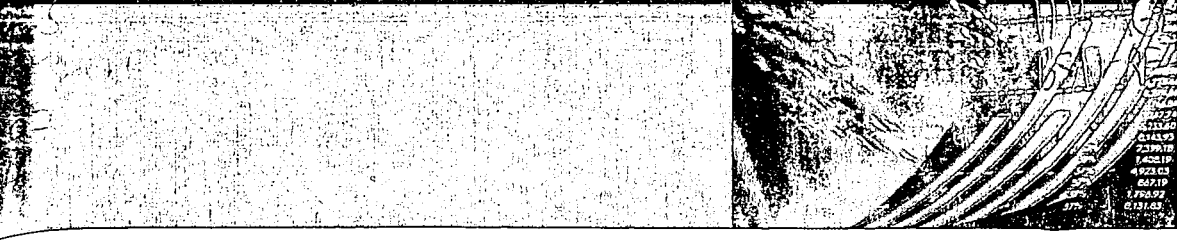


Introducción a la **Estadística**

Sheldon M. Ross

EDITORIAL REVERTÉ, S. A.

Barcelona • Bogotá • Buenos Aires • Caracas • México

Título de la obra original:
Introductory Statistics. Second Edition

Edición original en lengua inglesa publicada por
Elsevier Inc. of 525B Street, Suite 1900, San Diego, CA 92101-4495, USA

Versión española:

Equipo de traducción coordinado por

Prof. Dr. Teófilo Valdés Sánchez
Departamento de Estadística e Investigación Operativa
Facultad de Matemáticas
Universidad Complutense de Madrid

Propiedad de:
EDITORIAL REVERTÉ, S. A.
Loreto, 13-15. Local B
08029 Barcelona. ESPAÑA
Tel: (34) 93 419 33 36
Fax: (34) 93 419 51 89
reverte@reverte.com
www.reverte.com

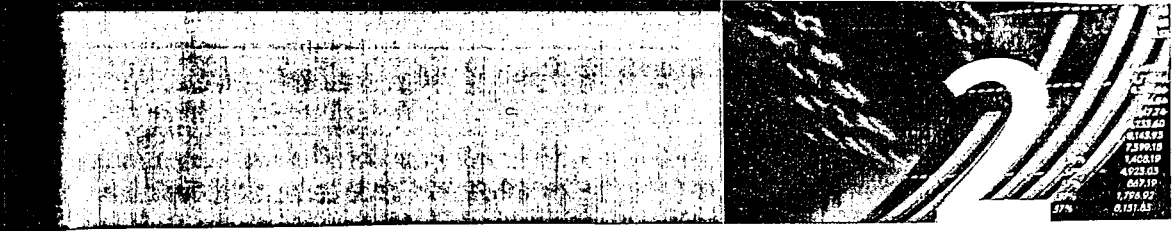
Edición en español:

© Editorial Reverté, S. A., 2007
ISBN: 978-84-291-5039-1

Depósito Legal: B-10495-2007

Impreso en España - Printed in Spain

Impreso por Alvagraf, S. L.
La Llagosta (Barcelona)



Descripción de los conjuntos de datos

Los números constituyen el único lenguaje universal.

Nathaniel West

La gente que no cuenta no cuenta.

Anatole France

2.1	Introducción	15
2.2	Tablas y gráficos de frecuencias	16
2.3	Datos agrupados e histogramas	28
2.4	Gráficos de tallos y hojas	40
2.5	Conjuntos de datos apareados	49
2.6	Comentarios históricos	56
	Términos clave	57
	Resumen	58
	Problemas de repaso	61

En este capítulo se aprenderán métodos para presentar y describir conjuntos de datos. Se introducirán distintos tipos de tablas y gráficos, que permitirán ver fácilmente las características clave de un conjunto de datos.

2.1 Introducción

Es muy importante que los resultados numéricos de cualquier estudio se presenten en forma clara y concisa, de modo que rápidamente se pueda tener una idea de las características esenciales de los datos. Esto es particularmente necesario cuando se trata de un amplio conjunto de datos, como frecuentemente ocurre en las encuestas o en los experimentos controlados. Realmente, una presentación efectiva de los datos a menudo revela con rapidez elementos tales como su categoría, su grado de simetría, lo concentrados o dispersos que están, dónde se concentran, etcétera. En este capítulo se tratarán distintas técnicas de presentación de datos, tanto tabulares como gráficas.

Las tablas y los gráficos de frecuencias que se presentan en la sección 2.2 incluyen una gran variedad de tablas y gráficos –gráficos de línea, gráficos de barras, y gráficos de polígono– que son útiles para describir conjuntos de datos que tienen un relativamente pequeño número de valores distintos. A medida que el número de valores distintos crece, éstos van dejando de ser efectivos, y es más conveniente dividir los datos en clases distintas para considerar solamente el número de valores que pertenecen a cada una de las clases. Esto se hace en la sección 2.3, donde se estudian los histogramas, un tipo de gráfico de barras que resulta de representar gráficamente las frecuencias de las clases. En la sección 2.4 se estudia una variación del histograma, conocida como gráfico de tallos y hojas, variación que utiliza los propios valores de los datos para representar los tamaños de las clases. En la sección 2.5 se considera la situación en la que los datos corresponden a pares de valores, como por ejemplo la población y la tasa de criminalidad de distintas ciudades, y se introduce el diagrama de dispersión como método efectivo de presentación de dichos datos. Finalmente, en la sección 2.6 se exponen algunos comentarios históricos.

2.2 Tablas y gráficos de frecuencias

Los siguientes datos representan los días de baja por enfermedad en las últimas 6 semanas de un grupo de 50 trabajadores de una cierta compañía.

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

Puesto que este conjunto de datos contiene un número relativamente pequeño de valores diferentes, conviene representarlo en una *tabla de frecuencias*, la cual incluye cada valor distinto junto con su frecuencia de ocurrencia. La tabla 2.1 es la tabla de frecuencias de los datos anteriores. En dicha tabla, la columna de frecuencias representa el número de ocurrencias de cada valor distinto del conjunto de datos. Observe que la suma de todas las frecuencias es 50, el número total de datos observados.

Utilice la tabla 2.1 para contestar a las preguntas siguientes:

(a) ¿Cuántos trabajadores han estado de baja por enfermedad al menos 1 día por enfermedad?

(b) ¿Cuántos trabajadores han estado de baja entre 3 y 5 días, ambos inclusive?

(c) ¿Cuántos trabajadores han han estado de baja más de 5 días?

Tabla 2.1 Tabla de frecuencias de los días de baja por enfermedad

Valor	Frecuencia	Valor	Frecuencia
0	12	5	8
1	8	6	0
2	5	7	5
3	4	8	2
4	5	9	1

Solución

(a) Puesto que 12 de los 50 trabajadores no estuvieron ningún día de baja, la respuesta es $50 - 12 = 38$.

(b) La respuesta es la suma de las frecuencias de los valores 3, 4 y 5; esto es, $4 + 5 + 8 = 17$.

(c) La respuesta es la suma de las frecuencias de los valores 6, 7, 8 y 9. Por tanto, la respuesta es $0 + 5 + 2 + 1 = 8$. ∎

2.2.1 Gráficos de líneas, gráficos de barras y polígonos de frecuencias

Se pueden mostrar gráficamente los datos de una tabla de frecuencias mediante un *gráfico de líneas*, en el que los valores sucesivos se representan sobre el eje horizontal y sus correspondientes frecuencias se representan mediantea la altura de una línea vertical. La figura 2.1 muestra el gráfico de líneas para los datos de la tabla 2.1.

En ocasiones, las frecuencias se representan no se representan mediante líneas sino mediante barras de una cierta anchura. Estos gráficos, llamados *gráficos de barras*, se utilizan muy a menudo. La figura 2.2 presenta un gráfico de barras que se corresponde con los datos de la tabla 2.1.

Otro tipo de gráfico utilizado para representar una tabla de frecuencias es el *polígono de frecuencias*, en el que se muestran gráficamente las frecuencias de los diferentes valores de los datos y luego se conectan los puntos del gráfico mediante líneas rectas. La figura 2.3 presenta el polígono de frecuencias de los datos de la tabla 2.1.

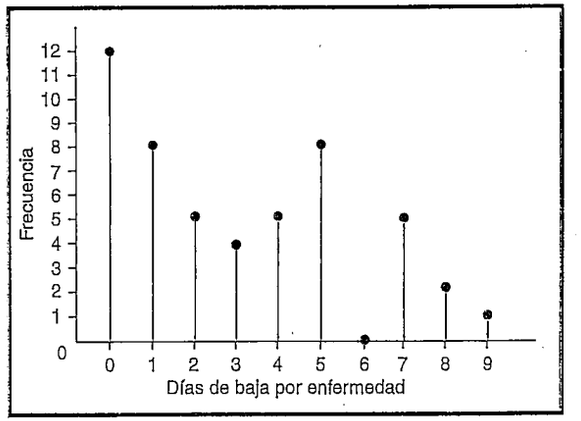


Figura 2.1 Gráfico de líneas.

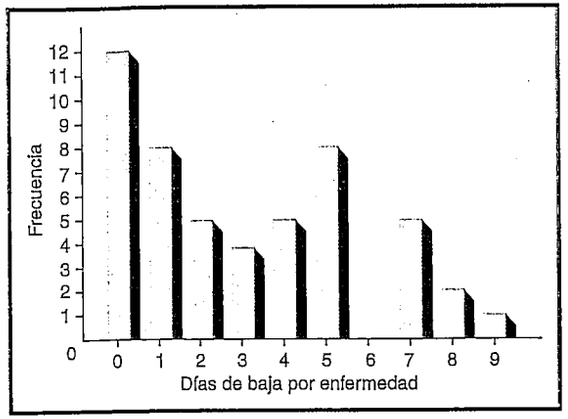


Figura 2.2 Gráfico de barras.

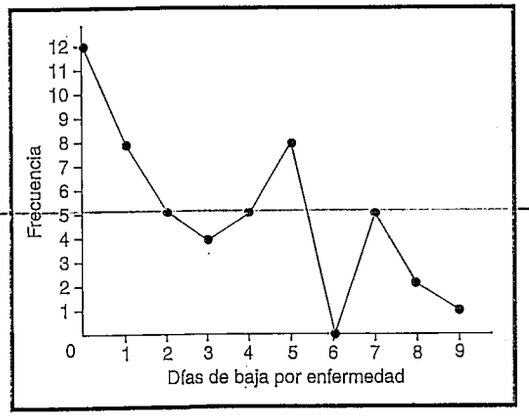


Figura 2.3 Polígono de frecuencias.

Se dice que un conjunto de datos es simétrico con respecto al valor x_0 si las frecuencias de los valores $x_0 - c$ y $x_0 + c$ son iguales para todo c. Es decir, para cada constante c, existe el mismo número de datos con un valor igual a c unidades por debajo de x_0 que con un valor igual a c unidades por encima de x_0. El conjunto de datos reflejado en la tabla de frecuencias de la tabla 2.2 es simétrico con respecto al valor $x_0 = 3$.

Los datos "próximos" a ser simétricos se dice que son *aproximadamente simétricos*. La forma más fácil de determinar si un conjunto de datos es aproximadamente simétrico consiste en representarlos gráficamente. La figura 2.4 incluye tres gráficos de barras: un conjunto de datos simétrico, un conjunto de datos aproximadamente simétrico, y el último, un conjunto de datos asimétrico.

Tabla 2.2 Tabla de frecuencias de un conjunto de datos simétrico

Valor	Frecuencia	Valor	Frecuencia
0	1	4	2
2	2	6	1
3	3		

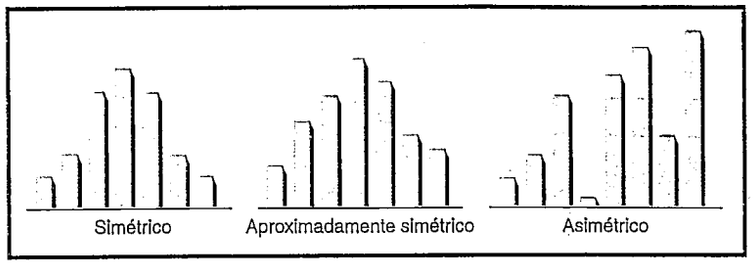


Figura 2.4 Gráfico de barras y simetría.

2.2.2 Gráficos de frecuencias relativas

En ocasiones, es más conveniente considerar y representar gráficamente las frecuencias *relativas* que las frecuencias *absolutas* de los datos. Si f representa la frecuencia de ocurrencia del valor x, se puede mostrar gráficamente la *frecuencia relativa* f/n frente a x, donde n representa el número total de observaciones del conjunto de datos. Para los datos de la tabla 2.1, $n = 50$ y las frecuencias relativas vienen reflejadas en la tabla 2.3. Observe que, mientras que la suma de la columna de frecuencias es igual al número total de observaciones del conjunto de datos, la suma de la columna de frecuencias relativas es 1.

En la figura 2.5 se presenta un polígono de frecuencias para las citadas frecuencias relativas. Un gráfico de frecuencias relativas tiene la misma apariencia que el gráfico análogo de frecuencias absolutas, aunque los valores del eje vertical se han dividido entre el número total de observaciones del conjunto de datos.

Para construir una tabla de frecuencias relativas de un conjunto de datos

Ordene el conjunto de datos de forma creciente en valores. Determine los valores distintos y sus frecuencias de ocurrencia. Liste los citados valores distintos junto con sus frecuencias f y sus frecuencias relativas f/n, donde n es el número total de observaciones del conjunto de datos.

Tabla 2.3 Frecuencias relativas de los datos de días de baja por enfermedad, $n = 50$.

Valor x	Frecuencia f	Frecuencia relativa f/n
0	12	$\frac{12}{50} = 0{,}24$
1	8	$\frac{8}{50} = 0{,}16$
2	5	$\frac{5}{50} = 0{,}10$
3	4	$\frac{4}{50} = 0{,}08$
4	5	$\frac{5}{50} = 0{,}10$
5	8	$\frac{8}{50} = 0{,}16$
6	0	$\frac{0}{50} = 0{,}00$
7	5	$\frac{5}{50} = 0{,}10$
8	2	$\frac{2}{50} = 0{,}04$
9	1	$\frac{1}{50} = 0{,}02$

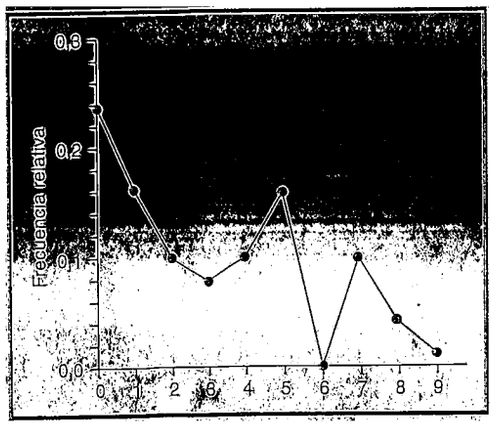


Figura 2.5 Polígono de frecuencias relativas.

Ejemplo 2.2 El Torneo de Maestros de Golf se juega cada año en Augusta (Georgia), en el Club Nacional de Golf. Para analizar las puntuaciones que han tenido los vencedores de este torneo, se han registrado las puntuaciones ganadoras desde 1968 hasta 2004.

Vencedores del Torneo de Maestros de Golf

Año	Vencedor	Puntuación	Año	Vencedor	Puntuación
1968	Bob Goalby	277	1987	Larry Mize	285
1969	George Archer	281	1988	Sandy Lyle	281
1970	Billy Casper	279	1989	Nick Faldo	283
1971	Charles Coody	279	1990	Nick Faldo	278
1972	Jack Nicklaus	286	1991	Ian Woosnam	277
1973	Tommy Aaron	283	1992	Fred Couples	275
1974	Gary Player	278	1993	Bernhard Langer	277
1975	Jack Nicklaus	276	1994	J.M. Olazabal	279
1976	Ray Floyd	271	1995	Ben Crenshaw	274
1977	Tom Watson	276	1996	Nick Faldo	276
1978	Gary Player	277	1997	Tiger Woods	270
1979	Fuzzy Zoeller	280	1998	Mark O'Meara	279
1980	Severiano Ballesteros	275	1999	J.M. Olazabal	280
1981	Tom Watson	280	2000	Vijay Singh	278
1982	Craig Stadler	284	2001	Tiger Woods	272
1983	Severiano Ballesteros	280	2002	Tiger Woods	276
1984	Ben Crenshaw	277	2003	Mike Weir	281
1985	Bernhard Langer	282	2004	Phil Nickelson	279
1986	Jack Nicklaus	279			

(a) Organice el conjunto de puntuaciones ganadoras mediante una tabla de frecuencias relativas.

(b) Represente estos datos mediante un gráfico de barras de frecuencias relativas.

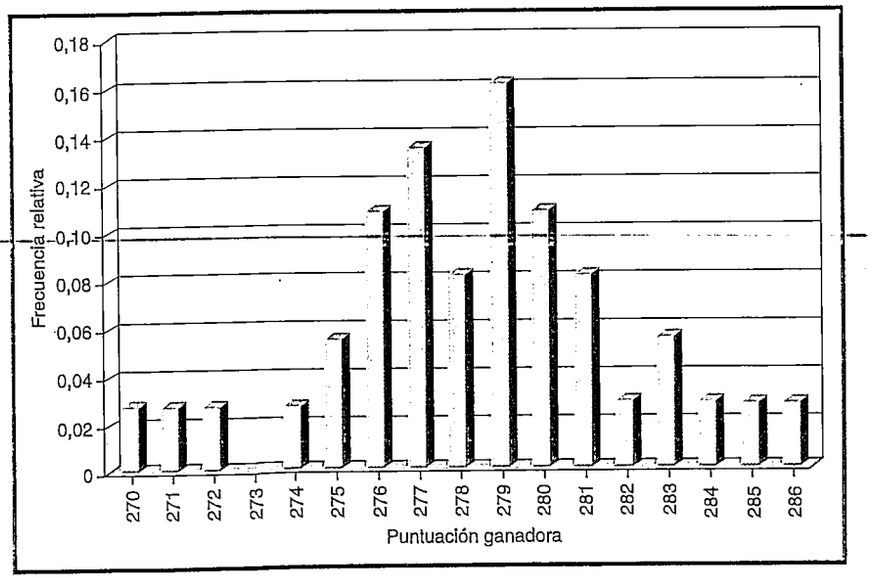
Solución

(a) Las 37 puntuaciones ganadoras varían desde la más baja de 270 hasta la más alta de 289. Su tabla de frecuencias relativas es la siguiente:

Puntuación ganadora	Frecuencia f	Frecuencia relativa $f/37$
270	1	0,027
271	1	0,027
272	1	0,027
274	1	0,.027

Puntuación ganadora	Frecuencia f	Frecuencia relativa $f/37$
275	2	0,054
276	4	0,108
277	5	0,135
278	3	0,081
279	6	0,162
280	4	0,108
281	3	0,081
282	1	0,027
283	2	0,054
284	1	0,027
285	1	0,027
286	1	0,027

(b) Un gráfico de barras de los datos anteriores es el siguiente:



2.2.3 Gráficos de tarta

Los *gráficos de tarta* se suelen utilizar para representar frecuencias relativas cuando los datos no son numéricos. Se construye un círculo que luego se divide en sectores, uno por cada valor diferente de datos. El área de cada sector, con la que se pretende representar la

Tabla 2.4 Armas utilizadas en los asesinatos.

Tipo de arma	Porcentaje de asesinatos causados con esta arma
Pistola de mano	52
Cuchillo	18
Escopeta	7
Rifle	4
Herramienta personal	6
Otras	13

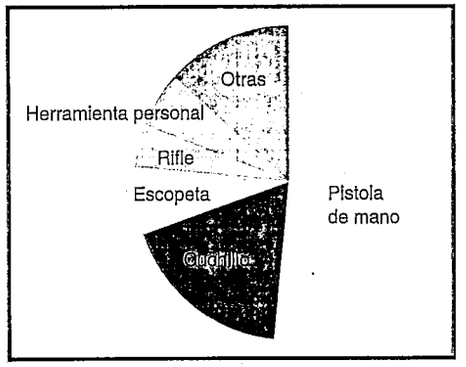


Figura 2.6 Gráfico de tarta.

frecuencia relativa de un valor, se determina como sigue. Si la frecuencia relativa del valor es f/n, el área de su sector debe coincidir con la fracción f/n del área total del círculo. Por ejemplo, los datos de la tabla 2.4 muestran las frecuencias relativas a las armas usadas en los asesinatos producidos en una gran ciudad durante 1985. Estos gráficos se representan mediante un gráfico de tarta en la figura 2.6.

Si un determinado valor tiene una frecuencia relativa f/n, su sector correspondiente puede obtenerse con la selección de un ángulo igual a $360f/n$ grados. Por ejemplo, en la figura 2.6, el ángulo del sector correspondiente al cuchillo como arma debe ser $360(0,18) = 64,8°$.

Problemas

1. Los siguientes datos representan los tamaños de 30 familias que residen en una pequeña ciudad de Guatemala.

   5, 13, 9, 12, 7, 4, 8, 6, 6, 10, 7, 11, 10, 8, 15,
   8, 6, 9, 12, 10, 7, 11, 10, 8, 12, 9, 7, 10, 7, 8

(a) Construya una tabla de frecuencias para estos datos.

(b) Represente estos datos mediante un gráfico de líneas.

(c) Represente gráficamente los datos mediante un polígono de frecuencias.

2. La siguiente tabla de frecuencias representa las ventas semanales de bicicletas de un comercio durante un periodo de 42 semanas.

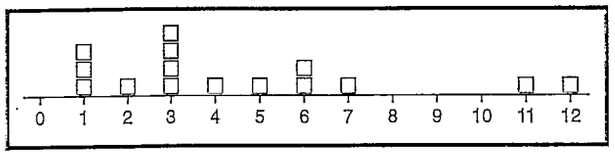
Valor	0	1	2	3	4	5	6	7
Frecuencia	3	6	7	10	8	5	2	1

(a) ¿En cuántas semanas se vendieron al menos 2 bicicletas?

(b) ¿En cuántas semanas se vendieron al menos 5 bicicletas?

(c) ¿En cuántas semanas se vendió un número par de bicicletas?

3. A 15 alumnos de cuarto curso se les preguntó a cuántas manzanas vivían de la escuela. Los resultados aparecen en el siguiente gráfico.



(a) ¿A qué número máximo de manzanas hasta la escuela se encuentra el domicilio de un alumno?

(b) ¿Cuál es el número mínimo de manzanas?

(c) ¿Cuántos alumnos viven a menos de 5 manzanas de la escuela?

(d) ¿Cuántos alumnos viven a más de 4 manzanas de la escuela?

4. Determine cuáles de los siguientes conjuntos de datos son simétricos, aproximadamente simétricos, o totalmente asimétricos.

A: 6, 0, 2, 1, 8, 3, 5
B: 4, 0, 4, 0, 2, 1, 3, 2
C: 1, 1, 0, 1, 0, 3, 3, 2, 2, 2
D: 9, 9, 1, 2, 3, 9, 8, 4, 5

5. La tabla siguiente lista los valores pero sólo algunas de sus frecuencias, para un conjunto de datos simétrico. Rellene las frecuencias que faltan.

Valor	Frecuencia
10	8
20	
30	7
40	
50	3
60	

6. Los siguientes valores representan las calificaciones de 32 estudiantes que se presentaron a un examen de Estadística.

55, 70, 80, 75, 90, 80, 60, 100, 95, 70, 75, 85, 80, 80, 70, 95,
100, 80, 85, 70, 85, 90, 80, 75, 85, 70, 90, 60, 80, 70, 85, 80

Represente estos datos mediante una tabla de frecuencias, y dibuje después un gráfico de barras.

7. Dibuje una tabla de frecuencias relativas para los datos del problema 1. Dibuje estas frecuencias relativas mediante un gráfico de líneas.

8. Los siguientes datos representan los tiempos de progresión, medidos en meses, de un tipo particular de tumor cerebral, llamado *glioblastoma*, en 65 pacientes:

6, 5, 37, 10, 22, 9, 2, 16, 3, 3, 11, 9, 5, 14, 11, 3, 1, 4, 6, 2, 7,
3, 7, 5, 4, 8, 2, 7, 13, 16, 15, 9, 4, 4, 2, 3, 9, 5, 11, 3, 7, 5, 9,
3, 8, 9, 4, 10, 3, 2, 7, 6, 9, 3, 5, 4, 6, 4, 14, 3, 12, 6, 8, 12, 7

(a) Construya una tabla de frecuencias relativas para este conjunto de datos.

(b) Represente gráficamente las frecuencias relativas mediante un polígono de frecuencias.

(c) ¿Es este conjunto de datos aproximadamente simétrico?

9. La siguiente tabla de frecuencias relativas se ha obtenido a partir de los datos registrados sobre el número mensual de operaciones de emergencia de apendicitis que se han llevado a cabo en un determinado hospital.

Valor	0	1	2	3	4	5	6	7
Frecuencia relativa	0,05	0,08	0,12	0,14	0,16	0,20	0,15	0,10

(a) ¿En qué proporción de meses ha habido menos de 2 operaciones de apendicitis de emergencia?

(b) ¿En qué proporción de meses ha habido más de 5 operaciones?

(c) ¿Es este conjunto de datos simétrico?

10. Las tablas y los gráficos de frecuencias relativas son particularmente útiles cuando se quieren comparar distintos conjuntos de datos. Los dos siguientes conjuntos de datos se refieren al número de meses que transcurrieron, en los primeros años de la epidemia, entre el diagnóstico y la muerte para dos muestras de pacientes con SIDA, varones y mujeres.

Hombres	15	13	16	10	8	20	14	19	9	12	16	18	20	12	14	14
Mujeres	8	12	10	8	14	12	13	11	9	8	9	10	14	9	10	

Represente en la misma gráfica los dos grupos de datos mediante polígonos de frecuencias. Utilice un color diferente para cada conjunto de datos. ¿Qué conclusión se puede sacar respecto a qué conjunto de datos que tiende a tener valores mayores?

11. Con los datos del ejemplo 2.2, determine la proporción de puntuaciones ganadoras del Torneo de Maestros de Golf que:

    (a) Son inferiores a 280

    (b) Son iguales o superiores a 282.

    (c) Están comprendidas entre 278 y 284, ambos inclusive

La tabla siguiente muestra el número medio de días de cada mes con al menos 0,01 pulgadas de lluvia en varias ciudades. Los problemas del 12 al 14 se refieren a ella.

Número medio de días con una precipitación de 0,01 pulgadas o más

Estado	Ciudad	Longitud de registro	Ene.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ago.	Sep.	Oct.	Nov.	Dic.	Anual
AL	Mobile	46	11	10	11	7	8	11	16	14	10	6	8	10	123
AK	Juneau	43	18	17	18	17	17	16	17	18	20	24	19	21	220
AZ	Phoenix	48	4	4	4	2	1	1	4	5	3	3	3	4	36
AR	Little Rock	45	9	9	10	10	10	8	8	7	7	7	8	9	103
CA	Los Angeles	52	6	6	6	3	1	1	1	0	1	2	4	5	36
	Sacramento	48	10	9	9	5	3	1	0	0	1	3	7	9	58
	San Diego	47	7	6	7	5	2	1	0	1	1	3	5	6	43
	San Francisco	60	11	10	10	6	3	1	0	0	1	4	7	10	62
CO	Denver	53	6	6	9	9	11	9	9	9	6	5	5	5	89
CT	Hartford	33	11	10	11	11	12	11	10	10	9	8	11	12	127
DE	Wilmington	40	11	10	11	11	11	10	9	9	8	8	10	10	117
DC	Washington	46	10	9	11	10	11	10	10	9	8	7	8	9	111
FL	Jacksonville	46	8	8	8	6	8	12	15	14	13	9	6	8	116
	Miami	45	6	6	6	6	10	15	16	17	17	14	9	7	129
GA	Atlanta	53	11	10	11	9	9	10	12	9	8	6	8	10	115
HI	Honolulu	38	10	9	9	9	7	6	8	6	7	9	9	10	100
ID	Boise	48	12	10	10	8	8	6	2	3	4	6	10	11	91
IL	Chicago	29	11	10	12	12	11	10	10	9	10	9	10	12	127
	Peoria	48	9	8	11	12	11	10	9	8	9	8	9	10	114
IN	Indianapolis	48	12	10	13	12	12	10	9	9	8	8	10	12	125
IA	Des Moines	48	7	7	10	11	11	11	9	9	9	8	7	8	107
KS	Wichita	34	6	5	8	8	11	9	7	8	8	6	5	6	86
KY	Louisville	40	11	11	13	12	12	10	11	8	8	8	10	11	125
LA	New Orleans	39	10	9	9	7	8	11	15	13	10	6	7	10	114
ME	Portland	47	11	10	11	12	13	11	10	9	8	9	12	12	128
MD	Baltimore	37	10	9	11	11	11	9	9	10	7	7	9	9	113

(Continuación)

Número medio de días con una precipitación de 0,01 pulgadas o más *(Continuación)*

Estado	Ciudad	Longitud de registro	Ene.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ago.	Sep.	Oct.	Nov.	Dic.	Anual
MA	Boston	36	12	10	12	11	12	11	9	10	9	9	11	12	126
MI	Detroit	29	13	11	13	12	11	11	9	9	10	9	12	14	135
	Sault Ste. Marie	46	19	15	13	11	11	12	10	11	13	13	17	20	165
MN	Duluth	46	12	10	11	10	12	13	11	11	12	10	11	12	134
	Minneapolis-St. Paul	49	9	7	10	10	11	12	10	10	10	8	8	9	115
MS	Jackson	24	11	9	10	8	10	8	10	10	8	6	8	10	109
MO	Kansas City	15	7	7	11	11	11	11	7	9	8	8	8	8	107
	St. Louis	30	8	8	11	11	11	10	8	8	8	8	10	9	111
MT	Great Falls	50	9	8	9	9	12	12	7	8	7	6	7	8	101
NE	Omaha	51	6	7	9	10	12	11	9	9	9	7	5	6	98
NV	Reno	45	6	6	6	4	4	3	2	2	2	3	5	6	51
NH	Concord	46	11	10	11	12	12	11	10	10	9	9	11	11	125
NJ	Atlantic City	44	11	10	11	11	10	9	9	9	8	7	9	10	112
NM	Albuquerque	48	4	4	5	3	4	4	9	9	6	5	3	4	61
NY	Albany	41	12	10	12	12	13	11	10	10	10	9	12	12	134
	Buffalo	44	20	17	16	14	12	10	10	11	11	12	16	20	169
	New York	118	11	10	11	11	11	10	10	10	8	8	9	10	121
NC	Charlotte	48	10	10	11	9	10	10	11	9	7	7	8	10	111
	Raleigh	43	10	10	10	9	10	9	11	10	8	7	8	9	111
ND	Bismarck	48	8	7	8	8	10	12	9	9	7	6	6	8	97
OH	Cincinnati	40	12	11	13	13	11	11	10	9	8	8	11	12	129
	Cleveland	46	16	14	15	14	13	11	10	10	10	11	14	16	156
	Columbus	48	13	12	14	13	13	11	11	9	8	9	11	13	137
OK	Oklahoma City	48	5	6	7	8	10	9	6	6	7	6	5	5	82
OR	Portland	47	18	16	17	14	12	9	4	5	8	13	18	19	152
PA	Philadelphia	47	11	9	11	11	11	10	9	9	8	8	9	10	117
	Pittsburgh	35	16	14	16	14	12	12	11	10	9	11	13	17	154
RI	Providence	34	11	10	12	11	11	11	9	10	8	8	11	12	124
SC	Columbia	40	10	10	11	8	9	9	12	11	8	6	7	9	109
SD	Sioux Falls	42	6	6	9	9	10	11	9	9	8	6	6	6	97
TN	Memphis	37	10	9	11	10	9	8	9	8	7	6	9	10	106
	Nashville	46	11	11	12	11	11	9	10	9	8	7	10	11	119
TX	Dallas-Fort Worth	34	7	7	7	8	9	6	5	5	7	6	6	6	78
	El Paso	48	4	3	2	2	2	4	8	8	5	4	3	4	48
	Houston	18	10	8	9	7	9	9	9	10	10	8	9	9	106
UT	Salt Lake City	59	10	9	10	9	8	5	5	6	5	6	8	9	91
VT	Burlington	44	14	12	13	12	14	13	12	12	12	12	14	15	154
VA	Norfolk	39	10	10	11	10	10	9	11	10	8	8	8	9	114
	Richmond	50	10	9	11	9	11	9	11	10	8	7	8	9	113
WA	Seattle	43	19	16	17	14	10	9	5	6	9	13	18	20	156
	Spokane	40	14	12	11	9	9	8	4	5	6	8	12	15	113
WV	Charleston	40	16	14	15	14	13	11	13	11	9	10	12	14	151
WI	Milwaukee	47	11	10	12	12	12	11	10	9	9	9	10	11	125
WY	Cheyenne	52	6	6	9	10	12	11	11	10	7	6	6	5	99
PR	San Juan	32	16	13	12	13	17	16	19	18	17	17	18	19	195

Fuente: Administración Atmosférica y Oceánica de Estados Unidos, *Datos climáticos comparativos.*

12. Construya una tabla de frecuencias relativas para el número medio de días de lluvia en enero en las diferentes ciudades. A continuación, represente gráficamente los datos mediante un polígono de frecuencias relativas.
13. Usando solamente los datos relativos a las 12 primeras ciudades de la lista, construya una tabla de frecuencias para el número medio de días de lluvia en los meses de noviembre y diciembre.
14. Usando sólo los datos que se refieren a las 24 primeras ciudades, construya una tabla de frecuencias relativas para el mes de junio y, por separado, otra para el mes de diciembre. Posteriormente, represente en un mismo gráfico los dos conjuntos de datos mediante polígonos de frecuencias relativas.
15. La tabla siguiente muestra el número de muertes que hubo en las carreteras británicas durante 1987 distribuidas por clases.

Clases	Número de muertes
Peatones	1699
Ciclistas	280
Motoristas	650
Conductores de automóviles	1327

Represente estos datos mediante un gráfico de tarta.

16. Los siguientes datos, sacados del *New York Times*, representan los porcentajes, por kilos de peso, de los distintos componentes de la basura de Nueva York. Represéntelos mediante un gráfico de tarta.

Materia orgánica (comida, desperdicios del jardín, madera, etc.)	37,3
Papel	30,8
Bultos (mobiliario, neveras, etc.)	10,9
Plástico	8,5
Cristal	5
Metal	4
Inorgánicos	2,2
Aluminio	0,9
Desperdicios peligrosos	0,4

2.3 Datos agrupados e histogramas

Como se ha visto en la sección 2.2, el uso de gráficos de barras o líneas es una forma bastante efectiva de representar las frecuencias de los diferentes valores. Sin embargo, en algunos conjuntos de datos el número de valores distintos es demasiado grande para que se puedan utilizar los gráficos citados. En su lugar, es posible clasificar dichos valores en grupos, o *intervalos de clase*, para luego representar gráficamente el número de datos que corresponden a cada clase. En la elección del número de intervalos de clase se debe ponderar entre: (i) elegir pocos a costa de perder mucha información sobre los datos reales de cada intervalo de clase, o (ii) elegir muchos, con lo que las frecuencias resultantes de cada

intervalo de clase pueden ser demasiado pequeñas para que se reconozcan los patrones de forma. Aunque lo más habitual suele ser entre 5 y 10 intervalos de clase, el número apropiado es una elección subjetiva, y uno puede, como es natural, probar distintos números de intervalos de clase para ver cuál de los gráficos resultantes revela más información sobre los datos. Es corriente, aunque no esencial, elegir intervalos de clase de igual longitud.

Los puntos inicial y final de cada intervalo de clase se llaman extremos del mismo. Nosotros utilizaremos el convenio de *inclusión por la izquierda*, lo que significa que el intervalo de clase incluye el extremo de la izquierda pero no el de la derecha. Por ejemplo, el intervalo 20-30 incluye los valores que son mayores *o iguales* que 20 y menores que 30.

Los datos de la tabla 2.5 representan los niveles de colesterol en la sangre de 40 estudiantes de primer curso de una cierta universidad. Antes de determinar las clases y sus frecuencias, es útil ordenar los datos de forma creciente, así se consiguen los 40 valores de la tabla 2.6.

Puesto que los datos varían entre el valor mínimo (171) y el máximo (227), el extremo de la izquierda de la primera clase debe ser menor o igual a 171, y el extremo de la derecha de la última clase debe ser mayor que 227. Podría elegirse tomar como primera clase el intervalo de 170 a 180, lo que nos lleva a tomar seis clases. La tabla 2.7 nos muestra las frecuencias (y también las frecuencias relativas) de los valores de datos que caen dentro de cada intervalo de clase.

Observación: Debido al convenio de inclusión por la izquierda, los valores iguales a 200 se colocarán dentro del intervalo con extremos 200 y 210, y no en el intervalo comprendido entre 190 y 200.

Un gráfico de barras en el que las barras sean adyacentes se llama *histograma*. El eje vertical de un histograma puede representar bien las frecuencias de los intervalos de clase o bien sus frecuencias relativas. En el primer caso, el histograma se llama *histograma de frecuencias*; en el segundo, se trata de un *histograma de frecuencias relativas*. La figura 2.7 presenta un histograma de frecuencias para los datos de la tabla 2.7.

Es importante saber que una tabla de frecuencias de intervalos de clase o un histograma basado en tal tabla no contiene toda la información del conjunto de datos originales. Ambas representaciones utilizan sólo el número de valores dentro de cada intervalo de clase, y no los valores reales de los datos. Así pues, aunque las tablas y los gráficos citados son un útil reflejo de los datos, el conjunto de datos originales se debe mantener *siempre*.

Tabla 2.5 Niveles de colesterol en la sangre

213	174	193	196	220	183	194	200
192	200	200	199	178	183	188	193
187	181	193	205	196	211	202	213
216	206	195	191	171	194	184	191
221	212	221	204	204	191	183	227

Tabla 2.6 Niveles de colesterol en la sangre en orden creciente

171, 174, 178, 181, 183, 183, 183, 184,187, 188, 191, 191, 191, 192, 193, 193, 193, 194, 194, 195, 196, 196, 199, 200, 200, 200, 202, 204, 204, 205, 206, 211, 212, 213, 213, 216, 220, 221, 221, 227

Tabla 2.7 Tabla de frecuencias de los niveles de colesterol en la sangre

Intervalos de clase	Frecuencias	Frecuencias relativas
170–180	3	$\frac{3}{40} = 0{,}075$
180–190	7	$\frac{7}{40} = 0{,}175$
190–200	13	$\frac{13}{40} = 0{,}325$
200–210	8	$\frac{8}{40} = 0{,}20$
210–220	5	$\frac{5}{40} = 0{,}125$
220–230	4	$\frac{4}{40} = 0{,}10$

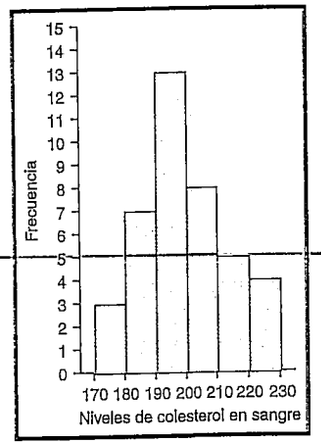


Figura 2.7 Histograma de frecuencias para los datos de la tabla 2.7.

Para construir un histograma a partir de un conjunto de datos

1. Ordene los datos en forma creciente.
2. Elija los intervalos de clase de manera que todos los datos aparezcan en alguno de ellos.
3. Construya una tabla de frecuencias.
4. Dibuje las barras adyacentes con alturas iguales a las frecuencias del paso 3.

La importancia de un histograma estriba en que permite organizar y presentar los datos gráficamente para que se pueda prestar atención a determinadas características importantes de los datos. Por ejemplo, un histograma puede a menudo indicar:

1. La simetría de los datos.
2. La dispersión de éstos.
3. Si existen intervalos que tienen un alto nivel de concentración de datos.
4. Si existen brechas entre los datos.
5. Si algunos valores de datos están muy separados de otros.

Por ejemplo, el histograma presentado en la figura 2.7 indica que las frecuencias de las sucesivas clases primero crecen y luego decrecen, y alcanzan un máximo en el intervalo de clase comprendido entre 190 y 200. Los histogramas de la figura 2.8 proporcionan una

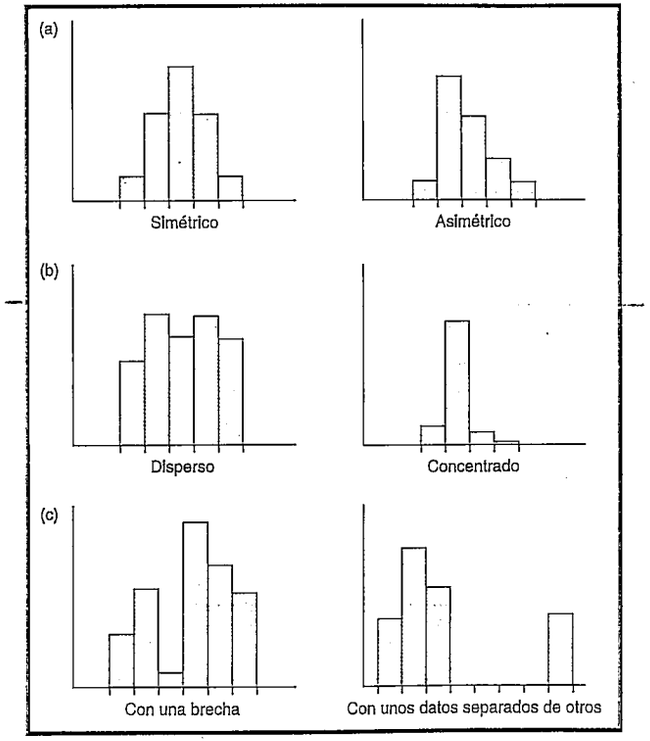


Figura 2.8 Características de los datos detectadas por los histogramas: (a) simetría, (b) grado de dispersión y dónde se concentran los valores, y (c) brechas en los datos y datos muy separados de otros.

información valiosa sobre los conjuntos de datos que representan. El conjunto de datos cuyo histograma se encuentra a la izquierda de la figura 2.8(a) es simétrico, mientras que el que se representa a la derecha no lo es. El conjunto de datos representado a la izquierda de la figura 2.8(b) se encuentra bastante disperso, mientras que el que se muestra a la derecha está más concentrado. El conjunto de datos representado a la izquierda de la figura 2.8(c) presenta una brecha, mientras que el representado al lado derecho tiene ciertos valores alejados del resto.

Ejemplo 2.3 La tabla 2.8 muestra las tasas de natalidad (por 1000 habitantes) en cada uno de los Estados de Estados Unidos. Represente gráficamente estos datos en un histograma.

Solución Puesto que los datos varían entre el valor más bajo, (12,4), y el más alto, (21,9), usaremos intervalos de clase de longitud 1,5, comenzando en el valor 12. Con estos intervalos de clase se obtiene la siguiente tabla de frecuencias.

Intervalos de clase	Frecuencias	Intervalos de clase	Frecuencias
12,0–13,5	2	18,0–19,5	2
13,5–15,0	15	19,5–21,0	0
15,0–16,5	22	21,0–22,5	2
16,5–18,0	7		

Un gráfico de histograma para estos datos se presenta en la figura 2.9.

Tabla 2.8 Tasas de natalidad por cada 100 habitantes

Estado	Tasa	Estado	Tasa	Estado	Tasa
Alabama	14,2	Louisiana	15,7	Ohio	14,9
Alaska	21,9	Maine	13,8	Oklahoma	14,4
Arizona	19,0	Maryland	14,4	Oregon	15,5
Arkansas	14,5	Massachusetts	16,3	Pennsylvania	14,1
California	19,2	Michigan	15,4	Rhode Island	15,3
Colorado	15,9	Minnesota	15,3	South Carolina	15,7
Connecticut	14,7	Mississippi	16,1	South Dakota	15,4
Delaware	17,1	Missouri	15,5	Tennessee	15,5
Florida	15,2	Montana	14,1	Texas	17,7
Georgia	17,1	Nebraska	15,1	Utah	21,2
Hawaii	17,6	Nevada	16,5	Vermont	14,0
Idaho	15,2	New Hampshire	16,2	Virginia	15,3
Illinois	16,0	New Jersey	15,1	Washington	15,4
Indiana	14,8	New Mexico	17,9	West Virginia	12,4
Iowa	13,1	New York	16,2	Wisconsin	14,8
Kansas	14,2	North Carolina	15,6	Wyoming	13,7
Kentucky	14,1	North Dakota	16,5		

Fuente: Departamento de Salud y Servicios Sociales.

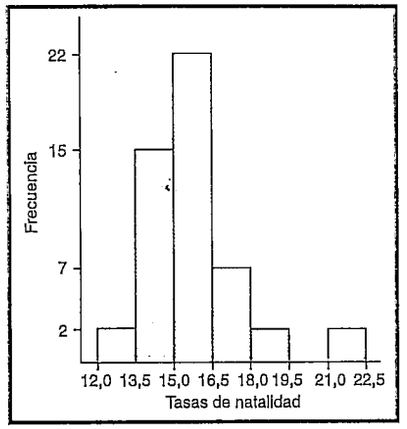


Figura 2.9 Histograma para las tasas de natalidad de los 50 Estados.

Un histograma es, en esencia, un diagrama de barras que muestra gráficamente las frecuencias o las frecuencias relativas de los datos que aparecen dentro de los distintos intervalos de clase. Dichas frecuencias de clase también se pueden representar gráficamente mediante polígonos de frecuencias (o de frecuencias relativas). Cada intervalo de clase es identificado por un valor, que generalmente coincide con el punto medio del intervalo. Después, estos valores se representan gráficamente frente a las frecuencias de los intervalos de clase que representan y los puntos del gráfico se conectan mediante líneas rectas para conseguir el polígono de frecuencias. Estos gráficos son especialmente útiles para comparar conjuntos de datos, puesto que en un mismo gráfico se pueden mostrar varios polígonos de frecuencias. □

Ejemplo 2.4 Los datos de la tabla 2.9 representan las frecuencias de los intervalos de clase para las presiones sistólicas sanguíneas de dos grupos de trabajadores industriales: aquellos cuya edad está comprendida entre 30 y 40 años, y aquellos cuya edad se encuentra entre 50 y 60 años.

Resulta difícil comparar directamente las presiones sanguíneas de los dos grupos de edad dado que el número total de trabajadores de cada grupo es diferente. Para salvar esta dificultad, se pueden computar y representar gráficamente las frecuencias *relativas* de cada una de las clases. Es decir, todas las frecuencias referidas a los trabajadores cuya edad varía entre 30 y 39 años se dividen entre 2540 (el número de dichos trabajadores) y todas las frecuencias referidas a los trabajadores con edades entre 50 y 59 años se dividen entre 731. Los resultados se muestran en la tabla 2.10.

La figura 2.10 es un gráfico de los polígonos de frecuencias relativas para ambos grupos de edad. Si se visualizan ambos polígonos de frecuencia en un mismo gráfico resulta fácil comparar los dos conjuntos de datos. Por ejemplo, aparentemente las presiones sanguíneas del grupo de mayor edad tienden a extenderse sobre valores más altos que los del grupo más joven.

Tabla 2.9 Frecuencias de clase de la presión sanguínea sistólica de dos grupos de trabajadores varones.

	Número de trabajadores	
Presión sanguínea	Edad entre 30-40 años	Edad entre 50-60 años
Menos de 90	3	1
90–100	17	2
100–110	118	23
110–120	460	57
120–130	768	122
130–140	675	149
140–150	312	167
150–160	120	73
160–170	45	62
170–180	18	35
180–190	3	20
190–200	1	9
200–210		3
210–220		5
220–230		2
230–240		1
Total	2540	731

Tabla 2.10 Frecuencias relativas de clase para las presiones sanguíneas.

	Porcentaje de trabajadores	
Presión sanguínea	Edad entre 30-40 años	Edad entre 50-60 años
Menos de 90	0,12	0,14
90–100	0,67	0,27
100–110	4,65	3,15
110–120	18,11	7,80
120–130	30,24	16,69
130–140	26,57	20,38
140–150	12,28	22,84
150–160	4,72	9,99
160–170	1,77	8,48
170–180	0,71	4,79
180–190	0,12	2,74
190–200	0,04	1,23
200–210		0,41
210–220		0,68
220–230		0,27
230–240		0,14
Total	100,00	100,00

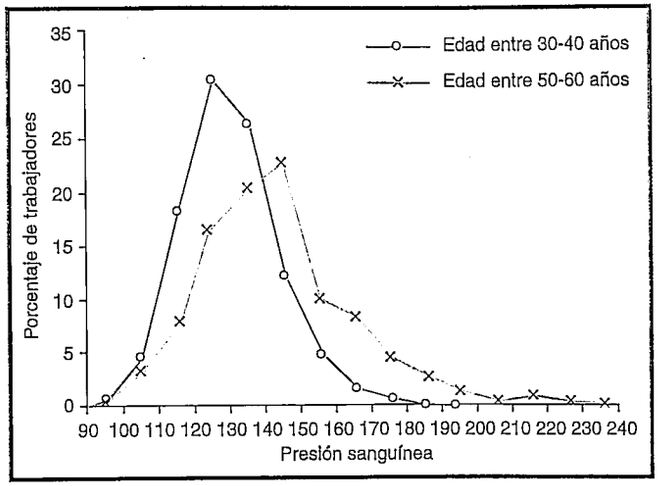


Figura 2.10 Polígonos de frecuencias relativas.

Problemas

1. Los siguientes conjuntos de datos representan las puntuaciones obtenidas por 40 estudiantes de sexto curso en un test de cociente de inteligencia (IQ, *Intelligence Quotient*) en una determinada escuela:

   114, 122, 103, 118, 99, 105, 134, 125, 117, 106, 109, 104, 111, 127,
   133, 111, 117, 103, 120, 98, 100, 130, 141, 119, 128, 106, 109, 115,
   113, 121, 100, 130, 125, 117, 119, 113, 104, 108, 110, 102

   (a) Presente este conjunto de datos en un histograma de frecuencias.

   (b) ¿Qué intervalo de clase contiene el mayor número de valores de datos?

   (c) ¿Existe, grosso modo, el mismo número de datos en cada uno de los intervalos de clase?

   (d) ¿El histograma parece aproximadamente simétrico?

2. Los siguientes datos muestran las temperaturas máximas (en grados Celsius), de los días 4 de julio de 30 años sucesivos, en la ciudad de San Francisco:

   22,8, 26,2, 31,7, 31,1, 26,9, 28,0, 29,4, 28,8, 26,7, 27,4, 28,2,
   30,3, 29,5, 28,9, 27,5, 28,3, 24,1, 25,3, 28,5, 27,7, 24,4,
   29,2, 30,3, 33,7, 27,5, 29,3, 30,2, 28,5, 32,2, 33,7

   (a) Presente este conjunto de datos en un histograma de frecuencias.

   (b) ¿Cuál diría que es la temperatura máxima "típica" de un 4 de julio en San Francisco?

   (c) ¿Qué otras conclusiones pueden extraerse del histograma?

3. Los siguientes datos (en miles de dólares) representan las rentas netas anuales de una muestra de contribuyentes:

47, 55, 18, 24, 27, 41, 50, 38, 33, 29, 15, 77, 64, 22, 19, 35, 39, 41,
67, 55, 121, 77, 80, 34, 41, 48, 60, 30, 22, 28, 84, 55, 26, 105, 62,
30, 17, 23, 31, 28, 56, 64, 88, 104, 115, 39, 25, 18, 21, 30, 57, 40,
38, 29, 19, 46, 40, 49, 72, 70, 37, 39, 18, 22, 29, 52, 94, 86, 23, 36

(a) Represente gráficamente estos datos mediante un histograma de frecuencias con 5 intervalos de clase.

(b) Represente gráficamente estos datos mediante un histograma de frecuencias con 10 intervalos de clase.

(c) ¿Cuál de los dos histogramas cree que es más informativo? ¿Por qué?

4. Un conjunto de 200 datos puntuales se dividió en 8 clases, todas de tamaño 3 (en las unidades de los datos). Después se determinaron las frecuencias de cada clase y se construyó una tabla de frecuencias. Sin embargo, ciertas entradas de esta tabla se perdieron. Supongamos que la parte de la tabla de frecuencias que se conservó es la siguiente:

Intervalo de clase	Frecuencia	Frecuencia relativa
		0,05
	14	
	18	
15–18	38	
		0,10
	42	
	11	

Rellene los valores perdidos de la tabla y dibuje un histograma de frecuencias relativas.

5. Los siguientes valores muestran las concentraciones de ozono (medidas en partes por 100 millones) en el aire del centro de la ciudad de Los Ángeles durante 25 días consecutivos del verano de 1984:

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

(a) Construya un histograma de frecuencias para este conjunto de datos, uno de cuyos intervalos de clase vaya de 3 a 5.

(b) Construya un histograma de frecuencias para este conjunto de datos, uno de cuyos intervalos de clase vaya de 2 a 3.

(c) ¿Qué histograma de frecuencias crees que es más informativo?

6. Los siguientes datos reflejan las producciones de carne, en miles de toneladas métricas, del año 2002 en 11 países distintos:

País	Producción	País	Producción
Argentina	2,748	Japón	520
Australia	2,034	México	1,450
Brasil	7,150	España	592
China	5,616	Reino Unido	1,390
Francia	1,666	Estados Unidos	12,424
Italia	1,161		

(a) Represente estos datos en un histograma de frecuencias.

(b) Un valor de dato que se encuentra muy separado del resto se denomina un *outlier*, o valor extremo. ¿Existe algún *outlier* en el conjunto de datos dado?

7. Considere los niveles de colesterol en la sangre de los primeros 100 estudiantes que aparecen en el conjunto del Apéndice A. Divida estos estudiantes en dos grupos por sexo, y construya una tabla de frecuencias relativas para cada uno de ellos. Dibuje en un mismo gráfico los polígonos de frecuencias relativas para los estudiantes varones y hembras. ¿Se pueden extraer conclusiones sobre la relación existente entre el sexo y el nivel de colesterol?

8. Utilice la siguiente tabla para construir un histograma de frecuencias de las cantidades que, por cada dólar recaudado por impuestos, el gobierno federal devuelve a los diferentes Estados.

Devolución federal a los Estados por cada dólar recaudado mediante impuestos (cifras del año fiscal 2002, ordenadas de mayor a menor)

Estado	Devolución	Estado	Devolución	Estado	Devolución	Estado	Devolución	Estado	Devolución
District of Columbia	$6,44	Arkansas	$1,55	Maryland	$1,22	Ohio	$1,03	Delaware	$0,85
New Mexico	2,37	Oklahoma	1,52	Arizona	1,21	Georgia	1,01	Colorado	0,78
North Dakota	2,07	Virginia	1,50	Nebraska	1,19	Florida	1,01	Minnesota	0,77
Alaska	1,91	Kentucky	1,50	Utah	1,14	Indiana	1,00	Illinois	0,77
Mississippi	1,89	Louisiana	1,48	Kansas	1,13	Oregon	0,98	California	0,76
West Virginia	1,82	South Carolina	1,34	Vermont	1,13	Texas	0,92	Massachusetts	0,75
Montana	1,67	Maine	1,34	Pennsylvania	1,09	Wisconsin	0,88	Nevada	0,74
Alabama	1,64	Missouri	1,34	Rhode Island	1,08	Michigan	0,88	New Hampshire	0,66
South Dakota	1,61	Idaho	1,31	North Carolina	1,07	Washington	0,87	Connecticut	0,65
Hawaii	1,57	Tennessee	1,26	Wyoming	1,06	New York	0,85	New Jersey	0,62
		Iowa	1,23						

Fuente: Administración Fiscal.

La tabla siguiente muestra los datos relativos a las tasas de mortalidad accidental en Estados Unidos durante varios años. Utilícela para contestar a los problemas de 9 a 12.

Tasas de defunción por cada 100 000 habitantes para los principales tipos de muerte accidental en los Estados Unidos, 1970-2002

Año	Vehículos a motor	Caídas	Envene-namientos	Ahogamientos	Fuegos, incendios, humo	Ingestión de comida u objetos	Armas de fuego
1970	26,8	8,3	2,6	3,9	3,3	1,4	1,2
1980	23,4	5,9	1,9	3,2	2,6	1,4	0,9
1985	19,3	5,0	2,2	2,2	2,1	1,5	0,7
1990	18,8	4,9	2,3	1,9	1,7	1,3	0,6
1991	17,3	5,0	2,6	1,8	1,6	1,3	0,6
1992	16,1	5,0	2,7	1,4	1,6	1,2	0,6
1993	16,3	5,1	3,4	1,5	1,5	1,2	0,6
1994	16,3	5,2	3,5	1,5	1,5	1,2	0,5
1995	16,5	5,3	3,4	1,7	1,4	1,2	0,5
1996	16,5	5,6	3,5	1,5	1,4	1,2	0,4
1997	16,2	5,8	3,8	1,5	1,3	1,2	0,4
1998	16,1	6,0	4,0	1,6	1,2	1,3	0,3
1999	15,5	4,8	4,5	1,3	1,2	1,4	0,3
2000	15,7	4,8	4,6	1,3	1,2	1,6	0,3
2001	15,7	5,1	5,0	1,2	1,2	1,4	0,3
2002	15,7	5,2	5,6	1,1	1,0	1,5	0,3

Fuente: Consejo de Seguridad Nacional.

9. Construya un histograma de frecuencias relativas para las tasas de mortalidad anuales producidas con vehículos a motor.

10. Construya un histograma de frecuencias relativas para las tasas de mortalidad anuales debidas a caídas.

11. Construya un histograma de frecuencias relativas para las tasas anuales de mortalidad para el total de causas citadas.

12. ¿Diría que las tasas de mortalidad accidental se mantienen relativamente constantes?

13. A partir de la tabla que antecede al problema 12 de la sección 2.2, construya un histograma del número medio de días de lluvia en las ciudades de la lista.

14. Considere la tabla siguiente.

Edad del conductor, en años	Porcentaje de conductores	Porcentaje de conductores con accidentes fatales
15–20	9	18
20–25	13	21
25–30	13	14
30–35	11	11
35–40	9	7
40–45	8	6
45–50	8	5
50–55	7	5
55–60	6	4
60–65	6	3
65–70	4	2
70–75	3	2
Más de 75	3	2

Por el criterio de inclusión por la izquierda de las clases, el 13% del total de los conductores tienen como mínimo 25 años y menos de 30, y un 11% de los conductores muertos en accidentes de coche tienen por lo menos 30 años y menos de 35.

(a) Dibuje un histograma de frecuencias relativas para las clases de edad de los conductores.

(b) Dibuje un histograma de frecuencias relativas para las clases de edad de los conductores muertos en accidentes de coche.

(c) ¿Cuál es el grupo que tiene un mayor número de accidentes fatales?

(d) ¿Cuál es el grupo que merecería mayores descuentos en los seguros? Explique su razonamiento.

15. Las tablas de frecuencias relativas acumuladas muestran el porcentaje de valores de datos menores que un valor dado, para una sucesión creciente de valores. Dichas tablas se pueden construir a partir de tablas de frecuencias relativas mediante la suma de las frecuencias relativas de forma acumulada. La tabla siguiente muestra los valores iniciales de la tabla acumulada citada para los dos conjuntos de datos incluidos en la tabla 2.9. Muestra, por ejemplo, que el 5,44% de los hombres con unas edades comprendidas entre 30 y 40 años tiene presiones sanguíneas inferiores a 110, y que sólo un 3,56% de los que tienen entre 50 y 60 años tienen una presión inferior a la citada.

Tabla de frecuencias relativas acumuladas para los conjuntos de datos de la tabla 2.9

	Porcentaje de trabajadores	
Presión sanguínea menor de	Con edad de 30-40	Con edad de 50-60
90	0,12	0,14
100	0,79	0,41
110	5,44	3,56
120		
130		
.		
.		
.		
240	100	100

(a) Explique por qué la frecuencia relativa acumulada de la última clase debe ser 100.

(b) Complete la tabla.

(c) ¿Qué indica la tabla sobre los dos conjuntos de datos? (Esto es, ¿cuál tiende a tener valores menores?)

(d) Represente en un mismo gráfico los polígonos de frecuencias relativas acumuladas de los datos dados. Tales gráficos se denominan *ojivas*.

2.4 Gráficos de tallos y hojas

Una forma eficiente de representar un conjunto de datos de tamaño pequeño o moderado consiste en utilizar los *gráficos de tallos y hojas*. Tales gráficos se obtienen dividiendo cada valor de dato en dos partes –su tallo y su hoja–. Por ejemplo, si todos los datos son de dos dígitos, el tallo de un valor podría ser el dígito de las decenas y su hoja, el dígito de las unidades. Es decir, el valor 84 se expresaría como

Tallo	Hoja
8	4

y los dos valores 84 y 87 se representarían conjuntamente de la siguiente manera

Tallo	Hoja
8	4, 7

Ejemplo 2.5 La tabla 2.11 presenta las rentas per cápita para los 50 Estados de Estados Unidos y para el Distrito de Columbia.

Tabla 2.11 Rentas per cápita (en dólares por persona), en 2002

Estado		Estado	
Estados Unidos	30 941	Missouri	28 936
Alabama	25 128	Montana	25 020
Alaska	32 151	Nebraska	29 771
Arizona	26 183	Nevada	30 180
Arkansas	23 512	New Hampshire	34 334
California	32 996	New Jersey	39 453
Colorado	33 276	New Mexico	23 941
Connecticut	42 706	New York	36 043
Delaware	32 779	North Carolina	27 711
District of Columbia	42 120	North Dakota	26 982
Florida	29 596	Ohio	29 405
Georgia	28 821	Oklahoma	25 575
Hawaii	30 001	Oregon	28 731
Idaho	25 057	Pennsylvania	31 727
Illinois	33 404	Rhode Island	31 319
Indiana	28 240	South Carolina	25 400
Iowa	28 280	South Dakota	26 894
Kansas	29 141	Tennessee	27 671
Kentucky	25 579	Texas	28 551
Louisiana	25 446	Utah	24 306
Maine	27 744	Vermont	29 567
Maryland	36 298	Virginia	32 922
Massachusetts	39 244	Washington	32 677
Michigan	30 296	West Virginia	23 688
Minnesota	34 071	Wisconsin	29 923
Mississippi	22 372	Wyoming	30 578

Los datos que se muestran en la tabla 2.11 se pueden representar mediante el siguiente gráfico de tallos y hojas. Observe que los valores de las hojas aparecen en el gráfico en orden creciente.

22	372
23	512, 688, 941
24	706
25	020, 057, 128, 400, 446, 575, 579
26	183, 894, 982
27	671, 711, 744
28	240, 280, 551, 731, 821, 936
29	141, 405, 567, 596, 771, 923
30	001, 180, 296, 578
31	319, 727
32	151, 677, 779, 922, 996

33	276, 404
34	071, 334
36	043, 298
39	244, 453
42	120, 706

La elección de los tallos siempre se debe hacer de modo que el diagrama de tallos y hojas proporcione información sobre los datos. Como modelo, considere el ejemplo 2.6.

Ejemplo 2.6 Los siguientes datos representan la proporción de estudiantes de las escuelas públicas de primaria en 18 ciudades distintas.

55,2, 47,8, 44,6, 64,2, 61,4, 36,6, 28,2, 57,4, 41,3,
44,6, 55,2, 39,6, 40,9, 52,2, 63,3, 34,5, 30,8, 45,3

Si se hace que el tallo identifique el dígito de las decenas y que la hoja incluya las cifras restantes de cada valor, el gráfico de tallos y hojas resultante para los datos dados es el siguiente:

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

Se podría haber elegido que, para cada valor, el tallo viniera representado por su parte entera-y-la-hoja-por-su-parte-decimal,-de-modo que el valor 28,2 apareciera como

28 | ,2

Sin embargo, esta elección produciría demasiados tallos (con muy pocas hojas cada uno), con lo que el conjunto de datos no quedaría representado con claridad. □

Ejemplo 2.7 Los siguientes gráficos de tallos y hojas muestran los pesos de 80 asistentes a una convención de deportes. Los tallos representan las cifras de las decenas y las hojas las cifras de las unidades.

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

16	0, 1, 1, 1, 2, 4, 5, 6, 8, 8	(10)
17	1, 1, 3, 5, 6, 6, 6	(7)
18	1, 2, 2, 5, 5, 6, 6, 9	(8)
19	0, 0, 1, 2, 4, 5	(6)
20	9, 9	(2)
21	7	(1)
22	1	(1)
23		(0)
24	9	(1)

Los números que están entre paréntesis a la derecha representan la cantidad de valores que aparecen en cada clase de tallo; son valores que habitualmente resultan útiles. Nos indican, por ejemplo, que existen 10 valores dentro del tallo 16; esto es, hay 10 individuos cuyos pesos oscilan entre 160 y 169. Observe que un tallo sin hojas (tal como el tallo con valor 23) indica que no existen ocurrencias en dicha clase.

De este gráfico resulta evidente que casi todos los valores de datos se encuentran entre 100 y 200, y que su dispersión es bastante uniforme dentro de esta región, a excepción de los escasos valores que caen dentro de los intervalos con extremos 100 y 110, y 190 y 200. □

Los gráficos de tallos y hojas son bastante útiles para mostrar todos los valores de datos mediante una representación clara que puede ser un primer paso en la descripción, el resumen y el aprendizaje a partir de los datos. Resulta más adecuado para conjuntos de datos de tamaño moderado. (Si el tamaño del conjunto de datos fuera muy grande, desde un punto de vista práctico, los valores de las hojas podrían ser excesivos y puede que los gráficos de tallos y hojas no fueran más informativos que un histograma.) En cuanto a su forma, este gráfico se parece a un histograma girado, con el plus adicional de que presenta todos los valores existentes dentro de cada clase. Estos valores dentro de cada clase pueden ser de gran utilidad para detectar patrones en los datos, (tales como ver que todos los datos son múltiplos de algún valor), o para encontrar qué valores suceden con mayor frecuencia dentro de cada tallo.

En ocasiones, si un gráfico de tallos y hojas tiene demasiadas hojas por tallo resulta excesivamente desordenado. Una posible solución es la de duplicar el número de tallos, generando dos tallos nuevos por cada uno de los antiguos. Para los tallos del gráfico anterior, los pares de tallos nuevos podrían incluir todas las hojas con valores entre 0 y 4, por un lado, y los valores entre 5 y 9, por otro. Por ejemplo, supongamos que un tallo del gráfico antiguo fuera:

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

Éste se podría partir en los dos tallos siguientes:

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

1. Para los siguientes datos, dibuje los gráficos de tallos y hojas teniendo (a) 4 tallos y (b) 8 tallos.

   124, 129, 118, 135, 114, 139, 127, 141, 111, 144, 133, 127,
   122, 119, 132, 137, 146, 122, 119, 115, 125, 132, 118, 126,
   134, 147, 122, 119, 116, 125, 128, 130, 127, 135, 122, 141

2. Los datos siguientes muestran qué porcentajes de personas, con edad mayor o igual a 25 años, eran graduados universitarios, en el año 2002, para los distintos Estados de Estados Unidos y para el Distrito de Columbia. Represente estos datos mediante un gráfico de tallos y hojas.

Porcentajes estatales de personas de 25 años o más que son titulados universitarios, año 2002

Estado	Porcentaje	Ordinal*
Estados Unidos	**26,7**	(X)
Alabama	22,7	38
Alaska	25,6	26
Arizona	26,3	22
Arkansas	18,3	49
California	27,9	15
Colorado	35,7	2
Connecticut	32,6	5
Delaware	29,5	11
District of Columbia	44,4	(X)
Florida	25,7	25
Georgia	25,0	29
Hawaii	26,8	19
Idaho	20,9	45
Illinois	27,3	16
Indiana	23,7	33
Iowa	23,1	37
Kansas	29,1	12
Kentucky	21,6	43
Louisiana	22,1	41
Maine	23,8	32
Maryland	37,6	1
Massachusetts	34,3	4
Michigan	22,5	39
Minnesota	30,5	8
Mississippi	20,9	45
Missouri	26,7	21
Montana	23,6	34
North Dakota	25,3	28

Porcentajes estatales de personas de 25 años o más que son titulados universitarios, año 2002 (*Continuación*)

Estado	Porcentaje	Ordinal*
Ohio	24,5	31
Oklahoma	20,4	47
Oregon	27,1	17
Pennsylvania	26,1	24
Rhode Island	30,1	9
South Carolina	23,3	36
South Dakota	23,6	34
Tennessee	21,5	44
Texas	26,2	23
Utah	26,8	19
Vermont	30,8	7
Virginia	34,6	3
Washington	28,3	14
West Virginia	15,9	50
Wisconsin	24,7	30
Wyoming	19,6	48

* Cuando varios Estados comparten el mismo ordinal, se omite el siguiente valor ordinal. Puesto que se incluyen datos redondeados, los Estados pueden mostrar valores idénticos, siendo ligeramente distintos.
Fuente: Resumen Estadístico de Estados Unidos.

3. Los siguientes datos representan las edades, redondeadas al año más próximo, de 43 pacientes de emergencia en un hospital de adultos:

   23, 18, 31, 79, 44, 51, 24, 19, 17, 25, 27, 19, 44, 61, 22, 18,
   14, 17, 29, 31, 22, 17, 15, 40, 55, 16, 17, 19, 20, 32, 20, 45,
   53, 27, 16, 19, 22, 20, 18, 30, 20, 33, 21

   Construya un gráfico de tallos y hojas para este conjunto de datos. Utilice este gráfico para determinar el intervalo de edad de 5 años de amplitud que contiene el mayor número de datos.

4. Un psicólogo registró los 48 tiempos de reacción (en segundos) siguientes a un cierto estímulo.

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

   (a) Construya un gráfico de tallos y hojas para estos datos.

   (b) Construya un segundo gráfico de tallos y hojas usando tallos adicionales.

   (c) ¿Cuál de ellos parece más informativo?

   (d) Supóngamos que un artículo del periódico mantiene que "El tiempo típico de reacción es de_____ segundos". Rellene el hueco que falta con el número que debería figurar.

5. Los siguientes valores representan los ingresos diarios de los parquímetros de la ciudad de Nueva York (en unidades de 5000 dólares) en 30 días del año 2002.

   108, 77, 58, 88, 65, 52, 104, 75, 80, 83, 74, 68, 94, 97, 83,
   71, 78, 83, 90, 79, 84, 81, 68, 57, 59, 32, 75, 93, 100, 88

   (a) Represente estos datos mediante un gráfico de tallos y hojas.

   (b) ¿Parece algún dato "sospechoso"? ¿Por qué?

6. La volatilidad de una acción es una propiedad importante en la teoría de precios futuros. Representa un indicativo de la magnitud del cambio que, día a día, tiende a existir en los precios de la acción. Una volatilidad de 0 significa que el precio de la acción se mantiene constante. Cuanto más alta es la volatilidad, mayor es la tendencia al cambio del precio de las acciones. La lista siguiente muestra las volatilidades de 32 compañías cuyas acciones están incluidas en el mercado de valores de Estados Unidos:

   0,26, 0,31, 0,45, 0,30, 0,26, 0,17, 0,33, 0,32, 0,37, 0,38, 0,35, 0,28, 0,37,
   0,35, 0,29, 0,20, 0,33, 0,19, 0,31, 0,26, 0,24, 0,50, 0,22, 0,33, 0,51,
   0,44, 0,63, 0,30, 0,28, 0,48, 0,42, 0,37

   (a) Represente estos datos mediante un gráfico de tallos y hojas.

   (b) ¿Cuál es el dato de mayor magnitud?

   (c) ¿Cuál es el dato de menor valor?

   (d) ¿Cuál es el valor de dato "típico"?

7. La tabla siguiente muestra los resultados de los 25 primeros partidos de la Super Copa de fútbol profesional. Utilícela para construir un gráfico de tallos y hojas para

   (a) las puntuaciones ganadoras,

   (b) las puntuaciones perdedoras,

   (c) las cantidades que los puntos de los equipos ganadores sobrepasan a los de los equipos perdedores.

Super Bowls I-XXV

Partido	Fecha	Ganador	Perdedor
XXV	Jan. 27, 1991	Giants (NFC) 20	Buffalo (AFC) 19
XXIV	Jan. 28, 1990	San Francisco (NFC) 55	Denver (AFC) 10
XXIII	Jan. 22, 1989	San Francisco (NFC) 20	Cincinnati (AFC) 16
XXII	Jan. 31, 1988	Washington (NFC) 42	Denver (AFC) 10
XXI	Jan. 25, 1987	Giants (NFC) 39	Denver (AFC) 20
XX	Jan. 26, 1986	Chicago (NFC) 46	New England (AFC) 10
XIX	Jan. 20, 1985	San Francisco (NFC) 38	Miami (AFC) 16
XVIII	Jan. 22, 1984	Los Angeles Raiders (AFC) 38	Washington (NFC) 9
XVII	Jan. 30, 1983	Washington (NFC) 27	Miami (AFC) 17
XVI	Jan. 24, 1982	San Francisco (NFC) 26	Cincinnati (AFC) 21
XV	Jan. 25, 1981	Oakland (AFC) 27	Philadelphia (NFC) 10
XIV	Jan. 20, 1980	Pittsburgh (AFC) 31	Los Angeles (NFC) 19

Super Bowls I-XXV (*Continuación*)

Partido	Fecha	Ganador	Perdedor
XIII	Jan. 21, 1979	Pittsburgh (AFC) 35	Dallas (NFC) 31
XII	Jan. 15, 1978	Dallas (NFC) 27	Denver (AFC) 10
XI	Jan. 9, 1977	Oakland (AFC) 32	Minnesota (NFC) 14
X	Jan. 18, 1976	Pittsburgh (AFC) 21	Dallas (NFC) 17
IX	Jan. 12, 1975	Pittsburgh (AFC) 16	Minnesota (NFC) 6
VIII	Jan. 13, 1974	Miami (AFC) 24	Minnesota (NFC) 7
VII	Jan. 14, 1973	Miami (AFC) 14	Washington (NFC) 7
VI	Jan. 16, 1972	Dallas (NFC) 24	Miami (AFC) 3
V	Jan. 17, 1971	Baltimore (AFC) 16	Dallas (NFC) 13
IV	Jan. 11, 1970	Kansas City (AFL) 23	Minnesota (NFL) 7
III	Jan. 12, 1969	New York (AFL) 16	Baltimore (NFL) 7
II	Jan. 14, 1968	Green Bay (NFL) 33	Oakland (AFL) 14
I	Jan. 15, 1967	Green Bay (NFL) 35	Kansas City (AFL) 10

8. Considere el siguiente gráfico de tallos y hojas y el siguiente histograma referidos a un mismo conjunto de datos.

2	1, 1, 4, 7	2–3	x, x, x, x
3	0, 0, 3, 3, 6, 9, 9, 9	3–4	x, x, x, x, x, x, x, x
4	2, 2, 5, 8, 8, 8	4–5	x, x, x, x, x, x
5	1, 1, 7, 7	5–6	x, x, x, x
6	3, 3, 3, 6	6–7	x, x, x, x
7	2, 2, 5, 5, 5, 8	7–8	x, x, x, x, x, x

   A partir del gráfico de tallos y hojas, ¿qué se puede concluir que no sea visible desde el histograma?

9. Utilice los datos representados en el gráfico de tallos y hojas del problema 8 para contestar a las siguientes cuestiones:

   (a) ¿Cuántos datos se encuentran en los 40?

   (b) ¿Qué porcentaje de valores se encuentran por encima de 50?

   (c) ¿Qué porcentaje de valores tiene el dígito de las unidades igual a 1?

10. La tabla siguiente muestra las tasas del impuesto sobre la renta y las tasas de la Seguridad Social correspondientes a 2002 para un cierto grupo de países.

    (a) Represente los porcentajes pagados por el impuesto sobre la renta mediante un histograma.

    (b) Represente los porcentajes pagados a la Seguridad Social mediante un histograma.

    (c) Represente los porcentajes pagados a la Seguridad Social mediante un gráfico de tallos y hojas.

Cargas fiscales en los países seleccionados*

País	Impuesto sobre la renta (%)	Seguridad Social (%)	Pago total† (%)	País	Impuesto sobre la renta (%)	Seguridad Social (%)	Pago total† (%)	País	Impuesto sobre la renta (%)	Seguridad Social (%)	Pago total†† (%)
Denmark	33	11	43	France	13	13	27	New Zealand	20	0	20
Belgium	28	14	41	Canada	19	7	26	Slovak Republic	7	13	19
Germany	21	21	41	Australia	24	0	24	Spain	13	6	19
Finland	26	6	32	Czech Republic	11	13	24	Greece	1	16	17
Poland	6	25	31	United States	17	8	24	Portugal	6	11	17
Sweden	23	7	30	United Kingdom	16	8	23	Ireland	11	5	16
Turkey	15	15	30	Iceland	22	0	22	Japan	6	10	16
Netherlands	7	22	29	Luxembourg	8	14	22	Korea	2	7	9
Norway	21	8	29	Switzerland	10	12	22	Mexico	2	2	4
Austria	11	18	29								
Hungary	17	13	29								
Italy	19	9	28								

* No se incluyen los impuestos no indicados, tales como los impuestos sobre las ventas o sobre el valor añadido. Las tasas mostradas se aplican a las personas individuales con rentas medias.
† Es posible que los totales no coincidan con la suma debido a los redondeos.
Fuente: Organización para la Cooperación y el Desarrollo Económico, 2002.

11. Una forma útil de comparar dos conjuntos de datos consiste en colocar sus gráficos de tallos y hojas contiguamente. A continuación se representan las calificaciones obtenidas por los estudiantes de dos escuelas distintas en un examen estándar. En ambas escuelas, 24 estudiantes se presentaron al examen.

Escuela A		Escuela B
Hojas	Tallo	Hojas
0	5	3, 5, 7
8, 5	6	2, 5, 8, 9, 9
9, 7, 4, 2, 0	7	3, 6, 7, 8, 8, 9
9, 8, 8, 7, 7, 6, 5, 3	8	0, 2, 3, 5, 6, 6
8, 8, 6, 6, 5, 5, 3, 0	9	0, 1, 5
	10	0

(a) ¿Qué escuela obtuvo la mayor calificación?

(b) ¿Qué escuela obtuvo la menor calificación?

(c) ¿Qué escuela obtuvo los mejores resultados?

(d) Reúna los datos de las dos escuelas y represente las 48 calificaciones mediante un gráfico de tallos y hojas.

2.5 Conjuntos de datos apareados

En ocasiones, los conjuntos de datos consisten en pares de valores con algún tipo de relación entre ellos. Cada individuo del conjunto de datos presenta un valor x y un valor y. Por lo general, el par i-ésimo se denota mediante (x_i, y_i), $i = 1, \ldots, n$. Por ejemplo, en el conjunto de datos presentado en la tabla 2.12, x_i representa la puntuación obtenida en el test de coeficiente de inteligencia (IQ), e y_i representa el salario anual (redondeado en miles de dólares) del i-ésimo trabajador de una muestra de 30 trabajadores pertenecientes a una empresa. En este apartado, se mostrará cómo se pueden representar de manera efectiva conjuntos de datos con valores apareados.

Una posibilidad de representación de esos conjuntos de datos consiste en considerar separadamente cada uno de los datos apareados y en representar cada uno de ellos mediante histogramas o gráficos de tallos y hojas. Por ejemplo, las figuras 2.11 y 2.12 muestran los gráficos de tallos y hojas, respectivamente, de las puntuaciones del test IQ y de los salarios anuales correspondientes a los datos incluidos en la tabla 2.12.

Sin embargo, aunque las figuras 2.11 y 2.12 exponen amplia información sobre las puntuaciones del test IQ y sobre los salarios de los trabajadores, no nos dicen nada acerca de la relación existente entre ambas variables. Así por ejemplo, no son útiles por sí mismas para ayudar a discernir si las mayores puntuaciones en el test de inteligencia tienden a corresponderse con los salarios más elevados de la compañía. Para responder a cuestiones de este tipo, es preciso considerar simultáneamente los valores apareados de cada dato puntual.

Una forma útil de mostrar un conjunto de datos con valores apareados es la de representarlos mediante un gráfico cartesiano con dos ejes perpendiculares. En el eje x aparecerían los valores x de los datos, mientras que los valores y estarían en el eje y. Tales gráficos se denominan *diagramas de dispersión*. La figura 2.13 presenta un diagrama de dispersión para los datos de la tabla 2.12.

Tabla 2.12 Salarios frente a puntuaciones del test IQ

Trabajador i	Puntuación IQ x_i	Salario anual y_i (en miles de dólares)	Trabajador i	Puntuación IQ x_i	Salario anual y_i (en miles de dólares)
1	110	68	16	84	19
2	107	30	17	83	16
3	83	13	18	112	52
4	87	24	19	80	11
5	117	40	20	91	13
6	104	22	21	113	29
7	110	25	22	124	71
8	118	62	23	79	19
9	116	45	24	116	43
10	94	70	25	113	44
11	93	15	26	94	17
12	101	22	27	95	15
13	93	18	28	104	30
14	76	20	29	115	63
15	91	14	30	90	16

12	4	(1)
11	0, 0, 2, 3, 3, 5, 6, 6, 7, 8	(10)
10	1, 4, 4, 7	(4)
9	0, 1, 1, 3, 3, 4, 4, 5	(8)
8	0, 3, 3, 4, 7	(5)
7	6, 9	(2)

Figura 2.11 Gráfico de tallos y hojas para las puntuaciones del test IQ.

7	0, 1	(2)
6	2, 3, 8	(3)
5	2	(1)
4	0, 3, 4, 5	(4)
3	0, 0	(2)
2	0, 2, 2, 4, 5, 9	(6)
1	1, 3, 3, 4, 5, 5, 6, 6, 7, 8, 9, 9	(12)

Figura 2.12 Gráfico de tallos y hojas para los salarios anuales (en miles de dólares).

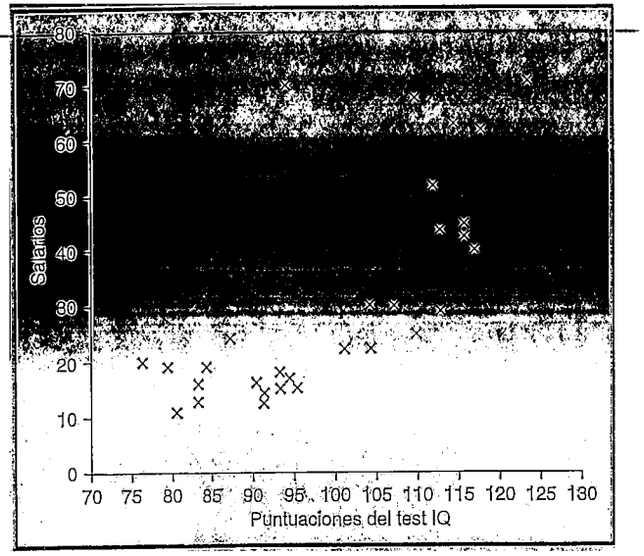


Figura 2.13 Diagrama de dispersión de puntuaciones del test IQ frente a los salarios.

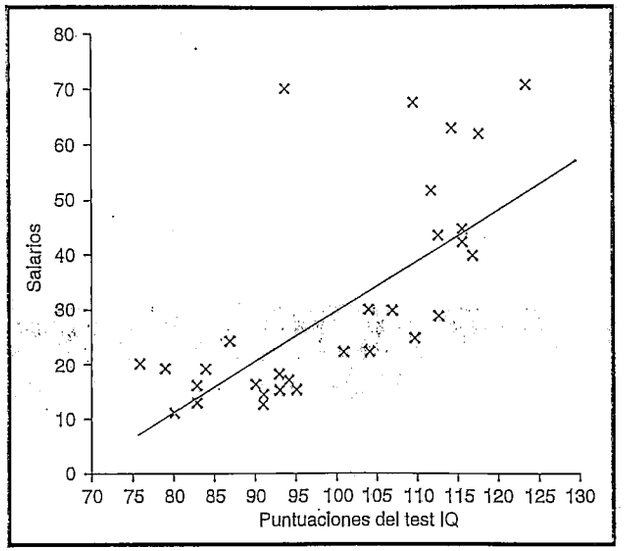


Figura 2.14 Diagrama de dispersión de puntuaciones del test IQ frente a los salarios; se ajusta a ojo una línea recta.

Resulta evidente, de la figura 2.13, que los salarios más altos se corresponden con las puntuaciones más altas del test IQ. Esto es, aunque no todos los trabajadores con puntuaciones IQ altas reciben un salario superior al que recibe otro trabajador con una puntuación menor (compare el trabajador 5 con el 29), generalmente, lo indicado resulta ser cierto.

El diagrama de dispersión de la figura 2.13 también puede resultar útil para establecer ciertos pronósticos. Por ejemplo, supongamos que pretendemos predecir el salario de un trabajador, similar a los considerados, cuya puntuación obtenida en el test de inteligencia fuera de 120. Una forma de hacerlo consiste en "ajustar a ojo" una línea recta al conjunto de datos, tal como se hizo en la figura 2.14. Puesto que el valor y en la recta correspondiente al valor x de 120 es más o menos de 45, este valor parece una predicción razonable para el salario de un trabajador cuya puntuación IQ sea 120.

Aparte de que representan los patrones conjuntos de dos variables y de que nos permiten hacer predicciones, los diagramas de dispersión resultan útiles para detectar *outliers*, los datos puntuales que aparentemente no siguen los patrones de los demás datos. [Por ejemplo, el punto (94,70) de la figura 2.13 parece que no siga la tendencia general.] Tras haber detectado los *outliers*, se puede decidir si el par de datos es significativo o si se debe a un error en la obtención de la información.

Problemas

1. Para determinar la relación entre la temperatura que hay al mediodía (medida en grados Celsius) y el número de piezas defectuosas producidas dicho día, una compañía registró los datos siguientes correspondientes a 22 días laborables.

Temperatura	Número de piezas defectuosas	Temperatura	Número de piezas defectuosas
24,2	25	24,8	23
22,7	31	20,6	20
30,5	36	25,1	25
28,6	33	21,4	25
25,5	19	23,7	23
32,0	24	23,9	27
28,6	27	25,2	30
26,5	25	27,4	33
25,3	16	28,3	32
26,0	14	28,8	35
24,4	22	26,6	24

(a) Dibuje un diagrama de dispersión.

(b) ¿Qué se puede concluir a partir del diagrama anterior?

(c) Si la temperatura al mediodía de mañana fuera de 24° C, ¿qué se podría conjeturar sobre el número de piezas defectuosas que se van a producir al día siguiente?

2. La tabla siguiente muestra, para cada Estado, el porcentaje de población que no dispone de seguro médico, en los años 1990, 2000 y 2002.

Cobertura* por seguros médicos en los Estados, 1990, 2000, 2002

	2002		2000		1990			2002		2000		1990	
	Sin seguro†	% sin seguro	Sin seguro†	% sin seguro	Sin seguro†	% sin seguro		Sin seguro†	% sin seguro	Sin seguro†	% sin seguro	Sin seguro†	% sin seguro
AL	564	12,7	582	13,3	710	17,4	MT	139	15,3	150	16,8	115	14,0
AK	119	18,7	117	18,7	77	15,4	NE	174	10,2	154	9,1	138	8,5
AZ	916	16,8	869	16,7	547	15,5	NV	418	19,7	344	16,8	201	16,5
AR	440	16,3	379	14,3	421	17,4	NH	125	9,9	103	8,4	107	9,9
CA	6 398	18,2	6 299	18,5	5 683	19,1	NJ	1 197	13,9	1 021	12,2	773	10,0
CO	720	16,1	620	14,3	495	14,7	NM	388	21,1	435	24,2	339	22,2
CT	356	10,5	330	9,8	226	6,9	NY	3 042	15,8	3 056	16,3	2 176	12,1
DE	79	9,9	72	9,3	96	13,9	NC	1 368	16,8	1 084	13,6	883	13,8
DC	74	13,0	78	14,0	109	19,2	ND	69	10,9	71	11,3	40	6,3
FL	2 843	17,3	2 829	17,7	2 376	18,0	OH	1 344	11,9	1 248	11,2	1 123	10,3
GA	1 354	16,1	1 166	14,3	971	15,3	OK	601	17,3	641	18,9	574	18,6
HI	123	10,0	113	9,4	81	7,3	OR	511	14,6	433	12,7	360	12,4
ID	233	17,9	199	15,4	159	15,2	PA	1 380	11,3	1 047	8,7	1 218	10,1
IL	1 767	14,1	1 704	13,9	1 272	10,9	RI	104	9,8	77	7,4	105	11,1
IN	797	13,1	674	11,2	587	10,7	SC	500	12,5	480	12,1	550	16,2

Cobertura* por seguros médicos en los Estados, 1990, 2000, 2002 (*Continuación*)

	2002		2000		1990			2002		2000		1990	
	Sin seguro†	% sin seguro	Sin seguro†	% sin seguro	Sin seguro†	% sin seguro		Sin seguro†	% sin seguro	Sin seguro†	% sin seguro	Sin seguro†	% sin seguro
IA	277	9,5	253	8,8	225	8,1	SD	85	11,5	81	11,0	81	11,6
KS	280	10,4	289	10,9	272	10,8	TN	614	10,8	615	10,9	673	13,7
KY	548	13,6	545	13,6	480	13,2	TX	5 556	25,8	4 748	22,9	3 569	21,1
LA	820	18,4	789	18,1	797	19,7	UT	310	13,4	281	12,5	156	9,0
ME	144	11,3	138	10,9	139	11,2	VT	66	10,7	52	8,6	54	9,5
MD	730	13,4	547	10,4	601	12,7	VA	962	13,5	814	11,6	996	15,7
MA	644	9,9	549	8,7	530	9,1	WA	850	14,2	792	13,5	557	11,4
MI	1 158	11,7	901	9,2	865	9,4	WV	255	14,6	250	14,1	249	13,8
MN	397	7,9	399	8,1	389	8,9	WI	538	9,8	406	7,6	321	6,7
MS	465	16,7	380	13,6	531	19,9	WY	86	17,7	76	15,7	58	12,5
MO	646	11,6	524	9,5	665	12,7	U,S,	43 574	15,2	39 804	14,2	34 719	13,9

* Para la población de todas las edades, incluidos los mayores de 65 años.

† En miles.

Fuente: Oficina de Censos, Estados Unidos. Departamento de Comercio.

(a) Dibuje un diagrama de dispersión en el que se relacionen las tasas correspondientes a los años 1990 y 2000.

(b) Dibuje un diagrama de dispersión para las tasas correspondientes a los años 2000 y 2002.

3. La tabla siguiente proporciona el número de habitantes de algunos de los condados más grandes de Estados Unidos.

Los 25 condados con mayor población, 2000–2002

Condado	2002 Población	2000 Población	Condado	2002 Población	2000 Población
Los Angeles, CA	9 806 577	9 519 330	Broward, FL	1 709 118	1 623 018
Cook, IL	5 377 507	5 376 741	Riverside, CA	1 699 112	1 545 387
Harris, TX	3 557 055	3 400 578	Santa Cara, CA	1 683 505	1 682 585
Maricopa, AZ	3 303 876	3 072 149	New York, NY	1 546 856	1 537 195
Orange, CA	2 938 507	2 846 289	Tarrant, TX	1 527 366	1 446 219
San Diego, CA	2 906 660	2 813 833	Clark, NV	1 522 164	1 375 738
Kings, NY	2 488 194	2 465 326	Philadelphia, PA	1 492 231	1 517 550
Miami-Dade, FL	2 332 599	2 253 362	Middlesex, MA	1 474 160	1 465 396
Dallas, TX	2 283 953	2 218 899	Alameda, CA	1 472 310	1 443 741
Queens, NY	2 237 815	2 229 379	Suffolk, NY	1 458 655	1 419 369
Wayne, MI	2 045 540	2 061 162	Bexar, TX	1 446 333	1 392 927
San Bemardino, CA	1 816 072	1 709 434	Cuyahoga, OH	1 379 049	1 393 845
King, WA	1 759 604	1 737 032			

Fuente: Oficina de Censos, Estados Unidos. Departamento de Comercio.

(a) Represente estos datos mediante un diagrama de dispersión.

(b) ¿Qué conclusiones se pueden sacar?

4. La tabla siguiente muestra el número de días en que, en los años comprendidos entre 1993 y 2002, no hubo los niveles de calidad aceptables en el aire de una muestra de distintas áreas metropolitanas de Estados Unidos.

Calidad del aire de las áreas metropolitanas de Estados Unidos seleccionadas, 1993–2002

Área metropolitana muestreada	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
Atlanta, GA	36	15	36	28	33	52	67	34	18	24
Bakersfield, CA	97	105	107	110	58	78	144	132	125	152
Baltimore, MD	48	40	36	28	30	51	40	19	32	42
Boston, MA–NH	2	6	7	4	7	8	10	1	12	16
Chicago, IL	4	13	24	7	10	12	19	2	22	21
Dallas, TX	12	24	29	10	27	33	25	22	16	15
Denver, CO	6	3	5	2	0	9	5	3	8	8
Detroit, MI	5	11	14	13	11	17	20	15	27	26
El Paso, TX	7	6	3	6	2	6	5	4	9	13
Fresno, CA	59	55	61	70	75	67	133	131	138	152
Houston, TX	27	41	66	28	47	38	52	42	29	23
Las Vegas, NV–AZ	3	3	3	14	4	5	8	2	1	6
Los Angeles–Long Beach, CA	134	139	113	94	60	56	56	87	88	80
Miami, FL	6	1	2	1	3	8	7	2	1	1
Minneapolis–St. Paul, MN–WI	0	2	5	0	0	1	1	2	2	1
New Haven–Meriden, CT	12	13	14	8	19	9	19	9	15	25
New York, NY	11	16	21	14	23	18	25	19	19	31
Orange County, CA	25	15	9	9	3	6	14	31	31	19
Philadelphia, PA–NJ	62	37	38	38	38	37	32	22	29	33
Phoenix–Mesa, AZ	14	10	22	15	12	14	10	10	8	8
Pittsburgh, PA	14	22	27	12	21	39	40	29	52	53
Riverside–San Bernardino, CA	168	150	125	118	107	96	123	145	155	145
Sacramento, CA	20	37	41	44	17	29	69	45	49	69
St. Louis, MO–IL	9	33	38	23	15	24	31	18	17	34
Salt Lake City–Ogden, UT	5	17	5	14	2	19	8	15	15	18
San Diego, CA	59	46	48	31	14	33	33	31	31	20
San Francisco, CA	0	0	2	0	0	0	10	4	12	17
Seattle–Bellevue–Everett, WA	0	3	2	6	1	3	6	7	3	6
Ventura, CA	43	63	66	62	45	29	24	31	25	11
Washington, DC–MD–VA–WV	52	22	32	18	30	47	39	11	22	34

Nota: Los datos indican el número de días con niveles de calidad no aceptable en el aire de las áreas metropolitanas muestreadas. Todos los valores fueron revisados para ajustarse a los estándares de calidad establecidos en 1998. Son partículas aceptables las que tienen un diámetro menor o igual a 2,5 milímetros.
Fuente: Agencia de Protección Medioambiental de Estados Unidos, Oficina de Planificación y Estándares de Estados Unidos.

(a) Dibuje un diagrama de dispersión en el que se relacionen los valores de cada ciudad en los años 2000 y 2002.

(b) ¿Tienden a corresponderse los valores mayores del año 2002 con los valores mayores del año 2000?

5. Los datos siguientes relacionan el periodo de atención (en minutos) y la puntuación en un test de inteligencia (IQ) de 18 niños en edad preescolar.

Periodo de atención	Puntuación IQ	Periodo de atención	Puntuación IQ	Periodo de atención	Puntuación IQ
2,0	82	6,3	105	5,5	118
3,0	88	5,4	108	3,6	128
4,4	86	6,6	112	5,4	128
5,2	94	7,0	116	3,8	130
4,9	90	6,5	122	2,7	140
6,1	99	7,2	110	2,2	142

(a) Dibuje un diagrama de dispersión.

(b) Haga una inferencia plausible sobre la relación existente entre el periodo de atención y la puntuación IQ.

6. Los siguientes datos muestran los porcentajes de interés de los préstamos y las tasas de inflación durante 8 años de la década de 1970.

Tasa de inflación	Porcentaje de interés	Tasa de inflación	Porcentaje de interés
3,3	5,2	5,8	6,8
6,2	8,0	6,5	6,9
11,0	10,8	7,6	9,0
9,1	7,9		

(a) Dibuje un diagrama de dispersión.

(b) Ajuste a mano una recta para los pares de datos.

(c) A partir de la recta anterior, haga una predicción del porcentaje de interés de un año cuya tasa de inflación fuera del 7,2%.

7. Los datos siguientes muestran las rentas per cápita de los residentes de 12 áreas metropolitanas de Estados Unidos.

	Renta per cápita	
Área metropolitana	1994	1996
San Francisco–Oakland–San Jose, CA	28 990 $	32 933 $
Salt Lake City–Ogden, UT	18 731 $	21 271 $
Portland–Salem, OR–WA	22 508 $	25 343 $

Área metropolitana	Renta per cápita 1994	1996
Boston–Worcester–Lawrence–Lowell–Br ockton, MA–NH	27 095 $	30 366 $
Phoenix–Mesa, AZ	20 911 $	23 377 $
Seattle–Tacoma–Bremerton, WA	25 287 $	28 269 $
Denver–Boulder–Greeley, CO	25 657 $	28 650 $
Minneapolis–St. Paul, MN–WI	26 246 $	29 299 $
Tampa–St. Petersburg–Clearwater, FL	21 503 $	23 984 $
Charlotte–Gastonia–Rock Hill, NC–SC	22 819 $	25 446 $
Kansas City, MO–KS	23 281 $	25 949 $
Atlanta, GA	24 451 $	27 241 $

(a) Represente estos datos mediante un diagrama de dispersión.

(b) En 1994 la renta per cápita de los habitantes de San Diego, CA, fue de 22 111 dólares. Haga una predicción del valor correspondiente a 1996.

8. En el problema 7 de la sección 2.4 se muestran los resultados de los 25 primeros partidos de fútbol americano de la Super Copa. Para cada partido, y denota los puntos del equipo ganador y x denota el número de puntos en que este último equipo superó a su contrario. Dibuje un diagrama de dispersión en el que se relacionen x e y. ¿Tienden a corresponderse los valores mayores de una variable con los valores mayores de la otra?

2.6 Comentarios históricos

Probablemente el primer caso registrado de representaciones estadísticas –entiéndase, representaciones de datos mediante tablas y gráficos– se debe a Edmund Halley, con sus análisis gráficos de las presiones barométricas en función de la altitud. Se publicaron en 1686 y se utilizó el sistema de coordenadas cartesianas, introducido por el científico francés René Descartes en sus trabajos de geometría analítica. Halley presentó un diagrama de dispersión y también fue capaz de ajustar una curva a los puntos del gráfico.

A pesar del éxito que Halley consiguió con sus representaciones gráficas, hasta los últimos años del siglo XVIII la mayor parte de los científicos que trababan en esta materia prefirieron utilizar las tablas, en lugar de los gráficos, para presentar sus datos. En realidad, no fue hasta 1786, año en que William Playfair ideó el gráfico de barras como representación de una tabla de frecuencias, cuando se empezaron a utilizar regularmente las representaciones gráficas. En 1801, Playfair inventó los gráficos de tarta y, poco tiempo después, introdujo el uso de histogramas para visualizar datos.

El uso de gráficos para representar datos continuos –es decir, datos en los que todos los valores son distintos– no fue habitual hasta los años 1830. En 1833, el francés A. M. Guerry utilizó los gráficos de barras para representar datos sobre crímenes, tras haber clasificado los datos en intervalos para después reproducirlos en histogramas. En 1846, el estadístico y científico social Adolphe Quetelet hizo un uso sistemático de los histogramas. Quetelet y sus

estudiantes demostraron la utilidad del análisis gráfico en el desarrollo de las ciencias sociales. Tras ello, Quetelet popularizó la práctica, ampliamente extendida hoy día, de comenzar cualquier trabajo de investigación reuniendo primero los datos numéricos para representarlos después. Realmente, esta actuación, junto con los pasos adicionales de clasificación de los datos y de utilización de los métodos de la inferencia estadística para extraer conclusiones, se ha convertido en el paradigma aceptado para investigar en todas las áreas relacionadas con las ciencias sociales. Igualmente, se ha convertido en una técnica importante en otros campos, tales como la investigación médica (para contrastar nuevos medicamentos y terapias) y otras áreas tradicionalmente no numéricas como la literatura (para asignar autor) y la historia (tal como fue utilizada por el historiador francés F. Braudel).


(Princeton University Libraries)
John Tukey

El término histograma fue acuñado por Karlww Pearson en sus disertaciones sobre los gráficos estadísticos. En 1970, el estadístico de Estados Unidos John Tukey utilizó el diagrama de tallos y hojas, que puede interpretarse como una variante del histograma. En palabras de Tukey: "Mientras que el histograma utiliza una marca no cuantitativa para indicar un valor de datos, está claro que la mejor marca es un dígito."

Términos clave

Frecuencia: Número de veces en las que un valor dado aparece en un conjunto de datos.

Tabla de frecuencias: tabla que presenta, para un conjunto de datos dado, cada valor distinto junto con su frecuencia.

Gráfico de líneas: Gráfico de una tabla de frecuencias. La abscisa especifica un valor de dato, y la frecuencia de ocurrencia de tal valor se identifica con la altura de una línea horizontal.

Gráfico de barras (o diagrama de barras): Similar al gráfico de líneas, excepto en que la frecuencia de un valor coincide ahora con la altura de la barra.

Polígono de frecuencias: Gráfico de los valores distintos y sus frecuencias, en el que se conectan los puntos del gráfico mediante rectas.

Conjunto de datos simétrico: Un conjunto de datos es simétrico con respecto a un valor dado x_0 si las frecuencias de los valores $x_0 - c$ y $x_0 + c$ son iguales para todo valor de c.

Frecuencia relativa: Frecuencia de un valor dividida entre el número total de datos del conjunto de éstos.

Gráfico de tarta: Gráfico que representa las frecuencias relativas mediante la división de un círculo en sectores.

Histograma: Gráfico en el que los datos se dividen en intervalos de clase, cuyas frecuencias se muestran en un gráfico de barras.

Histograma de frecuencias relativas: Histograma en el que se muestran gráficamente las frecuencias relativas de cada dato del conjunto.

Gráfico de tallos y hojas: Similar a un histograma, con la excepción de que las frecuencias se indican en una lista con los últimos dígitos (las hojas) de los datos.

Diagrama de dispersión: Gráfico bidimensional de un conjunto de datos apareados.

Resumen

En este capítulo se han explicado distintas formas de representar gráficamente conjuntos de datos. Por ejemplo, consideremos el siguiente conjunto de 13 datos:

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

Se pueden representar estos valores mediante la siguiente tabla de frecuencias, que muestra cada valor distinto junto con el número de veces que aparece en el conjunto de datos:

Tabla de frecuencias

Valor	Frecuencia	Valor	Frecuencia
1	2	4	3
2	4	5	1
3	2	6	1

Los datos también se pueden visualizar gráficamente mediante un *gráfico de líneas*, o bien mediante un *gráfico de barras*. En ocasiones, los distintos valores de datos se representan mediante estos gráficos, y después los puntos resultantes se conectan mediante líneas rectas. Esto da lugar a un *polígono de frecuencias*.

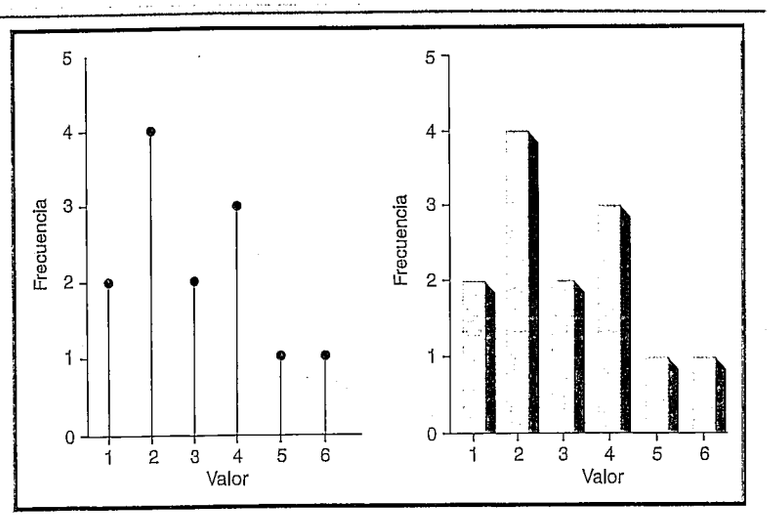
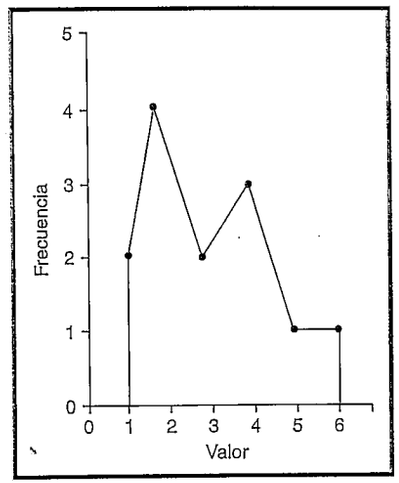
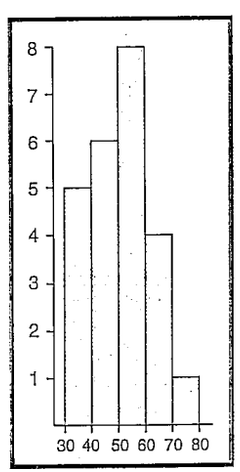


Gráfico de líneas. Gráfico de barras.



Polígono de frecuencias.

Cuando hay un gran número de valores de datos, éstos se suelen clasificar por intervalos de clase. Un gráfico de barras en el que se presentan los distintos intervalos de clase junto con el número de datos que aparecen dentro de cada intervalo se denomina *histograma*. En el eje *y* de este gráfico se pueden representar las frecuencias de clase (el número de valores dentro de cada intervalo de clase), o bien las proporciones de datos que aparecen dentro de cada clase. En el primer caso, el gráfico se denomina *histograma de frecuencias*; en el segundo, recibe el nombre de *histograma de frecuencias relativas*.



Histograma.

Considere el siguiente conjunto de datos:

41, 38, 44, 47, 33, 35, 55, 52, 41, 66, 64, 50, 49, 56,
55, 48, 52, 63, 59, 57, 75, 63, 38, 37

Si se usan los cinco intervalos de clase

30–40, 40–50, 50–60, 60–70, 70–80

junto con el convenio de inclusión por la izquierda (lo que significa que el intervalo contiene todos los puntos mayores o iguales que su extremo izquierdo y estrictamente menores que su extremo derecho), se consigue el histograma de la página 59 como representación del conjunto de datos citado.

Los conjuntos de datos también se pueden representar gráficamente mediante *gráficos de tallos y hojas*. El siguiente gráfico de tallos y hojas representa el anterior conjunto de datos.

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

Gráfico de tallos y hojas.

En ocasiones, los datos se presentan en pares. Es decir, para cada elemento del conjunto de datos existe un valor x y un valor y. Un gráfico de los valores de x e y se denomina *diagrama de dispersión*. El diagrama de dispersión puede ser de gran utilidad para comprobar cuestiones tales como si los valores altos de x aparecen junto con valores altos de y, o si los valores altos de x se corresponden con valores bajos de y, o si no existe aparentemente ninguna relación entre los valores x e y de cada par.

El siguiente conjunto de pares de datos

i	1	2	3	4	5	6	7	8
x_i	8	12	7	15	5	12	10	22
y_i	14	10	17	9	13	8	12	6

se puede representar mediante el siguiente diagrama de dispersión. El diagrama indica que los valores altos de cualquier variable del par están, por lo general, asociados con los valores bajos de la otra variable.

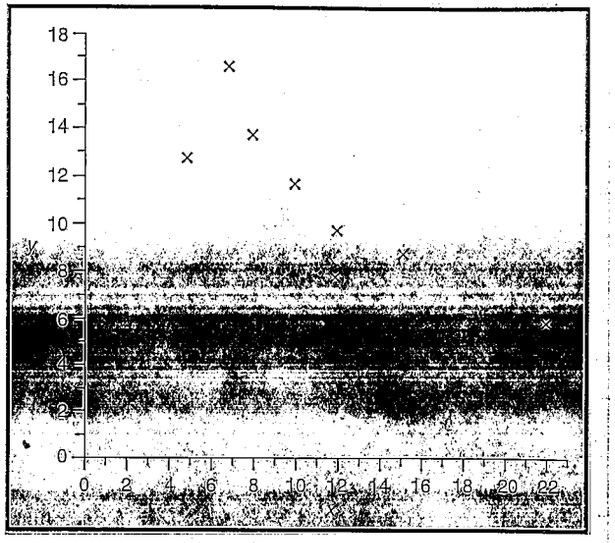


Diagrama de dispersión.

Normalmente, el uso de estas herramientas gráficas facilita que se reconozcan a primera vista las características relevantes de un conjunto de datos. Como resultado, se pueden poner de manifiesto aspectos que no resultan evidentes desde los propios datos en bruto. La elección de qué gráfico se va a utilizar depende de cuestiones tales como el tamaño del conjunto de datos, el tipo de datos o el número de valores distintos.

Problemas de repaso

1. Los siguientes datos muestran los tipos de sangre de 50 donantes voluntarios en cierta clínica:

   O A O AB A A O O B A O A AB B O O O A B A A O A A O
   B A O AB A O O A B A A A O B O O A O A B O AB A O B

   (a) Represente estos datos mediante una tabla de frecuencias.

   (b) Represéntelos también a través de una tabla de frecuencias relativas.

   (c) Represéntelos en un gráfico de tarta.

2. Los datos siguientes provienen de una muestra de precios, redondeados al céntimo más próximo, de un galón de gasolina estándar en el área de la Bahía de San Francisco en mayo de 1991.

   121, 119, 117, 121, 120, 120, 118, 124, 123, 139, 120,
   115, 117, 121, 123, 120, 123, 118, 117, 122, 122, 119

(a) Construya un histograma de frecuencias para este conjunto de datos.

(b) Construya un polígono de frecuencias.

(c) Construya un gráfico de tallos y hojas.

(d) ¿Existe algún dato aparentemente separado de los demás?

3. La siguiente tabla de frecuencias muestra el número de suicidios de mujeres, en ocho Estados alemanes durante 14 años.

Número de suicidios por año	0	1	2	3	4	5	6	7	8	9	10
Frecuencias	9	19	17	20	15	11	8	2	3	5	3

Así, por ejemplo, existieron 20 observaciones en las que ocurrieron 3 suicidios en los Estados y los años los correspondientes.

(a) ¿Cuántos suicidios se registraron a lo largo de los 14 años?

(b) Represente los datos anteriores mediante un histograma.

4. La tabla siguiente muestra las tasas de criminalidad (por 100 000 habitantes) de 1991 en los distintos Estados de Estados Unidos. Utilícela para construir:

(a) Un histograma de frecuencias de las tasas totales por crímenes violentos en los Estados nororientales.

(b) Un histograma de frecuencias relativas de las tasas totales por crímenes de propiedad en los Estados del sur.

(c) Un gráfico de tallos y hojas de las tasas por asesinato en los Estados del occidente.

(d) Un gráfico de tallos y hojas de las tasas por hurto en los Estados del oeste central.

		Crímenes violentos					Crímenes de propiedad			
Región, División, y Estado	Total	Total	Asesinato	Secuestro con violencia	Robo	Atraco grave	Total	Robo	Hurto	Robo de automóvil
Estados Unidos	5 898	758	9,8	42	273	433	5 140	1 252	3 229	659
Nororientales	5 155	752	8,4	29	352	363	4 403	1 010	2 598	795
New England	4 950	532	4,1	30	159	338	4 419	1 103	2 600	716
Maine	3 768	132	1,2	22	23	86	3 636	903	2 570	163
New Hampshire	3 448	119	3,6	30	33	53	3 329	735	2 373	220
Vermont	3 955	117	2,1	31	12	72	3 838	1 020	2 674	144
Massachusetts	5 332	736	4,2	32	195	505	4 586	1 167	2 501	919
Rhode Island	5 039	462	3,7	31	123	304	4 577	1 127	2 656	794
Connecticut	5 364	540	5,7	29	224	280	4 824	1 191	2 838	796
Atlántico Medio	5 227	829	9,9	29	419	372	4 398	978	2 598	823
New York	6 245	1 164	14,2	28	622	499	5 081	1 132	2 944	1 004
New Jersey	5 431	635	5,2	29	293	307	4 797	1 016	2 855	926
Pennsylvania	3 559	450	6,3	29	194	221	3 109	720	1 907	482

(Continuación)

		Crímenes violentos					Crímenes de propiedad			
Región, División, y Estado	Total	Total	Asesinato	Secuestro con violencia	Robo	Atraco grave	Total	Robo	Hurto	Robo de automóvil
Occidente central	5 257	631	7,8	45	223	355	4 626	1 037	3 082	507
Noreste central	5 482	704	8,9	50	263	383	4 777	1 056	3 151	570
Ohio	5 033	562	7,2	53	215	287	4 471	1 055	2 916	500
Indiana	4 818	505	7,5	41	116	340	4 312	977	2 871	465
Illinois	6 132	1 039	11,3	40	456	532	5 093	1 120	3 318	655
Michigan	6 138	803	10,8	79	243	470	5 335	1 186	3 469	680
Wisconsin	4 466	277	4,8	25	119	128	4 189	752	3 001	436
Noroeste central	4 722	457	5,4	34	129	288	4 265	991	2 918	356
Minnesota	4 496	316	3,0	40	98	175	4 180	854	2 963	363
Iowa	4 134	303	2,0	21	45	235	3 831	832	2 828	171
Missouri	5 416	763	10,5	34	251	467	4 653	1 253	2 841	558
North Dakota	2 794	65	1,1	18	8	38	2 729	373	2 229	127
South Dakota	3 079	182	1,7	40	19	122	2 897	590	2 192	115
Nebraska	4 354	335	3,3	28	54	249	4 020	727	3 080	213
Kansas	5 534	500	6,1	45	138	310	5 035	1 307	3 377	351
Sur	6 417	798	12,1	45	252	489	5 618	1 498	3 518	603
Atlántico sur	6 585	851	11,4	44	286	510	5 734	1 508	3 665	561
Delaware	5 869	714	5,4	86	215	408	5 155	1 128	3 652	375
Maryland	6 209	956	11,7	46	407	492	5 253	1 158	3 365	731
District of Columbia	10 768	2 453	80,6	36	1 216	1 121	8 315	2 074	4 880	1 360
Virginia	4 607	373	9,3	30	138	196	4 234	783	3 113	339
West Virginia	2 663	191	6,2	23	43	119	2 472	667	1 631	175
North Carolina	5 889	658	11,4	35	178	434	5 230	1 692	3 239	299
South Carolina	6 179	973	11,3	59	171	731	5 207	1 455	3 365	387
Georgia	6 493	738	12,8	42	268	415	5 755	1 515	3 629	611
Florida	8 547	1 184	9,4	52	400	723	7 363	2 006	4 573	784
Sureste central	4 687	631	10,4	41	149	430	4 056	1 196	2 465	395
Kentucky	3 358	438	6,8	35	83	313	2 920	797	1 909	215
Tennessee	5 367	726	11,0	46	213	456	4 641	1 365	2 662	614
Alabama	5 366	844	11,5	36	153	644	4 521	1 269	2 889	363
Mississippi	4 221	389	12,8	46	116	214	3 832	1 332	2 213	286
Suroeste central	7 118	806	14,2	50	254	488	6 312	1 653	3 871	788
Arkansas	5 175	593	11,1	45	136	402	4 582	1 227	3 014	341
Louisiana	6 425	951	16,9	41	279	614	5 473	1 412	3 489	573
Oklahoma	5 669	584	7,2	51	129	397	5 085	1 478	3 050	557
Texas	7 819	840	15,3	53	286	485	6 979	1 802	4 232	944
Oeste	6 478	841	9,6	46	287	498	5 637	1 324	3 522	791
Mountain	6 125	544	6,5	44	122	371	5 581	1 247	3 843	491
Montana	3 648	140	2,6	20	19	99	3 508	524	2 778	206
Idaho	4 196	290	1,8	29	21	239	3 905	826	2 901	178

(Continuación)

Región, División, y Estado	Total	Crímenes violentos					Crímenes de propiedad			
		Total	Asesinato	Secuestro con violencia	Robo	Atraco grave	Total	Robo	Hurto	Robo de automóvil
Wyoming	4 389	310	3,3	26	17	264	4 079	692	3 232	155
Colorado	6 074	559	5,9	47	107	399	5 515	1 158	3 930	426
New Mexico	6 679	835	10,5	52	120	652	5 845	1 723	3 775	346
Arizona	7 406	671	7,8	42	166	455	6 735	1 607	4 266	861
Utah	5 608	287	2,9	46	55	183	5 321	840	4 240	241
Nevada	6 299	677	11,8	66	312	287	5 622	1 404	3 565	652
Pacífico	6 602	945	10,7	47	345	542	5 656	1 351	3 409	896
Washington	6 304	523	4,2	70	146	303	5 781	1 235	4 102	444
Oregon	5 755	506	4,6	53	150	298	5 249	1 176	3 598	474
California	6 773	1 090	12,7	42	411	624	5 683	1 398	3 246	1 039
Alaska	5 702	614	7,4	92	113	402	5 088	979	3 575	534
Hawaii	5 970	242	4,0	33	87	118	5 729	1 234	4 158	336

Fuente: Oficina Federal de Investigación de Estados Unidos, Crimen en Estados Unidos, anuario.

5. Construya una tabla de frecuencias para un conjunto de datos de 10 valores que sea simétrico y tenga (a) 5 valores distintos y (b) 4 valores distintos. (c) ¿Con respecto a qué valores son simétricos los conjuntos de datos de los apartados (a) y (b)?

6. Los datos siguientes se refieren a las reservas de petróleo estimadas, en miles de millones de barriles, en cuatro regiones del hemisferio occidental. Represente estos datos mediante un gráfico de tarta.

Estados Unidos	38,7
América del Sur	22,6
Canadá	8,8
México	60,0

7. La siguiente tabla contiene las cantidades (en millones de dólares) invertidas en Estados Unidos procedentes de una selección de países europeos en los años 2000 y 2002.

Inversión extranjera en Estados Unidos, procedente de una selección de países (en millones de dólares)

	2000	2002
Europa	887 014	1 006 530
Austria	3 007	3 439
Bélgica	14 787	9 608
Dinamarca	4 025	1 924
Finlandia	8 875	7 212
Francia	125 740	170 619

Inversión extranjera en Estados Unidos, procedente de una selección de países (en millones de dólares)

	2000	2002
Alemania	122 412	137 036
Irlanda	25 523	26 179
Italia	6 576	6 695
Liechtenstein	319	259
Luxemburgo	58 930	34 349
Holanda	138 894	154 753
Noruega	2 665	3 416
España	5 068	4 739
Suecia	21 991	21 989
Suiza	64 719	113 232
Reino Unido	277 613	283 317

Fuente: Oficina de Análisis Económico, Departamento de Comercio de Estados Unidos.

(a) Represente los datos de 2000 y 2002 en dos gráficos de tarta contiguos.

(b) Dibuje su diagrama de dispersión.

8. Los datos siguientes se refieren a las edades (redondeadas al entero más próximo) en las que fallecieron cierto número de pacientes de un hospital (sin servicio de natalidad) de una gran ciudad:

1, 1, 1, 1, 3, 3, 4, 9, 17, 18, 19, 20, 20, 22, 24, 26, 28, 34,
45, 52, 56, 59, 63, 66, 68, 68, 69, 70, 74, 77, 81, 90

(a) Represente este conjunto de datos en un histograma.

(b) Represéntelo mediante un polígono de frecuencias.

(c) Represéntelo mediante un polígono de frecuencias relativas.

(d) Represéntelo en un gráfico de tallos y hojas.

Los problemas del 9 al 11 se refieren a los últimos 50 estudiantes del Apéndice A.

9. (a) Dibuje un histograma de los pesos de estos estudiantes.

(b) Comente ese histograma.

10. Dibuje un diagrama de dispersión que relacione los pesos con los niveles de colesterol. Comente qué se refleja en ese diagrama.

11. Dibuje un diagrama de dispersión que relacione los pesos y las presiones sanguíneas. ¿Qué le sugiere ese diagrama?

Los problemas 12 y 13 se refieren a la tabla siguiente, donde se muestran las calificaciones en Matemáticas y Lengua de varios estudiantes del último curso de educación secundaria.

Estudiante	Calificación en Lengua	Calificación en Matemáticas	Estudiante	Calificación en Lengua	Calificación en Matemáticas
1	520	505	8	620	576
2	605	575	9	604	622
3	528	672	10	720	704
4	720	780	11	490	458
5	630	606	12	524	552
6	504	488	13	646	665
7	530	475	14	690	550

12. Dibuje dos gráficos de tallos y hojas contiguos para las calificaciones de Matemáticas y Lengua. ¿Los estudiantes, como grupo, han obtenido las mejores calificaciones en una de estas asignaturas? Si es así, ¿en cuál?

13. Dibuje un diagrama de dispersión para las calificaciones de los estudiantes en ambas materias. ¿Tienden a aparecer las calificaciones altas en una asignatura junto a las calificaciones altas en la otra?

14. La tabla siguiente proporciona información acerca de las edades de los habitantes de Estados Unidos y México.

	Proporción de población (en porcentaje)	
Edad	México	Estados Unidos
0–9	32,5	17,5
10–19	24	20
20–29	14,5	14,5
30–39	11	12
40–49	7,5	12,5
50–59	4,5	10,5
60–69	3,5	7
70–79	1,5	4
Más de 80	1	2

(a) ¿Qué porcentaje de la población de México tiene menos de 30 años?

(b) ¿Qué porcentaje de la población de Estados Unidos tiene menos de 30 años?

(c) Dibuje dos polígonos de frecuencias relativas en un mismo gráfico. Utilice colores distintos para los datos de México y de Estados Unidos.

(d) En general, ¿cómo compararía las distribuciones de edad de ambos países?

15. Los datos siguientes se refieren a las precipitaciones anuales y mensuales (en pulgadas) que son habituales en varias ciudades.

Precipitaciones habituales, mensuales y anuales, en las ciudades seleccionadas

Estado	Ciudad	Ene.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ago.	Sep.	Oct.	Nov.	Dic.	Anual
AL	Mobile	4,59	4,91	6,48	5,35	5,46	5,07	7,74	6,75	6,56	2,62	3,67	5,44	64,64
AK	Juneau	3,69	3,74	3,34	2,92	3,41	2,98	4,13	5,02	6,40	7,71	5,15	4,66	53,15
AZ	Phoenix	0,73	0,59	0,81	0,27	0,14	0,17	0,74	1,02	0,64	0,63	0,54	0,83	7,11
AR	Little Rock	3,91	3,83	4,69	5,41	5,29	3,67	3,63	3,07	4,26	2,84	4,37	4,23	49,20
CA	Los Angeles	3,06	2,49	1,76	0,93	0,14	0,04	0,01	0,10	0,15	0,26	1,52	1,62	12,08
	Sacramento	4,03	2,88	2,06	1,31	0,33	0,11	0,05	0,07	0,27	0,86	2,23	2,90	17,10
	San Diego	2,11	1,43	1,60	0,78	0,24	0,06	0,01	0,11	0,19	0,33	1,10	1,36	9,32
	San Francisco	4,65	3,23	2,64	1,53	0,32	0,11	0,03	0,05	0,19	1,06	2,35	3,55	19,71
CO	Denver	0,51	0,69	1,21	1,81	2,47	1,58	1,93	1,53	1,23	0,98	0,82	0,55	15,31
CT	Hartford	3,53	3,19	4,15	4,02	3,37	3,38	3,09	4,00	3,94	3,51	4,05	4,16	44,39
DE	Wilmington	3,11	2,99	3,87	3,39	3,23	3,51	3,90	4,03	3,59	2,89	3,33	3,54	41,38
DC	Washington	2,76	2,62	3,46	2,93	3,48	3,35	3,88	4,40	3,22	2,90	2,82	3,18	39,00
FL	Jacksonville	3,07	3,48	3,72	3,32	4,91	5,37	6,54	7,15	7,26	3,41	1,94	2,59	52,76
	Miami	2,08	2,05	1,89	3,07	6,53	9,15	5,98	7,02	8,07	7,14	2,71	1,86	57,55
GA	Atlanta	4,91	4,43	5,91	4,43	4,02	3,41	4,73	3,41	3,17	2,53	3,43	4,23	48,61
HI	Honolulu	3,79	2,72	3,48	1,49	1,21	0,49	0,54	0,60	0,62	1,88	3,22	3,43	23,47
ID	Boise	1,64	1,07	1,03	1,19	1,21	0,95	0,26	0,40	0,58	0,75	1,29	1,34	11,71
IL	Chicago	1,60	1,31	2,59	3,66	3,15	4,08	3,63	3,53	3,35	2,28	2,06	2,10	33,34
	Peoria	1,60	1,41	2,86	3,81	3,84	3,88	3,99	3,39	3,63	2,51	1,96	2,01	34,89
IN	Indianapolis	2,65	2,46	3,61	3,68	3,66	3,99	4,32	3,46	2,74	2,51	3,04	3,00	39,12
IA	Des Moines	1,01	1,12	2,20	3,21	3,96	4,18	3,22	4,11	3,09	2,16	1,52	1,05	30,83
KS	Wichita	0,68	0,85	2,01	2,30	3,91	4,06	3,62	2,80	3,45	2,47	1,47	0,99	28,61
KY	Louisville	3,38	3,23	4,73	4,11	4,15	3,60	4,10	3,31	3,35	2,63	3,49	3,48	43,56
LA	New Orleans	4,97	5,23	4,73	4,50	5,07	4,63	6,73	6,02	5,87	2,66	4,06	5,27	59,74

Fuente: Administración Nacional Oceánica y Atmosférica de Estados Unidos, *Climatología de Estados Unidos*, Septiembre, 1982.

(a) Represente las precipitaciones habituales del mes de abril en un gráfico de tallos y hojas.

(b) Represente las cantidades anuales en un histograma.

(c) Dibuje un diagrama de dispersión que relacione las cantidades de abril con las anuales.

16. Un valor muy separado del resto de valores se llama *outlier* (o valor extremo). En los siguientes conjuntos de datos, especifique qué valores son *outliers*, si es que existen.

(a) 14, 22, 17, 5, 18, 22, 10, −17, 25, 28, 33, 12

(b) 5, 2, 13, 16, 9, 12, 7, 10, 54, 22, 18, 15, 12

(c) 18, 52, 14, 20, 24, 27, 43, 17, 25, 28, 3, 22, 6

17. En la siguiente tabla se presentan datos sobre el número de coches importados en Estados Unidos procedentes de Japón y Alemania, en los años comprendidos entre 1970 y 2002.

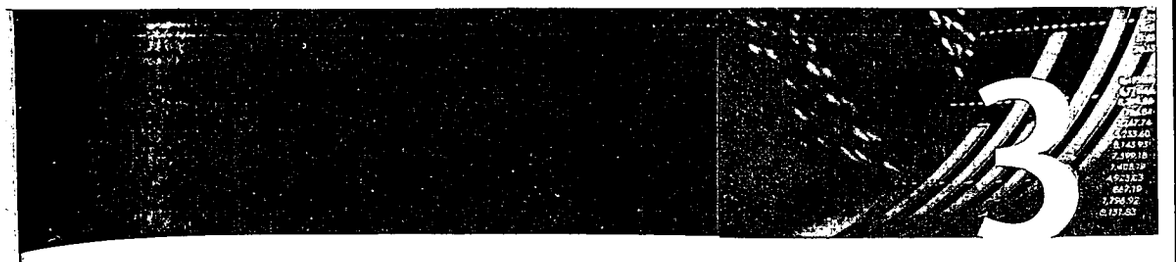
Coches nuevos importados en Estados Unidos

	Japón	Alemania
1970	381 338	674 945
1971	703 672	770 807
1972	697 788	676 967
1973	624 805	677 465
1974	791 791	619 757
1975	695 573	370 012
1976	1 128 936	349 804
1977	1 341 530	423 492
1978	1 563 047	416 231
1979	1 617 328	495 565
1980	1 991 502	338 711
1981	1 911 525	234 052
1982	1 801 185	259 385
1983	1 871 192	239 807
1984	1 948 714	335 032
1985	2 527 467	473 110
1986	2 618 711	451 699
1987	2 417 509	377 542
1988	2 123 051	264 249
1989	2 051 525	216 881
1990	1 867 794	245 286
1991	1 762 347	171 097
1992	1 598 919	205 248
1993	1 501 953	180 383
1994	1 488 159	178 774
1995	1 114 360	204 932
1996	1 190 896	234 909
1997	1 387 812	300 489
1998	1 456 081	373 330
1999	1 707 277	461 061
2000	1 839 093	488 323
2001	1 790 346	494 131
2002	2 046 902	574 455

Fuente: Oficina de Censos, División de Comercio Exterior.

(a) ¿Qué conclusiones se pueden sacar acerca del número de coches alemanes y japoneses importados en Estados Unidos desde 1990?

(b) Presente un diagrama de dispersión que relacione las importaciones de coches japoneses y alemanes desde 1990.



Uso de la Estadística para sintetizar conjuntos de datos

Odio los promedios. No se puede cometer mayor error que decir que la Aritmética es una ciencia exacta. Existen permutaciones y aberraciones discernibles para mentes perfectamente nobles como la mía; cambios sutiles que los contables normales no pueden descubrir, escondidas leyes de los números que sólo pueden ser percibidas por una mente como la mía. Por ejemplo, si se promedian números de abajo a arriba y después de arriba abajo, el resultado es siempre distinto.

Carta a la *Mathematical Gazette*
(revista matemática británica del siglo XIX)

La forma de dar sentido a los datos en bruto consiste en comparar y contrastar, para entender las diferencias.

Gregory Bateson, en *Pasos hacia una ecología de la mente*

3.1	Introducción	70
3.2	Media muestral	71
3.3	Mediana muestral	80
3.4	Moda muestral	96
3.5	Varianza muestral y desviación típica muestral	98
3.6	Conjuntos de datos normales y la regla empírica	108
3.7	Coeficiente de correlación muestral	121
	Términos clave	135
	Resumen	136
	Problemas de repaso	138

El objetivo de este capítulo es desarrollar medidas que se puedan usar para sintentizar un conjunto de datos. Estas medidas, formalmente llamadas *estadísticos*, son magnitudes numéricas cuyos valores vienen determinados por los datos. Se estudiarán la media muestral, la mediana muestral y la moda muestral, estadísticos que miden el centro o el valor central de un conjunto de datos. También se considerarán otros estadísticos que informan sobre la variación del conjunto de datos. Se aprenderá qué sig-

nifica que un conjunto de datos sea normal, y se presentará una regla empírica relativa a los conjuntos normales. También se estudiarán los conjuntos de datos compuestos por valores apareados, y se presentará un estadístico que sirve para medir el grado en el que un diagrama de dispersión de datos apareados se puede aproximar por una recta.

3.1 Introducción

En los experimentos actuales a menudo se hace un seguimiento de miles de individuos, y se observan algunas de sus características a lo largo del tiempo. Por ejemplo, en 1951, para conocer qué consecuencias en la salud se derivan de ciertas prácticas habituales, los médicos estadísticos R. Doll y A. B. Hill enviaron unos cuestionarios a todos los médicos de Reino Unido, y recibieron las respuestas de 40 000 médicos. En los cuestionarios se solicitaba información sobre la edad, los hábitos alimentarios y deportivos y sobre el consumo de tabaco. A esos médicos se les hizo un seguimiento durante 10 años, y se determinó la causa de muerte de los que fallecieron durante ese periodo de control. Como se puede imaginar, en ese estudio se utilizó un extensísimo conjunto de datos. Por ejemplo, aunque la atención se centrara en una única variable, tal como la edad de los doctores, en un determinado periodo de tiempo, el conjunto de datos resultante tendría un número de valores muy elevado: 40 000. Para que se pueda intuir la información contenida en un conjunto de datos tan grande es necesario resumir o sintetizar el conjunto de datos mediante una serie de medidas. En este capítulo se mostrarán los distintos estadísticos que se pueden utilizar para sintetizar determinadas características de un conjunto de datos.

Para empezar, supongamos que se dispone de un conjunto de datos muestrales procedentes de una población subyacente. Mientras que en capítulo 2 se mostró cómo describir y representar gráficamente los conjuntos de datos en toda su extensión, aquí nos interesaremos en determinar ciertas medidas sintéticas referidas a los datos. Estas medidas reciben el nombre de *estadísticos*. Se entiende por *estadístico* cualquier magnitud numérica cuyo valor se pueda determinar a partir de los datos.

Definición

Cualquier magnitud numérica calculable a partir de los datos se denomina *estadístico*.

Nos fijaremos en los estadísticos que describen la tendencia central del conjunto de datos; es decir, que describen el centro del conjunto de valores de datos. En las secciones 3.2, 3.3 y 3.4 se presentarán sucesivamente tres estadísticos de este tipo: la media muestral, la mediana muestral y la moda muestral. Una vez que se tenga una idea sobre el centro de un conjunto de datos, la cuestión que surge de manera natural es qué cantidad de *variación* existe. Esto es, ¿la mayor parte de los valores están próximos al centro, o, por el contrario, varían mucho alrededor de éste? En la sección 3.5 se analizarán la varianza muestral y la desviación típica muestral, que son estadísticos diseñados para medir la variación.

En la sección 3.6 se introducirá el concepto de conjunto normal de datos, áquel cuyo histograma tiene una forma acampanada. Para los conjuntos de datos próximos a la normalidad, se presentará una regla que se puede utilizar para aproximar la proporción de datos que distan de la media muestral menos de un cierto número de veces la desviación típica.

En las seis primeras secciones de este capítulo se tratan conjuntos de datos en los que cada dato está compuesto por un solo valor. Por otra parte, en la sección 3.7 se tratarán los datos apareados. Esto es, cada dato puntual consistirá en un valor x y un valor y. Por ejemplo, el valor x podría representar el número medio de cigarrillos que un fumador consume al día, mientras que el valor y podría identificarse con la edad de fallecimiento del individuo. Se introducirá un estadístico denominado *coeficiente de correlación muestral*, cuyo valor indica el grado en el que los datos puntuales con valores altos de x presentan, igualmente, valores altos de y; o, alternativamente, el grado en que valores bajos de x van unidos a valores bajos de y.

Del estudio de Doll y Hill se deduce que sólo alrededor del 1 por 1000 de los doctores no fumadores falleció de cáncer de pulmón. Entre los fumadores compulsivos, la cifra fue de 1 de cada 8. Adicionalmente, la tasa de mortalidad por ataque de corazón para los fumadores resultó ser un 50% más alta que para los no fumadores.

3.2 Media muestral

Supongamos que se dispone de una muestra de n datos cuyos valores serán designados por $x_1, x_2, \ldots, x_n$. Un estadístico usado para indicar el centro de este conjunto de datos es la *media muestral*, definida como la media aritmética de los valores de datos.

Definición

La *media muestral*, denotada por $\bar{x}$ (léase, "x barra"), se define por:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Ejemplo 3.1 Las eficiencias en el consumo de carburante (medidas en número de millas recorridas por galón de carburante) de los coches vendidos en Estados Unidos durante los años comprendidos entre 1999 y 2003 tuvieron como promedio:

28,2, 28,3, 28,4, 28,5, 29,0

Encuentre la media de este conjunto de datos.

Solución La media muestral $\bar{x}$ coincide con la media aritmética de los cinco valores de datos. Así pues,

$$\bar{x} = \frac{28{,}2 + 28{,}3 + 28{,}4 + 28{,}5 + 29{,}0}{5} = \frac{142{,}4}{5} = 28{,}48$$

Observe a partir de este ejemplo que, aunque la media muestral es la media aritmética de los distintos valores de datos, no tiene por qué coincidir con ninguno de éstos. □

Consideremos de nuevo el conjunto de datos $x_1, x_2, \ldots, x_n$. Si cada valor de dato se incrementa en una constante c, la media muestral se incrementa igualmente en el valor c. Matemáticamente, esto se puede expresar diciendo que si

$$y_i = x_i + c \quad \text{para} \quad i = 1, \ldots, n$$

se verifica que

$$\bar{y} = \bar{x} + c$$

donde $\bar{y}$ y $\bar{x}$ representan las medias muestrales de los valores y_i y de los valores x_i, respectivamente. Por consiguiente, cuando sea conveniente, se puede calcular $\bar{x}$ si se añade, primero, c a todos los valores de datos; después, se calcula la media muestral $\bar{y}$; y, finalmente, se resta c a $\bar{y}$ para obtener $\bar{x}$. Puesto que en ocasiones es más sencillo trabajar con los datos transformados en lugar de con los datos originales, el proceso indicado puede simplificar enormemente el cálculo de $\bar{x}$. El siguiente ejemplo ilustra este hecho.

Ejemplo 3.2 Las puntuaciones obtenidas por los ganadores del Torneo de Maestros de Golf de Estados Unidos entre 1981 y 1990 fueron las siguientes:

280, 284, 280, 277, 282, 279, 285, 281, 283, 278

Encuentre la media muestral de las puntuaciones anteriores.

Solución En vez de sumar directamente todos los valores anteriores, restemos primero 280 (esto es, $c = -280$) de cada uno de ellos. Así se obtienen los siguientes datos transformados:

0, 4, 0, −3, 2, −1, 5, 1, 3, −2

La media muestral $\bar{y}$, de estos últimos valores es

$$\bar{y} = \frac{0 + 4 + 0 - 3 + 2 - 1 + 5 + 1 + 3 - 2}{10} = \frac{9}{10}$$

Si a $\bar{y}$ le añadimos 280 se obtiene que la media de los datos originales es

$$\bar{x} = 280{,}9 \quad \blacksquare$$

Si cada valor de dato se multiplica por c, igualmente queda multiplicada por c la media resultante. Esto es, si

$$y_i = cx_i \quad i = 1, \ldots, n$$

se verifica que

$$\bar{y} = c\bar{x}$$

Por ejemplo, supongamos que la media de las alturas de un conjunto de individuos es de 5 pies. Si se quisiera cambiar la unidad de medida de pies a pulgadas, cada nuevo valor coincidiría con el antiguo multiplicado por 12. Se sigue que la media muestral de los datos nuevos coincide con $12 \cdot 5 = 60$. Es decir, la media muestral es de 60 pulgadas.

En el siguiente ejemplo se aborda el cálculo de la media muestral cuando los datos vienen dados mediante una tabla de frecuencias.

Ejemplo 3.3 El número de vestidos vendidos diariamente en una boutique de señoras durante los seis últimos días viene expresado en la tabla de frecuencias siguiente:

Valor	Frecuencia
3	2
4	1
5	3

¿Cuál es la media muestral?

Solución Dado que el conjunto de datos originales se compone de los siguientes 6 valores

3, 3, 4, 5, 5, 5

la media muestral resultante será

$$\bar{x} = \frac{3 + 3 + 4 + 5 + 5 + 5}{6}$$

$$= \frac{3 \times 2 + 4 \times 1 + 5 \times 3}{6}$$

$$= \frac{25}{6}$$

Esto es, la media muestral del número de vestidos vendidos diariamente es de 4,25. ■

En el ejemplo 3.3 se ha visto que, cuando los datos se dan mediante una tabla de frecuencias, la media muestral se puede expresar como la suma de los productos de los valores distintos y sus frecuencias dividida por el tamaño del conjunto de datos. Este resultado se verifica siempre. Para verlo, supongamos que los datos vienen dados en una tabla de frecuencias, donde se incluyen los k valores distintos, $x_1, x_2, \ldots, x_k$, junto con sus respectivas frecuencias, $f_1, f_2, \ldots, f_k$. El conjunto de datos consistirá, pues, en n observaciones, donde $n = \sum_{i=1}^{k} f_i$ y donde cada valor x_i aparece f_i veces, para $i = 1, \ldots, k$. Por consiguiente, la media muestral de este conjunto de datos será

$$\bar{x} = \frac{x_1 + \cdots + x_1 + x_2 + \cdots + x_2 + \cdots + x_k + \cdots + x_k}{n}$$

$$= \frac{f_1 x_1 + f_2 x_2 + \cdots + f_k x_k}{n} \tag{3.1}$$

Si $w_1, w_2, \ldots, w_k$ son números positivos que sumen 1, la suma

$$w_1x_1 + w_2x_2 + \cdots + w_kx_k$$

se dice que es la *media ponderada* de los valores $x_1, x_2, \ldots, x_k$; siendo w_i el peso de x_i. Por ejemplo, supongamos que $k = 2$. Si $w_1 = w_2 = 1/2$, la media ponderada

$$w_1x_1 + w_2x_2 = \frac{1}{2}x_1 + \frac{1}{2}x_2$$

coincide exactamente con la media ordinaria de x_1 y x_2. Si, por otra parte, $w_1 = 2/3$ y $w_2 = 1/3$, la media ponderada resultante

$$w_1x_1 + w_2x_2 = \frac{2}{3}x_1 + \frac{1}{3}x_2$$

asigna a x_1 un peso que es el doble del asignado a x_2.

Si se escribe la ecuación (3.1) en la forma

$$\bar{x} = \frac{f_1}{n}x_1 + \frac{f_2}{n}x_2 + \cdots + \frac{f_k}{n}x_k$$

se ve que la media muestral $\bar{x}$ es una media ponderada del conjunto de valores distintos. Los pesos dados al valor distinto x_i es f_i/n, la proporción de valores iguales a x_i. Así, por ejemplo, en el ejemplo 3.3 se podría escribir que

$$\bar{x} = \frac{2}{6} \times 3 + \frac{1}{6} \times 4 + \frac{3}{6} \times 5 = \frac{25}{6}$$

Ejemplo 3.4 En un artículo titulado "Los efectos del uso del casco sobre la gravedad de los daños craneales producidos en los accidentes de moto", publicado en el *Journal of the American Statistical Association* (1992, p. 48-56), A. Weiss analizó una muestra de 770 accidentes de moto similares ocurridos en el área de Los Ángeles en 1976 y 1977. Cada accidente se clasificó según la gravedad del daño sufrido por el conductor. La clasificación utilizada fue la siguiente:

Clasificación del accidente	Interpretación
0	Sin daño craneal
1	Daño menor
2	Daño moderado
3	Grave, sin peligro de muerte
4	Grave, con peligro de muerte
5	Crítico, supervivencia incierta en el momento del accidente
6	Fatal

En 331 de los accidentes el conductor llevaba casco, mientras que en los restantes 439 accidentes no fue así. La siguiente tabla muestra las frecuencias de las distintas gravedades de los accidentes en los que el conductor llevaba puesto el casco y en los que no lo llevaba.

Clasificación	Frecuencia entre los conductores con casco	Frecuencia entre los conductores sin casco
0	248	227
1	58	135
2	11	33
3	3	14
4	2	3
5	8	21
6	1	6
	331	439

Encuentre la media muestral de las clasificaciones de gravedad para los conductores que llevaban casco y para los que no lo llevaban.

Solución La media muestral para los conductores que llevaban casco es

$$\bar{x} = \frac{0 \cdot 248 + 1 \cdot 58 + 2 \cdot 11 + 3 \cdot 3 + 4 \cdot 2 + 5 \cdot 8 + 6 \cdot 1}{331} = \frac{143}{331} = 0{,}432$$

La media muestral para aquellos que no llevaban casco es

$$\bar{x} = \frac{0 \cdot 227 + 1 \cdot 135 + 2 \cdot 33 + 3 \cdot 14 + 4 \cdot 3 + 5 \cdot 21 + 6 \cdot 6}{439} = \frac{396}{439} = 0{,}902$$

Por consiguiente, los datos indican que aquellos motoristas que llevaban el casco sufrieron, como media, daños craneales menos graves que aquellos que no lo llevaban. ■

3.2.1 Desviaciones

Supongamos de nuevo que los datos muestrales consisten en los n valores $x_1, x_2, \ldots, x_n$, y que $\bar{x} = \sum_{i=1}^{n} x_i/n$ es la media muestral. Las diferencias entre cada uno de los valores y la media muestral se denominan *desviaciones*.

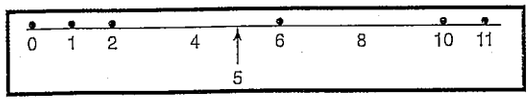


Figura 3.1 El centro de gravedad de 0, 1, 2, 6, 10, 11 es $(0 + 1 + 2 + 6 + 10 + 11)/6 = 30/6 = 5$.

Definición

Las *desviaciones* son las diferencias entre los valores de datos y la media muestral. El valor de la i-ésima desviación es $x_i - \bar{x}$.

Una identidad que puede resultar útil es que la suma de todas las desviaciones debe ser igual a 0. Es decir,

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

La certeza de esta igualdad se puede comprobar como sigue:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x}) &= \sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\bar{x}\\ &= n\bar{x} - n\bar{x}\\ &= 0\end{aligned}$$

Esta igualdad establece que la suma de todas las desviaciones positivas de la media muestral debe compensar exactamente la suma de todas las desviaciones negativas. En términos físicos, esto significa que si se colocan n pesos de igual masa en una varilla (sin peso) en los puntos x_i, $i = 1, \ldots, n$, $\bar{x}$ es el punto en el que la varilla se mantiene en equilibrio. Este centro de equilibrio se conoce con el nombre de *centro de gravedad* (figura 3.1).

Perspectiva histórica

En los primeros viajes marinos era bastante común que gran parte de la carga de un barco resultara dañada debido a las tormentas. Para compensar esta pérdida potencial, existía un acuerdo estándar, en el sentido de que todos aquellos que tenían mercancía a bordo deberían contribuir a pagar por el valor de los artículos perdidos o dañados. La cantidad que se reclamaba a cada uno de ellos se denominaba *havaria*, y de esta palabra latina se deriva el término inglés *average* (media, en español). [De hecho, si existían n personas que transportaban mercancías y las pérdidas sufridas por cada una de ellas fueran $x_1, \ldots, x_n$, la pérdida total sería $x_1 + \cdots + x_n$ y la havaria de cada uno se fijaba en $(x_1 + \cdots + x_n)/n$.]

Ejemplo 3.5 Con los datos del ejemplo 3.1, las desviaciones a la media muestral, 28,48, son

$$x_1 - \bar{x} = 28{,}2 - 28{,}48 = -0{,}28$$

$$x_2 - \bar{x} = 28{,}3 - 28{,}48 = -0{,}18$$

$$x_3 - \bar{x} = 28{,}4 - 28{,}48 = -0{,}08$$

$$x_4 - \bar{x} = 28{,}5 - 28{,}48 = -0{,}02$$

$$x_5 - \bar{x} = 29{,}0 - 28{,}48 = -0{,}52$$

Como comprobación, observe que la suma de las desviaciones es

$$-0{,}28 - 0{,}18 - 0{,}08 + 0{,}02 + 0{,}52 = 0$$ ■

Problemas

1. Los siguientes datos representan las calificaciones en un examen de Estadística para una muestra de estudiantes:

   87, 63, 91, 72, 80, 77, 93, 69, 75, 79, 70, 83, 94, 75, 88

   ¿Cuál es la media muestral?

2. Los siguientes datos (procedentes del Departamento de Agricultura, *Consumo de alimentos, precios y gastos*) muestran el consumo de queso (en libras) per cápita en Estados Unidos durante una muestra de años.

Año	1965	1975	1985	1995	2001
Consumo per cápita	10,0	14,8	23,4	26,4	30,1

   Encuentre la media muestral de los datos anteriores.

3. Los datos siguientes muestran los promedios anuales de lluvia caída (en pulgadas) y de días con precipitación en una muestra de ciudades.

Ciudad	Promedio de lluvia	Promedio de días con precipitación
Albany, NY	35,74	134
Baltimore, MD	31,50	83
Casper, WY	11,43	95
Denver, CO	15,31	88
Fargo, ND	19,59	100
Houston, TX	44,76	105
Knoxville, TN	47,29	127
Los Angeles, CA	12,08	36
Miami, FL	57,55	129
New Orleans, LA	59,74	114
Pittsburgh, PA	36,30	154
San Antonio, TX	29,13	81
Wichita, KS	28,61	85

Fuente: Administración Oceánica y Atmosférica Nacional.

(a) Encuentre la media muestral de los promedios de lluvia en pulgadas.

(b) Encuentre la media muestral de los promedios de los días con precipitación.

4. Considere cinco números y suponga que la media de los cuatro primeros es 14.

(a) Si el quinto número es 24, ¿cuál es la media de los cinco números?

(b) Si la media de los cinco números es 24, ¿cuál es el valor del quinto número?

5. Los siguientes datos, sacados del *Resumen estadístico de Estados Unidos,* 1993, muestra el número de policías fallecidos en actos de servicio en Estados Unidos durante los años comprendidos entre 1979 y 1990. Encuentre la media muestral de estos datos.

164, 165, 157, 164, 152, 147, 148, 131, 147, 155, 145, 132

6. Suponga que la media muestral de un conjunto de 10 datos puntuales es $\bar{x} = 20$.

(a) Si se descubre que se ha leído incorrectamente un dato con valor 15 y que se le ha dado el valor 13, ¿cuál será el valor revisado de la media muestral?

(b) Si existiera un dato adicional con valor 22, ¿aumentaría o disminuiría el valor de $\bar{x}$?

(c) Con los datos originales [y no con los datos revisados en el apartado (a)], ¿cuál sería el nuevo valor de $\bar{x}$ del apartado (b)?

7. La tabla siguiente lista la cantidad anual de casos de tétano que han sido notificados en Estados Unidos para una muestra de años. Calcule la media muestral.

Año	1970	1975	1980	1982	1984	1985	1987	2001
Número de casos	148	102	95	88	74	83	48	62

Fuente: Centro de control de enfermedades de Estados Unidos. *Sumario de enfermedades notificables, morbosidad y mortalidad.*

8. El siguiente gráfico de tallos y hojas refleja los puntos obtenidos por el autor de este texto en 15 juegos de bolos. Calcule la media muestral.

18	2, 4, 7
17	0
16	1, 9
15	2, 2, 4, 8, 8
14	
13	2, 1, 5, 5

9. Encuentre la media muestral de este conjunto de datos:

1, 2, 4, 7, 10, 12

Calcule, ahora, las medias de los conjuntos de datos

3, 6, 12, 21, 30, 36 y 6, 7, 9, 12, 15, 17

10. Suponga que $\bar{x}$ es la media muestral de un conjunto de datos compuesto por los valores $x_1, \ldots, x_n$. Si los datos se transforman de acuerdo con la expresión

$$y_i = ax_i + b \qquad i = 1, \ldots, n$$

¿Cuál es la media muestral del conjunto de datos $y_1, \ldots, y_n$? (En la expresión anterior, a y b son constantes dadas.)

11. Los datos siguientes reflejan el número total de incendios en Ontario (Canadá), ocurridos en los sucesivos meses del año 2002.

6, 13, 5, 7, 7, 3, 7, 2, 5, 6, 9, 8

Encuentre la media muestral de este conjunto de datos.

12. El siguiente conjunto de datos muestra el número total de coches vendidos en Estados Unidos en una muestra de años. Los datos están dados en unidades de miles de coches. Encuentre la media muestral del número de coches vendidos anualmente en dichos años.

Año	1980	1985	1990	1995	2000	2002
Número de ventas	8010	11 653	9783	11 985	12 832	12 326

Fuente: Resumen estadístico de Estados Unidos, 1990.

13. La mitad de los valores de una muestra son iguales a 10, y los de la otra mitad son todos iguales a 20. ¿Cuál es la media muestral?

14. La siguiente tabla de frecuencias refleja las edades de los componentes de una joven orquesta sinfónica.

Edades	Frecuencias
16	9
17	12
18	15
19	10
20	8

Encuentre la media muestral de las edades dadas.

15. En una muestra de datos, la mitad de los valores son iguales a 10, una sexta parte son iguales a 20 y una tercera parte son iguales a 30. ¿Cuál es la media muestral?

16. Existen dos entradas a un aparcamiento. El estudiante 1 contabiliza el número de coches que pasan diariamente a través de la entrada 1, y el estudiante 2 hace lo mismo en la entrada 2. A lo largo de 30 días, los datos del estudiante 1 tienen una media muestral de 122, mientras la media muestral de los datos del estudiante 2 es igual a 160. Sobre los 30 días citados, ¿cuál fue el número medio de coches que entraron en el aparcamiento?

17. Una compañía tiene dos plantas de producción. El salario medio de una muestra de 30 ingenieros de la planta 1 fue de 33 600 dólares, mientras que el salario medio de una muestra de 20 ingenieros de la planta 2 resultó ser de 42 400 dólares. ¿Cuál es el salario medio muestral de de los 50 ingenieros seleccionados?

18. Supongamos que se dispone de dos muestras distintas, de tamaños n_1 y n_2. Si la media muestral de la primera muestra es $\bar{x}_1$ y la de la segunda muestra es $\bar{x}_2$, ¿cuál es la media de la muestra conjunta, de tamaño $n_1 + n_2$?

19. Encuentre las desviaciones para cada uno de los tres conjuntos de datos del problema 9, y contraste la veracidad de sus respuestas mediante la comprobación de que, en cada caso, la suma de las desviaciones es 0.

20. Calcule las desviaciones de los datos del problema 14 y compruebe que su suma es 0.

3.3 Mediana muestral

Los siguientes datos representan el número de semanas que, tras completar un curso para aprender a conducir, tardó cada miembro de una muestra de siete personas en obtener su carné de conducir:

$$2, 110, 5, 7, 6, 7, 3$$

La media muestral de este conjunto de datos es $\bar{x} = 140/7 = 20$; así pues, seis de los siete valores de datos son menores que la media muestral, mientras que el séptimo valor es muy

superior a ésta. Lo que apunta una debilidad de la media muestral como indicador del centro de un conjunto de datos: a saber, su valor se encuentra muy afectado por los valores de datos extremos.

Un estadístico que se utiliza también para representar el centro de un conjunto de datos es la *mediana muestral*, definida como el valor medio cuando los datos están ordenados de menor a mayor. La mediana muestral será denotada por m.

Definición

Ordene los valores de datos de menor a mayor. Si el número de datos es impar, la *mediana muestral* coincide con el valor que se encuentra en la posición central en la lista ordenada; si el número de datos es par, la *mediana muestral* es la media de los dos valores que ocupan las posiciones centrales.

De esta definición se deduce que, si existen tres valores de datos, la mediana muestral coincide con el segundo valor más pequeño; mientras que, si existen cuatro valores, coincide con la media de los valores más pequeños segundo y tercero.

Ejemplo 3.6 Los siguientes datos representan el número de semanas que siete individuos tardaron en obtener su carné de conducir. Encuentre la mediana muestral.

$$2, 110, 5, 7, 6, 7, 3$$

Solución Ordenemos primero los datos en orden creciente.

$$2, 3, 5, 6, 7, 7, 110$$

Puesto que el tamaño de la muestra es 7, la mediana muestral será el cuarto valor más pequeño. Esto es, la mediana muestral del número de semanas que se tardó en obtener el carné de conducir es $m = 6$ semanas. □

Ejemplo 3.7 Los siguientes datos representan el número de días que a seis individuos les costó dejar de fumar tras completar un cursillo diseñado para este propósito.

$$1, 2, 3, 5, 8, 100$$

¿Cuál es la mediana muestral?

Solución Puesto que el tamaño muestral es 6, la mediana muestral es la media de los dos valores centrales una vez ordenados; así pues,

$$m = \frac{3 + 5}{2} = 4$$

Es decir, la mediana muestral es de 4 días. □

En general, para un conjunto de datos de n valores, la mediana muestral coincide con el $[(n + 1)/2]$ menor valor ordenado, cuando n es impar, y coincide con la media de los valores ordenados que ocupan las posiciones $(n/2)$ y $(n/2 + 1)$, cuando n es par.

Tanto la media muestral como la mediana muestral son estadísticos útiles para describir la tendencia central de un conjunto de datos. La media muestral, siendo la media aritmética, utiliza todos los valores de datos. La mediana muestral, puesto que sólo utiliza un único valor central o bien un par de valores centrales, no se ve afectada por los valores extremos.

Ejemplo 3.8 Los datos siguientes proporcionan los nombres de los máximos encestadores individuales de la Asociación de Baloncesto Nacional (NBA) junto con su promedio de puntos por partido en las temporadas comprendidas entre 1953 y 1991.

Temporada	Jugador, equipo	Promedio de puntos
1953–54	Neil Johnston, Philadelphia Warriors	24,4
1954–55	Neil Johnston, Philadelphia Warriors	22,7
1955–56	Bob Pettit, St. Louis Hawks	25,7
1956–57	Paul Arizin, Philadelphia Warriors	25,6
1957–58	George Yardley, Detroit Pistons	27,8
1958–59	Bob Pettit, St. Louis Hawks	29,2
1959–60	Wilt Chamberlain, Philadelphia Warriors	37,6
1960–61	Wilt Chamberlain, Philadelphia Warriors	38,4
1961–62	Wilt Chamberlain, Philadelphia Warriors	50,4
1962–63	Wilt Chamberlain, San Francisco Warriors	44,8
1963–64	Wilt Chamberlain, San Francisco Warriors	36,9
1964–65	Wilt Chamberlain, San Francisco Warriors–Phila. 76ers	34,7
1965–66	Wilt Chamberlain, Philadelphia 76ers	33,5
1966–67	Rick Barry, San Francisco Warriors	35,6
1967–68	Dave Bing, Detroit Pistons	27,1
1968–69	Elvin Hayes, San Diego Rockets	28,4
1969–70	Jerry West, Los Angeles Lakers	31,2
1970–71	Lew Alcindor, Milwaukee Bucks	31,7
1971–72	Kareem Abdul-Jabbar, Milwaukee Bucks	34,8
1972–73	Nate Archibald, Kansas City–Omaha Kings	34,0
1973–74	Bob McAdoo, Buffalo Braves	30,8
1974–75	Bob McAdoo, Buffalo Braves	34,5
1975–76	Bob McAdoo, Buffalo Braves	31,1
1976–77	Pete Maravich, New Orleans Jazz	31,1
1977–78	George Gervin, San Antonio Spurs	27,2
1978–79	George Gervin, San Antonio Spurs	29,6
1979–80	George Gervin, San Antonio Spurs	33,1
1980–81	Adrian Dantley, Utah Jazz	30,7
1981–82	George Gervin, San Antonio Spurs	32,3
1982–83	Alex English, Denver Nuggets	28,4
1983–84	Adrian Dantley, Utah Jazz	30,6
1984–85	Bernard King, New York Knicks	32,9
1985–86	Dominique Wilkins, Atlanta Hawks	30,3

Temporada	Jugador, equipo	Promedio de puntos
1986–87	Michael Jordan, Chicago Bulls	37,1
1987–88	Michael Jordan, Chicago Bulls	35,0
1988–89	Michael Jordan, Chicago Bulls	32,5
1989–90	Michael Jordan, Chicago Bulls	33,6
1990–91	Michael Jordan, Chicago Bulls	31,5
1991–92	Michael Jordan, Chicago Bulls	30,1

(a) Encuentre la mediana muestral del promedio de puntos.

(b) Calcule la media muestral del promedio de puntos.

Elimine las temporadas que comienzan en 1961 y en1962, en las que Wilt Chamberlain tuvo un promedio de 50,4 y 44,8 puntos por partido, respectivamente, y encuentre

(c) la mediana muestral

(d) la media muestral

Solución

(a) Puesto que existen 39 valores de datos, la mediana muestral coincide con el 20° valor menor. Existen 11 valores entre 20 y 29, por tanto, la mediana será el noveno valor menor cuando se eliminen todos los promedios inferiores a 30. Si se ordenan los restantes valores se obtiene

$$30{,}1,\ 30{,}3,\ 30{,}6,\ 30{,}7,\ 30{,}8,\ 31{,}1,\ 31{,}1,\ 31{,}2,\ 31{,}5,\ \ldots$$

En consecuencia, la mediana muestral es

$$m = 31{,}5$$

(b) La suma de los 39 valores es 1256,9 y, por tanto, la media muestral es

$$\bar{x} = \frac{1256{,}9}{39} \approx 32{,}228$$

Perspectiva histórica

El matemático holandés Christian Huygens fue uno de los primeros científicos que se dedicaron a la Teoría de la Probabilidad. En 1669, su hermano Ludwig, después de estudiar las tablas de mortalidad de la época, escribió a su famoso hermano mayor: "He estado confeccionando una tabla que muestra cuánto tiempo puede vivir la gente... ¡Vivir bien! De acuerdo con mis cálculos, tú deberías vivir unos $56\frac{1}{2}$ años, y yo 55." Christian, intrigado, analizó las tablas de mortalidad, pero llegó a unos estimadores, respecto a los años que ambos deberían vivir, distintos de los de su hermano. ¡Ludwig basó sus estimadores en la mediana muestral, mientras que Christian se basó en la media muestral!

(c) Si se eliminan los dos años especificados, la mediana es el 19° valor menor de los 37 valores restantes. A partir de la ordenación dada en (a), que comienza en el 12° valor menor, se obtiene que la mediana muestral es ahora

$$m = 31{,}2$$

(d) Si se eliminan los dos años citados, la suma de todos los valores de datos restantes se reduce a

$$1256{,}9 - 50{,}4 - 44{,}8 = 1161{,}7$$

Por tanto, la media muestral es ahora

$$\bar{x} = \frac{1161{,}7}{37} = 31{,}397$$

Así pues, se ve que eliminar los dos mayores valores del conjunto de datos tiene un efecto relativamente pequeño sobre la mediana, y la reduce de 31,50 a 31,20; mientras que su efecto sobre la media es mucho mayor, pues la reduce de 32,23 a 31,40. □

Para los conjuntos de datos aproximadamente simétricos sobre su valor central, la media muestral y la mediana muestral tienen valores próximos. Por ejemplo, los datos

$$4, 6, 8, 8, 9, 12, 15, 17, 19, 20, 22$$

son grosso modo simétricos alrededor del valor 12, que es su mediana muestral. La media muestral es $\bar{x} = 140/11 = 12{,}73$, que se encuentra próxima a 12.

La respuesta a la pregunta sobre cuál de los dos estadísticos sumariales es más informativo depende de qué es lo que se pretende conocer del conjunto de datos. Por ejemplo, si el gobierno establece un impuesto sobre la renta con tarifa plana (proporcional) y se pretende averiguar qué recaudación cabe esperar, la renta media de los ciudadanos será más interesante que la mediana (¿por qué?). Por el contrario, si el gobierno estuviera interesado en determinar un valor central de la cantidad de renta que los ciudadanos dedican a la vivienda, la mediana muestral podría ser más informativa (¿por qué?).

Aunque es interesante analizar si la media muestral o la mediana muestral es mas informativa, en una situación concreta, obseve que no debemos restringir nuestro conocimiento a sólo una de dichas magnitudes. Ambas son importantes y, por tanto, las dos se han de calcular cuando se está sintetizando un conjunto de datos.

Problemas

1. Los siguientes datos exponen las distancias que se recorren en una muestra de cursos de golf municipales.

   7040, 6620, 6050, 6300, 7170, 5990, 6330, 6780, 6540, 6690, 6200, 6830

   (a) Encuentre la mediana muestral.

   (b) Encuentre la media muestral.

2. (a) Determine la mediana muestral del conjunto de datos.

   14, 22, 8, 19, 15, 7, 8, 13, 20, 22, 24, 25, 11, 9, 14

   (b) Incremente cada valor de (a) en 5 unidades, y encuentre la nueva mediana muestral.

   (c) Multiplique por 3 cada valor de (a), y encuentre la nueva mediana muestral.

3. Si la mediana de un conjunto de datos x_i, $i = 1, \ldots, n$, es 10, ¿cuál es la mediana del conjunto de datos $2x_i + 3$, $i = 1, \ldots, n$?

4. Los siguientes datos reflejan las velocidades de 40 coches, medidas por radar en una calle de cierta ciudad.

   22, 26, 31, 38, 27, 29, 33, 40, 36, 27, 25, 42, 28, 19, 28, 26, 33, 26, 37, 22,
   31, 30, 44, 29, 25, 17, 46, 28, 31, 29, 40, 38, 26, 43, 45, 21, 29, 36, 33, 30

   Encuentre la mediana muestral.

5. Las tablas siguientes muestran las tasas de suicidio, de hombres y mujeres, por 100 000 individuos para un conjunto de países.

   Tasas de suicidios por 100 000 individuos

Sexo	Estados Unidos	Australia	Austria	Canadá	Dinamarca	Francia
Mujeres	5,4	5,1	15,8	5,4	20,6	12,7
Hombres	19,7	18,2	42,1	20,5	35,1	33,1

Sexo	Italia	Japón	Holanda	Polonia	Suecia	Reino Unido	Alemania
Mujeres	4,3	14,9	8,1	4,4	11,5	5,7	12,0
Hombres	11,0	27,8	14,6	22,0	25,0	12,1	26,6

   Fuente: Organización Mundial de la Salud, *Estadística sobre la Salud Mundial.*

   (a) Encuentre la mediana muestral para las tasas de suicidio de los hombres.

   (b) Encuentre la mediana muestral para las tasas de suicidio de las mujeres.

   (c) Encuentre la media muestral para las tasas de suicidio de los hombres.

   (d) Encuentre la media muestral para las tasas de suicidio de las mujeres.

6. Obtenga la mediana muestral del número medio anual de días de lluvia para las ciudades incluidas en el problema 3 de la sección 3.2.

7. Encuentre la mediana muestral del número medio anual de pulgadas de precipitación para las ciudades incluidas en el problema 3 de la sección 3.2.

8. Busque la mediana muestral de los datos presentados en el problema 8 de la sección 2.3.

9. Utilice la tabla sobre tasas de mortalidad que antecede al problema 9 de la sección 2.3 para calcular la mediana de las tasas de mortalidad debidas a:

   (a) caídas

   (b) envenenamientos

   (c) ahogamientos

10. La mediana muestral de 10 valores distintos es 5. Deduzca qué se podría decir acerca de la nueva mediana muestral si:

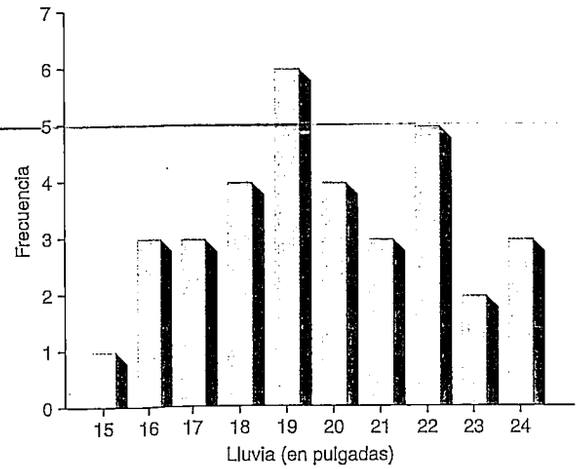
    (a) Al conjunto de datos se le añade un nuevo dato con valor 7.

    (b) Se añaden dos nuevos datos con valores 3 y 42.

11. El histograma de la figura de abajo representa las lluvias anuales, en pulgadas, caídas en una ciudad occidental durante los últimos 34 años. Puesto que los datos originales no son recuperables a partir del histograma, no se puede utilizar éste para calcular el valor de la media muestral y la mediana muestral. Aún así, basándonos en este histograma, diga cuál es el mayor valor posible de:

    (a) la media muestral

    (b) la mediana muestral



Diga, cuál es el menor valor posible de:

(c) la media muestral

(d) la mediana muestral

(e) Los datos reales se muestran a continuación:

15,2, 16,1, 16,5, 16,7, 17,2, 17,5, 17,7, 18,3, 18,6, 18,8, 18,9, 19,1,
19,2, 19,2, 19,6, 19,8, 19,9, 20,2, 20,3, 20,3, 20,8, 21,1, 21,4, 21,7,
22,2, 22,5, 22,5, 22,7, 22,9, 23,3, 23,6, 24,1, 24,5, 24,9

Determine la media muestral y la mediana muestral, y compruebe si concuerdan con sus respuestas anteriores.

12. Los datos siguientes muestran las temperaturas máxima y mínima (en grados Fahrenheit) del 4 de julio de 1993 en varias ciudades.

Ciudad	Temperaturas máxima/mínima del 4 de julio de 1993
Atlanta	96/75
Boise	75/53
Cleveland	90/68
Jacksonville	95/75
Norfolk	89/73
Providence	89/68
Rochester	85/59
Seattle	68/55
Toledo	93/71
Wilmington	95/71

Fuente: Periódico *New York Times*, 5 de julio de 1993.

(a) Encuentre la mediana muestral de las temperaturas máximas.

(b) Encuentre la mediana muestral de las temperaturas mínimas.

(c) Encuentre la mediana muestral de las diferencias entre las temperaturas máxima y mínima.

13. Con los datos del ejemplo 3.4, calcule las medianas muestrales de las lesiones craneales graves sufridas por los conductores de moto que llevaban puesto el casco y por los que no lo llevaban.

13. En las situaciones siguientes, ¿cuál de los estadísticos media muestral o mediana muestral piensa que es más informativo?

    (a) Para analizar si se debe cerrar una línea de autobús entre Rochester y Nueva York, un ejecutivo ha recopilado el número de viajeros en una muestra de días.

    (b) Para comparar a los estudiantes universitarios actuales con los de años anteriores, se consultan muestras de las calificaciones obtenidas en los exámenes de acceso a la universidad durante varios años.

    (c) El abogado defensor de un proceso judicial con jurado popular está analizando las puntuaciones de un test de inteligencia (IQ) obtenidas por los miembros del jurado.

    (d) Usted ha comprado su casa hace 6 años en una pequeña comunidad por un precio de 105 000 dólares, que coincidía con el precio medio y mediano de todas las casas que se vendieron aquel año en dicha comunidad. Sin embargo, en los dos últimos

años, se han construido varias casas nuevas mucho más caras que las anteriores. Para obtener una idea del valor actual de su casa, usted decide analizar los precios de venta de las casas vendidas recientemente en su comunidad.

15. Las mujeres suponen los siguientes porcentajes de fuerza laboral en las ocupaciones que se listan a continuación.

Ocupación	Porcentaje de mujeres	Ocupación	Porcentaje de mujeres
Ejecutivas de empresa	36,8	Médicos	17,6
Enfermeras	94,3	Abogadas	18,0
Supervisoras de ventas	30,5	Profesoras de enseñanza básica	85,2
Vendedoras	68,6	Empleadas de correos	43,5
Bomberas	1,9	Policías	10,9
Empleadas de la limpieza	41,5	Supervisoras de construcción	1,6
Trabajadoras de la construcción	2,8	Conductoras de camiones	2,1

Encuentre para estos porcentajes:

(a) la media muestral

(b) la mediana muestral

Adicionalmente, resulta que las mujeres representan el 44,4% de la fuerza laboral total en todas las ocupaciones anteriores. ¿Esto resulta coherente con sus respuestas a los apartados (a) y (b)? Explique por qué.

16. Con los datos relativos a los 30 primeros estudiantes del Apéndice A, encuentre la mediana muestral y la media muestral de:

(a) los pesos

(b) los niveles de colesterol

(c) las presiones sanguíneas

17. La tabla siguiente muestra las edades medianas de hombres y mujeres en el momento de contraer su primer matrimonio, correspondientes a las bodas celebradas entre los años 1992 y 2002.

(a) Encuentre la mediana muestral de la edad mediana de los hombres.

(b) Encuentre la mediana muestral de la edad mediana de las mujeres.

Edad mediana en el primer matrimonio, en Estados Unidos

Año	Hombres	Mujeres	Año	Hombres	Mujeres
2002	26,9	25,3	1996	27,1	24,8
2001	26,9	25,1	1995	26,9	24,5
2000	26,8	25,1	1994	26,7	24,5
1999	26,9	25,1	1993	26,5	24,5
1998	26,7	25,0	1992	26,5	24,4
1997	26,8	25,0			

3.3.1 Percentiles muestrales

La mediana muestral es un caso particular de los estadísticos conocidos como *percentiles muestrales de orden 100p por ciento*, donde p puede ser cualquier valor comprendido entre 0 y 1: Grosso modo, el percentil muestral de orden $100p$ por ciento es aquel valor que verifica que el $100p$ por ciento de los valores de los datos son menores que él y que el $100(1—p)\%$ de los valores de los datos son mayores que él.

Definición

El *percentil muestral de orden 100p por ciento* es aquel valor de dato que tiene la propiedad de que al menos el $100p$ por ciento de los valores de datos son menores o iguales que él y que al menos el $100(1—p)$ por ciento de los valores de datos son mayores o iguales que él. Si existen dos valores de datos que cumplen las condiciones anteriores, el percentil muestral de orden $100p$ por ciento se define como la media aritmética de ambos valores de datos.

Observe que la mediana muestral se corresponde con el percentil muestral de orden 50%. Es decir, coincide con el percentil muestral de orden $100p$ por ciento cuando $p = 0{,}50$.

Supongamos que se han ordenado de menor a mayor todos los valores de datos de una muestra de tamaño n. Para determinar el percentil muestral de orden $100p$ por ciento se debe encontrar aquel valor que verifica que:

1. Al menos np valores de datos son menores o iguales que él.
2. Al menos $n(1—p)$ valores de datos son mayores o iguales que él.

Ahora bien, si np no es un entero, el único valor de dato que cumple ambos puntos es aquel cuya posición de orden coincide con el primer entero superior a np. Por ejemplo, supongamos que se quiere determinar el percentil muestral de orden 90% en una muestra de tamaño $n = 12$. Puesto que $p = 0{,}9$, se tiene que $np = 10{,}8$ y $n(1—p) = 1{,}2$. Así pues, se estarán buscando aquellos valores para los que:

1. Al menos 10,8 valores de datos sean menores o iguales que él (por consiguiente, el valor de datos debe estar en la posición de orden 11 o mayor).
2. Al menos 1,2 valores de datos sean mayores o iguales que él (por tanto, debe ocupar la posición de orden 11 o menor).

Evidentemente, el único valor de dato que cumple ambos puntos es aquél que ocupa la posición de orden 11, y, en consecuencia, éste será el percentil muestral de orden 90%.

Por otro lado, si np es un entero, tanto el valor de dato que ocupa la np posición de orden como el valor de dato que ocupa la posición de orden $np + 1$ cumplen las condiciones de las definiciones 1. y 2.; en este caso, el percentil muestral de orden $100p$ por ciento será igual a la media aritmética de los dos valores de datos anteriores. Por ejemplo, supongamos que se desea encontrar el percentil muestral de orden 95% en un conjunto de datos con $n = 20$ valores. En este caso, tanto el 19° valor como el 20° valor (los dos valores mayores) serán mayores o iguales que al menos $np = 20(0{,}95) = 19$ de los valores de datos, y serán menores o iguales que al menos $n(1—p) = 1$ de dichos valores. El percentil muestral de orden 95% será, pues, la media aritmética de los valores que ocupan las posiciones de orden 19 y 20 (es decir, los dos mayores).

Resumiendo, se ha demostrado lo siguiente.

Para encontrar el percentil muestral de orden $100p\%$ de un conjunto de datos de tamaño n

1. Ordene los datos en sentido creciente.
2. Si np no es un entero, determine el menor entero mayor que np. El valor de dato que ocupa la posición de orden igual a este último entero será el percentil muestral de orden $100p$ por ciento.
3. Si np es un entero, el percentil muestral de orden $100p$ por ciento coincidirá con la media aritmética de los valores que ocupan las posiciones de orden np y $np + 1$.

Ejemplo 3.9 ¿Cómo se calcula el percentil muestral de orden 90% cuando el tamaño muestral es (a) 8, (b) 16 y (c) 100?

Solución

(a) Puesto que $0{,}9 \times 8 = 7{,}2$ no es un entero, si se ordenan los datos de menor a mayor, el percentil muestral de orden 90% coincidirá con el octavo valor menor (es decir, el valor mayor).

(b) Puesto que $0{,}9 \times 16 = 14{,}4$ no es un entero, el percentil muestral de orden 90% será el 15° valor menor.

(c) Puesto que $0{,}9 \times 100 = 90$ es un entero, el percentil muestral de orden 90% coincidirá con la media aritmética de los valores que ocupan las posiciones 90 y 91 una vez que los datos han sido ordenados de menor a mayor. ■

Ejemplo 3.10 La tabla 3.1 lista las primeras 20 universidades de Estados Unidos en una clasificación basada en los activos que han generado. Utilice estos datos para encontrar:

(a) el percentil muestral de orden 90%

(b) el percentil muestral de orden 20%

Solución

(a) Puesto que el tamaño muestral es 20 y $20 \times 0{,}9 = 18$, el percentil muestral de orden 90% coincide con la media aritmética entre los valores más pequeños 18° y 19° o, equivalentemente, la media aritmética entre los valores más grandes 2° y 3°. De donde:

$$\text{Percentil muestral de orden } 90\% = \frac{10\,523\,600 + 8\,630\,679}{2} = 9\,977\,140$$

Es decir, el percentil muestral de orden 90% para este conjunto de datos es aproximadamente igual a 9,98 miles de millones de dólares.

Tabla 3.1 Las 20 universidades más altas en la clasificación de becas generadas, 2002*

Universidad	Activos†	Universidad	Activos†
1. Harvard University	17 169 757 $	11. Washington University	3 517 104 $
2. Yale University	10 523 600	12. University of Pennsylvania	3 393 297
3. University of Texas System	8 630 679	13. University of Michigan	3 375 689
4. Princeton University	8 319 600	14. University of Chicago	3 255 368
5. Stanford University	7 613 000	15. Northwestern University	3 022 733
6. Massachusetts Institute of Technology	5 359 423	16. Rice University	2 939 804
		17. Duke University	2 927 478
7. Emory University	4 551 873	18. Cornell University	2 853 742
8. Columbia University	4 208 373	19. University of Notre Dame	2 554 004
9. University of California	4 199 067	20. Dartmouth College	2 186 610
10. The Texas A&M University System and Foundations	3 743 442		

Observación: Valor de mercado de los activos generados, excluyendo las donaciones privadas y el capital de trabajo.
* Con fecha de 30 de junio de 2002.
† En miles.
Fuente: Asociación Nacional de Agentes de Negocios Universitarios (NACUBO).

(b) Puesto que $20 \times 0{,}2 = 4$, el percentil muestral de orden 20% es el promedio entre los valores menores 4° y 5°, se obtiene el resultado:

$$\text{(Percentil muestral de orden } 20\%) = \frac{2\,927\,478 + 2\,939\,804}{2} = 2\,823\,641$$ ■

Los percentiles muestrales de órdenes 25, 50 y 75% se conocen como *cuartiles*.

Definición

El percentil muestral de orden 25% se llama *primer cuartil*. El percentil muestral de orden 50% se denomina *mediana* o *segundo cuartil*. El percentil muestral de orden 75% se llama *tercer cuartil*.

Los cuartiles dividen el conjunto de datos en cuatro partes, de forma que, aproximadamente, un 25% de los valores de datos se encuentran por debajo del primer cuartil, otro 25% de los valores se encuentra entre el primer y el segundo cuartil, un tercer 25% se encuentra entre el segundo y el tercer cuartil y, por último, el 25% restante de los valores supera al tercer cuartil.

Ejemplo 3.11 Encuentre los cuartiles muestrales para los siguientes 18 valores de datos, que se muestran ordenados y representan las puntuaciones de una liga de bolos.

122, 126, 133, 140, 145, 145, 149, 150, 157, 162, 166, 175, 177, 177, 183, 188, 199, 212

Solución Puesto que $0{,}25 \times 18 = 4{,}5$, el percentil muestral de orden 25% coincide con el quinto valor menor, que es 145.

Dado que $0{,}50 \times 18 = 9$, el segundo cuartil (o mediana muestral) es igual a la media del noveno y décimo valor menor, es decir, su valor es:

$$\frac{157 + 162}{2} = 159{,}5$$

Finalmente, puesto que $0{,}75 \times 18 = 13{,}5$, el tercer cuartil coincide con el 15° valor menor, que es 177. □

Problemas

1. Se han ordenado 75 valores en sentido creciente. ¿Cómo se determinarían los percentiles muestrales siguientes del conjunto de datos?
   (a) percentil de orden 80%
   (b) percentil de orden 60%
   (c) percentil de orden 30%
2. La siguiente tabla muestra las exportaciones de plátanos, en toneladas métricas, en una selección de países iberoamericanos y caribeños. Encuentre los cuartiles.

Exportaciones de plátanos, año 2000, en toneladas métricas

Ecuador	4 095 191
Costa Rica	2 113 652
Colombia	1 710 949
Guatemala	857 164
Panamá	489 805
Honduras	183 400
México	81 044
Santa Lucía	72 795
Brasil	72 468
Belice	64 400
República Dominicana	62 429
Nicaragua	44 402
San Vicente y las Granadinas	43 810
Jamaica	40 900
Surinam	34 000
Venezuela	33 543
Dominica	29 810
Bolivia	9 377

Exportaciones de plátanos, año 2000, en toneladas métricas (*Continuación*)

Perú	856
Granada	707
Argentina	412
Trinidad y Tobago	87
El Salvador	72
Paraguay	66
Chile	18
Guayana	10
Total	10 041 367

Fuente: FAO.

3. Considere un conjunto de datos con n valores. Diga cómo se calcula el percentil muestral de orden 95% cuando
   (a) $n = 100$
   (b) $n = 101$

La tabla siguiente muestra el número de médicos y dentistas por cada 100 000 habitantes para 12 Estados del occidente medio de Estados Unidos en el año 2000. Los problemas 4 y 5 se basan en ella.

Estado	Tasa de médicos	Tasa de dentistas
Ohio	188	56
Indiana	146	48
Illinois	206	61
Michigan	177	64
Wisconsin	177	70
Minnesota	207	70
Iowa	141	60
Missouri	186	55
North Dakota	157	55
South Dakota	129	54
Nebraska	162	71
Kansas	166	52

Fuente: Asociación Médica Americana, *Características y Distribuciones Médicas en Estados Unidos.*

4. Encuentre, para las tasas de médicos por cada 100 000 habitantes:
   (a) el percentil muestral de orden 40%
   (b) el percentil muestral de orden 60%
   (c) el percentil muestral de orden 80%

5. Encuentre, para las tasas de dentistas por cada 100 000 habitantes:
   (a) el percentil muestral de orden 90%
   (b) el percentil muestral de orden 50%
   (c) el percentil muestral de orden 10%
6. Supongamos que el percentil muestral de orden $100p$ por ciento para un conjunto de datos es 120. Si se suma 30 a cada valor de dato, ¿cuál es el nuevo valor del percentil muestral de orden $100p$ por ciento?
7. Supongamos que el percentil muestral de orden $100p$ por ciento para un conjunto de datos es 230. Si se multiplica cada valor por una constante positiva c, ¿cuál es el nuevo valor del percentil muestral de orden $100p$ por ciento?
8. Encuentre el percentil muestral de orden 90% del siguiente conjunto de datos.

   75, 33, 55, 21, 46, 98, 103, 88, 35, 22, 29, 73, 37, 101, 121, 144, 133, 52, 54, 63, 21, 7
9. La tabla siguiente muestra los fallecimientos por accidentes de tráfico (por 100 millones de millas recorridas) en el año 2001 en los 50 Estados y en el distrito de Columbia de Estados Unidos. Encuentre los cuartiles.

Muertes por accidente de tráfico por 100 millones de millas recorridas, 200

Estado	Tasa	Rango de orden
Estados Unidos	1,51	(X)
Alabama	1,75	16
Alaska	1,80	15
Arizona	2,06	7
Arkansas	2,08	6
California	1,27	37
Colorado	1,71	19
Connecticut	1,01	47
Delaware	1,58	24
District of Columbia	1,81	(X)
Florida	1,93	10
Georgia	1,50	26
Hawaii	1,61	23
Idaho	1,84	13
Illinois	1,37	31
Indiana	1,27	37
Iowa	1,49	27
Kansas	1,75	16
Kentucky	1,83	14
Louisiana	2,32	1
Maine	1,33	34
Maryland	1,27	37
Massachusetts	0,90	50
Michigan	1,34	33

Muertes por accidente de tráfico por 100 millones de millas recorridas, 200 (*Continuación*)

Estado	Tasa	Rango de orden
Minnesota	1,06	46
Mississippi	2,18	4
Missouri	1,62	22
Montana	2,30	2
Nebraska	1,36	32
Nevada	1,71	19
New Hampshire	1,15	44
New Jersey	1,09	45
New Mexico	1,99	9
New York	1,18	43
North Carolina	1,67	21
North Dakota	1,45	29
Ohio	1,29	36
Oklahoma	1,55	25
Oregon	1,42	30
Pennsylvania	1,49	27
Rhode Island	1,01	47
South Carolina	2,27	3
South Dakota	2,00	8
Tennessee	1,85	12
Texas	1,72	18
Utah	1,25	41
Vermont	0,96	49
Virginia	1,27	37
Washington	1,21	42
West Virginia	1,91	11
Wisconsin	1,33	34
Wyoming	2,16	5

Observación: : Cuando dos o más Estados comparten el mismo rango de orden, los siguientes rangos de orden se omiten. Debido al redondeo de datos, varios Estados pueden tener valores idénticos, aunque su rango sea distinto.

10. Los cuartiles de un extenso conjunto de datos son los siguientes:

    Primer cuartil = 35
    Segundo cuartil = 47
    Tercer cuartil = 66

    (a) Indique un intervalo que contenga aproximadamente un 50% de los datos.
    (b) Determine un valor que aproximadamente sea mayor que un 50% de los datos.
    (c) Determine un valor para el que aproximadamente un 25% de los datos sean mayores que él.
11. La mediana de un conjunto de datos simétrico es igual a 40 y su tercer cuartil es igual a 55. ¿Cuál es valor del primer cuartil?

3.4 Moda muestral

Otro indicador de la tendencia central es la *moda muestral*, que se define como el valor de dato que aparece con mayor frecuencia en un conjunto de datos.

Ejemplo 3.12 Los siguientes datos se refieren a las tallas de los últimos 8 vestidos vendidos en una boutique de mujeres.

8, 10, 6, 4, 10, 12, 14, 10

¿Cuál es la moda muestral?

Solución La moda muestral es 10, puesto que este es el valor que ocurre con mayor frecuencia. ■

Si no existe un único valor que aparezca con mayor frecuencia en el conjunto de datos, aquellos valores que tengan la máxima frecuencia se denominan *valores modales*. En esta situación se dice que no existe un valor único de la moda muestral.

Ejemplo 3.13 Las edades de 6 niños de una guardería son las siguientes:

2, 5, 3, 5, 2, 4

¿Cuáles son los valores modales de este conjunto de datos?

Solución Puesto que las edades 2 y 5 son las que ocurren con mayor frecuencia, estos dos valores (2 y 5) son los modales. ■

Resulta muy sencillo obtener el valor modal a partir de una tabla de frecuencias, puesto que coincide con aquel valor que tenga mayor frecuencia.

Ejemplo 3.14 La siguiente tabla de frecuencias muestra los valores obtenidos en 30 lanzamientos de un dado.

Valor	Frecuencia
1	6
2	4
3	5
4	8
5	3
6	4

Para estos datos, encuentre:

(a) la moda muestral

(b) la mediana muestral

(c) la media muestral

Solución

(a) Puesto que el valor 4 aparece con la mayor frecuencia, la moda muestral es 4.

(b) Puesto que existen 30 valores de datos, la mediana muestral coincide con la media entre el 15° y el 16° valor menor. Puesto que el 15° valor menor es 3 y el 16° valor menor es 4, la mediana muestral es 3,5.

(c) La media muestral es

$$\bar{x} = \frac{1 \cdot 6 + 2 \cdot 4 + 3 \cdot 5 + 4 \cdot 8 + 5 \cdot 3 + 6 \cdot 4}{30} = \frac{100}{30} \approx 3{,}333 \quad \blacksquare$$

Problemas

1. Relacione cada sentencia de la izquierda con el conjunto de datos correcto entre los que figuran a la derecha.

1. La moda muestral es 9	A: 5, 7, 8, 10, 13, 14
2. La media muestral es 9	B: 1, 2, 5, 9, 9, 15
3. La mediana muestral es 9	C: 1, 2, 9, 12, 12, 18

2. Con los datos del ejemplo 2.2, encuentre la moda muestral de las puntuaciones ganadoras del Torneo de Maestros de Golf.

3. Con los datos de los primeros 100 estudiantes del Apéndice A, encuentre la moda muestral para:

   (a) los pesos

   (b) las presiones sanguíneas

   (c) los niveles de colesterol

4. Suponga que usted quiere descubrir el salario del vicepresidente de un banco, al que acaba de conocer. Si pretende tener la mayor posibilidad de acertar a menos de 1000 dólares, ¿le gustaría conocer la media muestral, la mediana muestral o la moda muestral de los salarios de los vicepresidentes de bancos?

5. Construya un conjunto de datos para el que la media muestral sea 10, la mediana muestral sea 8 y la moda muestral sea 6.

6. Si la moda muestral de un conjunto de datos x_i, $i = 1, \ldots, n$, es igual a 10, ¿cuál será la moda muestral de los datos $y_i = 2x_i + 5$, $i = 1, \ldots, n$?

7. Varios corredores amateurs utilizan una pista de atletismo de un cuarto de milla de longitud. En una muestra de 17 corredores, 1 corrió 2 vueltas, 4 corrieron 4 vueltas, 5 corrieron 6 vueltas, 6 corrieron 8 vueltas y 1 corrió 12 vueltas.

   (a) ¿Cuál es la moda muestral del número de vueltas que han hecho estos corredores?

   (b) ¿Cuál es la moda muestral de las distancias en millas recorridas por los corredores?

3.5 Varianza muestral y desviación típica muestral

Aunque hasta ahora se han introducido estadísticos que miden la tendencia central de un conjunto de datos, todavía no se han considerado aquellos que miden su dispersión o variabilidad. Por ejemplo, pese a que los siguientes conjuntos de datos A y B tienen las mismas media y mediana muestrales, claramente existe una mayor dispersión en los valores de B que en los de A.

$$A\text{: } 1, 2, 5, 6, 6 \qquad B\text{: } -40, 0, 5, 20, 35$$

Una forma de medir la variabilidad de un conjunto de datos consiste en considerar las desviaciones de los valores de datos a un valor central. El valor central que se utiliza más frecuentemente para este propósito es la media muestral. Si los valores de datos son $x_1, \ldots, x_n$ y la media muestral es $\bar{x} = \sum_{i=1}^{n} x_i/n$, la desviación a la media del valor x_i es , $i = 1, \ldots, n$.

Se podría suponer que una medida natural de la variabilidad de un conjunto de datos es la media de las desviaciones a la media. Sin embargo, como se ha visto en la sección 3.2, $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. Esto es, la suma de las desviaciones a la media muestral es siempre igual a 0 y, por consiguiente, la media de las desviaciones a la media muestral también será igual a 0. Ahora bien, tras una reflexión adicional, se verá claro que no se desea permitir que las desviaciones positivas y negativas se compensen. Por el contrario, se deberían considerar las desviaciones a la media sin tener en cuenta sus signos. Esto se puede conseguir si se consideran los valores absolutos de las desviaciones o, algo más útil, si se consideran sus cuadrados.

La varianza muestral se define como la "media" de los cuadrados de las desviaciones a la media muestral. Sin embargo, por cuestiones técnicas (que se verán claras en el capítulo 8), esta "media" divide la suma de las n desviaciones al cuadrado por $n - 1$, en lugar de dividirla por n, como es habitual.

Definición

La *varianza muestral*, denotada por s^2, de los datos $x_1, \ldots, x_n$ con media $\bar{x} = \left(\sum_{i=1}^{n} x_i\right)/n$ se define como

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}$$

Ejemplo 3.15 Encuentre la varianza muestral del conjunto de datos A.

Solución Se determinará como sigue:

x_i	1	2	5	6	6
$\bar{x}$	4	4	4	4	4
$x_i - \bar{x}$	−3	−2	1	2	2
$(x_i - \bar{x})^2$	9	4	1	4	4

De donde, para el conjunto de datos A,

$$s^2 = \frac{9 + 4 + 1 + 4 + 4}{4} = 5{,}5 \quad \blacksquare$$

Ejemplo 3.16 Encuentre la varianza muestral del conjunto de datos B.

Solución La media muestral del conjunto de datos B es también $\bar{x} = 4$. Por consiguiente, para este conjunto de datos, se tendrá

x_i	−40	0	5	20	35
$x_i - \bar{x}$	−44	−4	1	16	31
$(x_i - \bar{x})^2$	1936	16	1	256	961

Así pues,

$$s^2 = \frac{3170}{4} = 792{,}5 \quad \blacksquare$$

La siguiente identidad algebraica resulta útil cuando se desea calcular a mano la varianza muestral:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \tag{3.2}$$

Ejemplo 3.17 Compruebe que, para el conjunto de datos A, se verifica la identidad (3.2).

Solución Puesto que $n = 5$ y $\bar{x} = 4$,

$$\sum_{i=1}^{5} x_i^2 - n\bar{x}^2 = 1 + 4 + 25 + 36 + 36 - 5(16) = 102 - 80 = 22$$

Del ejemplo 3.15,

$$\sum_{i=1}^{5}(x_i - \bar{x})^2 = 9 + 4 + 1 + 4 + 4 = 22$$

y, por consiguiente, la identidad queda comprobada. ■

Supongamos que se suma una constante c a cada uno de los valores de datos $x_1, \ldots, x_n$ para así obtener un nuevo conjunto de datos $y_1, \ldots, y_n$, donde

$$y_i = x_i + c$$

Para ver cómo afecta esto a la varianza muestral, recuerde, de la sección 3.2, que

$$\bar{y} = \bar{x} + c$$

y, por tanto,

$$y_i - \bar{y} = x_i + c - (\bar{x} + c) = x_i - \bar{x}$$

Es decir, las desviaciones y son iguales a las desviaciones x; en consecuencia, sus respectivas sumas de cuadrados son iguales. Así pues, se ha demostrado el resultado siguiente:

La varianza muestral se mantiene constante cuando se suma una constante a cada valor de dato

Se puede utilizar el resultado anterior, junto con la identidad (3.2), para reducir enormemente el tiempo de cálculo de la varianza muestral.

Ejemplo 3.18 Los siguientes datos muestran el número de policías asesinados en actos de servicio en Estados Unidos a lo largo de 10 años.

164, 165, 157, 164, 152, 147, 148, 131, 147, 155

Encuentre la varianza muestral del número de policías asesinados en esos años.

Solución En vez de trabajar directamente con los datos dados, restemos el valor 150 de cada uno de ellos. (Esto es, sumemos $c = -150$ a cada valor de dato.) Así se obtiene el conjunto de datos nuevo:

14, 15, 7, 14, 2, −3, −2, −19, −3, 5

Su media muestral es

$$\bar{y} = \frac{14 + 15 + 7 + 14 + 2 - 3 - 2 - 19 - 3 + 5}{10} = 3{,}0$$

La suma de los cuadrados de los datos nuevos es

$$\sum_{i=1}^{10} y_i^2 = 14^2 + 15^2 + 7^2 + 14^2 + 2^2 + 3^2 + 2^2 + 19^2 + 3^2 + 5^2 = 1078$$

Así pues, si se usa la identidad (3.2), se llega a que

$$\sum_{i=1}^{10} (y_i - \bar{y})^2 = 1078 - 10(9) = 988$$

Por lo tanto, la varianza muestral de los datos nuevos, que coincide con la de los datos originales, es

$$s^2 = \frac{988}{9} \approx 109{,}78$$ ■

La raíz cuadrada positiva de la varianza muestral se denomina *desviación típica muestral.*

Definición

Al valor s, definido por

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$$

se le denomina *desviación típica muestral.*

La desviación típica muestral se mide en las mismas unidades que los datos originales. Es decir, si por ejemplo los datos originales están dados en pies de longitud, la varianza muestral vendrá expresada en pies al cuadrado; mientras que la desviación típica muestral vendrá dada en pies.

Si cada valor de dato x_i, $i = 1, \ldots, n$, se multiplica por una constante c para obtener el nuevo conjunto de datos

$$y_i = cx_i \quad i = 1, \ldots, n$$

la varianza muestral de los datos y coincide con la varianza muestral de los datos x multiplicada por c^2. Esto es,

$$s_y^2 = c^2 s_x^2$$

donde s_y^2 y s_x^2 son las varianzas muestrales de los datos nuevos y de los datos originales, respectivamente. Si se extrae la raíz cuadrada de los dos miembros de la igualdad anterior se obtiene que la desviación típica de los datos y es igual a la desviación típica de los datos x multiplicada por el valor absoluto de c; es decir,

$$s_y = |c| s_x$$

Otro indicador de la variabilidad de un conjunto de datos es el *rango intercuartílico*, que es igual a la diferencia entre el tercer y el primer cuartil. Esto es, hablando grosso modo, el rango intercuartílico es la longitud del intervalo que contiene la mitad central de los datos.

Tabla 3.2 Distintos percentiles del Test de Analogías de Miller para cinco tipos de estudiantes

Percentil de orden (en %)	Ciencias Físicas	Medicina	Ciencias Sociales	Lengua y Literatura	Derecho
99	93	92	90	87	84
90	88	78	82	80	73
75	80	71	74	73	60
50	68	57	61	59	49
25	55	45	49	43	37

Ejemplo 3.19 El Test de Analogías de Miller es un test estandarizado al que se someten distintos alumnos que intentan acceder a determinados estudios universitarios y profesionales. La tabla 3.2 muestra algunos de los percentiles de las puntuaciones obtenidas por algunos de los estudiantes que se han presentado al test, clasificados por el tipo de estudios que pretenden seguir. Por ejemplo, la tabla 3.2 indica que 68 es la puntuación mediana de los estudiantes de Ciencias Físicas, mientras que la mediana de los de Derecho es 49.

Determine los rangos intercuartílicos para los estudiantes de los cinco tipos de estudios especificados.

Solución Puesto que el rango intercuartílico es la diferencia entre los percentiles de órdenes 75 y 25%, se tendrá que su valor será

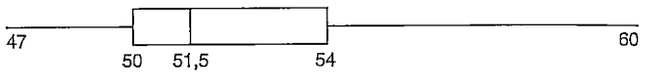
$80 - 55 = 25$ para las puntuaciones de los estudiantes de Ciencias Físicas
$71 - 45 = 26$ para las puntuaciones de los estudiantes de Medicina
$74 - 49 = 25$ para las puntuaciones de los estudiantes de Ciencias Sociales
$73 - 43 = 30$ para las puntuaciones de los estudiantes de Lengua y Literatura
$60 - 37 = 23$ para las puntuaciones de los estudiantes de Derecho ■

Los *diagramas de caja* se utilizan habitualmente para representar algunos de los estadísticos sintéticos de un conjunto de datos. En el eje x se dibuja un segmento entre los valores menor y mayor de los datos; superpuesta a este segmento, se coloca una "caja" que comienza en el primer cuartil y termina en el tercer cuartil, dentro de la cual se indica el valor del segundo cuartil mediante una línea vertical. Por ejemplo, en la siguiente tabla de frecuencias se muestran los salarios iniciales de una muestra de 42 graduados en Arte.

Salario inicial	Frecuencia
47	4
48	1
49	3
50	5
51	8

Salario inicial	Frecuencia
52	10
53	0
54	5
56	2
57	3
60	1

Los salarios oscilan entre los valores menor y mayor que coinciden con 47 y 60, respectivamente. El valor del primer cuartil (igual al 11° menor salario de la lista) es 50; el valor del segundo cuartil (igual a la media entre el 21° y 22° menores salarios) es 51,5; y el valor del cuartil tercero (que coincide con el 32° menor salario de la lista) es 54. El diagrama de caja para este conjunto de datos es el siguiente:



Un diagrama de caja.

Problemas

1. En la tabla siguiente se muestran los consumos per cápita de leche, en los años comprendidos entre 1983 y 1987, en Estados Unidos. Los datos provienen del Departamento de Agricultura de Estados Unidos, *Consumo de alimentos, precios y gastos*, anuario.

Año	Consumo (en galones per cápita)
1983	26,3
1984	26,2
1985	26,4
1986	26,3
1987	25,9

Encuentre la media muestral y la varianza muestral para estos datos.

2. Considere los dos conjuntos de datos siguientes:

A: 66, 68, 71, 72, 72, 75 B: 2, 5, 9, 10, 10, 16

(a) ¿Cuál parece tener mayor varianza muestral?

(b) Determine la varianza muestral del conjunto de datos A.

(c) Determine la varianza muestral del conjunto de datos B.

3. Los torneos de Maestros de Golf y Abierto de Estados Unidos son dos de los más prestigiosos torneos de golf de Estados Unidos. El torneo de Maestros se juega siempre en el campo de golf de Augusta, mientras que el Abierto se juega en diferentes campos cada año. Por ello, es probable que la varianza muestral de las puntuaciones ganadoras del Abierto de Estados Unidos sea mayor que la de las puntuaciones del torneo de Maestros. Para comprobar si es así se han recopilado las puntuaciones ganadoras de ambos torneos durante los años comprendidos entre 1981 y 1990.

	Puntuación ganadora									
Torneo	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
Abierto de EU	273	282	280	276	279	279	277	278	278	280
Maestros de Golf	280	284	280	277	282	279	285	281	283	278

(a) Calcule la varianza muestral de las puntuaciones ganadoras del Torneo Abierto de Estados Unidos.

(b) Calcule la varianza muestral de las puntuaciones ganadoras del Torneo de Maestros de Golf.

La tabla siguiente muestra el número de médicos y dentistas que había Japón en los años pares comprendidos entre 1984 y 2000. Los problemas 4 y 5 se refieren a esta tabla.

Número de médicos y dentistas (1984-2000)

	Médicos	Dentistas
1984	173 452	61 283
1986	183 129	64 904
1988	193 682	68 692
1990	203 797	72 087
1992	211 498	75 628
1994	220 853	79 091
1996	230 297	83 403
1998	236 933	85 669
2000	243 201	88 410

4. Determine la varianza muestral del número de médicos en los años citados.

5. Determine la varianza muestral del número de dentistas en dichos años.

6. Un individuo que necesitaba asegurar su coche preguntó cuáles eran las cuotas para idénticas coberturas en 10 compañías de seguros. Obtuvo los siguientes valores (correspondientes a las cuotas anuales, en dólares).

720, 880, 630, 590, 1140, 908, 677, 720, 1260, 800

Encuentre:

(a) la media muestral

(b) la mediana muestral

(c) la desviación típica muestral

La siguiente tabla muestra la población de 2003 en cada uno de los 50 Estados y en el Distrito de Columbia de Estados Unidos. Los problemas 7, 8 y 9 se refieren a esta tabla.

Población residente, 1 de julio de 2003

Estado	Número	Rango de orden
Estados Unidos	290 809 777	(X)
Alabama	4 500 752	23
Alaska	648 818	47
Arizona	5 580 811	18
Arkansas	2 725 714	32
California	35 484 453	1
Colorado	4 550 688	22
Connecticut	3 483 372	29
Delaware	817 491	45
District of Columbia	563 384	(X)
Florida	17 019 068	4
Georgia	8 684 715	9
Hawaii	1 257 608	42
Idaho	1 366 332	39
Illinois	12 653 544	5
Indiana	6 195 643	14
Iowa	2 944 062	30
Kansas	2 723 507	33
Kentucky	4 117 827	26
Louisiana	4 496 334	24
Maine	1 305 728	40
Maryland	5 508 909	19
Massachusetts	6 433 422	13
Michigan	10 079 985	8
Minnesota	5 059 375	21
Mississippi	2 881 281	31
Missouri	5 704 484	17
Montana	917 621	44
Nebraska	1 739 291	38
Nevada	2 241 154	35
New Hampshire	1 287 687	41

(*Continúa*)

Población residente, 1 de julio de 2003

Estado	Número	Rango de orden
New Jersey	8 638 396	10
New Mexico	1 874 614	36
New York	19 190 115	3
North Carolina	8 407 248	11
North Dakota	633 837	48
Ohio	11 435 798	7
Oklahoma	3 511 532	28
Oregon	3 559 596	27
Pennsylvania	12 365 455	6
Rhode Island	1 076 164	43
South Carolina	4 147 152	25
South Dakota	764 309	46
Tennessee	5 841 748	16
Texas	22 118 509	2
Utah	2 351 467	34
Vermont	619 107	49
Virginia	7 386 330	12
Washington	6 131 445	15
West Virginia	1 810 354	37
Wisconsin	5 472 299	20
Wyoming	501 242	50

Observación: Cuando varios Estados comparten el mismo rango de orden, el siguiente rango se omite. Debido al redondeo de los datos, varios Estados pueden tener valores idénticos pero rangos distintos.

7. Encuentre la varianza muestral de las poblaciones de los primeros 17 Estados.
8. Encuentre la varianza muestral de las poblaciones de los siguientes 17 Estados.
9. Encuentre la varianza muestral de las poblaciones de los últimos 17 Estados.
10. Si s^2 es la varianza muestral de los datos x_i, $i = 1, \ldots, n$, ¿cuál es la varianza muestral de los datos $ax_i + b$, $i = 1, \ldots, n$, donde a y b son constantes dadas?
11. Calcule la varianza muestral y la desviación típica muestral de los siguientes conjuntos de datos:
    (a) 1, 2, 3, 4, 5
    (b) 6, 7, 8, 9, 10
    (c) 11, 12, 13, 14, 15
    (d) 2, 4, 6, 8, 10
    (e) 10, 20, 30, 40, 50
12. En el lado estadounidense de la frontera con Canadá las temperaturas se miden en grados Fahrenheit, mientras que en el lado canadiense se miden en grados Celsius (o centígrados). Supongamos que la temperatura media diaria registrada durante el mes de enero en el lado de Estados Unidos fue de 40 °F y que la varianza muestral fue de 12.

    Utilice la fórmula siguiente, que permite transformar temperaturas Fahrenheit a Celsius

    $$C = \frac{5}{9}(F - 32)$$

    para encontrar

    (a) la media muestral registrada por los canadienses

    (b) la varianza muestral registrada por los canadienses

13. Calcule la media muestral y la varianza muestral de las presiones sistólicas sanguíneas de los primeros 50 estudiantes del conjunto de datos del Apéndice A. Haga lo mismo con los últimos 50 estudiantes del citado conjunto de datos. Compare las respuestas. Comente los resultados de esta comparación. ¿Resultan sorprendentes?
14. Si s es la desviación típica muestral de los datos x_i, $i = 1, \ldots, n$, ¿cuál es la desviación típica muestral de $ax_i + b$, $i = 1, \ldots, n$? En este problema, a y b son constantes dadas.
15. La siguiente tabla muestra el número de motos vendidas en Estados Unidos durante 8 años distintos. Utilícela para calcular la desviación típica muestral de las ventas de motos en los años citados.

Año	1975	1980	1983	1984	1985	1986	1987	1988
Ventas de motos (en miles)	940	1070	1185	1305	1260	1045	935	710

Fuente: Consejo de la Industria de Motocicletas.

16. Encuentre la desviación típica muestral del conjunto de datos reflejado en la siguiente tabla de frecuencias:

Valor	Frecuencia	Valor	Frecuencia
3	1	5	3
4	2	6	2

17. Los datos siguientes representan la acidez de 40 precipitaciones de lluvia sucesivas en el estado de Minnesota. La acidez se mide en una escala de pH que varía de 1 (muy ácida) a 7 (neutra).

    3,71, 4,23, 4,16, 2,98, 3,23, 4,67, 3,99, 5,04, 4,55, 3,24, 2,80, 3,44, 3,27, 2,66, 2,95, 4,70, 5,12, 3,77, 3,12, 2,38, 4,57, 3,88, 2,97, 3,70, 2,53, 2,67, 4,12, 4,80, 3,55, 3,86, 2,51, 3,33, 3,85, 2,35, 3,12, 4,39, 5,09, 3,38, 2,73, 3,07

(a) Encuentre la desviación típica muestral.

(b) Obtenga el rango muestral.

(c) Encuentre el rango intercuartílico.

18. Considere los dos siguientes conjuntos de datos.

A: 4,5, 0, 5,1, 5,0, 10, 5,2 B: 0,4, 0,1, 9, 0, 10, 9,5

(a) Determine el rango de cada conjunto de datos.

(b) Calcule la desviación típica muestral de cada conjunto de datos.

(c) Determine el rango intercuartílico de cada conjunto de datos.

3.6 Conjuntos de datos normales y la regla empírica

En la práctica, la mayoría de los conjuntos de datos grandes que uno encuentra tienen histogramas similares en cuanto a la forma. Por lo general, esos histogramas son simétricos con respecto al punto de máxima frecuencia y decrecen a ambos lados de ese punto siguiendo una forma acampanada. Tales conjuntos de datos se dice que son *normales*, y sus histogramas se denominan *histogramas normales*.

Definición

Se dice que un conjunto de datos es *normal* si el histograma que lo describe tiene las propiedades siguientes:

1. La máxima altura se alcanza en el intervalo central.
2. Si nos movemos desde el intervalo central en cualquier dirección, la altura decrece de tal modo que el histograma completo tiene una forma acampanada.
3. El histograma es simétrico con respecto al intervalo central.

La figura 3.2 muestra el histograma de un conjunto de datos normal.

Si el histograma de un conjunto de datos se aproxima al de un histograma normal, se dice que el conjunto de datos es *aproximadamente normal*. Por ejemplo, el histograma que aparece en la figura 3.3 proviene de un conjunto de datos aproximadamente normal; mientras que los histogramas presentados de las figuras 3.4 y 3.5 no lo son (puesto que cada uno de ellos es manifiestamente asimétrico). Cualquier conjunto de datos que no sea simétrico con respecto a su mediana se dice que es *asimétrico*. Se dice que es *asimétrico por la derecha* si tiene una cola alargada por la derecha, y se dice que es *asimétrico por la izquierda* si la cola alargada se encuentra a la izquierda. Así pues, el conjunto de datos representado en la figura 3.4 es asimétrico por la izquierda, mientras que el representado en la figura 3.5 es asimétrico por la derecha.

Se desprende de la simetría de los histogramas normales que, si un conjunto de datos es aproximadamente normal, su media muestral y su mediana muestral son aproximadamente iguales.

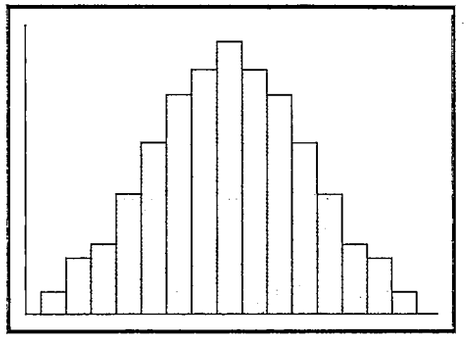


Figura 3.2 Histograma de un conjunto de datos normal.

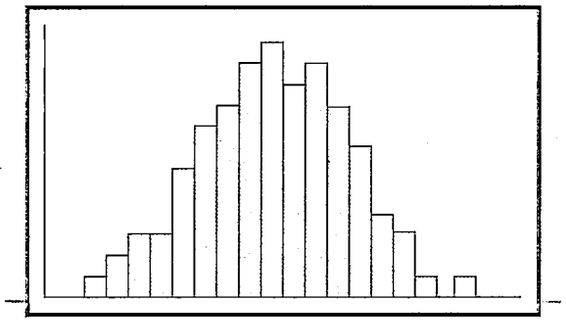


Figura 3.3 Histograma de un conjunto de datos aproximadamente normal.

Figura 3.4 Histograma de un conjunto de datos asimétrico por la izquierda.

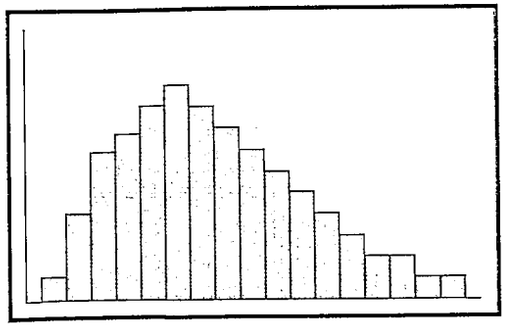


Figura 3.5 Histograma de un conjunto de datos asimétrico por la derecha.

Supongamos que $\bar{x}$ y s son, respectivamente, la media muestral y la desviación típica muestral de un conjunto de datos aproximadamente normal. La regla siguiente, conocida como *regla empírica*, especifica las proporciones aproximadas de datos observados que distan de la media muestral $\bar{x}$ en menos de s, $2s$ y $3s$.

Regla empírica

Si un conjunto de datos es aproximadamente normal con media muestral $\bar{x}$ y desviación típica muestral s, los siguientes puntos son ciertos:

1. Aproximadamente, un 68% de las observaciones caen dentro de

$$\bar{x} \pm s$$

2. Aproximadamente, un 95% de las observaciones caen dentro de

$$\bar{x} \pm 2s$$

3. Aproximadamente, un 99,7% de las observaciones caen dentro de

$$\bar{x} \pm 3s$$

Ejemplo 3.20 Las calificaciones obtenidas por 25 estudiantes en un examen de Historia aparecen representadas en el siguiente diagrama de tallos y hojas.

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

Si miramos esta figura de lado (o, equivalentemente, si giramos el libro) se puede ver que el histograma correspondiente es aproximadamente normal. Utilícela para evaluar la regla empírica.

Solución Si se hacen los cálculos pertinentes se obtiene que la media muestral y la desviación típica muestral son

$$\bar{x} = 73{,}68 \quad \text{y} \quad s = 12{,}80$$

La regla empírica establece que aproximadamente un 68% de los valores de datos se encuentran entre $\bar{x} - s = 60{,}88$ y $\bar{x} + s = 86{,}48$. Puesto que 17 observaciones caen realmente entre 60,88 y 86,48, el porcentaje real es del $100(17/25) = 68\%$. Del mismo modo, la regla empírica establece que aproximadamente un 95% de los datos se encuentran entre $\bar{x} - 2s = 48{,}08$ y $\bar{x} + 2s = 96{,}28$; mientras que realmente el 100% de los datos se encuentran dentro de este rango. ■

Un conjunto de datos que se ha obtenido muestreando una sobre población compuesta por varias subpoblaciones de tipos diferentes no es, por lo general, normal. Por el contrario, el histograma de tal conjunto de datos parece reflejar una combinación, o superposición, de histogramas normales y, en consecuencia, suele tener más de un pico, o chepa, local. Debido a que el histograma será más alto en esos picos locales que en otros valores próximos a ellos, esos picos son similares a las modas. Un conjunto de datos cuyo histograma tenga dos picos locales se dice que es *bimodal*. El conjunto de datos representado en la figura 3.6 es bimodal.

Puesto que un gráfico de tallos y hojas puede ser considerado como un histograma girado, aquél es útil para observar si un conjunto de datos es aproximadamente normal.

Ejemplo 3.21 El siguiente gráfico de tallos y hojas representa los pesos de 200 miembros de un club de salud.

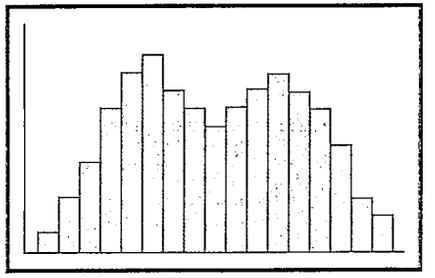


Figura 3.6 Histograma de un conjunto de datos bimodal.

24	9
23	
22	1
21	7
20	2, 2, 5, 5, 6, 9, 9, 9
19	0, 0, 0, 0, 0, 1, 1, 2, 4, 4, 5, 8
18	0, 1, 1, 2, 2, 2, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6, 6, 7, 9, 9, 9
17	1, 1, 1, 2, 3, 3, 4, 4, 4, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7, 9
16	0, 0, 1, 1, 1, 1, 2, 4, 5, 5, 6, 6, 8, 8, 8, 8
15	0, 1, 1, 1, 1, 1, 1, 5, 5, 5, 5, 6, 6, 6, 7, 7, 8, 9
14	0, 0, 0, 1, 2, 3, 4, 5, 6, 7, 7, 7, 8, 9, 9
13	0, 0, 0, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 4, 5, 5, 6, 6, 6, 6, 7, 7, 8, 8, 8, 9, 9, 9
12	1, 1, 1, 2, 2, 2, 3, 4, 4, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 8, 8, 9, 9, 9
11	0, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 5, 5, 6, 9, 9
10	0, 2, 3, 3, 3, 4, 4, 5, 7, 7, 8
9	0, 0, 9
8	6

Si se observa de lado, se ve que su histograma no parece aproximadamente normal. Sin embargo, es importante resaltar que estos datos consisten en los pesos de todos los miembros del club de salud, tanto mujeres como hombres. Puesto que estos dos grupos determinan dos poblaciones diferentes en cuanto a sus pesos, tiene sentido considerar separadamente los datos de cada sexo. Esto se hará a continuación.

Resulta que estos 200 valores de datos son los pesos de 97 mujeres y de 103 hombres. Si se separan los pesos de las mujeres de los pesos de los hombres, se obtienen los gráficos de tallos y hojas de las figuras 3.7 y 3.8.

Como se puede ver en estas figuras, parece que los datos separados por sexo son aproximadamente normales. Calculemos $\bar{x}_w$, s_w, $\bar{x}_m$ y s_m, las medias muestrales y las desviaciones típicas muestrales de las mujeres y los hombres, respectivamente.

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

Figura 3.7 Pesos de las 97 mujeres del club de salud.

24	9
23	
22	1
21	7
20	2, 2, 5, 5, 6, 9, 9, 9
19	0, 0, 0, 0, 0, 1, 1, 2, 4, 4, 5, 8
18	0, 1, 1, 2, 2, 2, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6, 6, 7, 9, 9, 9
17	1, 1, 1, 2, 3, 3, 4, 4, 4, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7, 9
16	0, 1, 1, 1, 1, 2, 4, 5, 6, 6, 8, 8, 8, 8
15	1, 1, 1, 5, 5, 5, 6, 6, 6, 7, 7, 8, 9
14	0, 5, 7, 7, 8, 9
13	0, 1, 2, 3, 7
12	9

Figura 3.8 Pesos de los 103 hombres del club de salud.

Los cálculos conducen a

$$\bar{x}_w = 125{,}70 \quad \bar{x}_m = 174{,}69$$
$$s_w = 15{,}58 \quad s_m = 21{,}23$$

Una comprobación adicional de la normalidad aproximada de los dos conjuntos de datos separados se obtiene si se observa la similitud entre los valores de la media muestral y de la mediana muestral de ambos casos. La mediana muestral de los pesos de las mujeres coincide con el 49° menor valor de dato, que es igual a 126; mientras que la mediana muestral de los datos de los hombres coincide con el 52° valor menor, que es igual a 174. Ambas medianas están próximas a sus correspondientes medias muestrales, cuyos valores son 125,7 y 174,69.

Dados los valores de la media muestral y de la desviación típica muestral, de la regla empírica se deduce que aproximadamente un 68% de las mujeres tendrán unos pesos comprendidos entre 110,1 y 141,3, y que aproximadamente un 95% de los hombres tendrán unos pesos comprendidos entre 132,2 y 217,2. De las figuras 3.7 y 3.8 se puede comprobar que los porcentajes reales son

$$100 \times \frac{68}{97} = 70{,}1 \quad \text{y} \quad 100 \times \frac{101}{103} = 98{,}1 \quad \square$$

Problemas

1. Los datos siguientes muestran el número de animales tratados diariamente en una clínica veterinaria a lo largo de un periodo de 24 días:

22, 17, 19, 31, 28, 29, 21, 33, 36, 24, 15, 28, 25, 28, 22, 27, 33, 19, 25, 28, 26, 20, 30, 32


(North Wind Picture Archives)

Adolphe Quetelet

Perspectiva histórica

Quetelet y el fraude descubierto mediante la curva normal

El estadístico y científico social belga Adolphe Quetelet fue un gran defensor de la hipótesis de que la mayor parte de los conjuntos de datos referidos a medidas humanas eran normales. En un estudio, midió el torso de 5738 soldados escoceses, representó gráficamente el conjunto de datos resultante en un histograma y concluyó que era normal.

En un posterior estudio, Quetelet utilizó la forma de los histogramas normales para descubrir la evidencia de un fraude relacionado con los reclutas del ejército francés. Estudió los datos relativos a las alturas de una extensa muestra de 100 000 reclutas. Representó gráficamente los datos en un histograma –con intervalos de clase de 1 pulgada– y encontró que, los datos parecían ser normales, con la excepción de tres intervalos de clase de alrededor de 62 pulgadas. En particular, existían menos valores en el intervalo comprendido entre 62 y 63 pulgadas; mientras que, en los intervalos de 60 a 61 y de 61 a 62 pulgadas, había ligeramente más de los que cabría esperar en un ajuste normal perfecto de los datos. Intentando averiguar por qué la curva normal no se ajustaba tan bien a los datos como él había supuesto que lo haría, Quetelet descubrió que 62 pulgadas era la altura mínima exigida a los soldados del ejército francés. Basándose en esto y en su idea sobre la muy extensa aplicabilidad de los datos normales, Quetelet llegó a la conclusión de que algunos reclutas, cuyas alturas eran ligeramente superiores a 62 pulgadas, "doblaban sus rodillas" para parecer más bajos y evitar, así, su reclutamiento.

Durante los siguientes 50 años posteriores a Quetelet, grosso modo entre 1840 y 1890, estuvo ampliamente extendida la idea de que la mayoría de los conjuntos de datos procedentes de poblaciones homogéneas (es decir, datos que claramente no provinieran de una mixtura de poblaciones diferentes) deberían ser normales, siempre que los tamaños muestrales fueran lo suficientemente grandes. Aunque los estadísticos actuales en cierto modo se muestran escépticos respecto a esa idea, es bastante corriente el que un conjunto de datos provenga de una población normal. Este fenómeno, que generalmente ocurre en los conjuntos de datos que surgen en las ciencias biológicas y físicas, se puede explicar en parte mediante un resultado matemático conocido como *teorema central del límite*. En realidad, el teorema central del límite (que se estudia en el capítulo 7) explicará por sí mismo por qué muchos de los conjuntos de datos que aparecen en las ciencias físicas son aproximadamente normales. Para explicar por qué, a menudo, los datos biométricos (es decir, datos generados en estudios de Biología) parecen ser normales, se utilizará una observación empírica debida a Francis Galton, conocida como regresión a la media, y que en la actualidad tiene una clara justificación científica. La *regresión a la media*, junto con el teorema central del límite y el paso de un gran número de generaciones, puede explicar por qué los conjuntos de datos biométricos son, habitualmente, normales. Esta explicación se presenta en el capítulo 12.

(a) Represente gráficamente estos datos en un histograma.

(b) Encuentre la media muestral.

(c) Encuentre la mediana muestral.

(d) ¿Son estos datos son aproximadamente normales?

2. Los datos siguientes reflejan las tasas de accidentalidad laboral por 100 000 horas trabajadas, para una muestra de empresas de semiconductores.

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

   (a) Represente gráficamente estos datos en un histograma.

   (b) ¿Este conjunto de datos es, a grandes rasgos, simétrico?

   (c) Si la respuesta a (b) es no, ¿es asimétrico por la izquierda o por la derecha?

   (d) Si la respuesta a (b) es sí, ¿es aproximadamente normal?

La tabla siguiente muestra las tasas de mortalidad infantil por 1000 nacimientos vivos en los 50 Estados y en el Distrito de Columbia de Estados Unidos. Los problemas 3 y 4 se refieren a esta tabla.

Tasa de mortalidad infantil, 2001

Estado	Tasa	Rango de orden
Estados Unidos	6,8	(X)
Alabama	9,4	4
Alaska	8,1	11
Arizona	6,9	26
Arkansas	8,3	10
California	5,4	45
Colorado	5,8	39
Connecticut	6,1	34
Delaware	10,7	1
District of Columbia	10,6	(X)
Florida	7,3	21
Georgia	8,6	8
Hawaii	6,2	32
Idaho	6,2	32
Illinois	7,7	14
Indiana	7,5	17
Iowa	5,6	43
Kansas	7,4	18
Kentucky	5,9	36
Louisiana	9,8	3
Maine	6,1	34
Maryland	8,1	11
Massachusetts	5,0	48
Michigan	8,0	13
Minnesota	5,3	47
Mississippi	10,5	2
Missouri	7,4	18
Montana	6,7	29

(*Continúa*)

Tasa de mortalidad infantil, 2001 (*Continuación*)

Estado	Tasa	Rango de orden
Nebraska	6,8	27
Nevada	5,7	42
New Hampshire	3,8	50
New Jersey	6,5	30
New Mexico	6,4	31
New York	5,8	39
North Carolina	8,5	9
North Dakota	8,8	6
Ohio	7,7	14
Oklahoma	7,3	21
Oregon	5,4	45
Pennsylvania	7,2	23
Rhode Island	6,8	27
South Carolina	8,9	5
South Dakota	7,4	18
Tennessee	8,7	7
Texas	5,9	36
Utah	4,8	49
Vermont	5,5	44
Virginia	7,6	16
Washington	5,8	39
West Virginia	7,2	23
Wisconsin	7,1	25
Wyoming	5,9	36

Observación: Representa las muertes de niños con una edad de menos de 1 año por cada 1000 nacimientos vivos, según el lugar de residencia. Excluye las muertes fetales. Cuando varios Estados comparten el mismo rango de orden, el siguiente rango se omite. Debido al redondeo de datos, varios Estados pueden tener valores idénticos pero rangos diferentes.

3. Para los datos sobre la mortalidad infantil.
   (a) Calcule la media muestral.
   (b) Calcule la mediana muestral.
4. Para los datos sobre la mortalidad infantil.
   (a) Represente estos datos mediante un gráfico de tallos y hojas.
   (b) ¿El conjunto de datos es aproximadamente normal?
5. Los siguientes datos son una muestra de precios de venta de casas en una comunidad de clase media de California. Los datos están dados en miles de dólares.

   166, 82, 175, 181, 169, 177, 180, 185, 159, 164, 170, 149, 188, 173, 170, 164, 158, 177, 173, 175, 190, 172

   (a) Encuentre la media muestral.
   (b) Encuentre la mediana muestral.
   (c) Represente gráficamente los datos en un histograma.
   (d) ¿El conjunto de datos es aproximadamente normal?
6. Los datos siguientes muestran la edad que tenían en el momento de su proclamación los 43 presidentes de Estados Unidos.

Presidente	Edad de proclamación
1. Washington	57
2. J. Adams	61
3. Jefferson	57
4. Madison	57
5. Monroe	58
6. J. Q. Adams	57
7. Jackson	61
8. Van Buren	54
9. W. Harrison	68
10. Tyler	51
11. Polk	49
12. Taylor	64
13. Fillmore	50
14. Pierce	48
15. Buchanan	65
16. Lincoln	52
17. A. Johnson	56
18. Grant	46
19. Hayes	54
20. Garfield	49
21. Arthur	50
22. Cleveland	47
23. B. Harrison	55
24. Cleveland	55
25. McKinley	54
26. T. Roosevelt	42
27. Taft	51
28. Wilson	56
29. Harding	55
30. Coolidge	51
31. Hoover	54

(*Continúa*)

Presidente	Edad de proclamación
32. F. Roosevelt	51
33. Truman	60
34. Eisenhower	62
35. Kennedy	43
36. L. Johnson	55
37. Nixon	56
38. Ford	61
39. Carter	52
40. Reagan	69
41. Bush, Sr.	64
42. Clinton	46
43. Bush, Jr.	54

(a) Encuentre la media muestral y la desviación típica muestral de este conjunto de datos.

(b) Dibuje un histograma de los datos dados.

(c) ¿Los datos parecen aproximadamente normales?

(d) Si la respuesta a (c) es sí, obtenga un intervalo para el que se pueda esperar que contiene el 95% de los datos observados.

(e) ¿Qué porcentaje de datos cae realmente dentro del intervalo que se ha obtenido en el apartado (d)?

7. Para los datos sobre los pesos de mujeres del club de salud presentados en la figura 3.7 se calcularon la media muestral y la desviación típica muestral, que resultaron ser 125,70 y 15,58, respectivamente. Basándose en la forma mostrada en la figura 3.7 y en los valores anteriores, calcule la proporción aproximada de las mujeres con unos pesos comprendidos entre 94,54 y 156,86 libras. ¿Cuál es la proporción real?

8. Con una muestra de 36 varones enfermos del corazón se obtuvieron los siguientes datos relativos a las edades en las que sufrieron el primer ataque de corazón.

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

(a) Determine $\bar{x}$ y s.

(b) A partir de la forma del gráfico de tallos y hojas, ¿qué porcentaje de valores de datos cabría esperar que estuvieran comprendidos entre $\bar{x} - s$ y $\bar{x} + s$? ¿Y entre $\bar{x} - 2s$ y $\bar{x} + 2s$?

(c) Encuentre los porcentajes reales para los intervalos dados en (b).

9. Si un histograma es asimétrico por la derecha, ¿qué estadístico será mayor: la media muestral o la mediana muestral? (*Sugerencia*: Si no está seguro, construya un conjunto de datos que sea asimétrico por la derecha y calcule después su media muestral y su mediana muestral.)

10. Los datos siguientes muestran las edades de 36 víctimas por crímenes violentos en una gran ciudad del este:

25, 16, 14, 22, 17, 20, 15, 18, 33, 52, 70, 38, 18, 13, 22, 27, 19, 23,
33, 15, 13, 62, 21, 57, 66, 16, 24, 22, 31, 17, 20, 14, 26, 30, 18, 25

(a) Determine la media muestral.

(b) Encuentre la mediana muestral.

(c) Determine la desviación típica muestral.

(d) ¿Este conjunto de datos parece aproximadamente normal?

(e) ¿Qué proporción de datos dista de la media muestral menos de una vez la desviación típica muestral?

(f) Compare la contestación dada en (e) con la aproximación derivada de la regla empírica.

La tabla siguiente lista las rentas per cápita, en 2002, para los 50 Estados de Estados Unidos. Los problemas de 11 a 13 se refieren a ella.

Rentas personales *per cápita* en dólares constantes (de 1996), año 2002

Estado	Renta	Rango de orden
Estados Unidos	27 857	(X)
Alabama	22 624	43
Alaska	28 947	14
Arizona	23 573	38
Arkansas	21 169	49
California	29 707	10
Colorado	29 959	9
Connecticut	38 450	1
Delaware	29 512	12
Florida	26 646	23
Georgia	25 949	28
Hawaii	27 011	20
Idaho	22 560	44
Illinois	30 075	8
Indiana	25 425	32

(*Continúa*)

Rentas personales per cápita en dólares constantes (de 1996), año 2002 (*Continuación*)

Estado	Renta	Rango de orden
Iowa	25 461	31
Kansas	26 237	26
Kentucky	23 030	39
Louisiana	22 910	41
Maine	24 979	33
Maryland	32 680	4
Massachusetts	35 333	3
Michigan	27 276	18
Minnesota	30 675	7
Mississippi	20 142	50
Missouri	26 052	27
Montana	22 526	45
Nebraska	26 804	22
Nevada	27 172	19
New Hampshire	30 912	6
New Jersey	35 521	2
New Mexico	21 555	47
New York	32 451	5
North Carolina	24 949	34
North Dakota	24 293	36
Ohio	26 474	25
Oklahoma	23 026	40
Oregon	25 867	29
Pennsylvania	28 565	15
Rhode Island	28 198	16
South Carolina	22 868	42
South Dakota	24 214	37
Tennessee	24 913	35
Texas	25 705	30
Utah	21 883	46
Vermont	26 620	24
Virginia	29 641	11
Washington	29 420	13
West Virginia	21 327	48
Wisconsin	26 941	21
Wyoming	27 530	17

Observación: Cuando varios Estados comparten el mismo rango de orden, el siguiente rango se omite. Debido al redondeo de datos, varios Estados pueden tener valores idénticos pero rangos diferentes.

11. Con los datos de los 25 primeros Estados:
    (a) Represente gráficamente los datos en un histograma.
    (b) Calcule la media muestral.
    (c) Calcule la mediana muestral.
    (d) Calcule la varianza muestral.
    (e) ¿Los datos son aproximadamente normales?
    (f) Utilice la regla empírica para obtener un intervalo que contenga aproximadamente el 68% de las observaciones.
    (g) Use la regla empírica para obtener un intervalo que contenga aproximadamente el 95% de las observaciones.
    (h) Determine la proporción real de observaciones que caen dentro del intervalo especificado en (f).
    (i) Determine la proporción real de observaciones que caen dentro del intervalo especificado en (g).
12. Repita el problema 11 utilizando, en esta ocasión, los datos de los 25 últimos Estados.
13. Repita el problema 11 utilizando ahora todos los datos de la tabla.

3.7 Coeficiente de correlación muestral

Consideremos el conjunto de datos apareados $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. En esta sección se presentará un estadístico, llamado *coeficiente de correlación muestral*, que mide el grado en el que valores grandes de x aparecen junto a valores grandes de y, y valores pequeños de x aparecen junto a valores pequeños de y.

Los datos de la tabla 3.3 reflejan los consumos medios de cigarrillos (variable x) y el número de radicales libres (variable y), medidos en las unidades adecuadas, que se han

Tabla 3.3 Consumo de cigarrillos y radicales libres

Persona	Número de cigarrillos consumidos	Radicales libres
1	18	202
2	32	644
3	25	411
4	60	755
5	12	144
6	25	302
7	50	512
8	15	223
9	22	183
10	30	375

encontrado en los pulmones de 10 fumadores. (Un radical libre es un solo átomo de oxígeno. Se cree que es potencialmente dañino porque es altamente reactivo y porque tiene una fuerte tendencia a combinarse con otros átomos dentro del cuerpo.) La figura 3.9 muestra un diagrama de dispersión de estos datos.

Si se examina la figura 3.9 se ve que cuando el consumo de cigarrillos es alto, el número de radicales libres tiende a ser igualmente alto, y que cuando el consumo de cigarrillos es bajo, el número de radicales libres también tiende a ser bajo. Cuando ocurre así, se dice que existe una *correlación positiva* entre las dos variables.

También estaremos interesados en determinar qué tipo de la relación existe entre dos variables cuando, en una de ellas, los valores altos tienden a estar asociados con los valores bajos en la otra. Por ejemplo, los datos de la tabla 3.4 representan los años de escolarización (variable x) y el pulso en situación de descanso, medido en latidos por minuto (variable y) para 10 individuos. En la figura 3.10 se incluye un diagrama de dispersión para estos datos. Se observa que los valores altos en el número de años de escolarización tienden a estar asociados con los valores bajos en el número de pulsaciones, y que los valores bajos en los años de escolarización tienden a estar asociados con los valores altos en el número de pulsaciones. Éste es un ejemplo de *correlación negativa.*

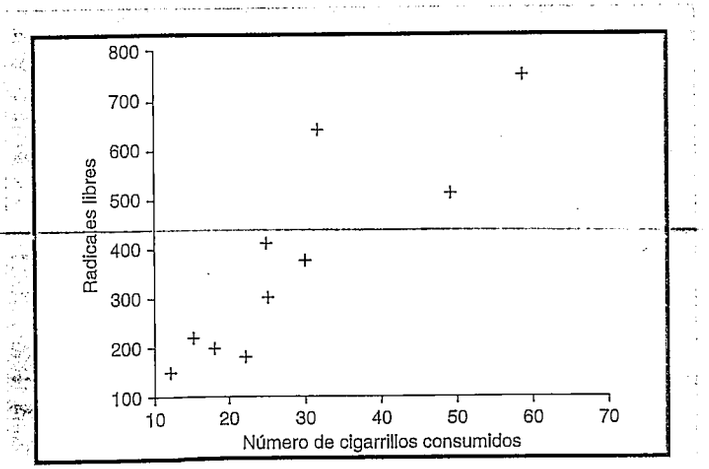


Figura 3.9 Consumo de cigarrillos frente a número de radicales libres.

Tabla 3.4 Pulsaciones por minuto y años de escolarización completados

	Persona									
	1	2	3	4	5	6	7	8	9	10
Años de escolarización	12	16	13	18	19	12	18	19	12	14
Pulsaciones	73	67	74	63	73	84	60	62	76	71

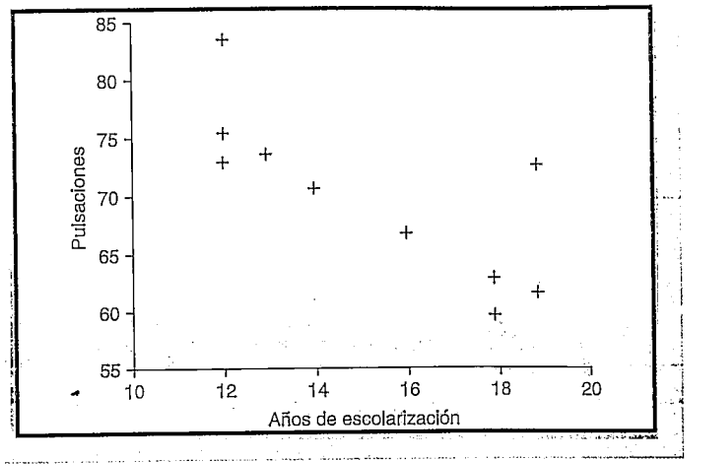


Figura 3.10 Diagrama de dispersión, de los años de escolarización y las pulsaciones por minutos.

Para obtener un estadístico que se pueda utilizar para medir la asociación entre los valores individuales de un conjunto de datos apareados, supongamos que los pares de valores del conjunto de datos son (x_i, y_i), $i = 1, \ldots, n$. Denotemos por $\bar{x}$ e $\bar{y}$ las medias muestrales de los valores x y de los valores y, respectivamente. Para cada par de valores i, consideremos $x_i - \bar{x}$, la desviación de su valor x de la media muestral, e $y_i - \bar{y}$, la desviación de su valor y de la media muestral. Ahora bien, si x_i es un valor x grande, será mayor que la media de todos los valores x y, por consiguiente, la desviación $x_i - \bar{x}$ será positiva. De igual forma, si x_i es un valor x pequeño, la desviación $x_i - \bar{x}$ será negativa. Puesto que lo mismo ocurre con las desviaciones y, se puede concluir lo siguiente:

Cuando los valores grandes de la variable x tienden a estar asociados con los valores grandes de la variable y, y si los valores pequeños de la variable x tienden a estar asociados con los valores pequeños de la variable y, los signos, positivos o negativos, de $x_i - \bar{x}$ e $y_i - \bar{y}$ de tienden a coincidir.

Ahora bien, si $x_i - \bar{x}$ e $y_i - \bar{y}$ tienen el mismo signo (positivo o negativo), su producto $(x_i - \bar{x})(y_i - \bar{y})$ será positivo. Por consiguiente, cuando los valores grandes de x tienden a estar asociados con los valores grandes de y, y si los valores pequeños de x tienden a estar asociados con valores pequeños de y, entonces $\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$ tenderá a tomar un valor positivo elevado.

La misma lógica implica que, cuando los valores grandes en una de las variables tienden a presentarse junto con los valores pequeños en la otra, los signos de $x_i - \bar{x}$ e $y_i - \bar{y}$ serán opuestos y, en consecuencia, $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ tomará un valor negativo elevado.

Para determinar qué significa que $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ tome un valor "elevado", se estandarizará esta suma y se dividirá por $n-1$; después se dividirá entre el producto de las dos desviaciones típicas muestrales. El estadístico resultante se conoce con el nombre de *coeficiente de correlación muestral.*

Definición

Denotemos por s_x y s_y las desviaciones típicas muestrales de los valores x y de los valores y, respectivamente. El *coeficiente de correlación muestral*, representado por r, de los pares de datos (x_i, y_i), $i = 1, \ldots, n$, se define por

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Cuando $r > 0$, se dice que los pares de datos muestrales están *correlacionados positivamente*; y cuando $r < 0$, se dice que están *correlacionados negativamente.*

A continuación se listan algunas de las propiedades del coeficiente de correlación muestral.

1. El coeficiente de correlación muestral siempre está comprendido entre –1 y +1.
2. El coeficiente de correlación muestral r será igual a +1 si, para alguna constante a, se verifica que

$$y_i = a + bx_i \quad \text{para } i = 1, \ldots, n$$

donde b es una constante positiva.

3. El coeficiente de correlación muestral r será igual a –1 si, para alguna constante a, se verifica que

$$y_i = a + bx_i \quad \text{para } i = 1, \ldots, n$$

donde b es una constante negativa.

4. Si r es el coeficiente de correlación muestral para los datos x_i, y_i, $i = 1, \ldots, n$, para cualquiera de las constantes a, b, c, d, el coeficiente de correlación para los datos

$$a + bx_i, \; c + dy_i \quad i = 1, \ldots, n$$

coincide con r, en el caso de que b y d tengan el mismo signo (es decir, si $bd \geq 0$).

La propiedad 1 indica que el coeficiente de correlación muestral r siempre está entre -1 y $+1$. La propiedad 2 refleja que r será igual a $+1$ si los datos apareados están alineados (es decir, si existe una relación *lineal* entre ellos), de forma que los valores grandes de y se corresponden con valores grandes de x. La propiedad 3 indica que r es igual a -1 cuando existe una relación lineal entre los pares de valores, para la que los valores grandes de y están unidos a los valores pequeños de x. La propiedad 4 establece que el valor de r se mantiene invariable cuando se añade una constante a cada valor de la variable x (o a cada valor de la variable y) o cuando cada valor de la variable x (o a cada valor de la variable y) se multiplica por una constante positiva. Esta propiedad implica que r no depende de las unidades en que se miden los datos. Por ejemplo, el coeficiente de correlación muestral para los pesos y las alturas de cierto número de personas no depende de si las alturas se miden en pies o en pulgadas o de si los pesos se miden en libras o kilogramos. De igual forma, si uno de los valores de cada par es la temperatura, el coeficiente de correlación muestral es idéntico tanto si la temperatura se mide en grados Fahrenheit como si se mide en grados Celsius.

Desde un punto de vista de eficiencia computacional, la siguiente fórmula del coeficiente de correlación resulta ser apropiada.

Fórmula computacional para r

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}$$

Ejemplo 3.22 La siguiente tabla muestra los consumos per cápita en Estados Unidos de leche entera y de leche desnatada durante tres años distintos.

	Consumo per cápita (en galones)		
	1980	1984	1987
Leche entera (x)	17,1	14,7	12,8
Leche desnatada (y)	10,6	11,5	13,2

Fuente: Departamento de Agricultura de Estados Unidos, *Consumo de alimentos, precios y gastos.*

Encuentre el coeficiente de correlación muestral para los datos dados.

Solución Para hacer que los cálculos sean más sencillos, empecemos restando 12,8 de cada valor x y restando 10,6 de cada valor y. Esto conduce al nuevo conjunto de datos:

	i		
	1	2	3
x_i	4,3	1,9	0
y_i	0	0,9	2,6

Ahora bien,

$$\bar{x} = \frac{4,3 + 1,9 + 0}{3} = 2,0667$$

$$\bar{y} = \frac{0 + 0,9 + 2,6}{3} = 1,1667$$

$$\sum_{i=1}^{3} x_i y_i = (1,9)(0,9) = 1,71$$

$$\sum_{i=1}^{3} x_i^2 = (4,3)^2 + (1,9)^2 = 22,10$$

$$\sum_{i=1}^{3} y_i^2 = (0,9)^2 + (2,6)^2 = 7,57$$

De donde,

$$r = \frac{1,71 - 3(2,0667)(1,1667)}{\sqrt{[22,10 - 3(2,0667)^2][7,57 - 3(1,1667)^2]}} = -0,97$$

Así pues, nuestros tres pares de datos muestran que existe una correlación negativa muy fuerte entre los consumos de leche entera y los de leche desnatada.

Para conjuntos de datos pequeños, tales como el del ejemplo 3.22, el coeficiente de correlación muestral puede obtenerse fácilmente a mano. Sin embargo, para conjuntos de datos grandes, su cálculo resulta tedioso y es conveniente usar una calculadora o un software estadístico. ■

Ejemplo 3.23 Calcule el coeficiente de correlación muestral para los datos de la tabla 3.3, en la que se relacionan los consumos de cigarrillos con el número de radicales libres en el interior de los pulmones de varios fumadores.

Solución El número de pares de datos es 10, cuyos valores son los siguientes:

18, 202
32, 644
25, 411
60, 755
12, 144
25, 302
50, 512

15, 223
22, 183
30, 375

Tras los cálculos pertinentes se llega a que el coeficiente de correlación muestral es 0,8759639. ■

El alto valor de este coeficiente de correlación muestral indica que existe una fuerte correlación positiva entre el consumo de cigarrillos de una persona y el número de radicales libres en el interior de sus pulmones.

Ejemplo 3.24 Calcule el coeficiente de correlación muestral para los datos de la tabla 3.4, donde se relacionan el número de pulsaciones por minuto de una persona con el número de años de escolarización que ha completado.

Solución Los pares de valores son los siguientes:

12, 73
16, 67
13, 74
18, 63
19, 73
12, 84
18, 60
19, 62
12, 76
14, 71

El coeficiente de correlación muestral es −0,763803.

El alto valor negativo de este coeficiente de correlación muestral indica que, para los datos en cuestión, un alto número de pulsaciones tiende a estar asociado a un bajo número de años de escolarización, y que un valor reducido en el número de pulsaciones tiende a corresponderse con un elevado número de años de escolarización. ■

El valor absoluto del coeficiente de correlación muestral r (esto es, $|r|$, su valor sin considerar el signo) es una medida de la fuerza de la relación lineal entre los valores x e y de cada par. Un valor de $|r|$ igual a 1 indica que existe una relación lineal perfecta; esto es, existe una recta que pasa por todos los puntos (x_i, y_i), $i = 1, \ldots, n$. Un valor de $|r|$ próximo a 0,8 indica que la relación lineal es relativamente fuerte; aunque no existe ninguna recta que pase a través de todos los puntos observados, existe una recta que pasa "cerca" de todos ellos. Un valor de $|r|$ próximo a 0,3 significa que la relación lineal es relativamente débil. El signo de r proporciona el sentido de la relación. Es positivo cuando la relación lineal es tal que los valores pequeños de y tienden a estar asociados con los valores pequeños de x, y cuando los valores grandes de y tienden a estar asociados con los valores igualmente grandes de x (por consiguiente, la relación lineal apunta hacia arriba); y es negativo cuando los valores grandes de y tienden a aparecer junto con los valores pequeños de x, y los valores pequeños de y tienden a aparecer junto con los valores grandes de x (por tanto, en este caso, la relación lineal apunta hacia abajo). En la figura 3.11 se reflejan los diagramas de dispersión de varios conjuntos de datos con distintos valores de r.

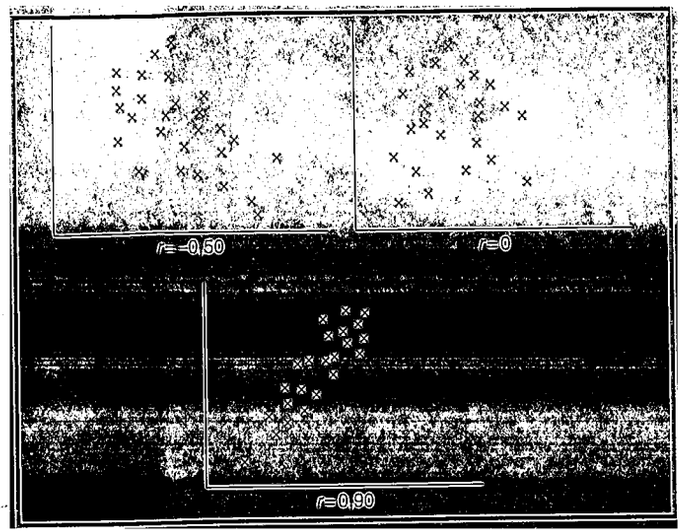


Figura 3.11 Coeficientes de correlación muestral.


(Bettmann)

Francis Galton

Perspectiva histórica

El desarrollo del coeficiente de correlación muestral y de su utilidad necesitó los esfuerzos de cuatro grandes estadísticos. La idea original fue de Francis Galton, quien estaba intentando estudiar las leyes de la herencia desde un punto de vista cuantitativo. Por este motivo, él quería ser capaz de cuantificar el grado en el que las características de un descendiente se relacionaban con las de sus padres. Ello le condujo a definir una forma de coeficiente de correlación muestral que difiere en cierta manera de la que se utiliza actualmente. Aunque originariamente pretendió utilizarlo para evaluar la influencia de la herencia de un padre sobre su descendencia, más tarde Galton se dio cuenta de que en realidad el coeficiente de correlación muestral era un método para evaluar la interrelación existente entre cualquier par de variables.

Aunque Francis Galton es considerado como el fundador de la Biometría –el análisis cuantitativo de la Biología–, Karl Pearson fue la figura más reconocida dentro de este ámbito, al menos con posterioridad a 1900. En ese año, la Real Sociedad de Londres aprobó una resolución en la que se indicaba que no se aceptarían más artículos que aplicaran las matemáticas a los estudios de Biología, y Pearson, con la ayuda financiera de Galton, fundó la revista estadística *Biometrika*, que todavía se edita hoy en día. La forma en que actualmente se utiliza el coeficiente de correlación muestral (que se ha presentado en este capítulo) se debe a Karl Pearson, por ello originalmente se conoció como *coeficiente de correlación del momento producto de Pearson.*

Las probabilidades asociadas a los posibles valores del coeficiente de correlación muestral r cuando los pares de datos provienen de poblaciones normales se deben a William Gosset. Sin embargo, en sus cálculos hubo ciertos errores técnicos que, posteriormente, fueron corregidos en un artículo de Ronald Fisher.

Problemas

1. Explique por qué el coeficiente de correlación muestral de los pares de datos

$$(121, 360), (242, 362), (363, 364)$$

es el mismo que el de los pares

$$(1, 0), (2, 2), (3, 4)$$

el cual, a su vez, coincide con el de los pares

$$(1, 0), (2, 1), (3, 2)$$

2. Calcule el coeficiente de correlación muestral para los pares de datos del problema 1.

Estadísticas en perspectiva

La correlación mide la asociación, no la causalidad

Los resultados del ejemplo 3.24 indican una fuerte correlación negativa entre los años de escolarización de los individuos y su número de pulsaciones cuando estaban en situación de descanso. Sin embargo, ello no implica que si aumenta el número de años de escolarización se reduzca directamente el número de pulsaciones por minuto. Es decir, el que los valores altos en el número de años de escolarización tiendan a estar asociados con los valores bajos en el número de pulsaciones no significa que los primeros sean la causa *directa* de los segundos. A menudo, la explicación de tal asociación se basa en un factor que no se ha tenido en cuenta, el que está relacionado con las dos variables que se consideren. En este ejemplo, podría ocurrir que una persona que hubiera estado escolarizada un alto número de años fuera más sensible a todo lo relacionado con el área de la salud y, en consecuencia, fuera más consciente de la importancia de hacer ejercicio y de tener buenos hábitos de alimentación; o quizá puede que no sea el conocimiento lo que establece la diferencia sino que, por el contrario, la gente con mayor educación acaba teniendo unos empleos que les permiten un mayor tiempo de ejercicio y mejores hábitos de nutrición. Probablemente, la fuerte correlación negativa encontrada entre los años de escolarización y el número de pulsaciones se deba a una combinación de estos y otros muchos factores subyacentes.

3. Los datos siguientes representan las puntuaciones obtenidas en un test de inteligencia (IQ) por 10 madres y por sus respectivas hijas mayores.

Puntuaciones de las madres	Puntuaciones de las hijas
135	121
127	131
124	112
120	115
115	99
112	118
104	106
96	89
94	92
85	90

(a) Dibuje un diagrama de dispersión.

(b) Haga una conjetura sobre el valor del coeficiente de correlación muestral r.

(c) Calcule el valor de r.

(d) ¿Qué conclusiones se pueden extraer acerca de la relación entre las puntuaciones de las madres y las de las hijas?

4. Los datos siguientes provienen de una muestra de 10 presos recientemente liberados que habían sido encarcelados por primera vez. Los datos incluyen el crimen cometido, su sentencia, y el tiempo real pasado en prisión.

Número	Crimen	Sentencia (en meses)	Tiempo en prisión (en meses)
1	Abuso de drogas	44	24
2	Falsificación	30	12
3	Abuso de drogas	52	26
4	Secuestro	240	96
5	Fraude de impuestos	18	12
6	Abuso de drogas	60	28
7	Robo	120	52
8	Desfalco	24	14
9	Robo	60	35
10	Robo	96	49

Dibuje un diagrama de dispersión de los tiempos de sentencia y de los tiempos reales en prisión. Calcule el coeficiente de correlación muestral. ¿Qué indica sobre la relación existente entre los tiempos sentenciados y los que se han cumplido realmente?

5. Con los datos del problema 4, determine el coeficiente de correlación muestral entre los tiempos de sentencia y las proporciones de estos tiempos que realmente se cumplieron. ¿Qué indica sobre la relación existente entre los tiempos sentenciados y dichas proporciones?

6. Los datos siguientes se refieren al número de adultos que están en prisión y de los que están en libertad condicional en los 12 Estados del occidente medio de Estados Unidos. Los datos están en miles de adultos.

Estado	En prisión	En libertad condicional
Illinois	18,63	11,42
Indiana	9,90	2,80
Iowa	2,83	1,97
Kansas	4,73	2,28
Michigan	17,80	6,64
Minnesota	2,34	1,36
Missouri	9,92	4,53
Nebraska	1,81	0,36
North Dakota	0,42	0,17
Ohio	20,86	6,51
South Dakota	1,05	0,42
Wisconsin	5,44	3,85

(a) Dibuje un diagrama de dispersión.

(b) Determine el coeficiente de correlación muestral entre el número de adultos en prisión y el número de adultos en libertad condicional.

(c) Rellene la palabra que falta. Los Estados que tienen un alto número de adultos en prisión tienden a tener un _____ número de adultos en libertad condicional.

7. Los siguientes datos relacionan el número de juicios criminales en varias ciudades de Estados Unidos y el porcentaje de sentencias de culpabilidad resultantes de ellos.

Ciudad	Porcentaje de juicios con sentencias de culpabilidad	Número de juicios criminales
San Diego, CA	73	11 534
Dallas, TX	72	14 784
Portland, OR	62	3 892
Chicago, IL	41	35 528
Denver, CO	68	3 772
Philadelphia, PA	26	13 796
Lansing, MI	68	1 358
St. Louis, MO	63	3 649
Davenport, IA	60	1 312
Tallahassee, FL	50	2 879
Salt Lake City, UT	61	2 745

Determine el coeficiente de correlación muestral entre el número de juicios criminales y el porcentaje de sentencias de culpabilidad. ¿Qué se puede decir sobre el grado de asociación entre estas dos variables que se han considerado?

8. Los siguientes datos relacionan los consumos per cápita de leche entera y de leche desnatada en Estados Unidos durante los años comprendidos entre 1980 y 1987, con la exclusión de 1981. (Algunos de estos datos fueron utilizados en el ejemplo 3.22.)

	Consumos (en galones)						
	1980	1982	1983	1984	1985	1986	1987
Leche entera	17,1	15,6	15,2	14,7	14,3	13,4	12,8
Leche desnatada	10,6	10,8	11,1	11,1	12,1	12,8	13,2

Fuente: Consumo de alimentos, precios y gastos.

Calcule el coeficiente de correlación muestral para los consumos de leche entera y de leche desnatada en los años citados.

9. Los siguientes datos muestran las rentas monetarias per cápita, en dólares, para 12 ciudades de Estados Unidos en los años 1979 y 1985.

Ciudad	Renta en 1979	Renta en 1985
New York	7 271	11 188
Baltimore	5 877	8 647
Denver	8 553	12 490
Austin	7 368	11 633
Cincinnati	6 874	10 247
Omaha	7 714	12 886
Detroit	6 215	8 852
Memphis	6 466	9 362
Milwaukee	7 029	9 765
St. Louis	5 877	8 799
Charlotte	7 952	12 259
Buffalo	5 929	8 840

Calcule el coeficiente de correlación muestral para las rentas per cápita de estas ciudades en 1979 y en 1985.

10. Los siguientes datos muestran el número de médicos y dentistas, por 100 000 habitantes, en Estados Unidos durante seis años diferentes

	1980	1981	1982	1983	1985	1986	2001
Médicos	211	217	222	228	237	246	253
Dentistas	54	54	55	56	57	57	59

Fuente: *Estadísticas de recursos sanitarios, anuario.*

(a) Compruebe si el número de médicos y el número de dentistas en los años citados están correlacionados positivamente.

(b) ¿Se puede pensar que un valor elevado en una de las dos variables causa por sí mismo un elevado valor en la otra? Si la respuesta es negativa, ¿cómo se podría explicar la correlación positiva existente?

En la tabla siguiente se incluyen las tasas de mortalidad, por una serie de causas seleccionadas, en diferentes países. Esta tabla será utilizada en los problemas del 11 al 13.

Tasas de mortalidad por 100 000 habitantes para las causas y los países seleccionados

				Neoplasia maligna de				
País	Año	Enfermedad isquémica de corazón	Enfermedad cerebro-vascular	Pulmón, tráquea, bronquios	Estómago	Pecho (mujeres)	Bronquitis, enfisema, asma	Enfermedades de hígado y cirrosis crónicas
Estados Unidos	1984	218,1	60,1	52,7	6,0	31,9	8,3	12,9
Alemania Occidental	1986	159,5	100,4	34,6	18,3	32,6	26,1	19,3
Australia	1985	230,9	95,6	41,0	10,1	30,0	16,9	8,7
Austria	1986	155,1	133,2	34,3	20,7	31,6	22,3	26,6
Bélgica	1984	120,6	95,0	55,9	14,7	36,8	22,6	12,4
Bulgaria	1985	245,9	254,5	30,6	24,2	21,5	28,6	16,2
Canadá	1985	200,6	57,5	50,6	9,0	34,5	9,7	10,1
Checoslovaquia	1985	289,4	194,3	51,3	22,4	27,3	33,8	19,6
Dinamarca	1985	243,8	73,4	52,2	10,9	39,7	37,1	12,2
España	1981	79,0	133,9	26,0	19,7	19,0	19,1	23,3
Finlanda	1986	259,8	105,0	36,4	17,3	23,9	19,8	8,8
Francia	1985	76,0	79,7	32,2	10,8	27,1	11,7	22,9
Holanda	1985	164,6	71,1	56,3	15,6	38,2	17,8	5,5
Hungría	1986	240,1	186,5	55,0	25,9	31,2	43,8	42,1
Italia	1983	128,9	121,9	42,1	23,9	28,9	30,9	31,5
Japón	1986	41,9	112,8	24,9	40,7	8,1	12,2	14,4
Noruega	1985	208,5	88,6	26,3	14,4	25,9	18,2	6,9
Nueva Zelanda	1985	250,5	98,4	42,0	11,2	37,7	25,8	4,8
Polonia	1986	109,4	75,3	47,2	24,2	21,1	33,4	12,0
Portugal	1986	76,6	216,4	18,7	26,5	22,6	17,8	30,0
Reino Unido:								
Escocia	1986	288,0	128,4	68,7	14,9	41,2	14,8	7,3
Inglaterra y Gales	1985	247,6	104,5	57,2	15,2	41,9	24,2	4,8
Suecia	1985	244,7	73,0	23,2	12,5	26,0	14,3	6,4
Suiza	1986	112,0	65,6	36,6	12,0	36,6	17,5	10,4

Fuente: Organización Mundial de la Salud, *Estadísticas de salud mundial.*

Si está ejecutando el programa 3-2 o se está usando algún paquete estadístico para resolver los problemas del 11 al 13, utilice todos los datos. Si está trabajando con una calculadora de mano, use sólo los datos referidos a los siete primeros países.

11. Encuentre el coeficiente de correlación muestral entre las tasas de mortalidad por enfermedad isquémica de corazón y por enfermedad de hígado crónica.

12. Encuentre el coeficiente de correlación muestral entre las tasas de mortalidad por cáncer de estómago y por cáncer de pecho en mujeres.

13. Encuentre el coeficiente de correlación muestral entre las tasas de mortalidad por cáncer de pulmón y por bronquitis, enfisema y asma.

14. En un famoso experimento, un investigador de la Universidad de Pittsburg solicitó la cooperación de los maestros de las escuelas públicas de Boston para conseguir un diente de leche de cada alumno. Después, serró todos los dientes que se habían recogido y determinó sus contenidos de plomo. Finalmente, hizo una representación gráfica de los contenidos de plomo frente a las puntuaciones de cada alumno en un test de inteligencia (IQ). Encontró una fuerte correlación negativa entre los contenidos de plomo y las puntuaciones citadas. Los periódicos resaltaron este hecho como una "prueba" de que las ingestiones de plomo producían un descenso en los niveles de inteligencia.

    (a) ¿Esta conclusión es necesariamente cierta?

    (b) Indique otras explicaciones posibles.

15. En un estudio reciente se encontró una fuerte correlación positiva entre los niveles de colesterol en adultos jóvenes y los tiempos que empleaban viendo la televisión.

    (a) ¿Era esperable tal resultado? ¿Por qué?

    (b) ¿Se puede pensar que ver televisión sea la causa de padecer mayores niveles de colesterol?

    (c) ¿Se puede pensar que tener niveles altos de colesterol hace que un joven adulto vea más televisión?

    (d) ¿Cómo se podría explicar el resultado del estudio?

16. Un análisis de los puntos conseguidos y de las faltas cometidas por los jugadores de baloncesto en la Conferencia del Pacífico estableció que existía una fuerte correlación positiva entre ambas variables. El analista difundió que este hecho prueba que los jugadores de baloncesto claramente ofensivos tienden a ser muy agresivos y que, en consecuencia, tienden a cometer un gran número de faltas. ¿Puede haber una explicación más simple para la correlación positiva encontrada? (*Sugerencia:* Piense en el número medio de minutos por juego que cada jugador está en pista.)

17. Un estudio publicado en octubre de 1993 en la revista *New England Journal of Medicine* encontró que la gente que tenía armas de protección en casa tenía tres veces más posibilidades de ser asesinados que aquellos que no tenían armas. ¿Prueba esto que las posibilidades de que un individuo sea asesinado se incrementan cuando decide comprar un arma para tenerla en casa? Explique su respuesta.

Términos clave

Estadístico: Magnitud numérica cuyo valor se puede determinar a partir de los datos.

Media muestral: Media aritmética de los valores de un conjunto de datos.

Desviación: Diferencia entre un valor de dato y la media muestral. Si x_i es el i-ésimo valor de dato y $\bar{x}$ es la media muestral, la diferencia $x_i - \bar{x}$ se denomina *desviación i-ésima.*

Mediana muestral: Valor central de un conjunto de datos ordenado. Para un conjunto de datos con n valores, la mediana muestral es el $(n + 1)/2$ valor menor, cuando n es impar; y es la media entre el $n/2$ y el $n/2 + 1$ menores valores, si n es par.

Percentil muestral de orden 100p por ciento: Valor de dato que cumple que al menos un 100p por ciento de los datos son menores o iguales que él y al menos un $100(1 - p)$ por ciento de los valores son mayores o iguales que él. Si existen dos valores de datos que cumplen estas condiciones, el percentil citado es igual a la media de ambos.

Primer cuartil: Percentil muestral de orden 25%.

Segundo cuartil: Percentil muestral de orden 50%, que también coincide con la mediana muestral.

Tercer cuartil: Percentil muestral de orden 75%.

Moda muestral: Valor de dato que ocurre con mayor frecuencia en un conjunto de datos.

Varianza muestral: Estadístico s^2, definido por

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}$$

Mide la media de las desviaciones al cuadrado.

Desviación típica muestral: Raíz cuadrada positiva de la varianza muestral.

Rango: Diferencia entre el mayor y el menor valor de dato.

Rango intercuartílico: Diferencia entre el tercer y el primer cuartil.

Conjunto de datos normal: Aquél cuyo histograma es simétrico con respecto a su intervalo central y que decrece a ambos lados de este intervalo siguiendo una forma acampanada.

Conjunto de datos asimétrico: Aquél cuyo histograma no es simétrico con respecto al intervalo de clase central. Se dice que es asimétrico por la derecha si su histograma presenta una cola alargada hacia la derecha, y se dice que es asimétrico por la izquierda si la cola alargada se sitúa hacia la izquierda.

Conjunto de datos bimodal: Aquél cuyo histograma presenta dos picos o chepas.

Coeficiente de correlación muestral: Para el conjunto de valores apareados x_i, y_i, $i = 1, \ldots, n$, se define por

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

donde $\bar{x}$ y s_x son, respectivamente, la media muestral y la desviación típica muestral de los valores x, y, de forma similar, se definen $\bar{y}$ y s_y. Un valor de r próximo a $+1$ indica que valores grandes de x tienden a estar apareados con valores grandes de y, y que valores pequeños de x tienden a estar apareados con valores pequeños de y. Un valor próximo a -1 indica que valores grandes de x tienden a estar apareados con valores pequeños de y, y que valores pequeños de x tienden a estar apareados con valores grandes de y.

Resumen

Se han visto tres estadísticos diferentes que describen el centro de un conjunto de datos: la media muestral, la mediana muestral y la moda muestral.

La media muestral de los datos $x_1 \ldots, x_n$ se define por

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

y es una medida del centro de los datos.

Si los datos vienen especificados mediante una tabla de frecuencias

Valor	Frecuencia
x_1	f_1
x_2	f_2
.	.
.	.
.	.
x_k	f_k

la media muestral de los $n = \sum_{i=1}^{k} f_i$ valores de datos puede expresarse como

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{n}$$

Una identidad de utilidad es

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

La mediana muestral es el valor central cuando los datos se encuentran ordenados de menor a mayor. Si existe un número par de datos, coincide con la media de los dos valores centrales. Es, también, una medida del centro de un conjunto de datos.

La moda muestral es el valor del conjunto de datos que ocurre con mayor frecuencia.

Supongamos que un conjunto de datos de tamaño n se ha ordenado de menor a mayor. Si np no es un entero, el percentil muestral de orden $100p$ por ciento se define como aquel valor que ocupa la posición que coincide con el menor entero que supera a np. Si np es un entero, el percentil muestral de orden $100p$ por ciento es la media entre los valores que ocupan las posiciones np y $np + 1$.

El percentil muestral de orden 25% es el *primer cuartil*. El percentil muestral de orden 50% (que coincide con la mediana muestral) se denomina *segundo cuartil*, y el percentil muestral de orden 75% se conoce como *tercer cuartil*.

La varianza muestral s^2 es una medida de la dispersión de los datos y se define por

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

donde n es el tamaño del conjunto. Su raíz cuadrada positiva se denomina *desviación típica muestral*, y se mide en las mismas unidades que los datos.

La identidad

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

resulta útil para calcular la varianza muestral con lápiz y papel o con una calculadora de mano.

El programa 3-1 permite computar la media muestral, la varianza muestral y la desviación típica muestral de cualquier conjunto de datos.

Otro estadístico que describe la dispersión de los datos es el *rango*, esto es, la diferencia entre el mayor y el menor valor de dato.

Los conjuntos de datos normales tienen su media muestral y su mediana muestral aproximadamente iguales. Sus histogramas son simétricos con respecto al intervalo central y tienen una forma acampanada.

El coeficiente de correlación muestral r mide el grado de asociación entre dos variables. Su valor está entre -1 y $+1$. Un valor de r próximo a $+1$ indica que cuando una de las variables es grande, la otra tiende a ser también grande, y cuando una de las variables es pequeña, la otra tiende igualmente a ser pequeña. Un valor de r cercano a -1 indica que cuando una de las variables es grande, la otra tiende a ser pequeña.

Un valor de $|r|$ grande indica la existencia de una fuerte asociación entre las dos variables. Asociación, sin embargo, no implica causalidad.

Problemas de repaso

1. Construya un conjunto de datos que sea simétrico con respecto a 0 y que contenga:
   (a) Cuatro valores distintos.
   (b) Cinco valores distintos.
   (c) En ambos casos, calcule la media muestral y la mediana muestral.
2. El siguiente gráfico de tallos y hojas refleja las presiones sanguíneas diastólicas de una muestra de 30 varones.

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

   (a) Calcule la media muestral .
   (b) Calcule la mediana muestral.
   (c) Obtenga la moda muestral.
   (d) Calcule la desviación típica muestral s.
   (e) ¿Los datos parecen ser aproximadamente normales?
   (f) ¿Qué proporción de valores de datos están comprendidos entre $\bar{x} + 2s$ y $\bar{x} - 2s$?
   (g) Compare la respuesta al apartado (f) con la proporción de datos entre ambos límites que se deduce de la regla empírica.

3. Los datos siguientes representan las edades medias de los residentes en cada uno de los 50 Estados de Estados Unidos.

29,3	27,7	30,4	31,1	28,5
32,1	28,0	31,3	26,6	25,8
25,9	33,0	31,5	30,0	28,4
24,9	31,6	26,6	25,4	29,2
29,3	27,9	31,8	31,5	30,3
28,5	29,3	26,6	31,2	32,1
31,4	30,1	27,0	28,5	27,6
28,9	29,4	30,5	31,2	29,4
29,3	30,1	28,8	27,9	30,4
32,3	30,4	25,8	27,1	26.9

   (a) Encuentre la mediana de estas edades.
   (b) Necesariamente, ¿ésta debe coincidir con la edad mediana de todos los habitantes de Estados Unidos?
   (c) Encuentre los cuartiles.
   (d) Encuentre el percentil muestral de orden 90%.
4. Utilice la tabla 3.2 (mostrada anteriormente) para completar la parte que falta en las frases siguientes:
   (a) Para que uno tenga una puntuación que esté dentro del 10% más alto de todos los estudiantes de Ciencias Físicas, debe ser de al menos ___.
   (b) Para que uno tenga una puntuación que esté dentro del 25% más alto de todos los estudiantes de Ciencias Sociales, debe ser de al menos ___.
   (c) Para que uno tenga una puntuación que esté dentro del 50% más bajo de todos los estudiantes de Medicina, debe ser de al menos ___.
   (d) Para que uno tenga una puntuación que esté dentro del 50% central de todos los estudiantes de Derecho, debe ser de al menos ___.
5. El número de crímenes violentos por 100 000 habitantes se muestra a continuación para cada uno de los 50 Estados de Estados Unidos. ¿Este conjunto de datos es aproximadamente normal?

Crímenes violentos por 100 000 habitantes, 2002

Estado	Tasa de criminalidad	Rango de orden
Estados Unidos	495	(X)
Alabama	444	21
Alaska	563	12
Arizona	553	13
Arkansas	424	22
California	593	10

Crímenes violentos por 100 000 habitantes, 2002 (*Continuación*)

Estado	Tasa de criminalidad	Rango de orden
Colorado	352	27
Connecticut	311	33
Delaware	599	9
Florida	770	2
Georgia	459	20
Hawaii	262	41
Idaho	255	42
Illinois	621	8
Indiana	357	26
Iowa	286	36
Kansas	377	24
Kentucky	279	38
Louisiana	662	6
Maine	108	48
Maryland	770	2
Massachusetts	484	18
Michigan	540	14
Minnesota	268	40
Mississippi	343	31
Missouri	539	15
Montana	352	27
Nebraska	314	32
Nevada	638	7
New Hampshire	161	47
New Jersey	375	25
New Mexico	740	4
New York	496	17
North Carolina	470	19
North Dakota	78	50
Ohio	351	29
Oklahoma	503	16
Oregon	292	34
Pennsylvania	402	23
Rhode Island	285	37
South Carolina	822	1
South Dakota	177	46
Tennessee	717	5
Texas	579	11
Utah	237	43

(*Continúa*)

Crímenes violentos por 100 000 habitantes, 2002 (*Continuación*)

Estado	Tasa de criminalidad	Rango de orden
Vermont	107	49
Virginia	291	35
Washington	345	30
West Virginia	234	44
Wisconsin	225	45
Wyoming	274	39

Observación: Los crímenes violentos se refieren a aquellos que fueron conocidos por la policía, incluyen asesinatos, secuestros forzados, robos y asaltos violentos. Cuando varios Estados comparten el mismo rango de orden, los siguientes rangos se omiten. Debido al redondeo de los datos, varios Estados pueden tener valores idénticos aunque un rango distinto.

6. Los datos siguientes representan los pesos de los recién nacidos en un hospital de una gran ciudad del este de Estados Unidos.

2,4, 3,3, 4,1, 5,0, 5,1, 5,2, 5,6, 5,8, 5,9, 5,9, 6,0, 6,1, 6,2, 6,3,
6,3, 6,4, 6,4, 6,5, 6,7, 6,8, 7,2, 7,4, 7,5, 7,5, 7,6, 7,6, 7,7, 7,8,
7,8, 7,9, 7,9, 8,3, 8,5, 8,8, 9,2, 9,7, 9,8, 9,9, 10,0, 10,3, 10,5

(a) Represéntelos gráficamente mediante un diagrama de tallos y hojas.

(b) Encuentre la media muestral $\bar{x}$.

(c) Encuentre la mediana muestral.

(d) Calcule la desviación típica muestral s.

(e) ¿Qué proporción de valores de datos están comprendidos entre $\bar{x} \pm 2s$?

(f) ¿Los datos parecen ser aproximadamente normales?

(g) Si su respuesta a (f) es sí, ¿qué proporción se estimaría para (e), si nos basamos en las respuestas a (b) y (d)?

*7. Sean a y b constantes. Demuestre que si $y_i = a + bx_i$, $i = 1, \ldots, n$, el coeficiente de correlación muestral, r, de los pares de datos x_i, y_i, $i = 1, \ldots, n$, viene dado por

(a) $r = 1$, si $b > 0$

(b) $r = -1$, si $b < 0$

(*Sugerencia:* Utilice la definición de r, y no su fórmula computacional.)

8. Los datos siguientes se han obtenido del libro *Investigaciones sobre la probabilidad de veredictos criminales y civiles,* publicado en 1837 por el matemático y probabilista francés Simeon Poisson. El libro enfatizaba las aplicaciones legales de la Probabilidad. Los datos se refieren al número de personas acusadas y condenadas por crímenes en Francia entre 1825 y 1830.

Año	Nº de acusados	Nº de condenados
1825	6652	4037
1826	6988	4348
1827	6929	4236
1828	7396	4551
1829	7373	4475
1830	6962	4130

(a) Determine la media muestral y la mediana muestral de los números de acusados.

(b) Determine la media muestral y la mediana muestral de los números de condenados.

(c) Determine la desviación típica muestral de los números de acusados.

(d) Determine la desviación típica muestral de los números de condenados.

(e) ¿Qué signo, positivo o negativo, se puede esperar que tenga el coeficiente de correlación muestral de las cifras de acusados y condenados?

(f) Determine el coeficiente de correlación muestral de los números de acusados y condenados.

(g) Determine el coeficiente de correlación muestral entre los números de acusados y los porcentajes de éstos que son condenados.

(h) Dibuje un diagrama de dispersión para los apartados (f) y (g).

(i) Haga una conjetura acerca del coeficiente de correlación muestral entre los números de condenados y los porcentajes de condenados sobre los acusados.

(j) Dibuje un diagrama de dispersión para las variables de (i).

(k) Determine el coeficiente de correlación muestral para las variables de (i).

9. Estudios recientes no han sido concluyentes sobre la posible conexión entre el consumo de café y la enfermedad coronaria de corazón. Un estudio indicó que los consumidores de grandes cantidades de café tenían mayores posibilidades de sufrir ataques de corazón que los consumidores moderados o los no consumidores, ¿prueba esto que el excesivo consumo de café incrementa el riesgo de sufrir un ataque de corazón? ¿Qué otras explicaciones son posibles?

10. Estudios recientes han indicado que las tasas de mortalidad de las personas casadas de mediana edad parecen ser menores que las de las personas solteras de mediana edad. ¿Significa esto que el matrimonio tiende a incrementar las longitudes de vida? ¿Qué otras explicaciones son posibles?

11. Un artículo del periódico *New York Times*, del 9 de junio de 1994, resaltaba un estudio en el que se mostraba que los años con bajos índices de inflación tendían a ser años con altos incrementos en la productividad media. En el artículo se argumentaba que este hecho apoyaba la tesis mantenida por la Reserva Federal en el sentido de que un bajo índice de inflación tiende a ocasionar un incremento en la productividad. Realmente, ¿se puede creer que el estudio proporciona una clara evidencia a favor de la tesis mantenida por la Reserva Federal? Explique la respuesta.