L = 13.6$, $U = 73.6$, $c = 12$, unidad de medida = 0.1

4. Realice lo mismo que en el inciso anterior.
   a) $L = 27$, $U = 87$, $c = 7$, unidad de medida = 1
   b) $L = 24.3$, $U = 112.5$, $c = 9$, unidad de medida = 0.1
   c) $L = 39$, $U = 130$, $c = 13$, unidad de medida = 1
   d) $L = 13.64$, $U = 75.24$, $c = 10$, unidad de medida = 0.01
   e) $L = 15.2$, $U = 75.2$, $c = 12$, unidad de medida = 0.1

5. Con referencia al ejercicio 3, determine las fronteras del primer intervalo de clase para cada conjunto de condiciones.

6. Basándose en el ejercicio 4, determine las fronteras del primer intervalo de clase para cada conjunto.

7. Use la tabla que se ilustra a continuación para construir:
   a) una tabla de frecuencia relativa;
   b) una tabla de frecuencia acumulada;
   c) una tabla de frecuencia relativa acumulada.

Clase	f
1-4	14
5-8	18
9-12	12
13-16	16
17-20	20

8. Con base en la tabla de frecuencia agrupada siguiente construya:
   a) una tabla de frecuencia relativa;
   b) una tabla de frecuencia acumulada;
   c) una tabla de frecuencia relativa acumulada.

Clase	f
10-15	13
16-21	10
22-27	9
28-33	17
24-39	22
40-45	6

9. Utilice los datos de este ejercicio para efectuar la misma actividad del inciso 8.
   a) una tabla de frecuencia relativa;
   b) una tabla de frecuencia acumulada;
   c) una tabla de frecuencia relativa acumulada.

x	f
4	1
7	3
8	6
9	4
10	2

10. La tabla de frecuencias no agrupadas de este inciso le servirá para construir:
    a) una tabla de frecuencia relativa;
    b) una tabla de frecuencia acumulada;
    c) una tabla de frecuencia relativa acumulada.

x	f
12	8
15	10
20	7
22	13
35	10
40	2

11. En la tabla adjunta, identifique:
    a) las marcas de clase;
    b) las fronteras de clase.

Clase	f
1-4	14
5-8	18
9-12	12
13-16	16
17-20	20

12. Igual que en el anterior, identifique:
    a) las marcas de clase;
    b) las fronteras de clase.

Clase	f
10-15	13
16-21	10
22-27	9
28-33	17
34-39	22
40-45	6

13. Utilice la regla de Sturges para determinar el número de clases para una colección de datos de tamaño igual que:
    a) 25.
    b) 50.
    c) 75.
    d) 100.
    e) 500.

14. Haga lo mismo que en el ejercicio anterior para una colección de datos de tamaño igual que:
   a) 35.
   b) 80.
   c) 95.
   d) 100.
   e) 1000.

Más aplicaciones

15. Los datos anotados en seguida representan los totales, en dólares, gastados en golosinas por una muestra de 25 estudiantes durante un periodo de exámenes.

57	28	63	38	29	89	77	72	39
47	64	84	88	42	36	72	69	
68	41	52	39	72	45	52	84	

Mediante seis clases construya una tabla de frecuencia agrupada.

16. Las observaciones siguientes representan las velocidades en millas por hora (mph), de 30 coches registrados por el radar de la policía en una carretera interestatal muy transitada:

57	63	70	53	61	60	67	79	64	62
66	73	71	78	84	53	48	80	54	60
67	65	62	55	52	69	73	72	66	58

Construya una tabla de frecuencia agrupada usando siete clases.

17. Use los datos del ejercicio 15 y la regla de Sturges para construir una tabla de frecuencia relativa agrupada.

18. Con los datos del ejercicio 16, use la regla de Sturges para construir una tabla de frecuencia relativa agrupada.

19. Las temperaturas del mediodía, en grados Farenheit, registradas el 1 de julio durante los últimos 28 años en un pueblo pequeño, son las siguientes:

66	83	77	90	78	84	83	80	77	79
75	88	72	66	83	85	94	88	79	79
72	78	76	84	81	73	80	90		

Use estos datos y seis clases para construir una tabla de frecuencia acumulada agrupada.

20. Las personas con casa propia en Estados Unidos parecen cambiar de domicilio frecuentemente. Los datos siguientes indican cuánto tiempo, en años, tenían en su domicilio en 1988 los propietarios de casas de los 50 estados.[9]

26.3	20.8	19.2	18.5	18.5	18.2	17.2	16.1
15.9	15.2	14.7	14.5	14.1	14.1	13.9	
13.7	13.5	13.3	13.3	13.2	12.3	12.3	
12.2	12.2	12.2	12.0	11.9	11.6	11.5	
11.5	11.4	11.4	11.4	11.1	11.0	10.9	
10.9	10.3	10.2	10.2	9.9	9.8	9.4	
9.1	8.8	8.5	8.5	8.4	8.0	7.9	

Con estos datos y siete clases construya una tabla de frecuencia acumulada agrupada.

21. Los datos adjuntos representan una muestra de precios, en centavos, de la gasolina con plomo en una cierta ciudad durante un mes en particular.

123.9	127.9	130.9	121.9	132.9	120.8	115.9
117.9	131.9	121.9	126.9	122.8	126.9	
137.9	115.9	115.9	121.9	126.9	119.9	
118.9	119.8	116.9	129.9	122.8	119.9	

Utilice estos datos y cinco clases para construir una tabla de frecuencia relativa acumulada agrupada.

22. Use los datos del ejercicio 20 y cinco clases para construir una tabla de frecuencias relativa acumulada agrupada.

23. Se preguntó a un grupo de 30 estudiantes cuántos libros habían comparado para el último semestre. Sus respuestas fueron:

5	6	5	5	4	5	4	5	3	6	4	4	4	6	2
9	5	4	3	3	8	11	7	8	7	4	10	4	3	6

   a) Construya una tabla de frecuencia relativa no agrupada, y
   b) una tabla de frecuencia relativa acumulada no agrupada.

24. Los datos anotados abajo representan el porcentaje de los ingresos familiares gastados en 1988, en alimentación en las áreas metropolitanas más grandes de Estados Unidos.[10]

14	13	29	22	14	12	15	12	15	16
16	16	16	17	17	12	12	11	14	12
12	9	19	17	16	11	13	15	14	12
14	15	13	13	11	13	14	13	13	12
12	14	15	11	14	11	12	13	12	11

   a) Haga una tabla de frecuencia relativa no agrupada.
   b) También una tabla de frecuencia relativa acumulada no agrupada.

25. Se clasificó a los estudiantes de una escuela pequeña de acuerdo con su categoría escolar y su preferencia musical. Los resultados están registrados en la tabla siguiente.

Preferencia musical	Categoría escolar Primer año	Segundo año	Penúltimo año	Último año	Total
Rock	16	11	7	6	40
Country	10	12	3	5	30
Clásica	3	1	2	4	10
Jazz	23	11	2	4	40
Folklor	3	0	6	1	10
Total	55	35	20	20	130

a) ¿Qué porcentaje de los estudiantes de primer año prefieren la música clásica?
b) ¿Qué porcentaje de los aficionados al rock son de segundo año?
c) ¿Qué tanto por ciento del total de los estudiantes prefieren la música country?
d) ¿Cuánto, en por ciento, de los estudiantes son de penúltimo año?
e) ¿Qué porcentaje del total de estudiantes son de penúltimo o de último año?
f) ¿Qué porcentaje prefiere la música country o la folklórica?

26. Una muestra de electores fue interrogada sobre su preferencia entre tres candidatos a alcalde. Los resultados están registrados por sexo en la tabla siguiente.

Sexo	Candidato A	B	C	Total
Hombre	15	16	4	35
Mujer	5	4	1	10
Total	20	20	5	45

a) ¿Qué porcentaje de electores prefiere al candidato B?
b) ¿Qué porcentaje de electores son hombres?
c) ¿Qué tanto por ciento de electores hombres prefiere al candidato C?
d) ¿Qué porcentaje de mujeres prefieren al candidato A?
e) ¿Qué porcentaje de electoras femeninas prefieren al candidato B?
f) ¿Qué tanto por ciento de electores prefieren a los candidatos A o C?

27. Los siguientes datos representan las cuentas telefónicas mensuales, en dólares, de 25 residentes de una pequeña comunidad:

19.80	36.05	28.50	21.48	21.15
25.12	23.47	27.81	26.66	20.35
30.22	25.49	20.80	23.83	25.35
23.48	25.81	26.83	20.77	19.98
35.87	22.02	21.07	30.96	33.38

a) ¿Qué porcentaje del grupo pagó más de 20 dólares?
b) ¿Qué porcentaje pagó más de 24 pero menos de 28 dólares?

28. Basándose en los datos del ejercicio 20, determine:
a) ¿Qué porcentaje de propietarios en Estados Unidos permanecieron en sus casas durante 10 años?
b) ¿Qué tanto por ciento permanecieron en sus casas más de 12 años pero menos de 20?

29. Considere la tabla de frecuencia agrupada siguiente:

Clases	f
4.5-9.4	2
9.5-14.4	3
14.5-19.4	4
19.5-24.4	1
24.4-29.4	8

a) Encuentre w, la amplitud de cada clase.
b) Las cinco marcas de clase.
c) Localice las fronteras para la primera clase.
d) ¿Qué porcentaje hay de datos mayores que 19.45?
e) ¿y de los datos menores de 24.45?
f) ¿Qué tanto por ciento cae en la clase 14.5-19.4?

Un paso más allá

30. Los datos siguientes representan los totales semanales, en dólares, gastados en comida por 50 parejas de recién casados.

57.10	70.89	59.17	60.08	49.16
56.17	66.94	67.08	58.10	71.28
50.25	46.39	55.01	68.81	58.70
69.48	46.02	54.16	65.07	48.09
58.32	50.82	45.43	57.20	62.30
51.42	58.76	46.37	47.16	63.51
55.45	58.14	57.14	58.63	55.14
60.37	52.41	74.13	62.38	51.15
64.00	47.75	52.59	42.73	60.32
48.10	59.62	59.46	57.16	58.19

Construya una tabla de frecuencia agrupada con el mínimo número de clases para las que la amplitud de clase es $w = 2.75$.

31. Los cálculos en millas por galón en 40 cargas del tanque de un automóvil nuevo son como sigue:

26.6	28.7	29.2	26.4	29.3	25.8	28.7	29.0
30.0	28.1	28.3	27.8	27.6	31.9	26.6	28.4
30.3	30.4	29.2	29.3	26.5	28.7	28.8	28.3
28.8	27.1	28.9	31.2	30.2	29.2	30.3	32.0
28.4	30.3	29.5	28.4	27.4	30.8	29.5	31.5

a) Construya una tabla de frecuencia relativa acumulada agrupada con ocho clases.
b) Use esa tabla para aproximar al quincuagésimo percentil.

32. En un supermercado se realizó un estudio de eficiencia cuyos resultados no representan los tiempos, en minutos, requerido para atender a 50 clientes en la caja:

3.5	1.8	2.3	0.7	5.2	0.9	0.9	0.9	3.0	1.1
1.2	2.3	1.7	3.2	1.7	0.4	1.4	0.7	1.2	0.7
1.6	0.3	1.0	1.0	0.5	0.6	2.8	2.4	0.3	3.1
0.8	1.2	1.7	1.2	0.2	4.0	2.5	1.9	0.8	1.2
0.2	1.3	0.6	0.6	1.8	0.7	1.5	1.3	1.4	1.1

a) Construya una tabla de frecuencia relativa con diez clases, que tenga el ancho de clase mínimo.
b) ¿Qué porcentaje de los datos cae en la primera clase?,
c) ¿y en la primera o en la última clase?

33. Para los datos del ejercicio 21, realice una tabla de frecuencia agrupada con siete clases que tenga el mínimo ancho posible, y en donde las marcas de clase terminen en 0.9 centavos.

34. Los datos siguientes representan los puntajes promedio, basados en un sistema de cinco puntos del último semestre para un grupo del penúltimo año de una escuela de psicología.

1.3	1.4	2.0	3.7	2.4	3.6	4.0	3.8
1.9	1.1	1.7	2.3	4.2	4.3	3.4	3.0
4.0	4.1	3.6	2.1	2.0	1.5	2.7	2.6
3.6	1.2	3.5	3.3	4.0	3.8	2.6	1.9

Construya una tabla de frecuencia agrupada que tenga ocho clases, con el mínimo ancho posible que concuerde con la unidad de medida y donde la primera clase empiece con 1.0.

SECCIÓN 2.3 Representación gráfica de datos

Una **gráfica** es una forma ilustrada de representar y resumir datos; a menudo, una representación de datos mediante ilustraciones hace más evidentes ciertas características que una tabla de frecuencia; un resultado de representar los datos en forma gráfica es que frecuentemente se descubren nuevas características de ellos; la presentación gráfica de los datos ha logrado un uso creciente en los medios de comunicación y eso se debe en parte, a la popularidad y uso de la graficación por computadora; hay gráficas de muchos tipos, las más usadas son la gráfica de pastel, la de barras, la lineal, el diagrama de tallo y hojas, el histograma y la ojiva. Discutiremos cada uno de ellos con algo de detalle.

Gráficas de barras y de pastel

Dos de los tipos de gráficas más comunes son las **gráficas de barras** y las **gráficas de pastel;** ambos tipos se usan generalmente para datos categóricos y nominales; las gráficas de pastel se usan sólo para representar partes de un total y son muy populares para visualizar información presupuestal.

Ambos tipos de gráficas se ilustrarán usando los datos de la tabla 2.16, que representa a los beneficiarios de los donativos hechos por ciudadanos estadounidenses en 1983.[11] La figura 2.3 nos muestra una gráfica de barras verticales para la información de la tabla 2.16.

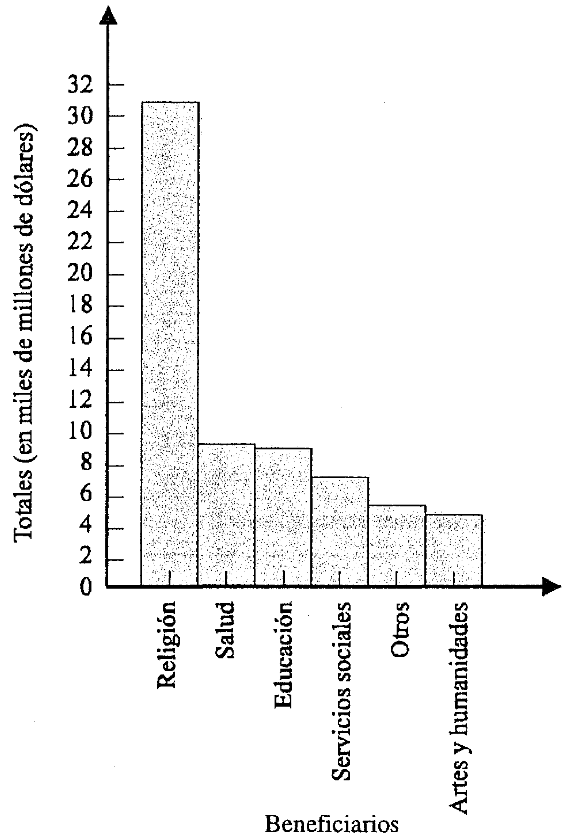
TABLA 2.16

Beneficiarios de donativos

Beneficiarios	Totales (en miles de millones de dólares)
Religión	31.0
Artes y humanidades	4.1
Servicios sociales	6.9
Educación	9.0
Salud	9.2
Otros	4.7

FIGURA 2.3

Gráfica de barras vertical de beneficiarios de donativos en 1983



La tabla 2.17 organiza los cálculos necesarios para construir una gráfica de pastel de los datos sobre los beneficiarios de los donativos mencionados; cada entrada en la columna de porcentajes se obtuvo dividiendo la cantidad entre el total (64.9) y multiplicando posteriormente por 100; las entradas en la columna de los grados se obtuvieron multiplicando las entradas en la columna de los porcentajes por 360, que es el número de grados en un círculo.

TABLA 2.17

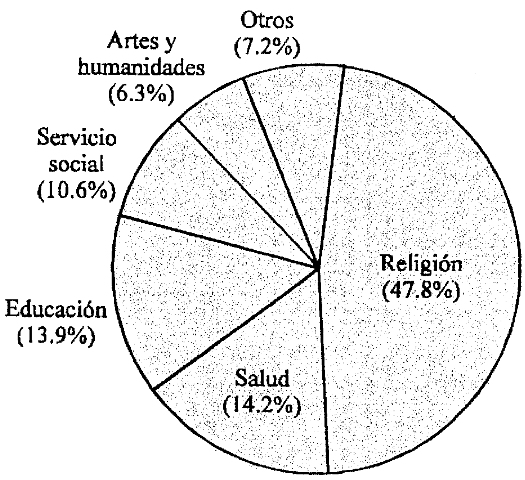
Cálculos para construir la gráficade pastel de los datos vistos en la tabla 2.16

Beneficiario	Cantidad	%	Grados	
Religión	31.0	47.8	$\boxed{172.1}$	← (0.478 × 360)
Salud	9.2	14.2	51.1	
Educación	9.0	13.9	50.0	
Servicios sociales	6.9	$\boxed{10.6}$	38.2	← $(\frac{6.9}{64.9})$ (100)
Artes y humanidades	4.1	6.3	22.7	
Otros	4.7	7.2	25.9	
Totales	64.9	100.0	360.0	

La gráfica de pastel en la figura 2.4 se construyó usando un transportador y la información de la tabla 2.17; en dicha gráfica podemos ver de una ojeada que la religión obtuvo la parte más grande, una cantidad aproximadamente igual al total de las cantidades restante, mientras que las artes y humanidades consiguieron la parte más pequeña.

FIGURA 2.4

Gráfica de pastel de los beneficiarios de donativos en 1983



Muchos programas computacionales para análisis de datos permiten trazar gráficas de pastel aun cuando se den porcentajes o datos sin agrupar.

Diagramas de tallo y hojas (d. de t. h.)

El uso de una tabla de frecuencia agrupada tiene una desventaja bastante obvia: los datos originales se pierden en el proceso de agrupamiento; para salvar esta limitación puede usarse **d. de t. h.** (**diagrama en forma de árbol**); los d. de t. h. ofrecen una forma novedosa y rápida de exhibir información numérica, si un numeral tiene dos o más dígitos, entonces se puede descomponer en una rama y una hoja. Un *tallo* es el primer dígito o parte, del numeral, mientras que una *hoja* está formada por él o los dígitos restantes. Por ejemplo, el numeral 278 puede descomponerse en dos formas:

2	78	27	8
↑	↑	↑	↑
tallo	hoja	tallo	hoja

La exhibición gráfica de datos es muy fácil de realizar usando tallos y hojas; cada dato aporta una hoja de algún tallo.

EJEMPLO 2.24

Construyamos un d. de t. h. para la colección de 25 calificaciones en un examen de álgebra:

78	67	65	87	75	65	71	54	94
64	84	82	81	68	85	76	89	
98	59	57	79	65	59	80	67	

Como todas las calificaciones caen entre 50 y 99, usemos los dígitos de las decenas en cada caso como el tallo y los de las unidades como la hoja.

Paso 1. Coloque los tallos en forma vertical usando un segmento de línea vertical, llamado tronco para separar los tallos de las hojas.

5	
6	
7	
8	
9	

Paso 2. Coloque cada hoja a la derecha de su tallo. Como la primera calificación es 78, colocamos la hoja 8 en su tallo 7.

	5		
	6		
Tallo ⟶	7	8 ⟵	Hoja
	8		
	9		

Si continuamos el proceso con cada calificación, obtendremos el d. de t. h.

5	9 7 4 9
6	4 5 7 5 7 8 5
7	8 6 1 9 5
8	5 4 2 9 7 1 0
9	8 4

Al observar el d. de t. h. anterior podemos concluir que:

1. La calificación más alta es 98.
2. La menor es 54.
3. Las calificaciones varían de 54 a 98.
4. El tallo 9 tiene menos hojas.
5. Los tallos 6 y 8 contienen más hojas, siete en cada uno.
6. El número total de hojas representa el tamaño de la muestra.

Note que no importa el orden en que las hojas se coloquen en un tallo; si se ordenan de menor a mayor, el d. de t. h. se llama un *d. de t. h. ordenado*. Un d. de t. h. ordenado para las calificaciones del examen de álgebra es como sigue:

5	4 7 9 9
6	4 5 5 5 7 7 8
7	1 5 6 8 9
8	0 1 2 4 5 7 9
9	4 8

Un d. de t. h. ordenado sirve para ordenar datos y calcular puntos de posición. Un **punto de posición** es un punto tal que un cierto porcentaje de datos cae antes que él. Por ejemplo, haciendo algunas sumas simples, uno puede observar rápidamente que el punto de posición del 40% para el conjunto de datos de las calificaciones de álgebra es 68; esto es: (0.40) (25) = 10, el punto de posición debajo del 68; un d. de t. h. ordenado constituye una representación gráfica correspondiente a una cierta tabla de frecuencia.

A veces puede ser deseable incluir menos de diez valores de hoja en una sola rama para desplegar los datos, sobre todo si pocas ramas contienen un gran número de hojas; cuando esto ocurre, tenemos una exhibición visual que corresponde a una tabla de frecuencia agrupada; cuando el número de clases crece, ciertas características relevantes de los datos pueden volverse más evidentes. Si cada tallo en un d. de t. h. se divide en dos tallos, llamados *subtallos*, y contienen el mismo número de valores correspondientes a hojas, el d. de t. h. que se obtiene se denomina un d. de t. h. de **doble tallo.** Las aplicaciones 2.10 y 2.11 ilustran modificaciones adicionales de los procedimientos de construcción para los d. de t. h.

EJEMPLO 2.25

Usemos la colección de 25 calificaciones del examen de álgebra para construir un d. de t. h. de doble tallo, donde cada tallo tenga cinco posibles valores de hojas; esto se lleva a cabo dividiendo cada tallo en dos subtallos, a y b; el subtallo a contendrá dígitos de 0 a 4 en calidad de hojas, y b incluirá dígitos de 5 a 9 como sus hojas. Por ejemplo, el subtallo 5a contendrá hojas del 0 al 4, y el 5b del 5 al 9. Ahora veamos el d. de t. h. de doble tallo resultante:

5a	4
5b	7 9 9
6a	4
6b	5 5 5 7 7 8
7a	1
7b	5 6 8 9
8a	0 1 2 4
8b	5 7 9
9a	4
9b	8

APLICACIÓN 2.10

Un estudio nacional sobre la utilidad de los reguladores de corriente, reveló que los costos de la energía eléctrica varían ampliamente a lo largo de Estados Unidos. Estos costos en las 25 ciudades más caras, medidos por el precio promedio en centavos y por kilowat/hora, en 1984 fueron:

16.5	14.3	14.3	13.9	13.8	11.2	11.1	11.1	10.8
13.1	12.8	12.1	12.0	11.8	10.8	10.8	10.7	
11.6	11.4	11.3	11.3	11.2	10.8	10.6	10.6	

Construya un d. de t. h. con estos datos.

Solución: Ignoraremos los puntos decimales; cada valor en el arreglo final puede llevar a su valor original multiplicando por 0.1. Así, trataremos los números como de tres dígitos comprendidos entre 106 y 165. Si usamos tallos de dos dígitos, obtenemos el siguiente diagrama ordenado:

10	6 6 7 8 8 8 8
11	1 1 2 2 3 3 4 6 8
12	0 1 8
13	1 8 9
14	3 3
15	
16	5

La hoja 5 en el tallo 16 representa 16.5 centavos. Podemos determinar fácilmente que un 20% de los costos promedio son superiores a 13.1 centavos; en esta aplicación no sería aconsejable usar hojas de dos dígitos y ramas de un dígito porque todas las hojas estarían en el mismo tallo, y ¿de qué serviría un d. de t. h. con un solo tallo? ■

La pantalla 2.1 ilustra el uso de MINITAB para construir un d. de t. h. para los datos de la aplicación 2.10.

Pantalla 2.1

```
MTB > SET C1
DATA > 16.5 14.5 14.3 13.9 13.8 13.1 12.8 12.1 12.0 11.8
DATA > 11.6 11.4 11.3 11.3 11.2 11.2 11.1 11.1 10.8 10.8
DATA > 10.8 10.8 10.7 10.6 10.6
DATA > END
MTB > STEM C1;
SUBC > INCREMENT = 1.

STEM-AND-LEAF OF C1    N=25
LEAF UNIT = 0.10

 7   10 6678888
(9)  11 112233468
 9   12 018
 6   13 189
 3   14 33
 1   15
 1   16 1
```

Note que STEM es una abreviatura del término inglés *stem-and-leaf diagram*, que corresponde a d. de t. h. Sólo es necesario que el usuario teclee las primeras cuatro letras de una orden. Vea también que la orden STEM-AND-LEAF tiene varias subórdenes; una de las subórdenes disponibles es INCREMENT, usada para especificar los tallos; para usar una de las subórdenes disponibles debemos escribir un punto y coma después de la orden; MINITAB responde con el símbolo de suborden, SUBC>. La suborden se escribe seguida de un punto, el cual informa a MINITAB que no siguen más subórdenes y que deben ejecutarse la orden principal y la suborden, si no se escribe el punto, MINITAB responde con otro símbolo de suborden, SUBC>.

MINITAB usa tres columnas en la respuesta; la columna del extremo izquierdo, llamada **de profundidad**, nos dice cuántas hojas están en la línea o acumuladas de esa línea hacia atrás o hacia adelante, dependiendo de qué extremo está más cercano. Por ejemplo, el 9 de la tercera línea contando de arriba a abajo, significa que hay nueve hojas en esa línea o debajo de ella; el 3, en la tercera línea partiendo de abajo significa que hay tres hojas en esa línea o debajo de ella. La línea con paréntesis contiene la observación central si el número total de observaciones es impar, y las dos observaciones centrales si el número total de observaciones es par. Los paréntesis encierran el número de hojas en esa línea. El segundo renglón de arriba hacia abajo contiene nueve valores y la mediana; siete valores caen arriba del segundo renglón y nueve valores caen después de éste. La segunda columna muestra los tallos, mientras que los números a la derecha del tallo son las hojas.

La pantalla 2.2 contiene un d. de t. h. de doble tallo para los datos de la aplicación 2.10. Note que la suborden INCREMENT = 1 no se usó en la pantalla 2.2 y que esto originó un d. de t. h. de doble rama.

Pantalla 2.2

```
MTB > STEM C1

STEM-AND-LEAF OF C1   N = 25
LEAF UNIT = 0.10

  7     10 6678888
 (7)    11 1122334
 11     11 68
  9     12 01
  7     12 8
  6     13 1
  5     13 89
  3     14 3
  2     14 5
  1     15
  1     15
  1     16
  1     16.5
```

APLICACIÓN 2.11 Los datos siguientes representan cambios porcentuales de un año, en el número de prisioneros en 25 prisiones federales y estatales.[12]

0.6	12.9	10.8	11.7	0.4	−11.1	0.6	2.5	0.2	−4.4	−1.4	−3.2	
−1.7	−1.2	7.0	−10.1	19.2	20.6	−0.5	9.8	2.1	16.3	8.8	20.8	4.1

Construya un d. de t. h. para los datos.

Solución: Si ignoramos el punto decimal, observamos que los datos corren de −111 a 208. Usemos valores de tallo de −1, −0, +0, 1 y 2; para lograr que todos los valores sean números de tres dígitos, pongamos un cero delante de los valores con dos dígitos; así 4.1 se representa por 041. Necesitamos dos tallos para cero para indicar los signos de los números. Por ejemplo, el tallo para el valor 0.6 es +0, el tallo para 7.0 es +0 y el tallo para −1.7 es −0. El valor 0.6 debe estar representado en el d. de t. h. como +006; 7.0 debe estar representado como +070 y −1.7 como −017. El d. de t. h. se muestra aquí:

−1	01 11
−0	05 12 14 17 32 44
+0	02 04 06 06 21 25 41 70 88 98
1	08 17 29 63 92
2	06 08

Recuerde que cada valor en el diagrama debe convertirse multiplicando por 0.1 antes de hacer la interpretación.

Las hojas en esta aplicación también pueden contener decimales; en ese caso, no necesitaríamos multiplicar los valores en el diagrama por 0.1 antes de realizar las interpretaciones que deban hacerse. El correspondiente d. de t. h. es:

−1	0.1 1.1
−0	0.5 1.2 1.4 1.7 3.2 4.4
+0	0.2 0.4 0.6 0.6 2.1 2.5 4.1 7.0 8.8 9.8
1	0.8 1.7 2.9 6.3 9.2
2	0.6 0.8

■

Los d. de t. h. también son útiles en otras aplicaciones, como a continuación se indica:

1. Se pueden comparar dos distribuciones similares si tienen los mismos tallos. En este caso, las hojas de un d. de t. h. pueden colocarse a la derecha de los tallos y las hojas del otro a la izquierda de las ramas, como se muestra aquí:

Hojas	Tallo	Hojas
8 6	5	3 6 8
9 8 7 3	6	2 7
8 6	7	3 5 5 7 6
5 4 3 0	8	3 3 4 5

2. Se pueden comparar más de dos distribuciones arreglando los diagramas en forma de columnas si comparten tallos comunes; los tallos se pueden colocar en el extremo izquierdo de un diagrama y las hojas asociarse como en el diagrama siguiente:

	Distribución 1	Distribución 2	Distribución 3
Tallo	Hojas	Hojas	Hojas
3	2 7	0	1 1
4	4 4 5	6 8	
5	7 9	3 4 7	5 6 8 9
6	3		3 4

Histogramas

Un **histograma** es un tipo de gráfica de barras para una distribución de frecuencia. Los histogramas pueden construirse para distribuciones de frecuencia agrupada y no agrupada. Consideraremos primero histogramas para distribuciones de frecuencia no agrupada.

Distribuciones de frecuencias no agrupadas

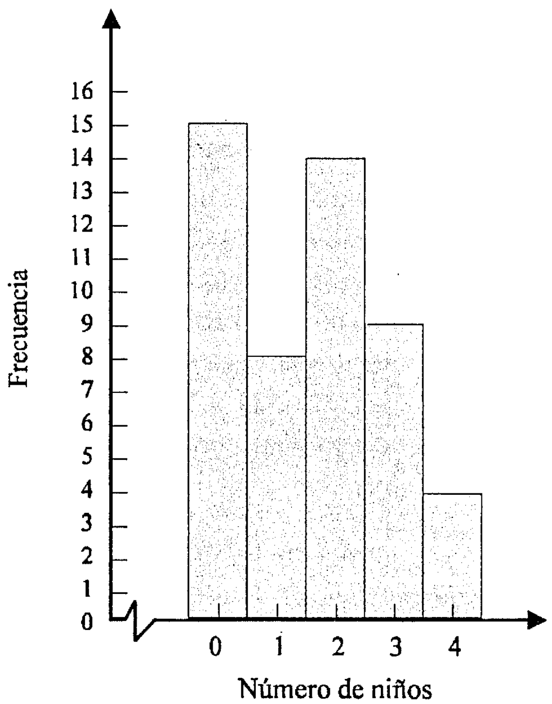
La idea de construir un histograma para frecuencia no agrupada de los datos, es representar cada frecuencia por una barra cuya área sea proporcional a ella. Típicamente, el ancho de cada barra se escoge como 1 y así el área de la barra es igual a la frecuencia de la medida.

APLICACIÓN 2.12

La tabla siguiente contiene el número de niños en edad escolar en cada una de las 50 familias de una muestra. Construya un histograma para datos.

Número de niños en edad escolar	Frecuencia (f)
0	15
1	8
2	14
3	9
4	4

FIGURA 2.5
Histograma para la frecuencia de los datos de la aplicación 2.12



Solución: Nuestro histograma contendrá cinco barras. Colocaremos el número de niños en edad escolar a lo largo del eje horizontal, la frecuencia a lo largo del eje vertical y el punto cero (0) en el eje horizontal, a la derecha de su posición usual, la intersección de los dos ejes. Esto nos permitirá centrar las barras sobre los valores de manera que el eje vertical no pase por la primera barra. Si escogemos el ancho de cada barra como 1 y la altura de cada barra como la frecuencia, entonces el área de cada barra será igual al producto de la frecuencia por 1. La suma de las áreas de las cinco barras será igual a la suma de las frecuencias; el histograma se ilustra en la figura 2.5. Advierta que el eje horizontal se rompe para llamar la atención sobre el hecho de que la escala horizontal no comienza en cero; rompemos el eje para indicar que no estamos tratando de distorsionar la perspectiva deliberadamente. ■

Histogramas para frecuencias agrupadas por datos

Para construir un histograma para datos medidos en una escala de intervalo o en una escala de razón, se acostumbra seguir dos pasos.

1. Se organizan los datos de una tabla de frecuencia agrupada.
2. Se construye una gráfica de barras usando las fronteras de clase para colocar las barras, y las frecuencias para indicar las alturas de las barras.

APLICACIÓN 2.13

La tabla de frecuencias agrupadas siguiente representa la tasa de desempleo, en porcentajes, para 27 ciudades del este.[13]

Tasa de desempleo (en porcentajes)	Número de ciudades
3.7-5.1	5
5.2-6.6	12
6.7-8.1	6
8.2-9.6	1
9.7-11.1	0
11.2-12.6	1
12.7-14.1	2
	27

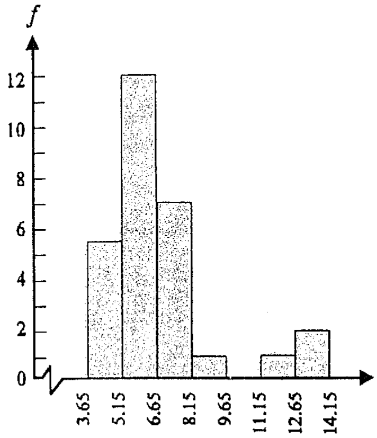
Realice un histograma con estos datos.

Solución: El histograma se construye colocando primero las fronteras de clase en el eje horizontal y las frecuencias en el eje vertical. Se traza una barra rectangular para cada clase usando las fronteras de clase para medir el ancho de la barra y la frecuencia para medir la altura; como todas las clases en una tabla de frecuencia agrupada tienen el mismo ancho, las áreas de las barras serán proporcionales a su altura; es decir, a las frecuencias de las clases. Para construir el histograma seguimos estos pasos:

Paso 1. Primero calculamos las fronteras de clase. Note que la unidad de medida es 0.1 de porcentaje, por lo tanto, para cada clase se resta: (0.5) (0.1) = 0.05 del límite inferior de clase para encontrar la frontera inferior de clase y se añade 0.05 al límite superior de clase para encontrar la frontera superior de clase.

FIGURA 2.6

Histograma para los datos de desempleo



Clase	Fronteras	f
3.7-5.1	3.65-5.15	5
5.2-6.6	5.15-6.65	12
6.7-8.1	6.65-8.15	6
8.2-9.6	8.15-9.65	1
9.7-11.1	9.65-11.15	0
11.2-12.6	11.15-12.65	1
12.7-14.1	12.65-14.15	2
		27

Paso 2. Se traza luego una gráfica de barras usando las fronteras de clase y las frecuencias; las fronteras se colocan a lo largo del eje horizontal y las frecuencias a lo largo del eje vertical, como lo muestra la figura 2.6. ■

Un histograma mejora nuestra habilidad para comparar las frecuencias de clase correspondientes; se puede comparar con facilidad la frecuencia de una clase con las de las clases vecinas; podemos ver inmediatamente que la segunda clase del histograma ilustrado en la figura 2.6 tiene la mayor frecuencia, y que la frecuencia de esa clase es el doble de la que está representada en la tercera clase; hay una declinación rápida en el número de ciudades representadas en las clases cuya tasa de desempleo está sobre el 8.15 y el 11.5 por ciento.

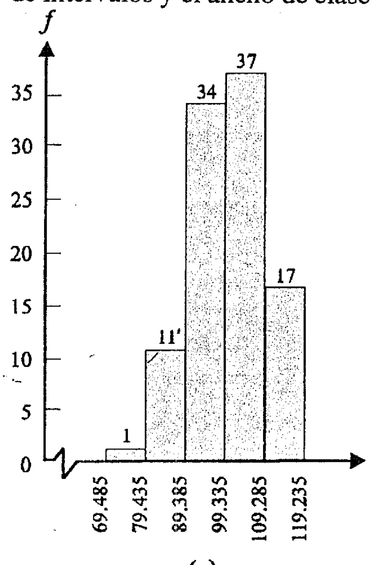
La forma de un histograma puede cambiar drásticamente con una variación en el número de intervalos n o en la amplitud de los intervalos w. Por esta razón, debemos ser cuidadosos al sacar conclusiones usando la forma de las distribuciones muestrales.

EJEMPLO 2.26

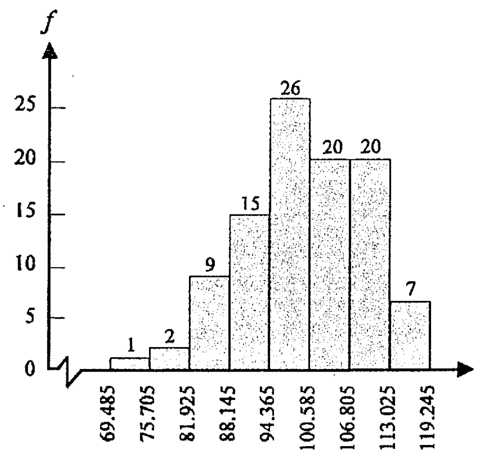
Los tres histogramas mostrados en la figura 2.7, representan una muestra de 100 medidas para valores distintos de n y w. El histograma en la parte (a) tiene $n = 5$ y $w = 9.95$; el de la parte (b) tiene $n = 8$ y $w = 6.22$; y el histograma en la parte (c) tiene $n = 5$ y $w = 4.60$. Advierta cómo cambia la apariencia cuando cambia el número de intervalos y el ancho de clase.

FIGURA 2.7

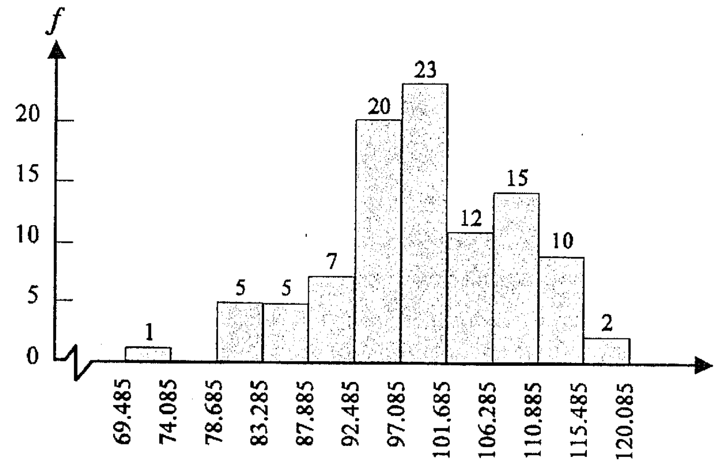
Efectos en la apariencia del histograma al cambiar el número de intervalos y el ancho de clase. El eje horizontal representa longitudes en pulgadas



(a)



(b)



(c)

Se puede usar MINITAB para construir un histograma de datos sin agrupar. La pantalla 2.3 contiene un histograma hecho con MINITAB para los datos adjuntos, que representan las edades en años de una muestra de 40 turistas que viajaron recientemente a Japón por American Airlines durante un periodo de un mes.

67	18	63	74	28	44	60	69	44	66
36	26	50	34	44	41	58	68	43	51
62	43	54	63	71	62	54	65	61	52
60	61	45	66	80	72	61	57	65	70

Pantalla 2.3

```
MTB > SET C1
DATA > 67 18 63 74 28 44 60 69 44 66
DATA > 36 26 50 34 44 41 58 68 43 51
DATA > 62 43 54 63 71 62 54 65 61 52
DATA > 60 61 45 66 80 72 61 57 65 70
DATA > END
MTB > HIST C1;
SUBC > INCREMENT = 8;
SUBC > START = 21.5.

HISTOGRAM OF C1  N=40

MIDPOINT    COUNT
  21.50         1 *
  29.50         2 **
  37.50         3 ***
  45.50         6 ******
  53.50         6 ******
  61.50        12 ************
  69.50         8 ********
  77.50         2 **
```

> Note que la suborden INCREMENT = 8 especifica que el ancho de cada intervalo de clase va a ser 8, y la suborden START = 21.5 indica que el punto medio de la primera clase va a ser 21.5, por lo tanto, la primera clase comienza con 17 y el primer intervalo de clase es 18-25. MINITAB no muestra los intervalos, sólo los puntos medios o marcas de cada uno.

Histogramas de frecuencia relativa

Se puede construir un **histograma de frecuencia relativa** cambiando la escala vertical de un histograma de frecuencias. En lugar de empezar con una tabla de frecuencia agrupada comenzamos con una tabla de frecuencia relativa agrupada; la altura de las barras en un histograma de esta naturaleza indicará la *proporción* del total representado por cada clase. Su forma básica se parece a la del histograma de frecuencia correspondiente.

EJEMPLO 2.27

La tabla de frecuencia relativa correspondiente a los datos de la aplicación 2.13 se muestra en la tabla 2.18, y el histograma de frecuencia correspondiente aparece en la figura 2.8.

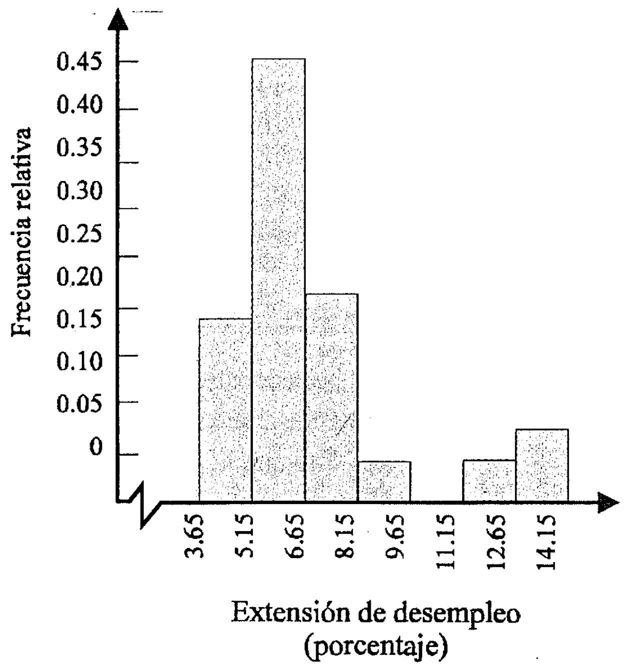
TABLA 2.18

Tabla de frecuencia relativa para los datos de desempleo

Clase	Fronteras	f	f relativa
3.7-5.1	3.65-5.15	5	0.19
5.2-6.6	5.15-6.65	12	0.44
6.7-8.1	6.65-8.15	6	0.22
8.2-9.6	8.15-9.65	1	0.04
9.7-11.1	9.65-11.15	0	0.00
11.2-12.6	11.15-11.65	1	0.04
12.7-14.1	12.65-14.15	2	0.07
		27	1.00

FIGURA 2.8

Frecuencias relativas para los datos de desempleo de la aplicación 2.13



Gráficas lineales o polígonos de frecuencia

Una **gráfica lineal** o **polígono de frecuencia** se construye usando una tabla de frecuencia agrupada con marcas de clase. La gráfica de líneas ofrece una alternativa útil respecto al histograma; la elección de cuál se usará es generalmente de tipo personal; una gráfica lineal crea la impresión de que las frecuencias cambian más suavemente, mientras que un histograma sugiere que las frecuencias cambian abruptamente; puede construirse una gráfica lineal o un polígono de frecuencia para los datos exhibidos, en una tabla de frecuencia agrupada identificando cada marca de clase y su correspondiente frecuencia (X, f) con un punto de la gráfica. Estos puntos se unen formando una sucesión de segmentos, como se ve en la aplicación 2.14.

APLICACIÓN 2.14

La tabla de frecuencia agrupada siguiente reporta los ingresos anuales promedio, hasta los 100 más cercanos, de los trabajadores fabriles en 27 ciudades del este de Estados Unidos.[14] Construya un polígono de frecuencia para esos datos.

Ingreso promedio	Número de ciudades
$12,500-14,300	1
14,400-16,200	5
16,300-18,100	3
18,200-20,000	7
20,100-21,900	6
22,000-23,800	1
23,900-25,700	3
25,800-27,600	1

Solución:

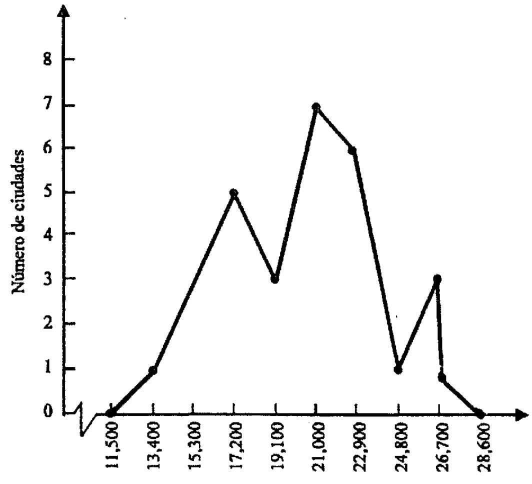
Paso 1. Encontramos primero las marcas de clase, designadas por X.

Ingreso promedio	f	X
$12,500-14,300	1	13,400
14,400-16,200	5	15,300
16,300-18,100	3	17,200
18,200-20,000	7	19,100
20,100-21,900	6	21,000
22,000-23,800	1	22,900
23,900-25,700	3	24,800
25,800-27,600	1	26,700

Paso 2. Ahora construimos la gráfica de líneas mostrada en la figura 2.9. Las marcas de clase se colocan en el eje horizontal y las frecuencias en el eje vertical. Note que la gráfica de líneas se "baja" en ambos extremos, concertando el primero y el último puntos a puntos del eje horizontal que distan w = 1900 de las marcas de clase más cercanas.

FIGURA 2.9

Ingresos promedio de trabajadores fabriles



Las características notables de los datos están exhibidas en la gráfica lineal de la figura 2.9.

1. La mayoría de las ciudades caen entre los extremos de la escala. Sólo una ciudad tiene trabajadores fabriles con un ingreso promedio anual de aproximadamente 13,400 dólares y sólo una ciudad posee trabajadores que ganan un promedio anual de alrededor de 26,700 dólares.
2. Los datos parecen tener su centro aproximadamente en 19,000 dólares. ■

Ojivas

Una gráfica lineal construida a partir de una tabla de frecuencia acumulada o de una tabla de frecuencia relativa acumulada, se llama **ojiva**. Las ojivas ofrecen un medio gráfico para interpolar o aproximar el número o porcentaje de observaciones menores o iguales que un valor específico.

Para localizar los puntos de una ojiva, se usa una frontera superior de clase y su correspondiente frecuencia acumulada o frecuencia relativa acumulada; después se unen los puntos consecutivos por segmentos de recta; las frecuencias acumuladas o las frecuencias relativas acumuladas, se colocan siempre en el eje vertical. La aplicación 2.15 ilustra la construcción de una ojiva.

APLICACIÓN 2.15

Trace una ojiva usando frecuencias acumuladas para los datos de la aplicación 2.14.

Solución:

Paso 1. Primero encontramos las frecuencias acumuladas.

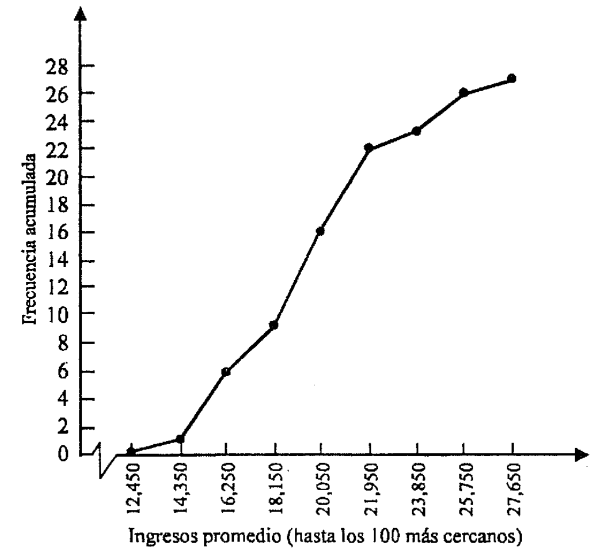
Ingreso promedio	Frontera superior	f	Frecuencias acumuladas
$12,500-14,300	14,350	1	1
14,400-16,200	16,250	5	6
16,300-18,100	18,150	3	9
18,200-20,000	20,050	7	16
20,100-21,900	21,950	6	22
22,000-23,800	23,850	1	23
23,900-25,700	25,750	3	26
25,800-27,600	27,650	1	27

Paso 2. Usamos las fronteras de las clases para marcar los puntos en el eje horizontal y las frecuencias para los puntos en el eje vertical.

Paso 3. Construimos la ojiva (figura 2.10). Vea que la frecuencia acumulada para la frontera inferior de la primera clase es 0. Podemos determinar de un vistazo el número de ciudades donde los trabajadores fabriles tienen ingresos promedio inferiores a una cantidad específica. ■

FIGURA 2.10

Una ojiva para los ingresos anuales de los trabajadores fabriles de 27 ciudades del este de Estados Unidos



Uso de ojivas para determinar percentiles

Se puede usar una ojiva de frecuencia relativa acumulada para determinar percentiles, como se describe en la aplicación 2.16.

APLICACIÓN 2.16

Construya una ojiva de frecuencia relativa acumulada para los datos de la aplicación 2.15, y úsela para aproximar el 50º percentil (P_{50}) y el 75º percentil (P_{75}). Recuerde que el 75º percentil es la medida por debajo de la cual cae el 75% de las medidas.

Solución:

Paso 1. Primero encontramos la frecuencia relativa acumulada usando la frecuencia acumulada.

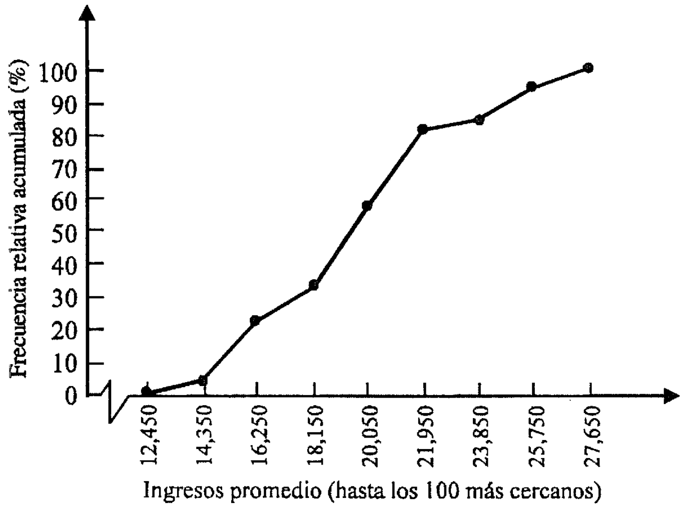
Ingreso promedio	Frontera superior	*f* acumulada	*f* relativa acumulada
$12,500-14,300	14,350	1	0.037
14,400-16,200	16,250	6	0.222
16,300-18,100	18,150	9	0.333
18,200-20,000	20,050	16	0.593
20,100-21,900	21,950	22	0.815
22,000-23,800	23,850	23	0.852
23,900-25,700	25,750	26	0.963
25,800-27,600	27,650	27	1.000

Paso 2. Usamos las fronteras de clase para colocar los puntos en el eje horizontal y la frecuencia relativa acumulada para los puntos en el eje vertical.

Paso 3. Construimos la ojiva (Fig. 2.11). Vemos que P_{50}, el 50º percentil, está entre \$18,150 y \$20,050 aproximadamente \$19,500 y que P_{75}, el 75º percentil, es un poco menos que \$22,000; por lo tanto, casi 50% de las ciudades tiene trabajadores fabriles que obtienen un ingreso promedio menor de \$19,500, y 75% de las ciudades cuentan con trabajadores fabriles con un ingreso promedio menor de \$22,000. Como resultado, aproximadamente 25% de los trabajadores fabriles ganan entre 19,500 y 22,000 dólares. ▪

FIGURA 2.11

Ojiva de ingresos de trabajadores fabriles



Histogramas, ojivas y formas de poblaciones

Los histogramas y las ojivas para datos muestrales proporcionan al investigador una idea de la forma de la población de la que se seleccionó la muestra. El histograma de una muestra sugiere la forma de la curva de frecuencia poblacional correspondiente; un histograma de frecuencia relativa para una muestra debe tener una forma asimilar a la de la distribución poblacional de frecuencia relativa, y una ojiva para una muestra debe tener aproximadamente la misma forma que la ojiva de la población. Como las poblaciones se representan a menudo por curvas de frecuencia relativa o por curvas de frecuencia relativa acumulada, es importante que entendamos sus contrapartes muestrales.

EJEMPLO 2.28

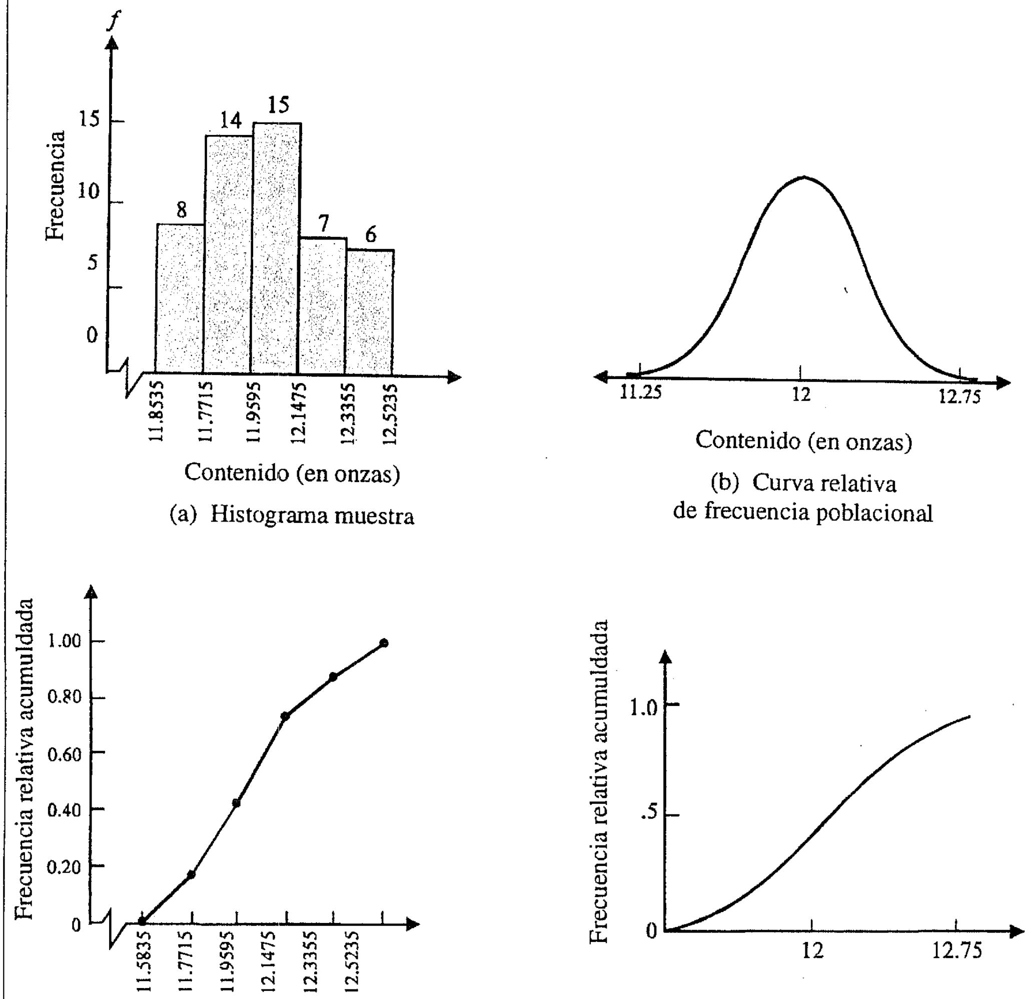
Suponga que un llenador automático de botellas en una fábrica de cerveza se programa para vertir 12 onzas del líquido en cada botella. Una muestra de 50 botellas proporciona los contenidos siguientes en onzas:

12.335	12.111	12.166	11.900	11.889	12.057	11.848	
12.151	11.717	11.584	12.497	12.083	12.018	11.704	
12.187	12.082	12.491	11.929	11.743	12.035	12.335	
12.520	11.988	12.080	12.001	11.990	11.748	12.103	
12.185	12.100	11.846	12.240	12.339	11.611	11.856	
11.629	11.912	11.786	11.853	11.655	12.101	11.886	
12.410	11.956	12.108	11.923	11.853	11.919	12.130	12.408

El histograma de la figura 2.12(a) ilustra la distribución del contenido de una botella para la muestra de 50 botellas de cerveza; este histograma nos aproxima a una población de forma acampanada, que es llamada *distribución normal;* esto lo estudiaremos con detalle en el capítulo 7 y se ilustra en la figura 2.12(b).

FIGURA 2.12

Histograma muestral, distribución poblacional y ojiva para los datos de contenido de cerveza



(a) Histograma muestra

(b) Curva relativa de frecuencia poblacional

(c) Curva de frecuencia relativa acumulada muestral

(d) Curva de frecuencia relativa acumulada poblacional

La ojiva de la muestra ilustrada en la figura 2.12(c), aproxima la forma de S de la distribución de frecuencia relativa acumulada para la distribución normal que apreciamos en la figura 2.12(d). Una distribución de frecuencia relativa acumulada para una población de forma acampanada o normal, tendrá siempre una forma con la apariencia de una S.

A menudo, los datos resultantes de un proceso o aplicación particular tendrán una forma conocida, como lo es una distribución acampanada; posteriormente veremos que esta información puede usarse para evaluar datos muestrales tomados de una población.

GRUPO DE EJERCICIOS 2.3

Habilidades básicas

1. Considere la muestra de calificaciones siguiente:

   A C D B C C C D F F
   D F A D C B C D D B

   Construya una
   a) gráfica de barras y
   b) una gráfica de pastel.

2. Considere la muestra de tasas de movimiento que siguen:

   X R X PG PG X X R PG13 G
   G G PG12 R R R G R G PG

   Trace:
   a) una gráfica de barras y
   b) una gráfica de pastel.

3. Construya un histograma de frecuencia para los datos listados a continuación; use seis barras:

17	14	16	8	31	16	14	9	17	11
25	24	28	10	48	24	12	13	43	24
32	37	33	42	11	34	16	41	21	15

4. Grafique un histograma de frecuencia que contenga 7 barras para los datos del inciso 3.

5. Dibuje un polígono de frecuencias para los datos del ejercicio 3 usando ocho puntos, incluidos los puntos finales.

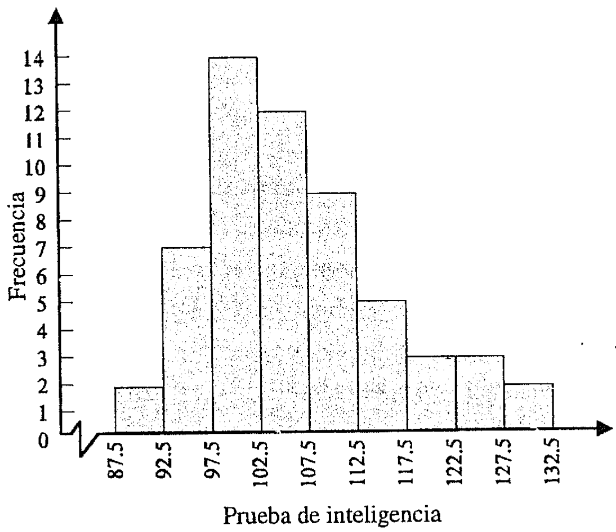
6. Construya un polígono de frecuencia para los datos del ejercicio 3, usando nueve puntos; incluya los puntos finales.

7. Trace un d. de t. h. para los datos adjuntos:

49	62	53	61	51	51	49	61	54	52
48	49	62	60	45	62	49	53	51	45
52	61	61	62	63	53	61	59	62	51
50	50	65	54	67	66	59	72	75	62

8. Haga un d. de t. h. con los datos del inciso 3.

9. Construya un histograma de frecuencias relativas para el histograma adjunto.



El puntaje es obtenido en una prueba de inteligencia de una selección aleatoria entre estudiantes de la escuela ABC.

10. Construya una ojiva correspondiente a la siguiente tabla de frecuencia agrupada:

Clase	f
1-4	14
5-8	18
9-12	12
13-16	16
17-20	20

11. Diga la clase de gráficas que son apropiadas para
    a) datos cualitativos y
    b) datos cuantitativos.

12. ¿Qué clase de gráficas son apropiadas para:
    a) datos nominales?,
    b) ordinales?,
    c) datos de intervalo y de razón?

Más aplicaciones

13. El número de calorías consumidas cada hora por una mujer de 130 libras en la realización de diez actividades, se muestran en la tabla siguiente:[15]

Actividad	Calorías consumidas por hora
Acostada, despierta	72
Sentada en reposo	95
Trabajando ante su escritorio	128
Vistiéndose, desvistiéndose	140
Caminando a 2.6 mph	190
En bicicleta a 5.5 mph	295
Jugando tenis	384
Nadando en crawl lento	450
Trotando, sin carga	425
Corriendo a 5.3 mph	550

Construya una gráfica de barras a partir de estos datos.

14. La tabla adjunta enlista el número de estudiantes de diez países distintos que estudiaron en universidades de Estados Unidos durante el año académico 1988-1989.[16]

País de origen	Número de estudiantes
China	29,040
Taiwán	28,760
Japón	24,000
India	23,350
Corea	20,610
Malasia	16,170
Canadá	16,030
Hong Kong	10,560
Irán	8,950
Indonesia	8,750

Trace una gráfica de barras con los datos de arriba.

15. De 100 pacientes hospitalizados: 30 tienen sangre tipo O, 38 tipo A, 22 tipo B y 10 tienen sangre tipo AB. Con estos datos construya:
a) una gráfica de barras;
b) una gráfica de pastel.

16. Los datos adjuntos indican las calificaciones finales en historia estadounidense para una muestra de 50 estudiantes.

A	B	B	B	C	D	D	D	A	A
C	C	C	B	B	C	D	D	D	B
C	C	D	D	D	D	F	A	A	C
B	C	B	B	A	A	F	C	C	D
D	F	F	A	A	C	C	B	B	C

Con estos datos, construya:
a) una gráfica de barras;
b) una gráfica de pastel.

17. La tabla siguiente muestra las diez operaciones de cirugía plástica más comunes.[17]

Operación	Número realizado
Cirugía de mano	160,000
Reparación de desgarres	150,000
Remoción de tumores	100,000
Aumento de senos	75,000
Accidentes industriales	70,000
Cirugía de párpados	57,000
Cirugía de nariz	55,000
Eliminación de quemaduras	45,000
Reconstrucciones	45,000
Faciales	40,000

Haga una gráfica de barras con la información dada.

18. Construya una gráfica de barras para los datos siguientes, que indican los precios de un pasaje, en dólares, para los sistemas de transporte subterráneo más utilizado en el mundo:[18]

Sistema subterráneo	Precio del pasaje
Moscú	1¢
Tokio	78¢
Nueva York	$1.15
México	11¢
París	87¢
Osaka, Japón	78¢
Leningrado	1¢
Londres	$1.17
Seúl	28¢
Hong Kong	32¢

19. Construya un d. de t. h. para los datos siguientes, que muestran los puntajes obtenidos por 20 estudiantes en un examen de inglés.

67	71	90	46	51	74	34	65	55	63
71	66	54	46	22	69	61	57	46	84

20. La información adjunta representa el rendimiento promedio en millas por galón estimado por la Agencia de Protección Ambiental (EPA) para 30 coches nuevos.

22	31	20	27	21	29	27	35	47	29
27	23	51	41	30	34	27	35	27	27
31	38	25	27	44	35	34	32	21	19

Efectúe un d. de t. h. con los datos citados.

21. Con respecto al ejercicio 20, construya un histograma de frecuencia relativa usando siete clases.

22. Construya un histograma de frecuencia relativa usando seis clases para los datos del inciso 19.

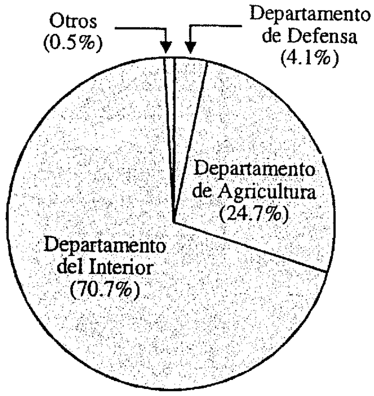
23. Realice un histograma para los datos que siguen; éstos presentan el número de automóviles vendidos por semana al año pasado por un agente de ventas.

Número de coches vendidos	Frecuencia (f)
0	4
1	17
2	12
3	13
4	2
5	3
6	0
7	1

24. Construya un histograma de frecuencia relativa con los datos del ejercicio anterior; ¿qué nota con respecto a las formas de este histograma y el del ejercicio 23? ¿En qué difieren?

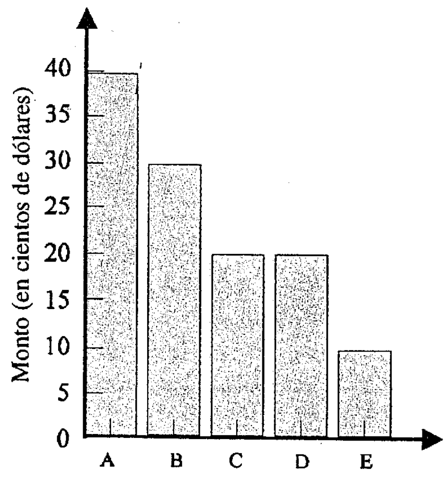
Un paso más allá

25. Convierta la gráfica de pastel adjunta en una gráfica de barras.



Territorio ocupado en 1975 por agencias federales, de un total de 761 millones de acres.

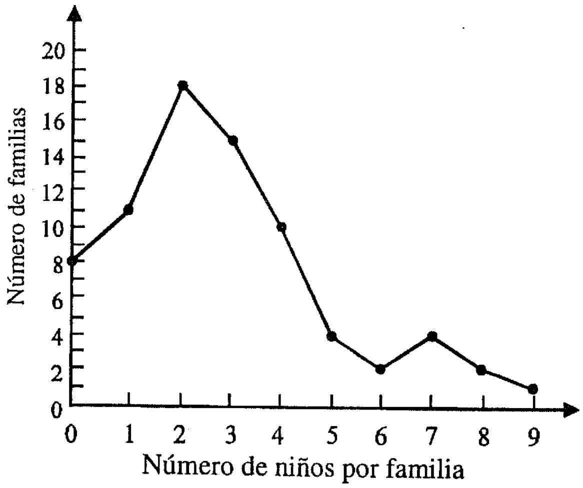
26. Convierta la gráfica de barras adjunta en una gráfica de pastel.



Ventas del último año (en cientos de dólares)

27. Remítase al histograma ilustrado en el ejercicio 9 y encuentre:
    a) El número total de puntajes de la prueba de inteligencia.
    b) Los puntajes de esa prueba que caen entre 97.5 y 102.5.
    c) El ancho de cada barra.
    d) El porcentaje de puntajes de pruebas que caen debajo de 117.5

28. Básese en la gráfica lineal adjunta para ubicar:
    a) El total de familias estudiadas.
    b) El total de niños estudiados.
    c) El porcentaje de familias que tienen cuatro niños.
    d) Familias que tienen menos de tres niños, en por ciento.



29. La tabla adjunta muestra el tiempo en horas y minutos que debieron trabajar los choferes de cinco ciudades para comprar hamburguesas, papas fritas y refrescos para una familia de cuatro miembros. Las cifras corresponden a los años 1979 y 1984.[19]

Ciudad	1979	1984
Chicago	0:58	1:02
Tokio	1:29	1:19
París	1:52	2:18
Dusseldorf	1:41	1.47
Londres	2:02	2:24

Construya una sola gráfica de barras que compare los tiempos de los choferes, para las cinco ciudades, y los dos años.

30. Construya una ojiva de frecuencia relativa para los datos descritos por la gráfica lineal del ejercicio 18.

RESUMEN DEL CAPÍTULO

En este capítulo aprendimos que los datos pueden clasificarse en cuantitativos y cualitativos; los primeros son susceptibles de clasificarse en discretos o continuos, dependiendo de que puedan contarse o no. También puede clasificarse a los datos según la escala de medición usada; las cuatro escalas usadas comúnmente son la nominal, la ordinal, la de intervalo y la de razón. Para organizar los datos se usan tablas y gráficas que permiten usarlos y entenderlos más fácilmente. Estudiamos también tablas de frecuencias agrupadas y no agrupadas; además, vimos que la distribución de los datos puede representarse gráficamente mediante gráficas de pastel, de barras, histogramas, d. de t. h., polígonos de frecuencias o gráficas lineales y ojivas.

REPASO DEL CAPÍTULO

■ TÉRMINOS IMPORTANTES ■

Los términos siguientes del capítulo estudiado se han mezclado para proporcionarle una mejor práctica de revisión. Dé una definición de cada uno con sus propias palabras y después verifique sus respuestas contra las definiciones del texto.

histograma de frecuencia relativa	frecuencia	datos ordinales
gráfica de barras	polígono de frecuencia	percentiles
tabla de frecuencia bivariada	tablas de frecuencia	datos cualitativos
gráfica de pastel	datos bivariados	datos cuantitativos
diagrama de doble tallo	gráfica	punto de posición
fronteras de clase	tabla de frecuencias agrupadas	datos de razón
datos	histograma	frecuencia relativa
marca de clase	datos continuos	tabla de frecuencia relativa
límites de clase	rango	diagrama de tallo y hojas
amplitud de clase	profundidad	regla de Sturges
frecuencia acumulada	datos de intervalo	marca de cuenta
datos sin agrupar	gráfica lineal	unidad
tabla de frecuencia relativa acumulada	datos nominales	datos univariados
tabla de frecuencia no agrupada	ojiva	datos discretos

■ SÍMBOLOS IMPORTANTES ■

f, frecuencia
w, amplitud de clase
c, número de clases
n, número total de medidas
R, rango
U, medida máxima
L, medida mínima
X, marca de clase
l_1, límite inferior de clase
l_2, límite superior de clase

■ HECHOS Y FÓRMULAS IMPORTANTES ■

Para una tabla de frecuencia agrupada todas las clases tienen la misma amplitud. Para una tabla de frecuencia agrupada, el número de clases debe estar entre 5 y 15, inclusive.

Regla de Sturges: el número de clases necesarias en una tabla de frecuencia agrupada es aproximadamente igual a $c = 3.3(\log n) + 1$, donde n es el número de medidas.

El ancho o amplitud de clase se encuentra dividiendo el rango entre el número de clases y redondeando el resultado al mínimo entero mayor que R/c.

Para una tabla de frecuencia agrupada, la primera clase siempre comienza con la medida mínima.

Las fronteras de clase se usan para construir histogramas y ojivas.

EJERCICIOS DE REPASO

1. Las estaturas en centímetros de 50 estudiantes mujeres de preparatoria, son las siguientes:

157	155	171	150	163	150	172	161
154	174	163	148	152	163	149	158
176	164	157	153	169	161	160	164
155	162	151	167	167	167	170	158
163	175	169	169	158	150	156	157
174	162	150	151	165	170	156	170
153	154						

a) Construya una tabla de frecuencia agrupada usando 10 clases.
b) Trace un d. de t. h.
c) Grafique una ojiva usando el resultado de la parte a.
d) Construya un histograma usando el resultado del inciso a.

2. Clasifique los datos siguientes como cuantitativos o cualitativos:
a) pesos en onzas de 20 manzanas;
b) colores de diez coches;
c) longitud en centímetros de una regla de 12 pulgadas;
d) preferencias religiosas de 15 personas;
e) calificaciones con letra de los estudiantes de equis clase;
f) calificaciones porcentuales de los estudiantes de una clase;
g) el sexo de 50 profesores;
h) la posición apagado/encendido de 30 interruptores de luz;
i) las calles en que viven 100 parientes;
j) talla de las camisetas de los miembros del equipo de futbol;
k) el número π (pi).

3. Para el ejercicio 2, clasifique los datos como nominales, ordinales, de intervalo o de razón.

4. En el ejercicio 2, clasifique los datos cuantitativos como discretos o continuos.

5. Se les pidió a 20 personas que identificaran su preferencia religiosa. Los resultados son:

C P P J J A J C P P
C J J C P P A P C J

donde C denota católico, P protestante, J judío y A ateo. Construya una
a) tabla de frecuencia;
b) gráfica de barras;
c) gráfica de pastel.

6. Dados sus límites, encuentre las amplitudes de las siguientes clases:

a) 7-16
b) 3.4-7.8
c) 1.3-4.5
d) 1.23-4.78
e) 0.03-0.09

7. Si una tabla de frecuencia agrupada debe contener ocho clases y la medida menor es 14 y la mayor 94, encuentre la amplitud de cada clase.

8. Considere la siguiente tabla de frecuencias bivariadas:

	Aprobado	Reprobado	Total
Hombre	11	15	26
Mujer	14	10	24
Total	25	25	50

Encuentre:
a) número de mujeres que aprobaron;
b) porcentaje de hombres que reprobaron;
c) porcentaje de aprobados que son hombres.

9. Se realizó un experimento para determinar el efecto de un cierto fármaco en los niveles de colesterol en la sangre, en mg/100 ml, en hombres de 30 años. Se obtuvieron las medidas:

245	185	230	225	265	210
245	165	195	170	205	225
160	240	285	175	260	225
235	120	145	185	195	210
190	220	140	215	195	

a) Construya un d. de t. h..
b) Haga una tabla de frecuencia agrupada con diez clases.
c) Trace un histograma de frecuencia relativa usando la tabla anterior.

10. Las estaturas, hasta la pulgada más cercana, de 33 estudiantes son las siguientes:

66	65	64	68	69	65	68	68	64	66	64
63	71	70	67	69	71	59	67	72	70	67
69	69	66	63	67	70	66	70	67	64	80

Construya un histograma de frecuencia agrupada que tenga ocho barras.

11. Grafique un d. de t. h. de doble tallo para los datos del ejercicio 10.

12. La siguiente tabla presenta la esperanza promedio de vida en Estados Unidos para los años 1950 y 1983.

	1950		1983	
	H	M	H	M
Recién nacidos	65.5	71.0	70.9	78.3
15 años de edad	68.6	73.5	72.2	79.3
25 años de edad	69.4	74.0	72.9	79.6
35 años de edad	70.2	74.5	73.7	80.0
45 años de edad	71.6	75.6	74.5	80.5
65 años de edad	77.7	80.0	79.5	83.8

a) Trace una gráfica de barras para la esperanza de vida promedio de los varones en 1983.
b) Construya una gráfica de barras para la esperanza de vida promedio de las mujeres en 1983.

13. Los datos siguientes representan los pesos en libras de una muestra de estudiantes en una preparatoria.

114	115	116	120	123	126	128	129	131
132	132	133	134	135	135	137	138	139
142	142	143	146	147	152	157	158	161
164	165	167	168	168	170	170	172	174
174	174	175	175	176	177	177	178	180
184	184	184	186	187	189	194	195	195
200	201	202	206	207	209			

a) Construya un d. de t. h.
b) Construya un d. de t. h. de doble tallo. ¿Revela este d. de t. h. algunas características de los datos que no fueron reveladas por el d. de t. h. de la parte a? Ofrezca una explicación para la diferencia de las formas.
c) Use un diagrama de la parte a, para construir un histograma de los datos sin agrupar.

14. La tabla de frecuencia adjunta contiene las velocidades en millas por hora, de una muestra de 60 coches que recorren la 14a. Avenida en Nueva York, según el registro del radar de un policía.

Clase	f
28-33	1
34-39	3
40-45	6
46-51	28
52-57	14
58-63	8

Construya una ojiva de frecuencia relativa.

Aplicaciones de computación

1. Los datos siguientes presentan los puntajes en matemáticas en el PAA, de una muestra de 100 estudiantes novatos de una universidad.

411	606	425	444	507	300	548	387	432
527	508	294	578	469	640	444	261	436
442	508	520	423	556	546	363	569	457
554	624	515	527	450	509	506	374	316
566	415	576	298	401	589	474	571	455
615	439	404	447	676	333	496	559	430
660	494	449	421	690	682	349	485	505
648	475	309	531	499	503	400	550	522
553	555	473	372	505	460	550	653	560
327	458	490	557	337	513	579	403	489
454	470	495	552	600	651	519	698	568
408								

Use un programa computacional para construir:

a) una tabla de frecuencia agrupada que tenga diez clases;

b) un d. de t. h. para los datos;

c) el histograma correspondiente a la tabla de frecuencia de la parte a.

2. Los datos siguientes muestran los puntajes de pruebas de inteligencia de una muestra de 100 estudiantes de 10º grado de la University High School:

132	103	94	78	108	105	98	114
89	112	95	82	86	124	118	120
87	120	107	95	104	100	81	91
94	99	89	93	86	98	122	78
117	115	91	90	97	79	97	104
71	95	86	87	97	107	78	149
124	87	80	71	92	80	106	121
123	117	114	90	109	90	72	86
75	94	94	83	100	86	105	116
95	83	93	109	116	128	94	69
134	111	116	94	135	88	88	102
130	99	94	73	93	98	110	80
120	92	99	80				

Use un programa computacional para construir

a) una tabla de frecuencia agrupadas que tenga doce clases.

b) un d. de t. h.

c) el histograma correspondiente a la tabla de frecuencia del inciso.

■ EXAMEN DE CONOCIMIENTOS DEL CAPÍTULO ■

Los datos siguientes indican los pesos en libras rebajados por un grupo de mujeres en las dos primeras semanas de un programa de ejercicios diarios:

1	2	12	3	15	5	12	11	3	4
3	5	0	7	17	6	17	13	2	5
5	7	1	11	3	9	9	8	18	9
10	9	4	12	1	8	8	7	11	9
15	11	8	4	5	11	3	14	12	10

Use el conjunto de datos para cubrir los ejercicios del 1 al 4.

1. Construya un d. de t. h.

2. Construya una tabla de frecuencia agrupada con cinco clases.

3. Trace una ojiva usando frecuencias relativas y la tabla construida en punto 2.

4. Grafique un histograma de frecuencia con cinco barras usando la misma tabla del ejercicio 2.

5. a) Encuentre el ancho de la clase 10-20, donde 10 y 20 son los límites de clase.

   b) Si $L = 32.1$, $U = 89.7$, $c = 5$ y la unidad de medida es 0.1, determine el límite superior de la primera clase.

3 Estadística descriptiva: análisis de datos univariados

DESCRIPCIÓN

3.1 Medidas de tendencia central y colocación
3.2 Medidas de dispersión o variabilidad
3.3 Tendencia central y dispersión para datos contenidos en tablas de frecuencia agrupada
3.4 Puntajes estándar y observaciones aberrantes

OBJETIVOS DEL CAPÍTULO

En este capítulo estudiaremos:

- *Cuatro medidas de tendencia central.*
- *Cómo calcular las medidas de tendencia central para datos agrupados y no agrupados.*
- *Cómo encontrar percentiles para datos agrupados y no agrupados.*
- *Ventajas y desventajas del uso de cada medida de tendencia central.*
- *El concepto de sesgo.*
- *El concepto de suma de cuadrados.*
- *Cuatro medidas de dispersión.*
- *Cómo calcular las medidas de dispersión para datos agrupados y no agrupados.*
- *Teorema de Chebichev.*
- *Puntajes estándar.*
- *Cómo construir y usar gráficas de caja.*

MOTIVADOR 1

Se sospecha que una gran universidad discrimina a las mujeres en sus políticas de contratación. Para investigar los cargos, una Comisión de Derechos Humanos determina el número de hombres y mujeres que han respondido a la oferta de puestos docentes durante un periodo de tres años, y calcula los porcentajes de hombres y mujeres calificados. Los datos obtenidos son los de la tabla siguiente:

	Hombres			Mujeres		
Área	Núm. de solicitudes	Núm. de aceptados	Porcentaje aceptado	Núm. de solicitudes	Núm. de aceptados	Porcentaje aceptado
Matemáticas	75	15	20	20	8	40
Química	100	35	35	30	21	70
Física	50	9	18	50	18	36
Inglés	25	1	4	400	32	8
Total	250	50	20	500	79	16

¿Proporcionan los datos evidencia de la discriminación contra las mujeres? Note que, para cada área, la proporción de mujeres contratadas duplica la proporción de hombres aceptados, pero los datos combinados de todas las categorías indican que la proporción de hombres es más alta. Después de que usted concluya el estudio del capítulo, será capaz de responder la pregunta inicial.

Panorama del capítulo

En el capítulo 2 presentamos los métodos para organizar los datos mediante tablas y gráficas. Esas técnicas representan medios visuales de descubrir relaciones, modos de comportamiento y tendencias en los datos; en este capítulo queremos complementar las interpretaciones visuales, hechas posibles por tablas y gráficas, con medidas numéricas de características poseídas por muchas colecciones de datos cuantitativos; dichas características incluyen el centro, la dispersión y los puntos de posición de un conjunto de datos.

SECCIÓN 3.1 Medidas de tendencia central y de colocación

Medidas de tendencia central

La primera característica de un conjunto de datos que deseamos medir es el centro o la tendencia central. El propósito de una **medida de tendencia central** es resumir un conjunto de datos de forma que podamos tener un panorama general; una medida tal sirve como representante del resto de la información. Una medida de tendencia central de un conjunto de datos proporciona también una idea del valor central de un conjunto aparentemente desorganizado de observaciones. Considere los cuatro ejemplos siguientes:

1. Pesos en libras: 5, 6, 12, 15 y 20.
2. Calificaciones para un examen: 31, 73, 78, 79, 80 y 81.
3. Colores de coches: tres blancos, cuatro rojos, siete negros y uno azul.
4. Puestos académicos: siete profesores, tres profesores asociados, dos profesores asistentes y diez instructores.

En los ejemplos 1 y 2, la escala usada es de razón; en el 3, nominal y en el 4, ordinal. ¿Qué medidas usaría usted para describir el valor central o para representar el conjunto de datos de cada ejemplo? Hay muchas medidas de tendencia central que se usan para encontrar un centro de un conjunto de datos; cuatro son las más comunes: la media, la mediana, la moda y el rango medio.

La **media** es el promedio aritmético.

La **mediana** es el puntaje ordenado medio.

La **moda**, si existe, es el puntaje más frecuente.

El **rango medio** es el promedio aritmético de las medidas mayor y menor.

Para describir las medidas centrales en los cuatro ejemplos que acabamos de dar, usaríamos la media para el ejemplo 1, la mediana para el ejemplo 2 y la moda para los ejemplos 3 y 4. Examinemos ahora en detalle cada medida de tendencia central y conozcamos las razones de la elección en los cuatro ejemplos.

Media

La media o promedio aritmético de un conjunto de números se encuentra sumando los números y dividiendo después la suma entre n, el número de medidas.

EJEMPLO 3.1

Los diez puntajes siguientes representan el número de puntos anotados en diez juegos de basquetbol por el jugador A: 6, 10, 3, 7, 6, 6, 8, 5, 9 y 10. La media es

$$\frac{6+10+3+7+6+6+8+5+9+10}{10}=\frac{70}{10}=7$$

El valor 7 representa, en algún sentido, el número central o "medio" de los puntos anotados en diez juegos por el jugador A.

La media se puede calcular tanto para muestras como para poblaciones, del mismo modo, pero se denotan en forma diferente; la media muestral se denota por $\bar{x}$ y la media poblacional por la letra griega μ (pronúnciese mu). Una fórmula para calcular la media de una muestra de datos numéricos está dada por

Media muestral

$$\bar{x}=\frac{\Sigma x}{n} \tag{3.1}$$

donde $\bar{x}$ denota la media muestral, x denota una medida de la muestra, Σx denota la suma de las medidas de la muestra y n es el tamaño de la muestra. Una fórmula para encontrar la media poblacional está dada por

Media poblacional

$$\mu=\frac{\Sigma x}{N}$$

donde μ es la media de la población y N es el tamaño de la población. Para usar esta fórmula debe conocerse la medida de cada elemento de la población (véase la aplicación 3.1), sin embargo, la media muestral no tiene sentido para todos los tipos de datos, como veremos en el ejemplo 3.2.

APLICACIÓN 3.1

Los totales anuales, en miles de millones de dólares, para las exportaciones agrícolas de Estados Unidos de 1974 a 1983 son: 21.9, 21.9, 23.0, 23.6, 29.4, 34.7, 41.2, 43.3, 39.1 y 33.7.[20] Determine la media si los datos constituyen una población.

Solución: La suma de las medidas es $\Sigma x = 311.8$. En consecuencia, la media poblacional es

$$\mu = \frac{\Sigma\, x}{N}$$

$$= \frac{311.8}{10} = 31.18$$

Por lo tanto, el total promedio de las exportaciones agrícolas en el periodo de diez años es 31.18 miles de millones de dólares. ▪

EJEMPLO 3.2

Suponga que hemos registrado el color de cabello de diez estudiantes de un colegio; la frase "color promedio de cabello" no tiene sentido, los datos de esta situación son cualitativos y la media se puede calcular sólo para datos cuantitativos.

En ocasiones muchas observaciones comparten valores comunes, como en las distribuciones de frecuencia no agrupada. Suponga que tenemos la muestra siguiente de edades en año de principiantes de una universidad:

18 18 18 18 19 19 19 20 20 21

Si aplicamos la definición de media muestral, la fórmula (3.1), a estos datos obtenemos:

$$\bar{x} = \frac{\Sigma\, x}{n} = \frac{190}{10} = 19$$

Para encontrar Σx, es más simple sumar los cuatro productos (4)(18), (3)(19), (2)(20) y (1)(21). Cada producto puede escribirse como fx, donde f es la frecuencia con que aparece una edad x (véase la tabla 3.1); la suma de los valores de f es igual a n y la suma de los valores de fx es igual $\Sigma\, x$.

Tabla 3.1
Datos y tabla de frecuencia para edades de 10 estudiantes de recién ingreso

Datos		Tabla de frecuencia		
x	Talla	x	f	fx
18	||||	18	4	72
19	|||	19	3	57
20	||	20	2	40
21	|	21	1	21
			10	190

La media muestral también es igual a $\bar{x} = \Sigma fx/\Sigma f = 190/10 = 19$.