Para encontrar la media de datos muestrales exhibidos en una tabla de frecuencias, usamos la fórmula siguiente:

Media muestral para datos en una tabla de frecuencias

$$\bar{x} = \frac{\Sigma f x}{\Sigma f} \qquad (3.2)$$

Desventajas del uso de la media

La media tiene una seria desventaja: se ve afectada por los valores extremos del final de una distribución. Como depende del valor de cada medida, los valores extremos pueden llevarla a representar defectuosamente los datos.

EJEMPLO 3.3

Suponga que un corredor de maratón ha corrido en seis de los maratones más grandes del país, quedando en las posiciones siguientes (el orden es el de los maratones):

3 5 4 6 2 85

En la última carrera, en la que él ocupó el 85° lugar, fue todo el tiempo tratando de ganar la carrera. Corrió en primer lugar las primeras 22 millas, pero le dieron calambres y tuvo que caminar parte de las últimas cuatro millas. Si la media se usa para describir la habilidad del corredor, entonces debe usarse el valor 17.5, pero como terminó a lo más en sexto lugar en las cinco primeras carreras, no parece razonable usar la media para medir su capacidad de correr. Quizá la mediana proporcione una medida mejor, pues en este ejemplo la media se afecta mucho por el valor extremo 85.

Mediana

Para datos medidos en al menos una escala de intervalo, la mediana es el puntaje medio ordenado. Por ejemplo, la mediana de los puntajes ordenados de un examen 9, 22, 37, 45 y 57, es 37.

Cómo determinar la mediana

1. Ordene los datos.
2. Si el número de medidas es impar, entonces la mediana será la medida en el centro, pero si el número de medidas es par, la mediana es la media de las dos medidas que ocupan posiciones centrales.

La mediana de una población se denota por μ y la mediana de una muestra se denota por $\tilde{x}$.

EJEMPLO 3.4

Suponga que en los últimos siete juegos los Bobcats anotaron los números siguientes de puntos:

6 10 3 21 0 35 14

La mediana de los puntos anotados se encuentra ordenando primero los puntajes:

0 3 6 10 14 21 35

Se ve fácilmente que el puntaje correspondiente a la mediana es 10, pues sólo un puntaje ocupa la posición central. Si en el próximo juego los Bobcats anotaran 42 puntos, entonces los ocho puntajes formarían la secuencia siguiente:

0 3 6 10 14 21 35 42

Como ahora hay un número par de puntajes, los valores 10 y 14 ocupan las posiciones de en medio, y resulta que la mediana es 12, el promedio de 10 y 14.

Pantalla 3.1

La pantalla 3.1 ilustra el uso de MINITAB para determinar la media y la mediana para los datos de los Bobcats (6, 10, 3, 21, 0, 35 y 14) del ejemplo 3.4.

```
MTB > SET C1
DATA > 6 10 3 21 0 35 14
DATA >END
MTB > MEAN C1
  MEAN = 12.714
MTB > MEDIAN C1
  MEDIAN = 10.000
```

Las primeras tres líneas se usan para introducir los datos. Después del símbolo del sistema MTB>, el usuario escribe la orden SET C1 para crear una columna etiquetada C1 que contendrá los datos; el sistema entonces responde con el símbolo de datos, DATA, en la segunda línea; en seguida el usuario escribe los datos (los números se separan usando un espacio o una coma). La orden END debe ser introducida por el usuario para indicar el fin del conjunto de datos. En la cuarta línea, la orden MEAN C1 proporcionada por el usuario indica que se pide el valor de la media, el sistema responde en la línea siguiente con el valor de la media (MEAN = 12.714). Análogamente, en el siguiente renglón se pide el valor de la mediana mediante la orden MEDIAN C1 dada por el usuario. El sistema responde con MEDIAN = 10.000. Recuerde que, al finalizar cada orden, el usuario debe oprimir la tecla enter o return para registrar la orden en el sistema de la computadora.

EJEMPLO 3.5

Como la mediana es el valor de en medio para una distribución, puede no haber tanto valores por debajo como por encima de él. Por ejemplo, considere la muestra siguiente de cinco valores:

6 6 6 7 8

El valor 6 de la mediana no tiene valores por abajo de él, pero tiene dos valores que lo superan.

El uso de la mediana para datos de intervalo posee tanto ventajas como desventajas. Una ventaja es que la mediana no se ve afectada por puntajes

extremos al final de la distribución; ésta fue la razón de escoger la mediana para representar la medida "de en medio", para los datos de calificaciones ilustrados en el ejemplo del principio de la sección; la desventaja del uso de la mediana reside en que no es fácilmente determinable si el conjunto de datos es grande, puesto que las medidas deben ordenarse primero, ponerse en orden numérico de menor a mayor o al contrario. Para conjuntos grandes de datos que han sido organizados en una tabla de frecuencia donde los valores de x están ordenados, o un diagrama de tallo y hojas ordenado, la mediana se encuentra así:

> Si n es impar, la mediana es la medida en el lugar $(n + 1)/2$; y si n es par, la mediana es el promedio de las medidas en los lugares $n/2$ y $n/2 + 1$.

Note que $(n + 1)/2$ no representa una de las medidas, sino el número de valores que deben contarse para llegar a la mediana. Para los cinco valores ordenados 4, 8, 12, 13 y 14, la medida con rango $(5 + 1)/2 = 3$ es 12.

APLICACIÓN 3.2

Encuentre la mediana para los datos muestrales organizados en la tabla 3.2, una tabla de frecuencia que representa el número de faltas en cada periodo de clases durante la primavera de 1988 en un grupo de introducción a la filosofía.

TABLA 3.2
Datos de faltas para la aplicación 3.2

Número de faltas	Frecuencia	*f* acumulada
0	10	10
1	10	20
2	8	28
3	4	22
4	8	40

Solución: Como consecuencia de la regla dada y el hecho de que haya 40 medidas involucradas, la mediana es el promedio de las medidas vigésima y vigésima primera; note que como hay un número par de medidas, dos de las medidas ocupan las posiciones de en medio; para llegar a la mediana podemos contar ya sea en dirección de la medida menor a la mayor, o viceversa. Como el valor número 20 contando desde la medida menor, es el valor 21 contando desde la medida mayor, sólo necesitamos promediar esos valores contando desde el valor menor, por lo tanto, la mediana de los datos es $(1 + 2)/2 = 1.5$ faltas. ■

Moda

La moda, si se da, es la medida más frecuente; tiene dos ventajas: para ciertas muestras pequeñas, se le determina fácilmente y, en general, no se ve afectada por los valores extremos al final de un conjunto de datos ordenados, como

en el ejemplo 3.7; cuando se analizan datos cualitativos, como en el ejemplo 3.8, la moda es la única medida de tendencia central que puede utilizarse. Finalmente, la moda puede usarse como una medida de tendencia central para datos numéricos empleados en sentido cualitativo (véase el ejemplo 3.9).

EJEMPLO 3.6

Con las medidas

1 1 3 3 3 2 7 8

la moda es 3.

EJEMPLO 3.7

La moda no se ve afectada por medidas extremas, como se observa en las dos muestras siguientes, A y B, cada una con una moda de 2.

A: 1, 2, 2, 2, 3, 78

B: 1, 2, 2, 2, 3, 8

La medida extrema de 78 en la muestra A no tiene efectos en el valor de la moda.

EJEMPLO 3.8

Suponga que los tipos de sangre para un grupo de 12 estudiantes de enfermería son A, A, B, A, AB, O, O, B, O, A, B y AB. La moda, o el tipo de sangre más frecuente, es el tipo A, para estos datos, no tiene sentido usar la media o la mediana para localizar una observación central, ya que la moda es la única medida de tendencia central que tiene sentido aquí.

EJEMPLO 3.9

264	Mt. Savage
324	Wellersburg
463	Lonaconing
689	Frostburg
697	Cumberland
722	Cumberland
724	Cumberland
729	Cumberland
759	Cumberland
777	Cumberland
895	Grantsville

La C & P Telephone Company proporciona servicio local sólo a puntos ubicados en un área geográfica específica; cualquier llamada hecha desde un punto dentro de ese circuito de llamadas a un punto fuera de él tiene un cargo adicional que hace la compañía mencionada. Los primeros tres dígitos de un número telefónico de siete indican el área a la cual se llama, y los últimos cuatro el lugar dentro del área de la llamada; los lugares a los que se pueden llamar desde un teléfono localizado en una cierta área se identifican porque el número telefónico comienza con 689. Cada uno de los números siguientes representa los primeros tres dígitos de una muestra de llamadas hechas por un negocio desde un teléfono localizado en el área de llamadas de Frostburg, durante un periodo de una hora:

264 324 463 689 697 722 689 895 324 324

¿Cuál observación deberemos usar para representar el valor central de esta muestra? Aunque las observaciones son números, éstos representan datos medidos en una escala nominal y se usan en sentido cualitativo; sólo representan etiquetas y el orden no está involucrado, en consecuencia, la mediana no tiene sentido porque los datos no toman un orden particular. Por ejemplo, no tiene sentido preguntarse por la relación de orden entre 264 y 324, o entre Cumberland y Cumberland; tampoco tiene sentido promediar las diez observaciones porque los números no se usan en un sentido cuantitativo y no tiene caso sumarlos. ¿Qué interpretación le daríamos a (264 + 324) en el contexto de este ejemplo? La única medida de tendencia central que es apropiada

para esta aplicación es la moda; el valor de la moda es 324. Esto puede servir para representar el valor central de las diez observaciones.

Una moda para datos en una tabla de frecuencia, se encuentra localizando el valor de frecuencia máxima, si no todas las frecuencias son iguales. El valor de x que corresponde al valor de frecuencia máxima se toma como una moda. Para la aplicación 3.2, se ve fácilmente que las modas son 0 y 1.

Desventajas de la moda

La moda tiene varias desventajas como medida de tendencia central; una de ellas es que para un cierto conjunto de datos puede no haber moda; esta situación surge cuando todos los datos tienen la misma frecuencia; otra desventaja es que la moda puede existir pero no ser única, como en el ejemplo 3.11.

EJEMPLO 3.10

Las medidas

rojo negro café azul

no tienen moda. Las medidas

2 2 3 3 4 4 5 5

tampoco tienen moda.

EJEMPLO 3.11

Con las medidas: rojo, rojo, rojo, negro, azul, blanco, blanco y blanco; tanto rojo como blanco son modas. En este caso la colección de observaciones se llama **bimodal.**

Rango medio

El rango medio de un conjunto de datos es el promedio de las medidas mayor y menor.

APLICACIÓN 3.3

Los siguientes son los números de torceduras necesarios para romper ocho barras forjadas de una aleación: 32, 38, 45, 44, 27, 36, 40 y 38. Determine el rango medio.

Solución: El rango medio es el promedio de las medidas mayor y menor. La medida mayor es $U = 45$ y la medida menor es $L = 27$. El rango medio es

$$\text{Rango medio} = \frac{L + U}{2}$$

$$= \frac{27 + 45}{2} = 36 \quad \blacksquare$$

APLICACIÓN 3.4

¿Qué medida de tendencia central debe usarse para indicar el salario central de todos los trabajadores en Estados Unidos?

Solución: La medida preferible es la mediana. Debido a los salarios elevados en un extremo de la escala, ni la media ni el rango medio deben usarse; desde luego, la medida apropiada dependerá de cómo se le vaya a utilizar; para indicar el estado financiero en el mercado internacional, a los estadou-

nidenses les gustaría usar la media. Una razón para no usar la moda es que no hay garantía de que exista una única moda; puede no existir o haber un gran número de valores que ocurran con mayor frecuencia. ■

Medidas de colocación

Un **punto de posición,** para una distribución, es aquel valor para el cual una porción específica de la distribución queda en o debajo de él; la mediana es un ejemplo de punto de posición, y también lo son los percentiles, deciles y cuartiles.

EJEMPLO 3.12

Un 50% de la distribución es menor o igual que la mediana, y otro 50% es mayor o igual que la mediana, por lo tanto, la mediana es un punto de posición.

Percentiles

El n-ésimo **percentil**, denotado con P_n, es el valor para el cual al menos $n\%$ de la distribución cae en o por debajo de él y al menos $(100-n)\%$ cae en o por arriba de él.

Un conjunto de datos tiene 99 puntos percentiles que lo dividen en 100 partes; cada parte contiene aproximadamente 1% de las medidas. Estos puntos percentiles se etiquetan con $P_1, P_2, P_3, P_4, \ldots P_{99}$.

EJEMPLO 3.13

Supongamos que queremos encontrar el vigésimo quinto punto percentil, o percentil 25, de la muestra exhibida en el siguiente diagrama de tallo y hojas ordenado:

3	4	4	6	9				
4	3	6	7	[8]	[9]			
5	0	1	1	5	7	7	8	9
6	0	0	4	4	7			
7	1	5	8	8	8	9		
8	4	6	8	8				

El tamaño de la muestra es $n = 32$. El percentil 25 es aquella medida para la cual al menos 25% de la muestra cae en o debajo de él y al menos el 75% se ubica en o por encima de él.

(25%)(32) = al menos 8 valores en o debajo de él.

(75%)(32) = al menos 24 valores en o por encima de él

Si contamos ocho hojas desde la punta del tronco, llegamos a la hoja 8 en el tallo 4. El valor 48 tiene 8 valores en o debajo de él y 24 valores encima; el valor 49 también satisface esas condiciones porque 8 valores están debajo de él y 24 encima. El percentil 25 es el promedio de 48 y 49; por lo tanto, $P_{25} = 48.5$.

EJEMPLO 3.14

Suponga que queremos encontrar el trigésimo percentil de los datos del ejemplo 3.13. El percentil 30 será aquella medida que tenga al menos 30% de la muestra en o por debajo de ella y al menos 70% de la muestra en o por encima de ella.

(30%)(32) = al menos 9.6 valores en o por debajo de ella

(70%)(32) = al menos 22.4 de los valores en o por encima de ella

Como en el proceso de contar se obtienen números enteros, el trigésimo percentil debe tener al menos 10 valores en o debajo de él y 23 valores en o encima de él. Cuando menos en ambos casos hemos escogido el mínimo entero mayor que el producto, al examinar el diagrama de tallo y hojas del ejemplo 3.13, determinamos que 50 satisface ambas condiciones. Así, $P_{30} = 50$.

3	4 4 6 9
4	3 6 7 8 9
5	[0] 1 1 5 7 7 8 9
6	0 0 4 4 7
7	1 5 8 8 8 9
8	4 6 8 8

Cuartiles y deciles

Los **cuartiles** son números que dividen en cuatro partes a un conjunto ordenado de medidas, extendiéndose desde la mínima hasta la máxima medida, por lo que cada parte cuenta con aproximadamente 25% de las medidas.

Hay tres puntos cuartiles, denotados con Q_1, Q_2, Q_3. El *primer cuartil*, Q_1, es el percentil 25, el *segundo cuartil*, Q_2, es el percentil 50 o la mediana, y el *tercer cuartil*, Q_3, es el 75° percentil.

$$Q_1 = P_{25}$$
$$Q_2 = \tilde{x} = P_{50}$$
$$Q_3 = P_{75}$$

Los **deciles** son números que dividen en diez partes a un conjunto de medidas que van desde la menor a la mayor, de tal forma que cada parte contiene aproximadamente 10% de las medidas.

Hay nueve deciles, denotados con $D_1, D_2, D_3, \ldots$ y D_9; D_n es el *n-ésimo* decil, cada punto decil corresponde a un punto percentil. Por ejemplo, $D_4 = P_{40}$, $D_7 = P_{70}$, y así sucesivamente.

APLICACIÓN 3.5

Una muestra de doce trabajadores se probó en cuanto a su capacidad de sostener firmemente un objeto; las medidas, ordenadas de menor a mayor, fueron 80.6, 89.9, 101.4, 102.6, 115.0, 120.1, 123.4, 126.3, 131.8, 138.6, 151.6 y 160.5. Determine:

a) el primer cuartil.
b) el segundo cuartil.
c) el tercer cuartil.
d) el segundo decil.

Solución:

a) El primer cuartil es el vigésimo quinto percentil.
Q_1 tendrá cuando menos (0.25)(12) = 3 valores que caen en o debajo de él.
Q_1 tendrá, también al menos (0.75)(12) = 9 valores que caen en o por encima de él.

Al menos tres observaciones deben estar en o por debajo de Q_1 y al menos nueve en o por encima de Q_1; los valores 101.4 y 102.6 cumplen ambos estos requerimientos. El primer cuartil Q_1 es por esto el promedio de 101.4 y 102.6. De aquí que:

$$Q_1 = \frac{101.4 + 102.6}{2} = 102$$

b) El segundo cuartil es la mediana; la mediana es el promedio de la sexta y la séptima medidas; de esta manera, el segundo cuartil es:

$$Q_2 = \frac{120.1 + 123.4}{2} = 121.75$$

c) El tercer cuartil es el percentil 75. Del inciso a) podemos determinar que el número de observaciones en o encima de Q_3 es al menos 3, y su número de observaciones en o abajo, es 9; dos valores cumplen estos requerimientos: 131.8 y 138.6. Así, Q_3 es el promedio de estos valores:

$$Q_3 = \frac{131.8 + 138.6}{2} = 135.2$$

d) El segundo decil será el vigémiso percentil, ya que (0.2)(12) = 2.4 valores deben caer en o debajo de D_2 y al menos (0.8)(12) = 9.6 valores deben estar en o encima de éste; el valor 2.4 debe redondearse a 3 y el 9.6 a 10; el resultado de contar debe ser un número entero por lo que, siempre redondearemos para satisfacer el criterio "al menos"; entonces, al menos tres valores deben estar en o debajo de D_2 y al menos diez valores en o encima de D_2; la medida 101.4 satisface estas condiciones. Por lo tanto, $D_2 = 101.4$. ■

Sesgo

La forma de un histograma depende de la posición relativa de la media, la mediana y la moda. En un **histograma simétrico**, ambos lados, determinados por la media, son idénticos (véase el ejemplo 3.15); cuando los lados de un histograma no son idénticos, tenemos lo que se llama un **histograma sesgado**. Un histograma, o conjunto de datos, para el cual hay menos medidas debajo de la media que arriba de ella, se dice que está *sesgado a la izquierda,* como en el ejemplo 3.16. Por otro lado, como veremos en el ejemplo 3.17, un histograma o conjunto de datos, para el cual las medidas por arriba de la media aparecen con menor frecuencia que las medidas por debajo de ella, se dice que está *sesgado a la derecha.*

EJEMPLO 3.15

Como ilustración de una distribución simétrica, considere los datos siguientes y su correspondiente histograma de frecuencias:

x	f
1	1
2	3
3	5
4	7
5	5
6	3
7	1

Podemos ver de la tabla que $\bar{x} = 4$, $\tilde{x} = 4$ y la moda es igual a 4. El histograma de frecuencias correspondiente se muestra en la figura 3.1. Podemos ver que un histograma simétrico tiene su media igual a su mediana, de hecho, para el histograma de la figura 3.1, la media, la mediana y la moda son todas idénticas, pero esto no siempre ocurre con un histograma simétrico, como lo indica la figura 3.2; ahí la media y la mediana son iguales a 8, pero la distribución es bimodal, con modas 7 y 9.

Figura 3.1
Histograma simétrico frecuencias

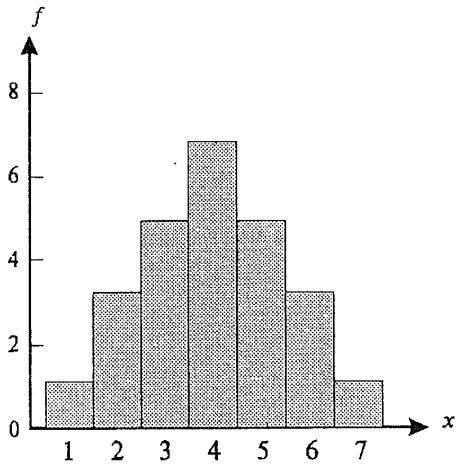
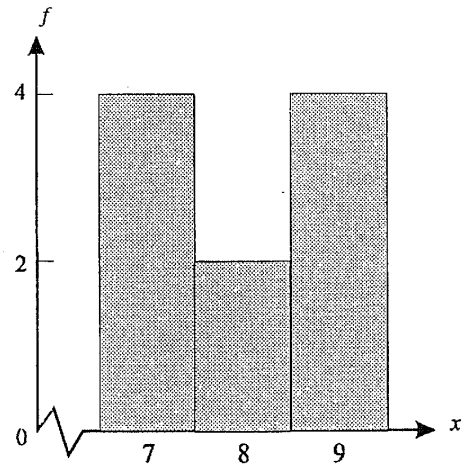


Figura 3.2
Histograma simétrico bimodal de frecuencias



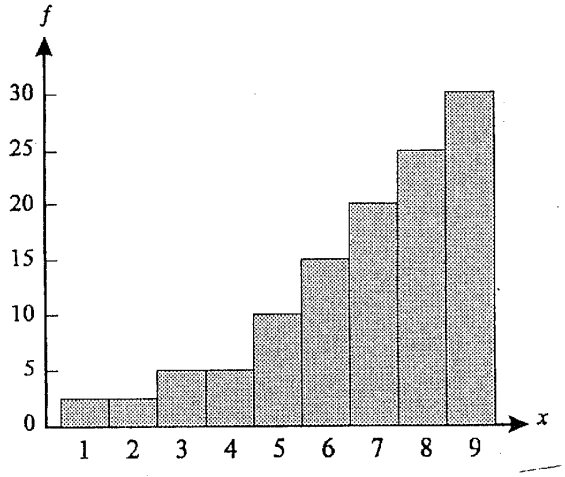
EJEMPLO 3.16

Como ejemplo de una distribución sesgada a la izquierda, consideremos los datos siguientes:

x	1	2	3	4	5	6	7	8	9
f	2	2	5	5	10	15	20	25	30

La media es 6.94, la mediana es 7 y la moda es 9; para esta distribución hay 75 valores arriba de $\bar{x} = 6.94$ y 39 valores abajo de $\bar{x} = 6.94$. El histograma para este conjunto de datos se muestra en la figura 3.3; advierta que se alarga en el lado izquierdo; de un histograma como éste, sesgado a la izquierda, se dice en ocasiones que está *negativamente sesgado*. Un histograma de este tipo tiene siempre su mediana más grande que su media.

Figura 3.3
Distrubución sesgada a la izquierda



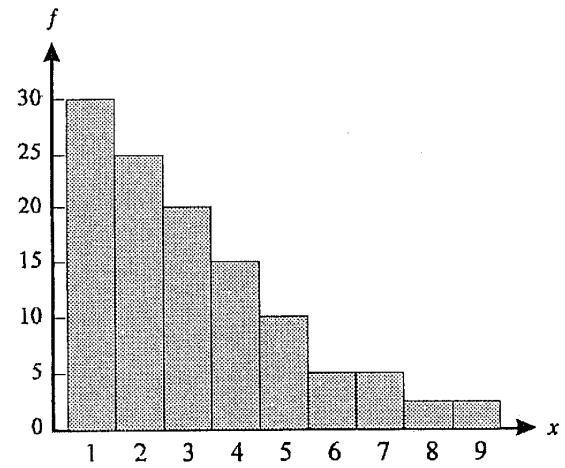
EJEMPLO 3.17

Como ilustración de una distribución sesgada a la derecha, observe:

x	1	2	3	4	5	6	7	8	9
f	30	25	20	15	10	5	5	2	2

La media es 3.06, la mediana es 3 y la moda es 1; para esta distribución hay 75 valores abajo de la media $\bar{x} = 3.03$ y 39 valores arriba de la media. El histograma de este conjunto de datos se muestra en la figura 3.4, note que se alarga en el extremo derecho; tal histograma se describe a veces como *positivamente sesgado*. Un histograma positivamente sesgado tiene siempre más grande su media que su mediana.

Figura 3.4
Distrubución sesgada a la derecha



GRUPO DE EJERCICIOS 3.1

Habilidades básicas

1. Calcule la media, la mediana, la moda y el rango medio para cada una de las muestras siguientes:
   a) 3, 9, 12, 7, 16, 20, 33, 3
   b) 5, 7, 22, 17, 5, 7, 20
   c) 8, 6, 0, 17, 12, 7, 5
   d) –4, 0, 13, 9, 4, 14, 20, 15

2. Calcule la media, la mediana, la moda y el rango medio para cada una de las muestras siguientes:
   a) 12, 7, 3, 20, 33, 2, 12
   b) 12, 15, 23, 7, 12, 40, 22, 16
   c) 5, 0, 7, 7, 13, 16, 9
   d) –5, 6, 13, 26, 0, 14, 25, 13

3. Calcule la media para una muestra donde:
   a) $\Sigma x = 37$ y $n = 12$
   b) $\Sigma x = 20.6$ y $n = 56$ y
   c) $\Sigma x = -12$ y $n = 33$

4. Calcule la media para una muestra donde:
   a) $\Sigma x = 12.5$ y $n = 16$
   b) $\Sigma x = 19$ y $n = 22$ y
   c) $\Sigma x = -43.2$ y $n = 50$

5. Calcule la media, la mediana y la moda para cada una de las muestras siguientes:
   a) 0, 0, 1, 1, 1, 0, 0, 0
   b) 3, 3, 3, 2, 2, 2, 4, 5, 3
   c) 0, 1, 1, 2, 2, 3, 3, 4, 4
   d) –1, 0, 0, 0, –1, 2, –2, 3

6. Calcule la media, la mediana y la moda para cada de las muestras siguientes:
   a) 0, 1, 2, 3, 8, 10
   b) 0, 1, 2, 3, 8, 12, 50
   c) –12, –6, –5, 0, 13, 16, 0
   d) 0, 0, 0, 1, 1, 1, 1, 1, 0

7. Determine el sesgo de estas muestras:
   a) 12, 7, 16, 22, 17, 13, 16, 7, 10
   b) 14, 17, 2, 7, 13, 17, 22, 37, 0, 15
   c) 5, 10, 15, 25, 40, 65, 100
   d) 5, 10, 95, 90, 50

8. Determine el sesgo de las muestras anotadas abajo:
   a) 22, 13, 15, 2, 18, 34, 16
   b) 5, 17, 17, 17, 3, 100
   c) 0, 0, 0, 1, 1, 1, 1
   d) 2, 2, 3, 4, 4, 1, 5

9. Un instructor borra accidentalmente la calificación de uno de sus seis estudiantes; las cinco calificaciones restantes son 76, 85, 43, 89 y 65, y la media de las seis es 70. Encuentre la calificación que se borró.

10. El salario medio anual que se paga a cuatro ejecutivos oficiales en jefe de una gran corporación es 125,000 dólares. ¿Puede llegar a ganar alguno de ellos 600,000 dólares?

Más aplicaciones

11. Si el ingreso medio de 20 trabajadores es de 40,000 dólares, ¿cuál es su ingreso total?

12. Si la estatura media de una muestra de 25 jugadores de basquetbol es 6.9 pies, ¿cuál es la suma de estaturas de los 23 jugadores?

13. En un esfuerzo por reducir su consumo de café, un trabajador de oficina registra los números siguientes de tazas de café consumidas durante un periodo de 20 días:

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

   ¿Qué medida de tendencia central le servirá mejor a su propósito? ¿Cuál es el valor numérico?

14. A continuación hay una colección de calificaciones del examen de estadística de 25 estudiantes, en un examen de 50 preguntas,

38	39	33	37	34	31	38	36	35	5

   ¿Cuál medida de tendencia central es más útil para describir el valor central? ¿Cuál es su valor numérico?

15. Un jugador de boliche ha estado jugando regularmente durante los últimos cinco años. Sus puntajes para los seis últimos juegos son: 201 187 162 234 208 198; para esta muestra calcule los valores de los estadísticos siguientes, si existen:
   a) media b) mediana
   c) moda d) rango medio
   e) Q_1, Q_2 y Q_3 f) D_4

16. En una investigación realizada por la secretaria de un médico para averiguar los tiempos de espera en minutos de los pacientes que acuden con el doctor, una muestra de pacientes de un día arrojó los resultados:

35	25	35	50	25	55	30	50	35	35
5	5	60	35	30	30	25	55	30	20
60	25	25	40	80	20	20	5	5	10

   a) Describa un tiempo típico de espera usando la media.
   b) Describa un tiempo típico de espera usando la mediana.
   c) ¿Cuál medida, media o mediana, considera usted que es más representativa del conjunto de datos? Explique.

d) Determine en tres cuartiles.
e) Determine en cuatro deciles.

17. La tabla siguiente contiene los salarios en cientos de dólares, de 25 trabajadores.

Salario anual	Frecuencia
55	7
60	5
70	6
80	4
300	3

a) ¿Cuál es la moda?
b) ¿Cuál es la media?
c) Diga la mediana y
d) el rango medio.
e) Determine el sesgo.
f) ¿Cuál medida de tendencia central usaría para determinar el valor central? Explique.
g) ¿Cuál es Q_1?
h) ¿Cuál es D_6?

18. Se escogió una muestra de 705 conductores de autobús y se registró en la tabla siguiente el número de accidentes de tránsito que tuvieron durante cuatro años.

Número de accidentes	Frecuencia
0	114
1	157
2	158
3	115
4	78
5	44
6	21
7	7
8	6
9	1
10	3
11	1

a)¿Cuál es la moda?
b) Señale la media,
c) la mediana y
d) el rango medio.
e) Determine el sesgo.
f) ¿Cuál medida de tendencia central usaría para determinar el valor central? Explique su respuesta.
g) ¿Cuánto vale Q_3?
h) ¿Cuánto vale D_4?

19. Diga y explique la medida de tendencia central que utilizaría para seleccionar un termómetro preciso que debe comprarse en una ferretería local.

Un paso más allá

20. Si 20 puntajes tienen una media de 15 y 30 puntajes, una media de 25, ¿cuál es la media del grupo total de 50 puntajes?

21. Suponga que 6 es la media de una muestra de cuatro puntajes.
a) Si se suma 5 a cada puntaje ¿cuál es la media del nuevo conjunto? *Sugerencia:* ensaye en un ejemplo.
b) Si cada puntaje se multiplica por 5, ¿cuál será la media entonces?

22. Si las medidas (x) en una muestra se transforman mediante la fórmula $y = ax + b$, determine una fórmula para la media de las medidas transformadas (y).

23. Una maestra hizo un examen con el mismo grado de dificultad en cada uno de sus tres grupos; con los resultados determinó las tres medianas y las promedió para estimar el punto central de su habilidad profesional. ¿Puede engañarse al hacer esto? Diga por qué.

24. Al promediar porcentajes, a menudo se utiliza la media geométrica $\bar{x}_g$. La media geométrica se define por

$$\bar{x}_g = \sqrt[n]{x_1 x_2 x_3 \ldots x_n}$$

donde $x_1, x_2, \ldots, x_n$ son números positivos. Encuentre $\bar{x}_g$ para los porcentajes: 95, 125, 140 y 100.

25. La *media armónica* $\bar{x}_h$, que a menudo se utiliza para promediar velocidades desarrolladas en distancias iguales, se define como el recíproco del promedio de los recíprocos de los datos, esto es,

$$\bar{x}_h = \frac{n}{\Sigma (1/x)}$$

donde los n valores de x son positivos. Suponga que se maneja a lo largo de 20 millas, a 30 millas por hora, y en un tramo de 20 millas, a 60 millas por hora. ¿Cuál es la razón promedio de la velocidad en un viaje de 40 millas?

26. ¿Cuál de los valores $\bar{x}$ o $\bar{x}_g$ sería apropiado para los datos el ejercicio 24?

27. Un piloto de coches de carreras quiere promediar 60 millas por hora (mph), en dos recorridos con un trayecto de una milla; en el primer recorrido, su tiempo fue de 30 mph debido a un problema eléctrico con el sistema de

carburación. ¿Qué tan rápido debe conducir en el segundo recorrido para lograr su cometido inicial?

28. Suponga que un interés fue fijado en 2 dólares en 1988, 4 en 1989 y 2 en 1990. El cambio de porcentaje de 1988 a 1990 es 200, y de 1989 a 1990 es de 50. Encuentre el porcentaje promedio del cambio en el tipo de interés para el periodo de los tres años y justifique su respuesta.

29. La *raíz de la media de los cuadrados* (rms) de un conjunto de datos se define como la raíz cuadrada del promedio de la suma de los cuadrados de las medidas:

$$\text{rms} = \sqrt{\frac{\Sigma x^2}{n}}$$

Esto es utilizado para describir picos de voltaje de corriente alterna en electrónica. Determine la raíz de la media de los cuadrados de la muestra de voltajes: 120, 130, 140, 110 y 105.

30. Si a cada medida de un conjunto de datos se le suma una constante C, demuestre que la media del nuevo conjunto es igual a la suma de la media del conjunto original más la constante C.

31. Si cada medida de un conjunto de datos se multiplica por una constante C, demuestre que la media del nuevo conjunto es igual a C veces la media del conjunto original.

32. Suponga que tenemos una muestra de n 1s y m 0s. Demuestre que la media es igual a la proporción de 1s en la muestra.

33. Suponga que una muestra consiste de todos los pares enteros entre 238 y 874 inclusive. Encuentre la media y la mediana.

34. Dos jugadores profesionales de beisbol tienen los porcentajes de carrera que muestra la tabla siguiente:

Jugador A				Jugador B			
Año	Veces al bat	Hits	Prom	Año	Veces al bat	Hits	Prom
1973	189	57	0.302	1973	85	27	0.318
1974	80	21	0.263	1974	144	42	0.292
1975	212	72	0.340	1975	53	19	0.358
1976	71	17	0.239	1976	207	52	0.251
1977	212	64	0.302	1977	55	19	0.345
1978	97	26	0.268	1978	263	74	0.281
1979	281	89	0.317	1979	107	35	0.327
1980	129	37	0.287	1980	175	52	0.297
1981	151	57	0.377	1981	75	29	0.387
1982	130	34	0.262	1982	163	45	0.276
Total	1552	474	0.305	Total	1327	394	0.297

Si sus otras habilidades en el juego son iguales y está negociando el contrato para la siguiente temporada, ¿cuál jugador debe recibir el salario más alto, con base en el mejor porcentaje de bateo? Explique.

35. La media de una muestra es muy sensible a la presencia de puntajes extremos, llamados puntajes aberrantes, mientras que la mediana no lo es. En estos casos, ninguna de estas medidas es satisfactoria como medida de tendencia central; una alternativa es una *media ajustada,* se afecta menos por los puntajes aberrantes que la media, y aún no tiene la insensibilidad de la mediana. Una media ajustada se encuentra ordenando las medidas de menor a mayor, borrando un cierto número de medidas en ambos extremos de la lista ordenada, y promediando las medidas restantes; a porcentaje de valores borrados en cada extremos de la lista se le llama *porcentaje de ajuste.* Por ejemplo, si $n = 20$ y se han borrado las medidas máxima y mínima, entonces el porcentaje de ajuste es $1/20 = 0.05 = 5\%$. Encuentre la media ajustando un 10% para las muestras de datos:

35	25	35	50	25	55	30	50	4	35
5	5	60	35	30	25	55	30	20	60
25	40	80	20	20	5	10	100	95	30

En este caso, ¿es más precisa la descripción que proporciona del centro de la muestra, la media ajustada, que la resultante de la media? Explique.

36. Considere esta pantalla de MINITAB:

```
MTB> SET C1
DATA> 9 14 12 17 11 20 13 18 22 12 15 16 5 7 9 19 8
DATA >END
MTB> DESCRIBE C1

        N   MEAN  MEDIAN  TREAM  STDEV  SEMEAN
C1     17  13.35   13.00  13.33   4.89    1.18
       MÍN   MÁX     Q1     Q3
C1    5.00 22.00   9.00  17.50
```

Determine el porcentaje de ajuste para la media ajustada (TREAM, por el término en inglés *trimmed mean*).

37. ¿Apoyan los datos del motivador 3 la existencia de discriminación contra las mujeres? Diga por qué.

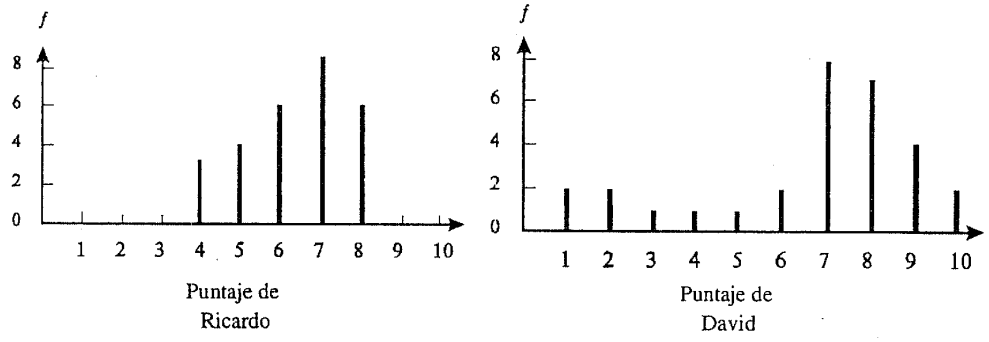
SECCIÓN 3.2 *Medidas de dispersión o variabilidad*

Es usual que las medidas de tendencia central solas no describan apropiadamente una característica en estudio. Por ejemplo, suponga que David y Ricardo lanzan, cada uno, 25 flechas a un blanco. Sus puntajes son como sigue:

	Frecuencia	
Puntaje	David	Ricardo
10	2	0
9	3	0
8	4	5
7	7	8
6	2	5
5	1	4
4	1	3
3	1	0
2	2	0
1	2	0

David y Ricardo tienen el mismo puntaje promedio, $\bar{x} = 6.32$. Pero, como lo ilustra la figura 3.5, el desempeño de David con el arco difiere del de Ricardo, las flechas de Ricardo están más dispersas que las de David. Necesitamos una medida que sea sensible a esta variabilidad; la media no lo es.

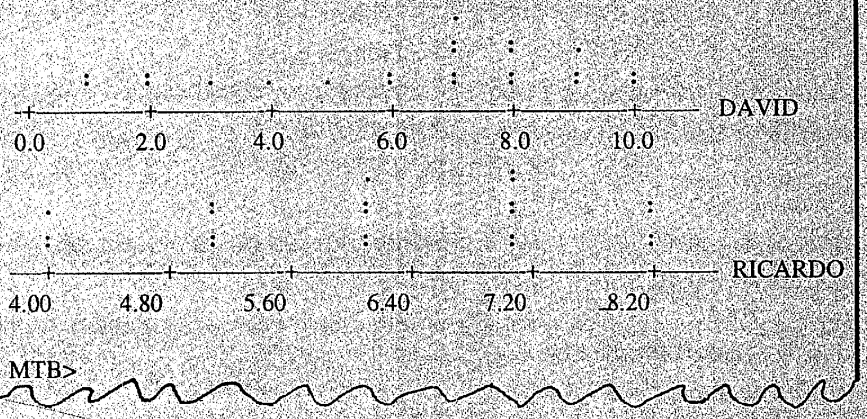
Figura 3.5
Variabilidad del puntaje de Ricardo y David



MINITAB incluye una función de graficación de funciones que usa la orden DOTPLOT. La pantalla de computadora 3.2 contiene dos gráficas hechas con la orden DOTPLOT, una para los puntajes de Ricardo y otra para los de David; cada una tiene escalas horizontales distintas que fueron asignadas por MINITAB. Es difícil comparar estas dos gráficas de puntos y analizar la dispersión de los datos debido a las medidas distintas de los intervalos (1.8 para los puntajes de Ricardo y 2 para los de David); la dificultad se supera fácilmente usando una suborden de MINITAB (SAME), como se muestra en la pantalla 3.3. Las gráficas de puntos usan la misma escala (2). Ahora es claro que los puntajes de David son más variables.

Pantalla 3.2

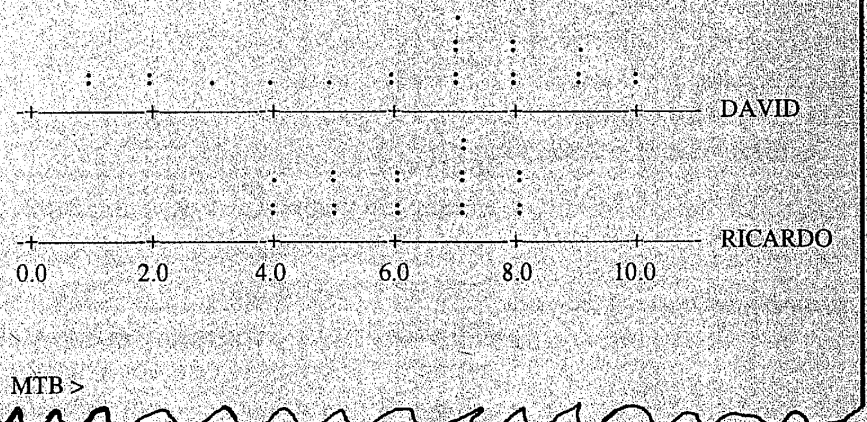
```
MTB> SET C1
DATA> 2(10) 3(9) 4(8) 7(7) 2(6) 5 4 3 2(2) 2(1)
DATA> END
MTB> SET C2
DATA> 5(8) 8(7) 5(6) 4(5) 3(4)
DATA> END
MTB> NAME C1 'DAVID'
MTB> NAME C2 'RICARDO'
MTB> DOTPLOT 'DAVID' 'RICARDO'
```




```
MTB>
```

Pantalla 3.3

```
MTB> DOTPLOT 'DAVID' 'RICARDO';
SUBC> SAME.
```




```
MTB >
```

¿Debe mandar el entrenador Wells a Jones como bateador emergente? Su porcentaje es 0.310, pero en algunos juegos lo ponchan todo el tiempo y en otros logra un hit en todas sus veces al bat. ¿O debe poner a Smith, quien tiene un porcentaje de bateo de 0.290 y logra al menos un hit en todos los juegos en que participa? La respuesta parece obvia: mandar a Smith porque su capacidad de bateo es menos variable. Cualquier colección de medidas hechas con una misma unidad variará según la precisión del instrumento de medición. Por ejemplo, en una caja de 24 barras de caramelo de 2 onzas, no todas las barras pesarán exactamente 2 onzas; si eso ocurre, la escala no es sensible o suficientemente precisa; si las mismas barras de caramelo se pesan en una balanza analítica sensible no tendrán todas el mismo peso, mostrarán

cierto grado de variabilidad y esto no es deseable, porque si los pesos exceden de 2 onzas, el fabricante perderá dinero en la producción y venta de las barras de caramelo; por otro lado, si los pesos de las barras son menores de 2 onzas, el consumidor estará siendo engañado, lo cual causará quejas del cliente y una pérdida potencial de negocios. En cualquier caso, una gran variabilidad en los pesos de las barras de dulce no puede ser tolerada administrativamente.

La **variabilidad** es un concepto fundamental en estadística. Hay muchas medidas de variabilidad o **medidas de dispersión** para una colección de datos cuantitativos. Entre estas medidas están incluidos:

a) el rango
b) el rango intercuartil
c) la varianza
d) la desviación estándar

Examinaremos ahora en detalle estas cuatro medidas de variabilidad.

Rango

Dada una distribución de medidas muestrales o poblacionales, el **rango** se define como la diferencia entre la medida máxima U y la medida mínima L; es decir:

$$R = U - L$$

EJEMPLO 3.18

Las edades en años de un grupo familiar son: 30, 2, 1, 7, 4, 32 y 10. El rango es

$$\begin{aligned} R &= U - L \\ &= 32 - 1 = 31 \end{aligned}$$

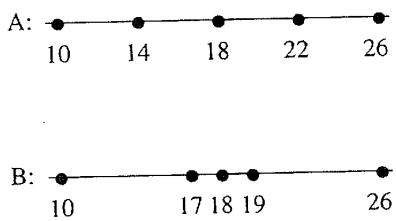
En la sección 2.2 usamos el rango R para determinar el ancho de los intervalos para una tabla de frecuencia agrupada. Como es fácil determinar el rango, a menudo se usa para estimar otras medidas de variabilidad, como la desviación estándar, que no se calcula fácilmente (véase el ejercicio 52 al final de esta sección). Sin embargo, el rango no siempre es una medida sensible para la dispersión de una colección de datos, como se ve en el ejemplo 3.19; y tiene otra desventaja: puede afectarse drásticamente por la presencia de valores extremos de los datos, llamado en ocasiones *observaciones aberrantes.*

EJEMPLO 3.19

Para los dos conjuntos de datos ilustrados en las rectas numéricas de la figura 3.6, ¿cuál es más disperso, A o B? La respuesta es, claramente, el conjunto A, pero note que A y B tienen el mismo rango; este ejemplo ilustra que el rango no es una medida sensible de dispersión, por esta razón, no se considera como una media de dispersión demasiado útil.

FIGURA 3.6

El rango como una mnedida de dispersión



Rango intercuartil

Una medida de dispersión que es indiferente de la presencia de observaciones aberrantes es el **rango intercuartil**, denotado por IQR (por el término en inglés *interquartile range*). Se define como:

Rango intercuartil

$$\text{IQR} = Q_3 - Q_1$$

donde Q_3 es el tercer cuartil y Q_1 es el primer cuartil.

EJEMPLO 3.20

Considere el siguiente conjunto ordenado de datos que representa los valores de oxígeno registrados (en mL/kg · min) de 21 corredores de mediana edad del sexo masculino, mientras pedalean en una bicicleta fija a 100 watts:[21]

12.81	14.95	15.83	15.97	17.90	18.27	18.34	19.82	19.94	20.62	
20.88	20.93	20.98	20.99	21.15	22.16	22.24	23.16	23.56	35.78	36.73

Los valores 35.78 y 36.73 aparecen como valores extremos u observaciones aberrantes, para este conjunto de datos. Por la definición de rango intercuartil, está claro que estos valores no tendrán efecto en el valor del rango intercuartil; esos dos valores pueden reemplazarse por otros dos cualesquiera que ocupen los lugares 20 y 21 del conjunto ordenado, lo que no afecta el valor del rango intercuartil. Calculemos el IQR usando los tres pasos siguientes:

1. Calcule el primer cuartil. El primer cuartil es aquel valor para el cual al menos (0.25)(21) = 5.25 de las medidas caen en él o debajo de él y al menos (0.75)(21) = 15.75 valores por arriba; así, 6 valores están situados en o debajo de 18.27 y 16 valores en o encima de 18.27. En consecuencia, el primer cuartil es

$$Q_1 = 18.27$$

2. Calcule el tercer cuartil. Contando seis valores desde el extremo derecho en el arreglo ordenado, determinamos que el tercer cuartil es

$$Q_3 = 22.16$$

3. Calcule el valor de IQR. El valor del rango intercuartil es

$$\begin{aligned}\text{IQR} &= Q_3 - Q_1 \\ &= 22.16 - 18.27 = 3.89\end{aligned}$$

El rango intercuartil no se afecta por las observaciones aberrantes 35.78 y 36.73, mientras que sí se afecta por 36.73.

Usaremos el rango intercuartil en la sección 3.4 para construir gráficas de caja, resúmenes de datos que proporcionan información sobre el centro, la dispersión, la simetría contra el sesgo y la presencia de observaciones aberrantes.

El rango y el rango intercuartil no son medidas sensibles de variación. El rango es dependiente sólo en los valores extremos L y U, mientras que el rango intercuartil no toma en cuenta las medidas debajo de Q_1 o arriba de Q_3. La varianza y la desviación estándar son ambas medidas más sensibles de variación que el rango o el rango intercuartil, pues toman en cuenta todas las medidas en un conjunto de datos, pero comparten una desventaja común consistente en que a ambas las influyen por puntajes extremos. Examinaremos estas medidas que se refieren al concepto de desviación de un valor más adelante.

Desviación de un valor

En estadística, la cantidad $(x - \bar{x})$ se llama el **valor de desviación**

$$\text{El valor de desviación} = x - \bar{x}$$

Una desviación positiva para una medida, indica que la medida está por encima de la media, mientras que una desviación negativa nos señala que está por debajo de la media; una desviación de 0 para una medida indica que la medida es igual a la media.

APLICACIÓN 3.6

Calcule la desviación de los puntajes para los datos siguientes, que representan el número de defectos encontrados por un inspector de automóviles en una línea de ensamblaje en los últimos cinco automóviles producidos: 1, 4, 6, 6 y 8.

Solución: Es fácil determinar que la media muestral sea $\bar{x} = 5$. Las desviaciones de los valores se presentan en la tabla siguiente:

x	$x - \bar{x}$
1	$1 - 5 = -4$
4	$4 - 5 = -1$
6	$6 - 5 = 1$
6	$6 - 5 = 1$
8	$8 - 5 = 3$

Podemos observar que:

a) Las medidas 6 y 8 están arriba de la media y sus desviaciones son positivas.
b) Las medidas 1 y 4 están debajo de la media y sus desviaciones son negativas.
c) La suma de las desviaciones es 0. ■

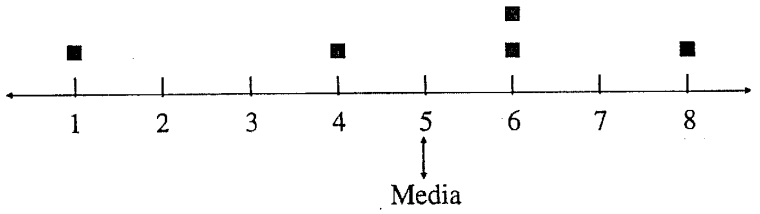
Se puede demostrar fácilmente que la suma de las desviaciones de los valores para cualquier conjunto de números es 0; esto es,

$$\Sigma\ (x - \bar{x}) = 0, \text{ para cualquier conjunto de datos} \qquad (3.3)$$

La ecuación 3.3 tiene una interpretación física interesante.

EJEMPLO 3.21

La media de un conjunto de números puede describirse geométricamente como el punto en la recta numérica que sirve como "centro de gravedad" para los números. Si imaginamos que la recta numérica está apoyada en un punto (punto de apoyo), localizado en la media y que en los números de la recta correspondientes a los números dados se colocan pesos de 1 unidad, entonces la ecuación 3.3 implica que los pesos debajo de la media compensarán perfectamente a los pesos arriba de la media; en otras palabras, la media sirve como centro de gravedad de los datos. Considere el diagrama siguiente para los datos sobre los defectos de los automóviles.



La recta numérica con un punto de apoyo en 5, la media del conjunto de los datos, estaría perfectamente balanceada si se colocaran pesos unitarios en los valores de los datos 1, 4, 6, 6 y 8.

APLICACIÓN 3.7

Los datos siguientes representan los totales anuales, en billones de dólares, erogados por Estados Unidos para exportaciones agrícolas desde países extranjeros entre 1974 y 1983, respectivamente: 10.2, 9.3, 11.0, 13.4, 14.8, 16.7, 17.4, 16.8, 15.4 y 16.2.[22] Encuentre la desviación para cada uno de los totales y verifique que la ecuación 3.3 es válida para el conjunto de datos.

Solución: Se encuentra que la media es $\bar{x} = 14.12$. Las desviaciones de los valores están contenidas en la tabla 3.3.

TABLA 3.3
Datos y desviación para la aplicación 3.7

Año	Total	Desviación
1974	10.2	–3.92
1975	9.3	–4.82
1976	11.0	–3.12
1977	13.4	–0.72
1978	14.8	0.68
1979	16.7	2.58
1980	17.4	3.28
1981	16.8	2.68
1982	15.4	1.28
1983	16.2	2.08
		0

Sumando los valores de las desviaciones tenemos

$$\Sigma (x - \bar{x}) = 0. \quad \blacksquare$$

Suma de cuadrados

La desviación de los valores puede usarse para describir la dispersión de una distribución dada de datos cuantitativos. Recuerde que la desviación de un

valor representa la distancia dirigida entre una medida y la media de un conjunto de datos; en consecuencia, podríamos pensar que el promedio de todas las desviaciones de los valores proporciona una medida de la dispersión de todas las medidas respecto a la media, pero eso no ocurre, pues la ecuación 3.3 dice que la suma de todas las desviaciones de los valores es 0. Al sumar, las desviaciones positivas de valores se cancelan con las desviaciones negativas. Para evitar este problema causado porque las desviaciones de valores negativas cancelan las positivas, podemos elevar primero al cuadrado cada desviación antes de sumar; la suma de los cuadrados de las desviaciones que se obtiene se llama la **suma de cuadrados** y se denota SS. Como veremos posteriormente, SS es muy útil en estadística para describir la dispersión de una colección de medidas respecto a su media.

Podemos calcular una suma de cuadrados ya sea para una muestra o para una población. Las fórmulas para ambos casos son las siguientes:

Fórmulas de suma de cuadrados

$$\underset{\text{Muestra}}{SS = \Sigma (x - \bar{x})^2} \qquad \underset{\text{Población}}{SS = \Sigma (x - \mu)^2} \tag{3.4}$$

Las fórmulas difieren, pero los procedimientos del cálculo son los mismos.

EJEMPLO 3.22

Encontremos la SS para la muestra siguiente de puntajes en los exámenes sobre la historia de América hechos por cinco estudiantes: 62, 80, 83, 72 y 73. Primero encontramos $\bar{x}$:

$$\bar{x} = \frac{62 + 80 + 83 + 72 + 73}{5} = 74$$

Entonces, usando la fórmula 3.4 tenemos:

$$\begin{aligned} SS &= \Sigma (x - \bar{x})^2 \\ &= (62 - 74)^2 + (80 - 74)^2 + (83 - 74)^2 + (72 - 74)^2 + (73 - 74)^2 \\ &= 144 + 36 + 81 + 4 + 1 = 266 \end{aligned}$$

En general, una suma de cuadrados SS se puede encontrar como sigue:

Cómo determinar SS

a) Determine la media.
b) Encuentre la desviación para cada medida.
c) Eleve al cuadrado cada una de las desviaciones.
d) Encuentre la suma de los cuadrados.

Para simplificar el procedimiento necesario en el cálculo de SS, serán útiles las fórmulas:

Fórmulas para el cálculo de SS

$$\text{SS} = \sum x^2 - \frac{(\sum x)^2}{n} \qquad \text{SS} = \sum x^2 - \frac{(\sum x)^2}{N} \tag{3.5}$$

Muestra Población

donde Σx^2 es la suma de los cuadrados de los datos, n es el tamaño de la muestra y N el tamaño de la población. Las dos fórmulas dadas en 3.5 pueden verificarse algebraicamente usando las propiedades de la suma que se encuentra en el apéndice A.

EJEMPLO 3.23

Note que $\Sigma x^2 \neq (\Sigma x)^2$. Esto puede demostrarse observando que

$$2^2 + 3^2 \neq (2+3)^2$$
$$13 \neq 25.$$

EJEMPLO 3.24

Con referencia al ejemplo 3.22, usemos la fórmula 3.5 para calcular SS de los puntajes del examen de historia de América. Organizamos primero los cálculos usando la tabla siguiente con dos columnas

x	x^2
62	3,844
80	6,400
83	6,889
72	5,184
73	5,329
370	27,646

Al usar la fórmula 3.5 para calcular SS, obtenemos:

$$\text{SS} = \sum x^2 - \frac{(\sum x)^2}{n}$$
$$= 27{,}646 - \frac{370^2}{5} = 266$$

Para propósitos de cálculo, se acostumbra preferir las fórmulas dadas en (3.5) a las dadas en (3.4). Por un lado, las fórmulas en (3.5) son más fáciles de usar con una calculadora, pues hay menos restas en ellas y no requieren encontrar la media. Si se usan las fórmulas (3.4) en situaciones donde la media no termina y se requiera redondearla, los cálculos llevarán a resultados faltos de precisión. Es fácil encontrar un ejemplo donde (3.5) da mayor precisión que (3.4).

EJEMPLO 3.25

Encontremos SS para los valores 0, 5 y 8. Si la media se redondea al décimo más próximo, entonces $\bar{x} = 13/3 \approx 4.3$. Entonces usando la fórmula (3.4) resulta:

$$\text{SS} = \Sigma (x - \bar{x})^2$$
$$= (0-4.3)^2 + (5-4.3)^2 + (8-4.3)^2 = 32.670$$

Usando la fórmula (3.5) tenemos:

$$SS = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$= 89 - \frac{13^2}{3} = 32.667$$

Al milésimo más próximo, las dos respuestas difieren por 0.003.

Varianza

La **varianza** de una población de medidas se define como el promedio de los cuadrados de las desviaciones de los valores y se denota por σ^2 (léase sigma cuadrada). El símbolo es la letra griega minúscula sigma. La varianza de la población está dada por la fórmula (3.6).

Varianza de una población

$$\sigma^2 = \frac{SS}{N} \tag{3.6}$$

La varianza de una muestra se denota por s^2 y se define por la fórmula siguiente:

Varianza de una muestra

$$s^2 = \frac{SS}{n-1} \tag{3.7}$$

En los capítulos del 8 al 15, usaremos la varianza muestral s^2 para estimar la varianza poblacional desconocida σ^2. Note que:

> Si fuéramos a calcular la varianza muestral s^2 dividiendo SS entre n en lugar de $n - 1$, estaríamos subestimando σ^2, en promedio.

Algunos estadísticos calculan la varianza muestral, únicamente con propósitos, dividiendo SS entre n; desde luego, para valores grandes de n hay poca diferencia entre los valores de SS/n y SS/$(n - 1)$. Si la varianza se usa por sí misma como medida descriptiva de la dispersión, es difícil interpretarla porque las unidades de la varianza son el cuadrado de las unidades de medida.

APLICACIÓN 3.8

Suponga que los puntajes de los exámenes de historia de América dados previamente: 62, 80, 83, 72 y 73 constituyen una población. Encuentre la varianza poblacional σ^2.

Solución: Al usar la fórmula (3.6), tenemos:

$$\sigma^2 = \frac{SS}{N}$$

$$= \frac{266}{5} = 53.2 \quad \blacksquare$$

APLICACIÓN 3.9

La tabla 3.4 muestra los costos por litro, en centavos de dólar, de la gasolina de alto octanaje en 19 ciudades del mundo. Determine la varianza muestral s^2.[23]

TABLA 3.4

Costo de la gasolina en 19 ciudades del mundo

Ciudad	Costo por litro
Amsterdam	57
Bruselas	53
Buenos Aires	38
Hong Kong	57
Johannesburgo	48
Londres	56
Madrid	59
Manila	46
México	25
Montreal	47
Nairobi	57
Nueva York	40
Oslo	65
París	58
Río de Janeiro	42
Roma	76
Singapur	59
Sidney	43
Tokio	79

Solución: Usaremos la fórmula (3.5) para calcular SS. Con este propósito, primero calculamos Σx y Σx^2; con la ayuda de una calculadora determinamos que $\Sigma x = 1005$ y $\Sigma x^2 = 56{,}171$. Así, la suma de cuadrados es:

$$SS = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

$$= 56{,}171 - \frac{1005^2}{19} = 3011.7895$$

Ahora aplicamos la fórmula (3.6) para obtener:

$$s^2 = \frac{SS}{n-1}$$

$$= \frac{3011.7895}{18} \approx 167.32$$

La varianza muestral de los 19 precios de gasolina es 167.32 centavos cuadrados. ■

EJEMPLO 3.26

Para los datos de los precios por litro de la gasolina de la aplicación 3.9, el conocimiento de que $s^2 = 167.32$ centavos cuadrados tiene muy poco significado por sí mismo, si es que tiene alguno. Sabemos que si el valor de la varianza es grande, entonces las medidas están muy dispersas, mientras que si es pequeño hay muy poca variabilidad en las medidas.

EJEMPLO 3.27

Si la varianza es 0, todas las medidas son iguales; esto es consecuencia de que SS es siempre mayor o igual que 0 y es igual a 0 sólo si cada medida es igual a la media.

EJEMPLO 3.28

Sin embargo, si al analizar dos muestras de datos A y B, hubiéramos encontrado que $s_A^2 = 10$ y $s_B^2 = 5$, sabríamos que las medidas de la muestra A están más dispersas respecto a su media de lo que lo están las medidas de la muestra B respecto a su media. La varianza se usa la mayoría de las veces y con propósitos descriptivos, para comparaciones como una medida relativa de variación.

Desviación estándar

Otra medida de dispersión, relacionada con la varianza, es la desviación estándar. La **desviación estándar** se define como la raíz cuadrada de la varianza. La desviación estándar poblacional se denota con σ y la desviación estándar muestral con s. En consecuencia, tenemos las fórmulas siguientes:

Desviación estándar muestral

$$s = \sqrt{s^2} = \sqrt{\text{varianza muestral}}$$

Desviación estándar poblacional

$$\sigma = \sqrt{\sigma^2} = \sqrt{\text{varianza poblacional}}$$

EJEMPLO 3.29

Para los datos de la aplicación 3.8, la desviación estándar poblacional es $\sigma = \sqrt{53.2} = 7.29$, y para los datos de la aplicación 3.9, la desviación estándar muestral es $s = \sqrt{167.32} = 12.94$ centavos.

Pantalla 3.4

La pantalla 3.4 ilustra el uso de MINITAB para los datos en la aplicación 3.9. Note que MINITAB no da la varianza directamente.

```
MTB> SET C1
DATA> 57 53 38 57 48 56 59 46 25 47 57 40 65 58 42 76 59 43 79
DATA> END
MTB> MEAN C1
   MEAN = 52.895
MTB> STDEV C1
   ST.DEV. = 12.935
MTB> LET K1 = STDEV(C1)**2
MTB> PRINT K1
K1 167.322
```

Note que K1 es el valor de la varianza. K1 se llama una constante en MINITAB; las constantes se llaman por su nombre, K1, K2, K3, ... cada una puede almacenar un número y pueden crearse usando la orden LET. La orden LET K1 = STDEV(C1)**2 almacena la varianza en la constante K1. Advierta también que el símbolo "**" significa exponenciación o elevar un número a una potencia.

¿Por qué necesitamos tanto la varianza como la desviación estándar, como medidas de dispersión? Una respuesta a esta pregunta involucra la unidad de

medida. Como vimos en la aplicación 3.9, si el conjunto de datos se refiere a medidas en centavos, entonces la unidad de la varianza es centavos al cuadrado y la unidad de la desviación estándar es centavos; por lo tanto, una expresión como $x - \bar{x}$ tendría significado, pero una expresión como $x - s^2$ no lo tendría porque en el primer caso las unidades coinciden, pero en el segundo no. En la sección 3.4, cuando estudiemos puntajes estándar, usaremos el hecho de que una cierta medida de una distribución y la media y la desviación estándar de la distribución tengan todas las mismas unidades de medida. Las aplicaciones 3.10 y 3.11 mostrarán el uso de la varianza muestral y la desviación estándar muestral para hacer comparaciones relativas.

APLICACIÓN 3.10

Los datos adjuntos representan el promedio de millas por galón diario por cinco días para los coches A y B, en condiciones similares.

A	20	25	30	15	35
B	15	27	25	23	35

a) Encuentre la media y el rango de millas por galón para cada coche.

b) ¿Cuál coche parece haber logrado un rendimiento más consistente si la consistencia se determina examinando las varianzas? Explique.

Solución:

a) Para el coche A tenemos:

$$R_A = 35 - 15 = 20$$
$$\bar{x}_A = 25$$

Para el coche B:

$$R_B = 35 - 15 = 20$$
$$\bar{x}_B = 25$$

Note que ambos coches tienen la misma media y el mismo rango en el registro de millas por galón.

b) Calculamos la varianza para el coche A, s_A^2.

x	$x - \bar{x}$	$(x - \bar{x})^2$
20	− 5	25
25	0	0
30	5	25
15	− 10	100
35	10	100
		SS = 250

Como consecuencia de la fórmula (3.7) tenemos:

$$s_A^2 = \frac{SS}{n - 1}$$
$$= \frac{250}{4} = 62.5$$

La varianza en el rendimiento de la gasolina para el coche A es de 62.5 millas cuadradas. Calculamos ahora la varianza para el coche B, s_B^2.

x	$x - \bar{x}$	$(x - \bar{x})^2$
15	− 10	100
27	2	4
25	0	0
23	− 2	4
35	10	100
		SS = 208

Como resultado de la fórmula (3.7) tenemos:

$$s_B^2 = \frac{SS}{n-1}$$
$$= \frac{208}{4} = 52$$

La varianza en el rendimiento de la gasolina para el coche B es de 52 millas cuadradas; como la varianza para el carro B es menor que para el carro A, el carro B resultó más consistente en el rendimiento. Si hubiéramos usado el rango, habríamos concluido que ambos coches tenían un rendimiento igualmente consistente. ■

APLICACIÓN 3.11

Los datos en la tabla 3.5 indican los precios, en dólares, por libra, de asado de cerdo y queso cheddar en 15 capitales del mundo.[24]

Tabla 3.5
Precios del asado de cerdo y del queso en capitales del mundo

Capital	Cerdo asado (sin hueso)	Queso cheddar
Berna	$6.61	$4.00
Bonn	2.38	2.74
Brasilia	1.27	1.08
Buenos Aires	1.36	2.03
Camberra	2.06	2.60
Londres	1.56	1.81
Madrid	2.33	3.15
México	1.08	2.29
Ottawa	1.99	3.98
París	2.47	2.37
Pretoria	1.95	1.76
Roma	2.46	2.96
Estocolmo	5.35	2.54
Tokio	4.19	2.38
Washington	3.29	2.69

¿Para cuál alimento, el asado de cerdo o el queso cheddar, son menos variables y más estables los precios?

Solución: Determinamos las cantidades siguientes con el uso de una calculadora:

Datos del asado de cerdo: $\Sigma x = 40.35$, $\Sigma x^2 = 143.01$, $n = 15$

Datos del queso cheddar: $\Sigma x = 38.38$, Σx^2 106.67, $n = 15$

Y como una consecuencia de las fórmulas (3.5) y (3.7) tenemos:

Datos del asado de cerdo

$$SS_p = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$
$$= 143.01 - \frac{(40.35)^2}{15} = 34.4685$$

La varianza de los datos del asado de cerdo es:

$$s_p^2 = \frac{SS_p}{n-1}$$
$$= \frac{34.4685}{14} \approx 2.46$$

Datos del queso

$$SS_c = 106.67 - \frac{(38.38)^2}{15} = 8.4684$$

Y la varianza de los datos del queso es:

$$s_c^2 = \frac{8.4684}{14} = 0.60$$

Por lo tanto, los precios del queso cheddar en el mundo son más estables que los del asado de cerdo. ■

Estimación de s

Es interesante notar que para muestras de un tamaño mínimo de 20 con una distribución de forma acampanada, tenemos la estimación siguiente de la desviación estándar muestral:

Estimación de *s*

$$s \approx \frac{R}{4} \tag{3.8}$$

donde R denota el rango; esta es una estimación conservadora que puede usarse para verificar nuestros cálculos de s y requiere muy poco esfuerzo. El significado de dividir el rango entre 4 se discutirá en el capítulo 7, cuando examinemos las distribuciones normales.

APLICACIÓN 3.12

Para los datos del queso cheddar en la aplicación 3.11, estime s usando la fórmula (3.8), y verifique la estimación calculando el valor de s.

Solución: El rango para los precios del queso cheddar es:

$$\begin{aligned} R &= U - L \\ &= 4.00 - 1.08 = 2.92 \end{aligned}$$

Como consecuencia de la fórmula 3.8, tenemos:

$$\begin{aligned} s_c &\approx \frac{R}{4} \\ &= \frac{2.92}{4} = 0.73 \end{aligned}$$

Como la desviación estándar es la raíz cuadrada de la varianza, podemos usar el resultado de la aplicación 3.11 para obtener:

$$\begin{aligned} s_c^2 &= 0.60 \\ s_c &= \sqrt{0.60} = 0.77 \end{aligned}$$

Como $R/4 = 0.73$ está en la misma "cancha" que $s_c = 0.77$, tenemos poca razón para sospechar que se ha cometido un error. ■

APLICACIÓN 3.13

Suponga que en una muestra la medida mayor es 90 y la menor 30; se ha calculado que la desviación estándar es 185. ¿Es razonable este valor? Explique.

Solución: No, el valor no parece razonable. El rango es 90 – 30 = 60 y la fórmula (3.8), vemos:

$$s \approx \frac{R}{4} = \frac{60}{4} = 15$$

Así, sospechamos que se ha cometido un error al calcular s como 185 y debe verificarse el procedimiento.

Varianza y desviación estándar para datos en tablas de frecuencia

A menudo tendremos ocasión de encontrar la varianza y la desviación estándar para datos desplegados en una tabla de frecuencia. Ambas medidas pueden calcularse una vez que se conoce SS; para encontrar SS en datos que tienen medidas con repetición, determinamos primero la frecuencia de cada medida.

EJEMPLO 3.30

Para encontrar la suma de cuadrados SS para los datos 2, 2, 2, 2 y 7, que representan el número de carreras concedidas por un pitcher de beisbol en los últimos cinco

juegos, sólo necesitamos encontrar la desviación de los valores 2 y 7. El cuadrado de la desviación del valor 2 puede entonces multiplicarse por su frecuencia, $f = 4$, para obtener la suma de los cuadrados de las desviaciones de los cuatro valores 2; esta suma se añade al cuadrado de la desviación del valor para 7 a fin de obtener SS. Como la media de los cinco datos es 3, tenemos:

$$\begin{aligned} SS &= \Sigma (x - \bar{x})^2 \\ &= 4(2 - 3)^2 + 1(7 - 3)^2 \\ &= 4 + 16 \\ &= 20 \end{aligned}$$

Con base en las ideas del ejemplo 3.30, tenemos las fórmulas siguientes para encontrar la suma de cuadrados cuando los datos se organizan en una tabla de frecuencia:

Suma de cuadrados para datos en una tabla de frecuencias

$$SS = \Sigma f(x - \bar{x})^2 \quad \text{Muestra} \qquad SS = \Sigma f(x - \mu)^2 \quad \text{Población} \tag{3.9}$$

APLICACIÓN 3.14

Las medidas siguientes representan los días que tarda el correo expreso, enviado desde la costa oeste, en llegar a su destino en la costa este en los pasados diez envíos: 2, 2, 2, 3, 3, 4, 4, 5, 5 y 10. Use las fórmulas (3.9) para determinar SS.

Solución: Primero construimos la tabla 3.6 que nos ayudará en los cálculos. Se encuentra fácilmente que la media muestral es $\bar{x} = 4$.

TABLA 3.6
Tabla de frecuencias para la aplicación 3.14

x	f	$x - \bar{x}$	$(x - \bar{x})^2$	$f(x - \bar{x})^2$
2	3	−2	4	12
3	2	−1	1	2
4	2	0	0	0
5	2	1	1	2
10	1	6	36	36
				SS = 52

El valor de SS es la suma de las entradas en la última columna, SS = 52. Para mayor ilustración, determinaremos también el valor de SS usando la fórmula (3.4) y la tabla 3.7.

TABLA 3.7
Cálculo de SS usando la fórmula (3.4)

x	$x - \bar{x}$	$(x - \bar{x})^2$	
2	−2	4	
2	−2	4	$f = 3$ y $(3)(4) = 12$
2	−2	4	
3	−1	1	
3	−1	1	$f = 2$ y $(2)(1) = 2$
4	0	0	
4	0	0	$f = 2$ y $(2)(0) = 0$
5	1	1	
5	1	1	$f = 2$ y $(2)(1) = 2$
10	6	36	$f = 1$ y $(1)(36) = 36$
		SS = 52	

Vemos que $SS = 52$, como se calculó usando la fórmula (3.9); note que la primera entrada de la quinta columna, 12, de la tabla 3.6 corresponde a la suma de las primeras tres entradas de 4 listadas en la última columna de la tabla 3.7, y así sucesivamente. ■

La fórmula de cálculo siguiente puede usarse para obtener la suma de cuadrados para datos desplegados en una tabla de frecuencia.

Fórmula para calcular SS usando frecuencias

$$SS = \Sigma f x^2 - \frac{(\Sigma f x)^2}{\Sigma f} \tag{3.10}$$

Muchas veces es más conveniente usar la fórmula (3.10) que las fórmulas (3.9). Vea que en la fórmula (3.10) soló aparece una resta y que no es necesario calcular primero la media.

APLICACIÓN 3.15

Encuentre la varianza muestral para los datos siguientes referentes al número de cigarros fumados durante un fin de semana por un grupo de 15 fumadores:

x	10	15	17	20	22
f	1	3	5	2	4

Solución: La tabla siguiente se usa para organizar los cálculos:

x	f	fx	x^2	fx^2
10	1	10	100	100
15	3	45	225	675
17	5	85	289	1445
20	2	40	400	800
22	4	88	484	1936
	15	268		4956

Como consecuencia de la fórmula (3.10), tenemos:

$$SS = \Sigma fx^2 - \frac{\Sigma (fx)^2}{\Sigma f}$$
$$= 4956 - \frac{268^2}{15} = 167.73$$

Luego, la varianza muestral es:

$$s^2 = \frac{SS}{n-1}$$
$$= \frac{167.73}{14} = 11.981$$

Advierta también que las entradas en la columna fx pueden encontrarse ya sea 1) multiplicando las entradas correspondientes en las columnas x y fx, o 2) elevando al cuadrado las entradas en la columna x y multiplicando después por los valores adecuados de f. ■

Desventajas de la varianza y de la desviación estándar

La varianza y la desviación estándar tienen una limitación seria: pueden verse gravemente afectadas en presencia de observaciones aberrantes, pues ambas dependen de la media, que se modifica por las medidas extremas. Cuando en un conjunto de datos están presentes observaciones aberrantes y se requiere una medida resistente a ellas, debe utilizarse el rango intercuartil.

Teorema de Chebichev

La desviación estándar muestral s indica la dispersión de los datos respecto a la media muestral. Si los valores de los datos se acumulan cerca de la media, entonces s es pequeña; si se dispersan considerablemente respecto a la media, entonces s es grande; pero, ¿cómo podemos determinar cuáles valores de s son grandes y cuáles son pequeños? Un teorema que lleva el nombre del matemático ruso Pafnuty Lvovich Chebichev (1821-1894), nos da alguna información útil sobre cómo la magnitud de la desviación estándar de cualquier conjunto de datos se relaciona con la concentración de éstos en torno a la media. Según el teorema de Chebichev, la afirmación siguiente es cierta para cualquier conjunto de datos cuantitativos, tanto poblacionales como muestrales:

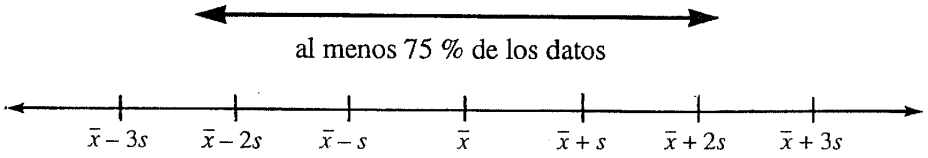
Teorema de Chebichev

La expresión $1 - 1/k^2$ representa la proporción mínima de los datos que dista no más de k desviaciones estándar de la media si $k \geq 1$.

Note que el resultado del cálculo $1 - 1/k^2$ es una fracción; al multiplicarla por 100 se obtiene el porcentaje mínimo de los datos que distan no más de k desviaciones estándar de la media, de acuerdo con el teorema de Chebichev, para cualquier conjunto de medidas.

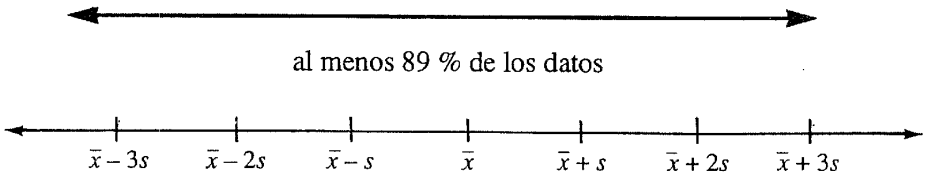
- Si $k = 1$, entonces $1 - 1/k^2 = 1 - 1/1^2 = 0$. Entonces, al menos 0% de los datos dista no más de una desviación estándar de la media (esto es, cae dentro de $\bar{x} \pm s$). Así, para $k = 1$, la interpretación no ofrece información útil respecto a la dispersión de los datos.
- Si $k = 3/2$, entonces $1 - 1/(3/2)^2 = 1 - 4/9 = 5/9 \approx 56\%$, por lo tanto, al menos el 56% de los datos distarán no más de 1.5 desviaciones estándar de la media (esto es, caerán dentro de $\bar{x} \pm 1.5s$).
- Si $k = 2$, al menos $1 - 1/2^2 = 3/4 = 75\%$. Entonces, al menos 75% de los datos deben distar no más de 2 desviaciones estándar de la media (esto es, caerán dentro $x \pm 2s$), como se ilustra en la figura 3.7.

FIGURA 3.7
Ilustración del teorema de Chebichev para $k = 2$



- Para $k = 3$, al menos $(1 - 1/3^2)\ 100\% = 89\%$ de los datos de cualquier muestra deben distar no más de 3 desviaciones estándar de la media (por ejemplo, caerán dentro de $\bar{x} \pm 3s$), como se muestra en la figura 3.8.

FIGURA 3.8
Ilustración del teorema de Chebichev para $k = 3$



APLICACIÓN 3.16

Aquí se recuerdan los datos del costo de la gasolina de la aplicación 3.9.

Ciudad	Costo por litro
Amsterdam	57
Bruselas	53
Buenos Aires	38
Hong Kong	57
Johannesburgo	48
Londres	56
Madrid	59
Manila	46
México	25
Montreal	47
Nairobi	57
Nueva York	40
Oslo	65
París	58
Río de Janeiro	42
Roma	76
Singapur	59
Sidney	43
Tokio	79

a) Determine el intervalo especificado por el teorema de Chebichev que contendrá al menos 75% de los datos.

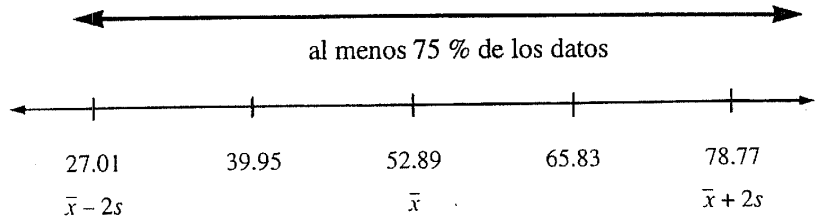
b) ¿Qué porcentaje de las medidas dista realmente menos de dos desviaciones estándar de la media?

Solución:

a) Con la ayuda de una calculadora, podemos determinar fácilmente que la media es $\bar{x} = 52.89$ centavos. Anteriormente determinamos el valor de la varianza muestral, $s^2 = 167.32$. Así, la desviación estándar es $s^2 = \sqrt{167.32} = 12.94$ centavos. De acuerdo con el teorema de Chebichev, al menos $1 - 1/4 = 3/4 = 75\%$ de los datos distará menos de dos desviaciones estándar de la media, para el conjunto de datos.

$$\bar{x} - 2s = 52.89 - 2(12.94) = 27.01$$
$$\bar{x} + 2s = 52.89 + 2(12.94) = 78.77$$

En consecuencia, el intervalo 27.01, 78.77 contendrá al menos 75% de los datos, como se ilustra en el diagrama.



b) Se encuentra que 17 de los 19 precios de gasolina (89.14%), cae entre 27.01 y 78.77. Esto es consistente con nuestros resultados en la parte a; el teorema de Chebichev especifica sólo una cota inferior para el porcentaje de datos que distan no más de dos desviaciones estándar de la media, como tal, proporciona una estimación conservadora, debido a que se tiene poca información sobre la forma de la muestra. ■

APLICACIÓN 3.17

Suponga que la asistencia promedio a un parte de beisbol de ligas mayores para juegos locales es de 35,500 personas, con una desviación estándar de 4,200. Use el teorema de Chebichev para determinar:

a) un intervalo que contenga al menos 80% de las asistencias a los juegos locales.

b) la proporción mínima de los juegos locales que tiene una asistencia de 25,000 a 46,000 personas.

Solución:

a) Establecemos $1 - 1/k^2$ igual a 0.80 y despejamos k.

$$1 - \frac{1}{k^2} = 0.80$$
$$\frac{1}{k^2} = 0.20$$

$$k^2 = \frac{1}{0.2} = 5$$
$$k = \sqrt{5} \approx 2.24$$

El intervalo es $\bar{x} \pm 2.24s = 35{,}500 \pm (2.24)(4200) = 35{,}500 \pm 9{,}408$ es decir (26,092, 44,908). Así, el teorema de Chebichev garantiza que al menos 80% de las asistencias está entre 26,092 y 44,908.

b) Note que los intervalos de Chebichev son simétricos respecto a la media. El ancho de un intervalo es:

$$w = (\bar{x} + ks) - (\bar{x} - ks) = 2ks$$

Primero determinamos el ancho del intervalo (25,000, 46,000). El ancho es:

$$w = 46{,}000 - 25{,}000 = 21{,}000$$

Planteamos $2ks$ igual a 21,000 y resolvemos la ecuación resultante para k:

$$2ks = 21{,}000$$
$$2k(4200) = 21{,}000$$
$$8400k = 21{,}000$$
$$k = \frac{21{,}000}{8400} = 2.5$$

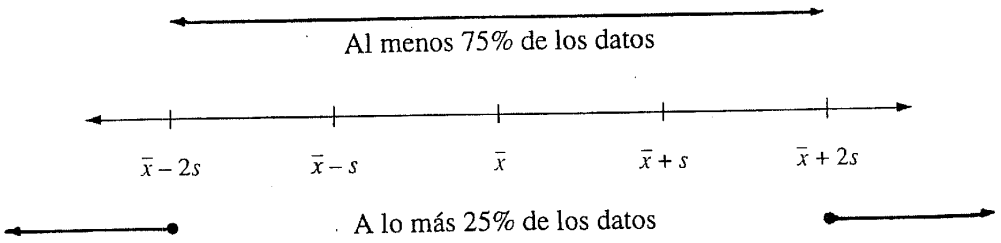
En consecuencia, al menos $1 - 1/(2.5)^2 = 1 - 1/6.25 = 0.84 = 84\%$ de los juegos locales tienen asistencias entre 25,000 y 46,000. ■

A veces es conveniente interpretar el teorema de Chebichev en términos distintos. La afirmación siguiente equivale al teorema de Chebichev.

Forma alternativa del teorema de Chebichev

A lo más $(1/k^2)100\%$ de los datos de cualquier conjunto, distan más de k desviaciones estándar de la media.

Para $k = 2$ tenemos el diagrama siguiente:



El teorema de Chebichev da una explicación de cómo la desviación estándar proporciona una medida de la variación para una sola muestra de población;

la validez del teorema no depende de la forma de la distribución, por eso resulta útil y poderoso.

Resumen de la notación usada

La carta siguiente resume la notación más frecuentemente usada en relación con muestras y poblaciones:

	Media	Mediana	Varianza	Desviación estándar	Tamaño
Muestra	$\bar{x}$	$\tilde{x}$	s^2	s	n
Población	μ	$\tilde{\mu}$	σ^2	σ	N

Note que $\bar{x}$, $\tilde{x}$, s^2, s y n son ejemplos de estadísticos, mientras que μ, $\tilde{\mu}$, σ^2, σ y N son ejemplos de parámetros. Recuerde del capítulo 1 que los estadísticos son los valores calculados a partir de una muestra, y que los parámetros son valores medidos a partir de una población; en estadística, el uso de letras griegas para denotar muchos parámetros es una convención generalizada; una excepción a la regla es la notación para el tamaño de la población.

GRUPOS DE EJERCICIOS 3.2

Habilidades básicas

1. Encuentre el rango, la varianza y la desviación estándar de la muestra:
   5 2 2 1 5 3 2 3 4
2. Encuentre el rango, la varianza y la desviación estándar de:
   9 6 4 6 5 8 7 6 7 0
3. Determine la varianza y la desviación estándar de la muestra 1, 3, 11, 15 y 20.
4. Determine la varianza y la desviación estándar la muestra 1, 2, 4, 10, 18 y 19.
5. Calcule x, s^2 y s para:
   a) $\Sigma x^2 = 232$, $\Sigma x = 25$, y $n = 15$.
   b) $\Sigma x^2 = 515$, $\Sigma x = 101$, y $n = 20$.
6. Calcule la media muestral, la varianza muestral y la desviación estándar muestral para la situación:
   a) $\Sigma x^2 = 52$, $\Sigma x = 7$, y $n = 9$.
   b) $\Sigma x^2 = 25$, $\Sigma x = 12$, y $n = 13$.
7. Para una muestra se han encontrado los siguientes valores:
   $\Sigma x^2 = 428$ $\Sigma x = 75$ $n = 10$
   ¿Son razonables?
8. Para una muestra se han encontrado los siguientes valores:
   $\Sigma x^2 = 48$ $\Sigma x = 7.5$ $n = 20$
   ¿Son razonables?
9. Suponga que hemos calculado la varianza de una muestra de tamaño 15, y obtenido 10 dividiendo SS entre 15 en lugar de 14. Encuentre el valor correcto de s^2.
10. Si una calculadora tiene interconstruido un programa para calcular la varianza, ¿cómo podría determinarse fácilmente cuál varianza, (s^2 o σ^2) se está calculando?
11. ¿Cuál es la suma de las desviaciones de los valores respecto a la media para cualquier conjunto de datos?
12. ¿Cuál es el promedio de cualquier conjunto de desviaciones de valores?
13. ¿Es siempre menor el valor de la desviación estándar que el de la varianza?
14. ¿Por qué carece de sentido la expresión $x - s^2$?
15. ¿Es posible que sean iguales el rango y la desviación estándar de una población? Si lo es, dé un ejemplo.
16. ¿Es posible que sean iguales el rango y la varianza? Si lo es, dé un ejemplo.
17. Si la desviación estándar de un conjunto de datos es 0, ¿qué puede afirmarse de dicho conjunto?
18. ¿Qué puede decirse si la desviación estándar de una muestra es negativa?

19. Suponga que una muestra tiene como media $\bar{x} = 25$ y como desviación estándar $s = 3.2$.
    a) Determine un intervalo que contenga al menos 90% de las medidas de la muestra.
    b) ¿Cuál es el porcentaje mínimo de la muestra que está contenido en el intervalo 17, 33?

20. Suponga que una muestra tiene como media $\bar{x} = 540$ y como desviación estándar $s = 10.5$.
    a) Determine un intervalo que contenga al menos 92% de las medidas de la muestra.
    b) ¿Cuál es el porcentaje mínimo de la muestra que está contenido en el intervalo 524.25, 566.25?

Más aplicaciones

21. El conjunto de datos siguiente representa las calificaciones del examen final para un grupo de 30 estudiantes de filosofía:

98	94	94	57	58	88	97	94	96	85
85	97	92	90	87	80	97	93	87	69
25	100	97	83	74	64	79	89	98	100

    Encuentre el porcentaje de calificaciones que distan menos de 2.1 desviaciones estándar de la media; use entonces el teorema de Chebichev para $k = 3.6$. ¿Los resultados son consistentes con el teorema?

22. Los datos siguientes representan los precios en centavos para una libra de flúor en 16 capitales del mundo:

41	28	10	16	35	18	21	5
40	30	25	18	14	30	33	24

    Encuentre el porcentaje de precios que distan menos de 1.5 desviaciones estándar de la media, luego use el teorema de Chebichev para $k = 1.5$. ¿Los resultados son consistentes con el teorema?

23. El total promedio gastado por los clientes en una tienda de comestibles es 8.34 dólares, y la desviación estándar del total de ventas es 8.33 dólares. ¿Qué puede decirse, usando la regla de Chebichev, de la proporción de clientes que gastan más de 25 dólares?

24. El número de pacientes que ingresan en el Memorial Hospital por día a la semana es en promedio 32, con una desviación estándar de 4; un día, ingresaron sólo 16 pacientes. Use la regla de Chebichev para decidir si éste es un número de ingresos poco usual para un día de la semana. Explique el resultado.

25. La tabla siguiente da una muestra de los tiempos de recorrido, en minutos, de un camino de 2.5 millas para dos coches, A y B.

A:	1.0	0.9	1.0	0.8	0.9	1.0	0.9	1.0
B:	1.3	1.3	1.0	0.9	1.1	0.9	1.4	1.3

    a) Encuentre el promedio de los tiempos de recorrido para cada uno de los coches, A y B.
    b) Calcule la varianza de los tiempos de recorrido para A y B, respectivamente.
    c) ¿Cuál coche tuvo un tiempo promedio menor de recorrido?
    d) ¿Qué coche tuvo un desempeño más consistente, si la consistencia se mide por la varianza?
    e) Encuentre el rango intercuartil para las muestras A y B.

26. La tabla siguiente da una muestra de tiempos de recorrido, en minutos, de un camino de 3 millas para dos coches, C y D.

C:	1.1	0.8	1.1	0.9	1.0	1.0	0.9	1.1
D:	1.2	1.4	1.3	0.9	1.1	0.8	1.5	1.4

    a) Encuentre el tiempo promedio de recorrido para cada uno de los coches, C y D.
    b) Localice la varianza de los tiempos de recorrido para cada uno de los carros.
    c) ¿Cuál coche tuvo un promedio menor de recorrido?
    d) ¿Cuál coche se desempeñó más consistentemente?
    e) Encuentre el rango intercuartil para las muestras C y D.

27. La tabla adjunta indica los salarios anuales, en dólares, para una muestra de 25 trabajadores.

Salario anual	Frecuencia
$5,500	7
6,000	5
7,000	6
8,000	4
30,000	3

    Encuentre:
    a) el rango.
    b) la media.
    c) la desviación estándar.
    d) el rango intercuartil.

28. La tabla adjunta muestra la distribución para el número de transistores defectuosos encontrados en 215 lotes producidos por un trabajador manual electrónico.

Número de transistores defectuosos	Número de lotes
0	25
1	78
2	54
3	33
4	16
5	7
6	2

a) Encuentre el rango.
b) Localice la varianza.
c) Ubique la desviación estándar.
d) ¿Cuál es el rango intercuartil?

29. Una gran lechería vigila continuamente el nivel de contenido graso en su producto; el porcentaje de grasa no debe desviarse mucho del 2% de la leche, siendo aceptable una desviación estándar del 10%; se obtuvo una muestra de 20 cartones de leche y se registró el porcentaje de grasa en cada uno. Los resultados se anotan a continuación.

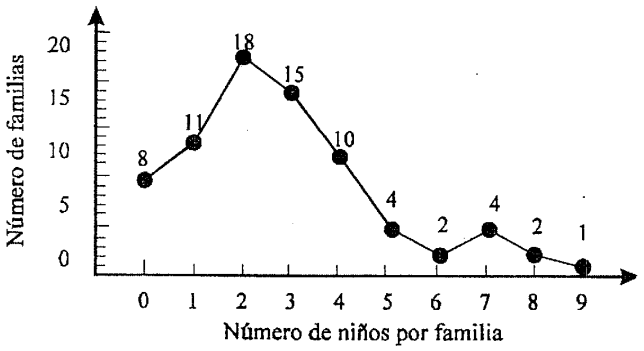
1.85	2.25	2.01	1.90	1.97
1.80	2.05	2.23	1.65	1.86
2.02	2.09	2.04	2.07	2.14
1.93	2.08	2.17	1.91	1.93

Calcule la media y la desviación estándar para la muestra de contenidos de grasa. ¿Hay evidencia de que el contenido de grasa es demasiado alto? Explique.

Un paso más allá

30. ¿Qué efecto tiene el tamaño de la muestra en la desviación estándar y en la varianza?

31. Encuentre $\bar{x}$ y s para el número de niños por familia de la muestra de datos ilustrada en la gráfica lineal adjunta.



32. En algunas situaciones, los datos son dicotómicos, consistentes sólo de dos valores distintos. Por ejemplo, datos dicotómicos son los obtenidos cuando las respuestas se registran como hombre-mujer, verdadero-falso, arriba-abajo, encendido-apagado, etc.; en tales casos, se acostumbra usar 0 para representar un valor de 1 para representar al otro. Si 1, 0, 0, 0, 1, 1, 1, 1, 1 y 0 representan una población de valores, encuentre μ y σ para la población de ceros y unos. Si p representa la proporción de unos demuestre que $\mu = p$ y $\sigma = \sqrt{p(1-p)}$.

33. Para cualquier colección finita de datos, determine el valor de c que hace $\Sigma(x - c)^2$ tan pequeña como sea posible.

34. Considere los tres conjuntos siguientes de datos.

A:	20	30	40	50	60
B:	–20	–10	0	10	20
C:	–2	–1	0	1	2

a) Encuentre SS para cada conjunto de datos; note que los valores del conjunto B se obtuvieron añadiendo –40 a cada medida en el conjunto A, y que los valores en el conjunto C se obtuvieron dividiendo cada medida del conjunto B entre 10.
b) ¿Qué relación existe entre SS_A y SS_B? ¿Y entre SS_A y SS_C?
c) ¿Qué relación hay entre s_A^2 y s_B^2? ¿Y entre s_A^2 y s_C^2?
d) ¿Qué relación se da entre s_A y s_B? ¿Y entre s_A y s_C?

35. Si se suma 3 a cada medida en un conjunto de diez que tienen una desviación estándar de 3, ¿cuál es la desviación estándar del nuevo conjunto de datos?

36. La calificación promedio en un examen de estadística fue 75 y la desviación estándar fue 10; después de devolver el examen a los estudiantes, el profesor determinó que una pregunta había sido mal calificada y que cada calificación debía aumentar en 5 puntos: Encuentre la media, la varianza y la desviación estándar para las calificaciones corregidas.

37. Considere la población de medidas X: 1.233, 1.236, 1.230, 1.236, 1.234, 1.237, 1.233, 1.235, 1.238 y 1.238. Suponga que cada medida se transforma usando $Y = 1000X - 1230$, y encuentre la media, la varianza y la desviación estándar de las medidas Y. Además, demuestre que:
a) $\mu_y = 1000\mu_x - 1230$. Como consecuencia, $\mu_x = (0.001)(\mu_y + 1230)$.
b) $\sigma_y^2 = (1000)^2\sigma_x^2$. Así, $\sigma_y^2 = (0.0000001)\sigma_y^2$.
c) $\sigma_y = (1000)\sigma_x$. Por lo que $\sigma_x = (0.001)\sigma_y$.

38. Si a cada medida de un conjunto de datos se les suma una constante C, demuestre que la varianza del nuevo conjunto es la misma que la del conjunto original.

39. Si cada medida de un conjunto de datos se multiplica por una constante C, demuestre que la suma de cuadrados del nuevo conjunto es igual a C^2 veces la suma de cuadrados del conjunto original.

40. Si cada medida de un conjunto de datos se multiplica por una constante C, ¿es igual la desviación estándar

del nuevo conjunto a C veces la desviación estándar del conjunto original?

41. Otra medida de dispersión es la *desviación absoluta promedio* (MAD). Se define por :

$$\text{MAD} = \frac{\Sigma |x - \bar{x}|}{n}$$

Calcule el valor de MAD para los datos del ejercicio 1.7

42. El coeficiente de variación proporciona una medida de variabilidad que es independiente de la unidad de medida; por ello, puede usarse para comparar la variabilidad de dos grupos de datos expresados en dos distintas unidades de medida. Por ejemplo, puede usarse para comparar la desviación estándar de la distribución de los ingresos anuales, y la desviación estándar de los años de servicio de todos los empleados de una compañía. El *coeficiente de variación* (CV) expresa la desviación estándar como un porcentaje de la media y se define como CV = $(s/\bar{x})(100)$. Suponga que un analista financiero de una firma de corredores de acciones quiere comparar la variación en las razones de precio-ganancia para un grupo de acciones comunes, con la variación en el rendimiento neto sobre la inversión; para las razones de precio-ganancia, la media es 9.8 y la desviación estándar 2.4, la media del rendimiento neto sobre la inversión es 20% y la desviación estándar es 4.3%. Use el coeficiente de variación para comparar la variación relativa de las raciones precio-ganancia respecto al rendimiento sobre la inversión.

43. Suponga que la planta de directores de una gran corporación, quiere comparar la dispersión de los ingresos de sus ejecutivos principales, contra la dispersión de los ingresos de sus empleados no especializados: para una muestra de los ejecutivos, el salario medio es 400,000 dólares y la desviación estándar es 50,000 dólares, mientras que para la muestra de empleados no especializados la media es 11,000 dólares y la desviación estándar es 1200. ¿En cuál grupo es mayor la dispersión relativa?

44. ¿Puede usarse el coeficiente de variación con datos que dan lugar a números negativos? Explique su respuesta.

45. El grado del sesgo de una distribución se mide generalmente por el *coeficiente de sesgo de Pearson*, denotado por CS. Para una muestra, se define por:

$$\text{CS} = \frac{3(\bar{x} - \tilde{x})}{s}$$

Para una distribución sesgada, el signo de CS corresponderá a la dirección del sesgo; una distribución simétrica tendrá CS = 0. Los datos siguientes representan los salarios iniciales, en miles de dólares, de una muestra de graduados de una gran universidad en el medio oeste: 29.2, 27.8, 29.0, 20.3, 16.9, 28.7, 19.6, 24.8, 17.4, 24.4, 20.8, 17.8, 16.2 y 17.8. Calcule su coeficiente de sesgo.

46. Encuentre un valor para la constante C que minimice $\Sigma |x - C|$ para la muestra siguiente de medidas: 2, 3, 7, 7 y 8.

47. Demuestre que $\Sigma (x - \bar{x})^2 = \Sigma x^2 - (\Sigma x)^2 / n$.

48. Si todas las medidas de una población distan menos de una desviación estándar de la media, caracterice la población; es decir, determine qué clase de números conforman la población.

49. Considere la muestra de medidas: 1.2, 2, 3, 4 y 4.9. Dé otra muestra de medidas que tenga una:
    a) media tres unidades mayor que la de la muestra original.
    b) varianza cuatro veces más grande que la original.
    c) media tres unidades mayor y una varianza cuatro veces más grande que la de la muestra original.

50. Dada una población, ¿puede ocurrir que la desviación estándar sea mayor que la mitad del rango? Explique.

51. Demuestre que para una muestra de dos medidas, $s = R/\sqrt{2}$.

52. Si s es la desviación estándar de una muestra, se puede demostrar que:

$$\frac{R}{2(n-1)} \leq s \leq \left(\frac{R}{2}\right)\sqrt{\frac{n}{n-1}}$$

donde n es el tamaño de la muestra y R es el rango. Los datos siguientes representan los niveles de colesterol en la sangre para una muestra de ocho personas: 239, 218, 227, 357, 161, 286, 310 y 245.
    a) Encuentre cotas superiores e inferiores para s.
    b) Estime s usando el punto medio del intervalo determinado por el resultado anterior.
    c) Calcule el valor de s y compare el resultado con el valor estimado en el inciso b.

SECCIÓN 3.3 *Tendencia central y dispersión para datos contenidos en tablas de frecuencia agrupada*

Es posible calcular las medidas de tendencia central y dispersión para datos exhibidos en una tabla de frecuencia agrupada, pero sus valores no son exactos sino únicamente aproximados; eso se debe al desconocimiento de las medidas en grupo, las cuales se han colocado en intervalos de clase. Antes de que las computadoras se volvieran de uso común, era necesario un gran trabajo para calcular las medidas de tendencia central y de dispersión para conjuntos grandes de datos; en un intento de manejar ese problema y de eliminar parte de los cálculos, los datos eran colocados en tablas de frecuencia agrupada y se debían hacer ciertas hipótesis antes de realizar los cálculos; la validez de estas hipótesis tenía un efecto directo en la precisión de los resultados.

Hoy en día, las computadoras de alta velocidad hacen posible procesar rápidamente listas enormes de datos proporcionando resultados altamente precisos, lo cual elimina las ventajas en los cálculos con tablas de frecuencia. Usted se preguntará entonces por qué nos interesa calcular valores aproximados de ciertos estadísticos a partir de tablas de frecuencia agrupada; existe una gran cantidad de datos resumidos en tablas de frecuencia agrupada construidas por otros y la única forma de calcular sus medidas de tendencia central es usar los datos agrupados.

Media para datos agrupados

Si debemos encontrar la media para datos proporcionados en tablas de frecuencia agrupada, usamos marcas de clase para representar las medidas para cada clase. Entonces la fórmula (3.2) se puede usar para determinar la **media muestral aproximada** $\bar{x}_a$, puesto que los datos originales se desconocen y cada observación está representada por su marca de clase.

APLICACIÓN 3.18

Los datos siguientes representan el número de discos vendidos cada día durante un periodo de 25 días en una tienda de música localizada en un centro comercial:

60	36	61	56	19	35	51	42	21	28	33	67	30
49	57	54	59	28	63	38	15	24	35	46	53	

Por conveniencia, los datos han sido exhibidos en la siguiente tabla de frecuencia agrupada:

Número de discos vendidos	Número de días
15-25	4
26-36	7
37-47	3
48-58	6
59-69	5

Encuentre:

a) $\bar{x}$, el número promedio de discos vendidos por día.

b) $\bar{x}_a$, el número promedio aproximado de discos vendidos por día.

Solución:

a) Con la ayuda de una calculadora manual, determinamos que la suma de las 25 medidas es $\Sigma x = 1060$. En consecuencia, la media muestral es:

$$\bar{x} = \frac{\Sigma x}{n} = \frac{1060}{25} = 42.4$$

Así, el número promedio de discos vendidos por día es 42.40.

b) Encontramos primero las marcas de clase X. Recuerde del capítulo 2 que una marca de clase es el punto medio de un intervalo de clase. Cada marca de clase se multiplica entonces por su frecuencia correspondiente, como lo muestra la tabla 3.8.

TABLA 3.8
Marcas de clase multiplicadas por las frecuencias para la aplicación 3.18

Clase	f	X	fX
15-25	4	20	80
26-36	7	31	217
37-47	3	42	126
48-58	6	53	318
59-69	5	64	320

Usando la fórmula (3.2), la media aproximada es:

$$\bar{x}_a = \frac{\Sigma (fX)}{\Sigma f} = \frac{1061}{25} = 42.44$$

Note que $\bar{x}_a = 42.44$ es sólo un valor aproximado para la media de las 25 medidas muestrales originales; la aproximación se considera buena comparada con el valor exacto $\bar{x} = 42.40$, obtenido en la parte a. ■

Mediana para datos agrupados

Hay dos métodos generales para calcular la mediana de datos previamente agrupados en clases; esos métodos difieren en la hipótesis relativa a la manera de agrupar los datos en clases.

Método I. Cualquier valor de la clase coincide con la marca de clase.
Método II. Los valores en cada clase se distribuyen uniformemente en la clase.

Esta hipótesis permite que la mediana tenga la propiedad especial siguiente para un histograma de frecuencias:

> Si se dibuja una recta vertical perpendicular al eje horizontal del histograma en el valor correspondiente a la mediana, entonces el área del histograma ubicada a la izquierda de la recta vertical, es igual al área del histograma ubicada a su derecha.

Considere la aplicación 3.19 y note que los valores aproximados de la mediana producidos por los dos métodos no coinciden; el método II es el usado típicamente para aproximar la mediana de datos agrupados en clases, debido a que las áreas del histograma, antes y después de la mediana, están igualmente distribuidas.

APLICACIÓN 3.19

La tabla 3.9 representa las velocidades, en millas por hora, para una muestra de 37 coches que recorren una zona escolar donde se permite circular hasta a 25 millas por hora. Encuentre la mediana aproximada de la velocidad.

TABLA 3.9
Datos para la aplicación 3.19

Velocidad	Número de coches	f acumulada
1–5	3	3
6–10	2	5
11–15	5	10
16–20	10	20
21–25	7	27
26–30	10	37

Solución:

Método I. Las marcas de clase, denotadas por X, están contenidas en la tabla siguiente. La marca de clase para la primera clase es $(1 + 5)/2 = 3$, y las otras marcas de clase se encuentran sumando 5, el ancho de clase, a la primera marca:

Velocidad	Número de coches	X	Acumulada f
1–5	3	3	3
6–10	2	8	5
11–15	5	13	10
16–20	10	18	20
21–25	7	23	27
26–30	10	28	37
	37		

Desde este punto, podemos determinar la mediana siguiendo la regla dada en la sección 3.1; como hay un número impar de medidas, la **mediana muestral aproximada** $\tilde{x}_a$ es la medida, marca de clase, que ocupa la 19ª posición en la tabla anterior. Así, la mediana aproximada es $\tilde{x}_a = 18$.

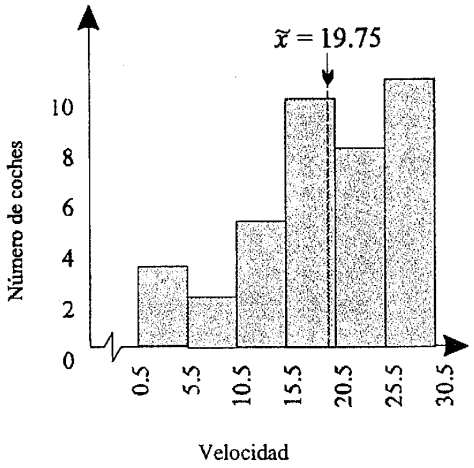
Método II. Como $n = 37$, queremos localizar el $n/2 = 37/2 = 18.5$-ésimo valor. Al observar la tabla notamos que tal valor cae en la clase 16-20, porque las tres primeras clases contienen un total de 10 valores y la cuarta 10 valores; por lo tanto, debemos contar $(18.5 - 10) = 8.5$ valores en la clase 16-20, bajo la hipótesis de que los 10 valores que caen en esta clase están distribuidos homogéneamente a lo largo de ella; en otras palabras, estamos buscando la medida en la clase 16-20 localizada en los 8.5/10 de la clase. Como el ancho de cada clase es $w = 5$, para encontrar el valor aproximado de la mediana $\tilde{x}_a$ sólo necesitamos sumar (8.5)/10 del ancho $w = 5$ a la frontera inferior de la cuarta clase. Así, el valor aproximado de la mediana es:

$$\tilde{x}_a = 15.5 + \frac{8.5}{10}(5)$$
$$= 15.5 + 4.25 = 19.75 \quad \blacksquare$$

EJEMPLO 3.31

Un histograma para los datos en la aplicación 3.19 está dado en la figura 3.9. Podemos verificar fácilmente que la suma de las áreas de los rectángulos anteriores al valor 19.75 es igual a la suma de las áreas de los rectángulos posteriores a 19.75.

FIGURA 3.9
Histograma para los datos de la aplicación 3.19



En general, si $\mathscr{L}$ es la frontera inferior de la clase en la cual cae la mediana, f es la frecuencia de la clase que contiene a la mediana, g es el número de valores que se deben contar para llegar a $\mathscr{L}$, contando desde el valor menor, y w es el ancho de clase, entonces, usando el método II, la mediana aproximada para los datos está dada por:

$$\tilde{x}_a = \mathscr{L} + \left(\frac{g}{f}\right)(w)$$

Para la aplicación 3.19, $\mathscr{L} = 15.5$, $g = 8.5$, $f = 10$ y $w = 5$. La sustitución de estos valores en la expresión anterior da:

$$15.5 + \left(\frac{8.5}{10}\right)(5) = 19.75$$

el mismo valor que obtuvimos en la aplicación.

Moda para datos agrupados

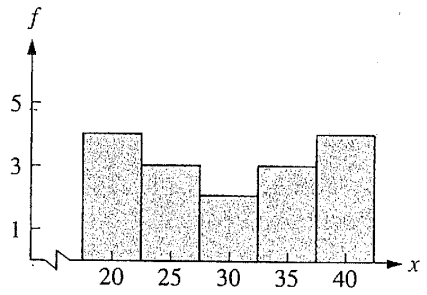
Una desventaja de usar la moda con una distribución de frecuencia agrupada es que el valor de la moda a menudo depende del agrupamiento arbitrario de los datos; por esta razón es que una moda para una distribución de frecuencia agrupada suele denominarse una **moda cruda** o **clase modal.**

Si los datos se organizan en una clase de frecuencia agrupada, una moda cruda o clase modal, si existe, puede identificarse fácilmente; corresponde a la marca de clase para una clase que contenga la frecuencia mayor y para datos desplegados en un histograma, una moda se asocia con la barra más alta.

EJEMPLO 3.32

Para el histograma ilustrado en la figura 3.10, se ve que las modas crudas son 20 y 40.

FIGURA 3.10
Histograma con dos modas



Rango promedio para datos agrupados

Para datos organizados en una tabla de frecuencia agrupada, el rango promedio es aproximadamente el promedio de la frontera inferior de clase de la primera clase y la frontera superior de clase de la última clase.

EJEMPLO 3.33

El rango promedio aproximado para los datos de la aplicación 3.19 es:

$$\frac{0.5 + 30.5}{2} = 15.5$$

Puntos de posición para datos de una tabla de frecuencia agrupada

El método II para encontrar el valor aproximado de la mediana para datos en una tabla de frecuencia agrupada, puede usarse también para encontrar puntos percentiles en una tabla del mismo tipo.

APLICACIÓN 3.20

Para los datos exhibidos en la tabla de frecuencia agrupada adjunta, encuentre P_{60}, el sexagésimo percentil, o 6° decil.

Velocidad	Número de coches	f Acumulada
1–5	3	3
6–10	2	5
11–15	5	10
16–20	8	18
21–25	7	25
26–30	10	35

Solución: Usamos la primera columna de la tabla para contar 60% de los datos, es decir, (0.60)(35) = 21 valores. Así, P_{60} debe caer en la clase que contenga la medida; esta clase es 21-25. El valor P_{60} se localiza dentro del intervalo 21-25 a una distancia de 2.14 de la frontera izquierda del intervalo. La distancia de 2.14 se obtuvo multiplicando (21-18)/7 por 5, el ancho de la clase. Así, el sexagésimo percentil es:

$$P_{60} = 20.5 + 2.14 = 22.64$$

Sesenta por ciento de los datos están por debajo del valor 22.64. ■

Varianza y desviación estándar

Las marcas de clase se usan típicamente para representar medidas que caen en las clases de una tabla de frecuencia agrupada cuando se necesita obtener la varianza o la desviación estándar aproximadas de los datos; al hacerse esto, se usan las fórmulas de la sección 3.2 para calcular la varianza y la desviación estándar, para el caso de distribuciones de frecuencia no agrupada.

GRUPO DE EJERCICIOS 3.3

Más aplicaciones

1. La tabla de frecuencias agrupadas exhibe las edades de una muestra de 36 personas asistentes a una película para adultos.

Clase	f
8–13	2
14–19	7
20–25	13
26–31	5
32–37	9

   a) Encuentre la edad media aproximada.
   b) Aproxime la mediana de las edades usando los métodos I y II.
   c) Encuentre P_{40}, el cuadragésimo percentil, y P_{65}, el percentil 65.
   d) Localice Q_3, el tercer cuartil, y D_3 el tercer decil.
   e) Determine el sesgo del histograma de frecuencia.
   f) Ubique la varianza aproximada.
   g) Estime la desviación estándar

2. La tabla ilustrada aquí da la distribución de la precipitación pluvial en un cierto condado de Maryland para el mes de junio durante los últimos 29 años.

Precipitación pluvial en pulgadas	Número de años
2.0-2.5	3
2.6-3.1	5
3.2-3.7	6
3.8-4.3	8
4.4-4.9	7

   a) Encuentre la media aproximada de precipitación pluvial.
   b) Aproxime la mediana usando el método II.
   c) Localice P_{40}, el cuadragésimo percentil, y P_{75}, el percentil 75.
   d) Ubique Q_1 el primer cuartil, y D_4, el cuarto decil.
   e) Determine el sesgo del histograma de frecuencias.
   f) ¿Cuál es la varianza aproximada?
   g) Estime la desviación estándar aproximada.

3. La tabla de frecuencia agrupada adjunta indica las edades de compradores de coches nuevos en una gran distribuidora. Encuentre:
   a) La edad media aproximada.
   b) La mediana aproximada de las edades usando los métodos I y II.
   c) La varianza muestral aproximada.
   d) La desviación estándar muestral aproximada.

e) P_{40} y P_{69}.
f) Q_1 y D_7.

Clase de edades	f
28-32	20
33-37	23
38-42	71
43-47	45
48-52	26

4. Los datos adjuntos indican los totales quincenales, en dólares, invertidos por una muestra de 50 empleados en un plan de beneficencia compartida:

Monto de la inversión	Número de empleados
30-34	5
35-39	11
40-44	14
45-49	8
50-54	5
55-59	7

Encuentre:
a) la media aproximada.
b) la mediana aproximada usando el método II.
c) la varianza aproximada.
d) la desviación estándar aproximada.
e) P_{60}, el percentil 60, y P_{65}, el sexagésimo quinto percentil.
f) Q_3, el tercer cuartil, y D_8, el octavo decil.

5. La tabla siguiente contiene los costos de reparación de un automóvil para los reclamos de categoría menor presentados ante una compañía de seguros:

Costo de la reparación	Frecuencia
0-99	12
100-199	35
200-299	75
300-399	84
400-499	125

Encuentre:
a) la media aproximada.
b) la mediana aproximada usando el método II.
c) la varianza aproximada.
d) la desviación estándar aproximada.
e) P_{20} el vigésimo percentil, y P_{35}, el trigésimo quinto percentil.
f) Q_3, el tercer cuartil, y D_9, el noveno decil.

6. La tabla adjunta contiene una distribución de frecuencia agrupada para la duración de 50 llamadas telefónicas de larga distancia, redondeadas al minuto más cercano, hechas por una agencia. Calcule la varianza aproximada y la desviación estándar aproximada para esta distribución.

Duración de la llamada	f
4-7	23
8-11	9
12-15	11
16-19	4
20-23	2
24-27	1

Un paso más allá

7. Para los datos del ejercicio 1, ¿cuál es el percentil correspondiente a una edad de 18 años? Este porcentaje se denomina usualmente el rango percentil de 18.

SECCIÓN 3.4 Puntajes estándar y observaciones aberrantes

Puntajes estándar como medidas de posición relativa

Suponga que después de hacer un examen de estadística usted obtiene su calificación; entonces, se interesa por saber cómo es su calificación respecto a la de los demás que hicieron el mismo examen, para saber si su calificación está por debajo o encima de la media y por cuánto. Un **puntaje estándar** le dará información sobre qué tan bien hizo el examen respecto al resto del grupo y le proporcionará una medida de su posición relativa dentro del mismo.

Roberto obtuvo 700 en la parte de matemáticas del SAT y Jaime 24 en habilidad matemática del examen de colocación en la universidad (*CPT* por

sus siglas en inglés). La media y la desviación estándar del SAT son 500 y 100, y del CPT 18 y 6, respectivamente. Si se supone que ambos exámenes miden algún tipo de habilidad, ¿cuál persona calificó más alto?; para responder esta pregunta necesitamos algún método que nos permita comparar puntajes de distribuciones distintas. Es claro que la desviación de cada puntaje respecto a su media no es una base de comparación correcta en este caso, pues la desviación de la calificación de Jaime es:

$$x - \bar{x} = 24 - 18 = 6$$

y la de la calificación de Roberto es:

$$x - \bar{x} = 700 - 500 = 200$$

Ninguna de ellas toma en cuenta la dispersión de los puntajes.

Si usamos puntajes estándar veremos que Roberto calificó más alto que Jaime en la habilidad medida por el examen. Un puntaje estándar toma en cuenta la variabilidad de las medidas respecto a su media.

Una medida que nos permite hacer comparaciones entre distribuciones distintas y toma en cuenta la dispersión de los puntajes es el puntaje estándar. Un puntaje estándar se define como:

$$\text{puntaje estándar} = \frac{\text{desviación del valor}}{\text{desviación estándar}}$$

y se denota comúnmente por z. Esta relación puede expresarse como:

Puntajes estándar

$$z = \frac{x - \mu}{\sigma} \qquad z = \frac{x - \bar{x}}{s}$$

Población Muestra

(3.11)

dependiendo de si lo que interesa es una población o una muestra.

> Puesto que un puntaje estándar se define como la razón de la desviación del valor entre la desviación estándar, representa el número de desviaciones estándar que un valor dista de la media.

Un puntaje estándar se denomina en ocasiones puntaje z. En relación con el ejemplo anterior, el puntaje estándar o **puntaje** z de Jaime es:

$$z = \frac{x - \mu}{\sigma}$$

$$= \frac{24 - 18}{6} = 1$$

y de Roberto es:

$$z = \frac{x - \mu}{\sigma}$$
$$= \frac{700 - 500}{100} = 2$$

La calificación de Jaime de 24 está una desviación estándar arriba de la media del examen CPT, y la calificación de Roberto de 700 está dos desviaciones estándar arriba de la media del SAT; como ambos puntajes z son positivos y el puntaje z de Roberto es superior al de Jaime, Roberto calificó más alto que Jaime en la habilidad medida por el examen.

APLICACIÓN 3.21

Suponga que un conjunto de puntajes tiene una media de 10 y una desviación estándar de 2.

a) Escriba los valores faltantes de la tabla siguiente.

x	4	6	8	10	12	14	16
z							

b) ¿Qué significa un puntaje z de 0 respecto al puntaje original?

c) ¿Qué indica un puntaje z positivo respecto al puntaje original?

d) ¿Qué quiere decir un puntaje z negativo respecto al puntaje original?

e) Además de indicar que un puntaje está arriba o debajo de la media, ¿qué información adicional proporciona un puntaje z?

Solución:

a) De la fórmula (3.11) obtenemos los siguientes puntajes z:

x	4	6	8	10	12	14	16
z	−3	−2	−1	0	1	2	3

b) Un puntaje z de 0 indica que el puntaje es la media.

c) Un puntaje z positivo quiere decir que el puntaje original está arriba de la media.

d) Un puntaje z negativo significa que el puntaje original está debajo de la media.

e) Un puntaje z también dice el número de desviaciones estándar que un puntaje dista de la media.

APLICACIÓN 3.22

Si una distribución de números obtenida de medir pesos de niños pequeños tiene una media de 20 libras y una desviación estándar de 2, ¿cuál es la unidad asociada con cada puntaje z?

Solución: Si x denota el peso de un niño en libras, entonces x libras menos 20 libras es $(x - 20)$ libras. Al dividir $x - 20$ libras entre 2 libras se obtiene un cociente de $(x - 20)/2$. En consecuencia, observamos que un puntaje z no tiene unidad de medida, es sólo un número. ■

APLICACIÓN 3.23

Repetimos aquí los datos de la aplicación 3.11 relativos a los precios del asado de cerdo y del queso cheddar.

Ciudad capital	Asado de cerdo (sin hueso)	Queso cheddar
Berna	$6.61	$4.00
Bonn	2.38	2.74
Brasilia	1.27	1.08
Buenos Aires	1.36	2.03
Camberra	2.06	2.60
Londres	1.56	1.81
Madrid	2.33	3.15
México	1.08	2.29
Ottawa	1.99	3.98
París	2.47	2.37
Pretoria	1.95	1.76
Roma	2.46	2.96
Estocolmo	5.35	2.54
Tokio	4.19	2.38
Washington	3.29	2.69

Use puntajes z para determinar cuál alimento tiene el precio relativo más alto en Washington con respecto a los precios en las otras capitales.

Solución: Se puede demostrar que $\bar{x}_p = 2.69$ dólares y $\bar{x}_c = 2.56$ dólares. Demostramos antes que $s_c = 0.77$ dólares y puede comprobarse fácilmente que $s_p = 1.57$ dólares. Como el asado de cerdo cuesta 3.29 dólares en Washington, su puntaje z_p es:

$$z_p = \frac{x - \bar{x}}{s}$$
$$= \frac{3.29 - 2.69}{1.57} = 0.38$$

El queso cheddar cuesta 2.69 dólares en Washington. Su puntaje z_c es:

$$z_c = \frac{x - \bar{x}}{s}$$
$$= \frac{2.69 - 2.56}{0.77} = 0.17$$

Así, el precio del asado es relativamente más alto en Washington que el del queso. ■

Suponga que μ y σ son la media y la desviación estándar, respectivamente, de una población finita; cada medida x tiene un puntaje z asociado. Los factores importantes siguientes, que se explican en la aplicación 3.24, ayudan a caracterizar la colección de puntajes estándar de una población:

La población de todos los puntajes estándar tiene una media de 0 y una desviación estándar de 1.

APLICACIÓN 3.24

a) Encuentre μ y σ para la población consistente en los valores 1, 2 y 3.
b) Localice los tres puntajes estándar.
c) Demuestre que la media de los puntajes estándar es 0 y que la desviación estándar es 1.

Solución:

a) La media poblacional es:

$$\mu_x = \frac{1 + 2 + 3}{3} = 2$$

Usamos la fórmula (3.6) para obtaner la varianza poblacional:

$$\sigma_x^2 = \frac{SS}{N} = \Sigma \frac{(x - \mu)^2}{N}$$
$$= \frac{(1 - 2)^2 + (2 - 2)^2 + (3 - 2)^2}{3} = \frac{2}{3}$$

Por tanto, la desviación estándar será:

$$\sigma_x = \sqrt{\frac{2}{3}} = 0.816$$

b) Encontramos los puntajes z usando la fórmula 3.11:
Para $x = 1$,

$$z = \frac{1 - 2}{0.816} = -1.225$$

Para $x = 2$,

$$z = \frac{2 - 2}{0.816} = 0$$

Para $x = 3$,

$$z = \frac{3 - 2}{0.816} = 1.225$$

c) La media de los puntajes z es cero. Para encontrar SS para los puntajes z, organizamos nuestros cálculos en la tabla siguiente y entonces usamos la fórmula 3.5.

z	z^2
−1.225	1.50
0	0
1.225	1.50
0	3

Dicha fórmula da:

$$SS = \Sigma z^2 - \frac{(\Sigma z)^2}{N}$$
$$= 3 - 0 = 3$$

Usando la fórmula (3.6) obtenemos:

$$\sigma_z^2 = \frac{SS}{N}$$
$$= \frac{3}{3} = 1$$

En consecuencia, la desviación estándar de los puntajes z es:

$$\sigma_z = \sqrt{\text{varianza}}$$
$$= \sqrt{1} = 1 \quad \blacksquare$$

MINITAB puede usarse para demostrar que la media y la desviación estándar de los puntajes z, en la aplicación 3.24, son 0 y 1, respectivamente. La pantalla 3.5 contiene las órdenes necesarias y las respuestas correspondientes.

Pantalla 3.5

```
MTB> SET C1
DATA> 1 2 3
DATA> END
MTB> LET C2 = (C1 - MEAN(C1))/STDEV(C1)
MTB> LET K1 = MEAN(C2)
MTB> LET K2 = STDEV (C2)
MTB> PRINT C1 C2

ROW C1 C2

  1   1  -1
  2   2   0
  3   3   1

MTB> PRINT K1 K2
K1       0
K2       1.00000
MTB>
```

La columna C2 en la pantalla 3.5 contiene los puntajes z de los datos en la columna C1. Las constantes K1 y K2 contienen la media y la desviación estándar, respectivamente, de los puntajes z en C2; la orden PRINT C1 C2 exhibe los valores en C1 y C2, y la orden PRINT K1 K2 exhibe la media y la desviación estándar, respectivamente, de los puntajes z. Note que los resultados concuerdan con los obtenidos en la aplicación 3.24.

Conversión de puntajes z a puntajes x

APLICACIÓN 3.25

Para algunas aplicaciones, es interesante revertir los puntajes z a sus **puntajes originales**. Por ejemplo, si $\bar{x} = 10$ y $s = 2$, encuentre el puntaje x correspondiente al z de $z = 16$.

Solución: Usaremos la fórmula del puntaje z y despejaremos x.

$$z = \frac{x - \bar{x}}{s}$$

$$16 = \frac{x - 10}{2}$$

Multiplicando ambos lados por 2, tenemos:

$$32 = x - 10$$

Si sumamos 10 a ambos miembros, resulta:

$$x = 42. \quad \blacksquare$$

Cuando de la fórmula del puntaje z se despeja x, obtenemos la fórmula (3.12), que puede usarse para encontrar el puntaje original x dado por un puntaje estándar z (véase la aplicación 3.26).

De puntajes z a puntajes originales

$$x = \mu + \sigma z \tag{3.12}$$

APLICACIÓN 3.26

Si una población tiene una media de 70 y una desviación estándar de 5, encuentre el puntaje original correspondiente al puntaje z de 1.5.

Solución: Por medio de la fórmula 3.12 obtenemos:

$$x = \mu + \sigma z$$
$$= 70 + (5)(1.5)$$
$$= 70 + 7.5 = 77.5 \quad \blacksquare$$

Gráficas de caja

Una **gráfica de caja** es un diagrama que proporciona información sobre el centro, la dispersión y la simetría o sesgo; utiliza cuartiles, y así, es resistente a las observaciones aberrantes; en ocasiones, a las gráficas de caja se les denomina **diagramas de caja y extensión**. Para construir una gráfica de caja se ejecutan los pasos siguientes:

Pasos para construir una gráfica de caja

1. Construya una recta numérica y marque en ella los tres cuartiles.
2. Dibuje una caja rectangular sobre la recta con los extremos localizados en el primer y tercer cuartiles; la altura de la caja no es importante.
3. Trace un segmento de recta vertical por el punto correspondiente a la mediana dentro de la caja.
4. Dibuje dos rectas horizontales, llamadas extensiones, una desde la mediana y la medida de la extrema izquierda y otra de la mediana a la medida del extremo derecho.

EJEMPLO 3.34

Usemos los datos siguientes, correspondientes a lecturas de ozono en partes por millón (ppm), tomadas al mediodía en una gran ciudad, para construir una gráfica de caja:

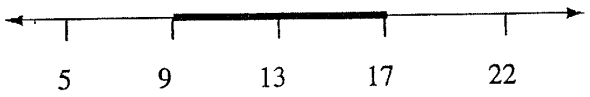
9 14 12 17 11 20 13 18 22 12 15 16 5 7 9 19 8

Primero arreglamos los datos en orden numérico, de menor a mayor:

5 7 8 9 9 11 12 12 13 14 15 16 17 18 19 20 22

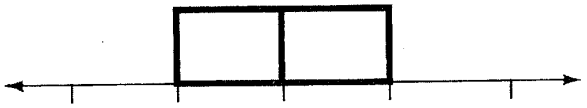
La mediana es $Q_2 = 13$, el cuartil inferior es $Q_1 = 9$ y el cuartil superior es $Q_3 = 17$.

Paso 1.

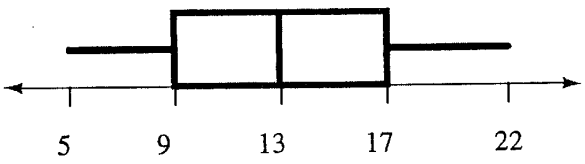


Paso 2.

Paso 3.



Paso 4.

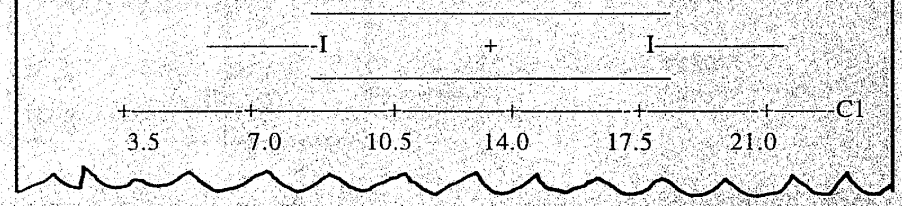


Como la mediana está un poco a la izquierda de la mitad de la caja y la extensión más larga está a la derecha, la distribución está sesgada a la derecha.

La pantalla 3.6 ilustra el uso de MINITAB para construir una gráfica de caja para los datos del ozono.

Pantalla 3.6

```
MTB> SET C1
DATA> 9 14 12 17 11 20 13 18 22 12 15 16 5 7 9 19 8
DATA> END
MTB> BOXPLOT C1
```



Los tres cuartiles, así como otros estadísticos descriptivos, se pueden encontrar usando la orden DESCRIBE C1 en la pantalla 3.7

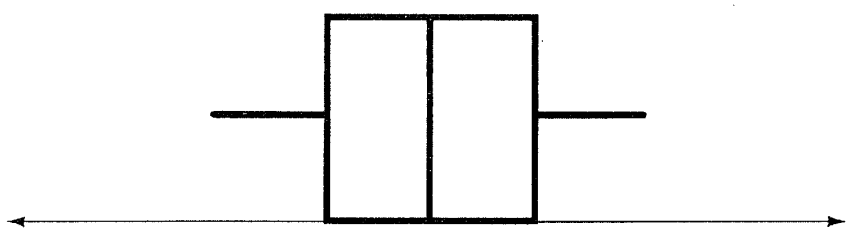
Pantalla 3.7

```
MTB> DESCRIBE C1

                N     MEAN   MEDIAN  TRMEAN   STDEV  SEMEAN
C1             17    13.35    13.00   13.33    4.89    1.18

               MÍN      MÁX       Q1       Q3
C1            5.00    22.00     9.00    17.50
```

EJEMPLO 3.35

La figura 3.11 muestra una gráfica de caja para un conjunto de datos que es simétrico; la recta mediana está exactamente en el centro de la caja y las dos extensiones son de la misma longitud. En la práctica, no esperaríamos una muestra de datos perfectamente simétrica, donde el lugar ocupado por la mediana es un buen indicador de la simetría; las longitudes de las extensiones dependen de los valores aislados y, por lo tanto, no son tan confiables como pronosticadores de simetría en la población como lo es la ubicación de la mediana dentro de la caja.

FIGURA 3.11
Gráficas de caja simétrica

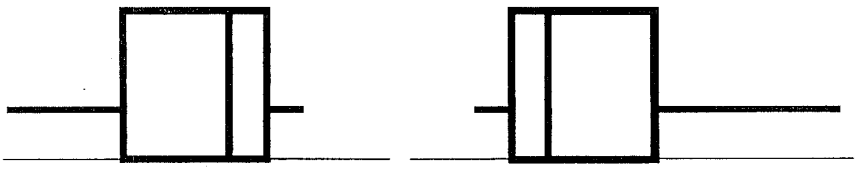


EJEMPLO 3.36

La figura 3.12 ilustra dos gráficas de caja para conjuntos sesgados; la gráfica de la izquierda está sesgada a la izquierda, y la de la derecha a la derecha. Note en cada caso la ubicación de la mediana; si la distribución está sesgada a la izquierda, la mediana queda a la derecha del centro de la caja, y si está sesgada a la derecha, la mediana estará a la izquierda. Igualmente, en la práctica cotidiana con datos reales,

las longitudes de las extensiones no son un buen indicador del sesgo en la población, porque dependen de valores aislados. Advierta que el ancho de la caja es el rango intercuartil y por ello, da una medida de la dispersión de los datos; si una extensión fuera especialmente larga, sería señal de que la medida extrema es una posible observación aberrante.

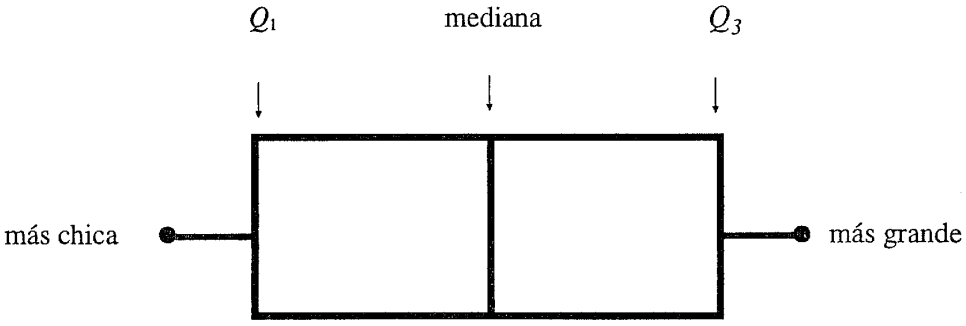
FIGURA 3.12
Gráfica de caja seguida



EJEMPLO 3.37

La figura 3.13 resume las características importantes de una gráfica de caja.

FIGURA 3.13
Rasgos importantes de una gráfica de caja



Detección de observaciones aberrantes

Una observación aberrante, como dijimos, es una medida extrema en un conjunto de datos; indica en ocasiones que se ha cometido un error, de anotación, por ejemplo, pero también puede representar una medida muy rara de la población. La investigación de observaciones aberrantes revela a menudo información útil y es bastante posible que una de ellas sea la "joya entre las piedras" en lugar de la "piedra entre las joyas". Estas observaciones pueden afectar tanto la media como la desviación estándar del conjunto de datos, distorsionando así el centro y la variabilidad; no hay consenso entre los investigadores sobre los que constituye una observación aberrante en un conjunto de datos. Una de las dos reglas prácticas siguientes son de uso típico para detectar observaciones aberrantes en un conjunto de datos.

Una medida es una observación aberrante de una muestra si se verifica una de estas reglas

Regla 1. El tamaño de la muestra es mayor de 10, la distribución de frecuencia tiene forma de campana y el puntaje z para la medida dista más de tres desviaciones estándar de la media.

Regla 2. La medida cae más de tres IQR debajo del cuartil menor, o más de tres IQR arriba del cuartil superior.

APLICACIÓN 3.27

Considere el siguiente conjunto ordenado (visto originalmente en el ejemplo 3.20), que representa las cantidades de oxígeno consumido (mL/kg · min), por 21 corredores hombres de mediana edad al pedalear en una bicicleta ergométrica de 100 watts:

12.81	14.95	15.83	15.97	19.90	18.27	18.34	19.82	19.94	20.62	36.73
20.88	20.93	20.98	20.99	21.15	22.16	22.24	23.16	23.56	35.78	

Determine si la medida 35.78 es una observación aberrante.

Solución: Usaremos las dos reglas prácticas.

1. Regla 1: un diagrama de tallo y hojas para los datos es como sigue:

12	81				
13					
14	95				
15	83	97			
16					
17	90				
18	27	34			
19	82	94			
20	62	88	93	98	99
21	15				
22	16	24			
23	16	56			
.					
.					
.					
35	78				
36	73				

Las medidas no siguen una forma de campana y por tanto la regla no debe usarse.

Pero, con el propósito de ilustrar, calculemos el puntaje z para la medida 35.78. La media es $\bar{x} = 443.01/21 = 21.06$, y la desviación estándar es $s = 5.75$. El puntaje z para 35.78 es:

$$z = \frac{35.78 - 21.06}{5.75} = 2.56$$

La medida 35.78 no es una observación aberrante, pues está sólo 2.56 desviaciones estándar arriba de la media. A causa de las medidas extremas, la media y la desviación estándar se han inflado y en consecuencia, el puntaje z ha sido reducido.

2. Regla 2: en el ejemplo 3.20 encontramos que IQR = 3.89, $Q_1 = 18.27$, $Q_3 = 22.16$. Como $22.16 + 3(3.89) = 33.83$ y $35.78 > 33.83$, podemos concluir que 35.78 es una observación aberrante. ■

GRUPO DE EJERCICIOS 3.4

Habilidades básicas

1. Si $\mu = 47$ y $\sigma = 15$, llene los valores faltantes en la tabla siguiente:

x	z
80	
	1.2
60	
	−2.37
47	
	3

2. Si $\mu = 35$ y $\sigma = 16$, llene los valores faltantes en la tabla siguiente:

x	z
50	
34	
	2.3
	−1.4
0	
	1

3. Considere la población: 4, 8, 12, 16 y 20.

   Encuentre:
   a) μ
   b) σ
   c) el puntaje z de cada uno de los puntaje en bruto
   c) la media y la desviación estándar de los puntos z en la parte c

4. Considere la siguiente muestra: 1, 2, 2, 6, 8 y 11 para localizar:

   a) $\bar{x}$.
   b) s.
   c) el puntaje z para cada medida.
   d) la media de los puntajes z.
   e) la desviación estándar de los puntajes z.

5. Construya una gráfica de caja con los datos siguientes:

   1.32 1.41 0.95 1.06 1.18
   1.26 0.99 1.26 1.10

6. Construya una gráfica de caja con estos datos:

   65 74 77 83 89 92 96 95 103 109

Más aplicaciónes

7. Susana obtuvo 625 puntos en el examen A en el cual $\mu = 600$ y $\sigma = 70$; María alcanzó 525 puntos en el examen B para el cual $\mu = 500$ y $\sigma = 25$. Si tanto Susana como María solicitan un trabajo y todos los otros factores son iguales, ¿a quién debe otorgársele el trabajo con base en los puntajes de los exámenes A, B? Use puntajes estándar para justificar su respuesta.

8. David y Ricardo están entrenando para el maratón de Boston. David está entrenando en un camino de Cumberland, mientras que Ricardo lo hace en uno de Frostburg; la media del tiempo para completar el recorrido del camino de Cumberland es 167.4 minutos y la desviación estándar es 25.9 minutos. La media del tiempo en el camino de Frostburg es 143.1 minutos y la desviación estándar de 20.7 minutos; David dice que su tiempo de recorrido del camino de Cumberland es 91.5 minutos y Ricardo declara que el suyo es 86.2 minutos. ¿Quién será el mejor en el maratón de Boston, según usted? Use puntajes estándar para justificar su respuesta.

9. Las medias y las desviaciones estándar de los puntajes de exámenes en cinco grupos se listan aquí; suponga que usted obtiene un puntaje de 75 en el examen. ¿En cuál grupo tendría la mejor posición relativa?
   a) $\mu = 65, \sigma = 10$
   b) $\mu = 70, \sigma = 5$
   c) $\mu = 55, \sigma = 15$
   d) $\mu = 75, \sigma = 2$
   e) $\mu = 70, \sigma = 3$

10. Las medias y las desviaciones estándar de los tiempos de carrera para cuatro carreras de distancia se anotan abajo suponga que usted obtiene un tiempo de 20 minutos en una carrera. ¿En cuál carrera tendría usted la mejor posición relativa?
    a) $\mu = 10, \sigma = 2$
    b) $\mu = 25, \sigma = 5$
    c) $\mu = 14, \sigma = 10$
    d) $\mu = 20, \sigma = 1$

11. Los trabajadores que utilizan la máquina A pueden producir cantidades diarias del producto C, con una media de 75 y una desviación estándar de 5, mientras que los trabajadores que utilizan la máquina B producen cantidades diarias del producto C, con una media de 80 y una desviación estándar de 8. Dick produjo 83 unidades con la máquina A y Juan 92 con la máquina

B. ¿Quién de ellos produjo la cantidad relativa mayor? ¿Por qué?

12. El salario medio anual de todos los programadores de cómputo hombres en una gran compañía es de 35,000 dólares, y la desviación estándar de 500 dólares. Una mujer programadora gana 20,000 dólares anuales y considera estar siendo discriminada. ¿Usted qué opina? ¿Por qué?

13. Los datos siguientes indican los montos, en centavos, del impuesto por galón de gasolina en diversas entidades de Estados Unidos:

9	9	13.5	7	6.5	11	9	11.7	11
11	12	9.8	5	13	8	11	9	9
8	9	10	13	13.7	8	13	8	
13	12	11	10.5	9	14	10	13	

a) Construya una gráfica de caja para los datos.
b) ¿Hay observaciones aberrantes?

14. La Nielson Company recaba información sobre los hábitos de atención a la televisión por parte de los estadounidenses. Los datos adjuntos indican el tiempo dedicado a la semana a ver televisión, en horas, para una muestra de 20 estudiantes universitarios:

16	36	22	27	38	51	30	25	10	5
29	21	26	31	11	25	33	25	15	16

a) Construya una gráfica de caja para los datos.
b) ¿Hay observaciones aberrantes?

Un paso más allá

15. Los datos en el diagrama de tallo y hojas adjunto indican las calificaciones logradas en un examen de estadística.

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

a) Construya una gráfica de caja.
b) ¿Encuentra observaciones aberrantes?

16. ¿Puede un puntaje de 5 tener un puntaje estándar de 3 si es miembro de una población con una media de 7? Explique.

17. Si un puntaje de 13 es miembro de una población con una media de 7 y tiene un puntaje estándar de 3, encuentre la varianza de la población.

18. Si un puntaje de 10 es miembro de una población con una varianza de 9 y tiene un puntaje estándar de 5, encuentre la media de la población.

19. Una población tiene una media igual a 7 y una varianza igual a 1. Encuentre el valor del puntaje que tiene un puntaje estándar igual al doble de su valor.

20. Si para definir una observación aberrante se usa sólo la restricción sobre el puntaje z, ¿es una observación aberrante el valor un millón en la muestra: {0, 0, 0, 0, 1,000,000}? Explique.

21. Como en el caso anterior, ¿es una observación aberrante el valor uno de la muestra {0, 0, 0, 1, 1, 1, 2, 2, 2, 5}?

22. ¿Puede encontrar una muestra de tamaño 4 cuyo valor máximo tenga un puntaje z *mayor que 3/2?*

RESUMEN DEL CAPÍTULO

En este capítulo introdujimos los conceptos de tendencia central, puntos de posición y variabilidad; estudiamos cuatro medidas de tendencia central: media, mediana, moda y rango promedio. Estas medidas proporcionan valores centrales para conjuntos de datos. Aprendimos que en una distribución las posiciones relativas de la media, la mediana y la moda determinan la simetría o sesgo de la distribución; después, estudiamos cuatro medidas de dispersión o variabilidad: rango, varianza, desviación estándar y rango intercuartil. Estas medidas se usan para describir la cantidad de dispersión en un conjunto de datos. Vimos que el teorema de Chebichev es importante para comprender el concepto de desviación estándar; finalmente, se introdujeron los puntajes estándar y las gráficas de caja: los primeros expresan las posiciones relativas de las medidas respecto a su media y también son útiles para hacer comparaciones relativas de datos de dos poblaciones o muestras diferentes; las gráficas de caja son útiles para exhibir el centro, la variabilidad y el sesgo o simetría en un diagrama, y para ayudar a identificar observaciones aberrantes en un conjunto de datos.

REVISIÓN DEL CAPÍTULO

■ TÉRMINOS IMPORTANTES ■

Los términos siguientes utilizados en el capítulo se han mezclado para proporcionarle una mejor práctica de revisión; dé una definición de cada uno con sus propias palabras y después verifique sus definiciones contra las dadas en el capítulo.

constante
moda cruda
diagrama de caja
bimodal
desviación de un valor
media
media de tendencia central
media de depresión
mediana
rango promedio
moda
rango
varianza
puntaje z
rango intercuartil
observación aberrante
gráfica de caja
cuartiles
deciles
variabilidad
mediana muestral aproximada
punto de posición
puntaje original
histograma sesgado
puntaje estándar
desviación estándar
suma de cuadrados
histograma simétrico
percentil
media muestral aproximada
clase modal

■ SÍMBOLOS IMPORTANTES ■

n, tamaño de la muestra
$\bar{x}$, medida muestral
μ, media poblacional
Σ, usada para indicar suma
N, tamaño de la población
$\tilde{\mu}$, mediana poblacional
$\tilde{x}$, mediana muestral
P_n, n-ésimo percentil
Q_1, primer cuartil
Q_2, segundo cuartil
Q_3, tercer cuartil
D_n, n-ésimo decil
R, rango
IQR, rango intercuartil
SS, suma de cuadrados
σ^2, varianza poblacional
s^2, varianza muestral
σ, desviación estándar pobalcional
s, desviación estándar muestral
$\bar{x}_a$, media muestral aproximada
$\tilde{x}_a$, mediana muestral aproximada

HECHOS Y FÓRMULAS IMPORTANTES

Media muestral: $\bar{x} = \frac{\Sigma x}{n}$ (3.1)

Media poblacional: $\mu = \frac{\Sigma x}{N}$

Media muestral para todos los datos agrupados en una tabla de frecuencia:

$\bar{x} = \frac{\Sigma f x}{\Sigma f}$ (3.2)

Rango intercuartil: $IQR = Q_3 - Q_1$

Para una colección finita de datos, la suma de desviaciones de los valores es 0; es decir, $\Sigma (x - \bar{x}) = 0$ (3.3)

Suma de cuadrados para una población: $SS = \Sigma (x - \mu)^2$ (3.4)

Suma de cuadrados para una muestra: $SS = \Sigma (x - \bar{x})^2$ (3.4)

Fórmula para calcular la suma de cuadrados de una muestra:

$SS = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$ (3.5)

Fórmula para calcular la suma de cuadrados de una población:

$SS = \Sigma x^2 - \frac{(\Sigma x)^2}{N}$ (3.5)

Varianza poblacional: $\sigma^2 = \frac{SS}{N}$ (3.6)

Varianza muestral: $s^2 = \frac{SS}{n-1}$ (3.7)

Desviación estándar muestral:

$s = \sqrt{\text{varianza muestral}}$

Desviación estándar poblacional:

$\sigma = \sqrt{\text{varianza poblacional}}$

Estimación de s: $s \approx \frac{R}{4}$ (3.8)

Suma de cuadrados para datos muestrales agrupados:
$SS = \Sigma f(x - \bar{x})^2$ (3.9)

Suma de cuadrados para datos poblacionales agrupados:
$SS = \Sigma f(x - \mu)^2$ (3.9)

Fórmula para calcular la suma de cuadrado de datos en una tabla de frecuencia:

$SS = \Sigma f x^2 - \frac{(\Sigma f x)^2}{\Sigma f}$ (3.10)

Teorema de Chebichev: al menos $(1 - 1/k^2)100\%$ de cualquier conjunto de datos dista menos de k desviaciones estándar de la media, si k es un número real mayor o igual a 1.

Puntaje z o estándar de una medida de una población:

$z = \frac{x - \mu}{\sigma}$ (3.11)

Puntaje z o estándar de una medida de una muestra:

$z = \frac{x - \bar{x}}{s}$ (3.11)

Una población de puntajes z tiene como media cero y como desviación estándar 1.

Una medida de una observación es aberrante de una muestra si:

Regla 1. El tamaño de la muestra es mayor que 10, la distribución de frecuencias tiene forma acampanada y el puntaje z dista más de tres desviaciones estándar de la media.

Regla 2. La medida cae más de tres IQR debajo del cuartil inferior, o más de tres IQR arriba del cuartil superior.

EJERCICIOS DE REPASO

1. Calcule la media, la mediana, la moda, el rango promedio, el rango, la varianza y la desviación estándar para cada una de las poblaciones siguientes:

a) 3, 7, 4, 6, 8, 2
b) 7, 8, 5, 2, 3
3) 9, 6, 0, 1, 4
d) 3, 3, 3

2. Calcule la media, la mediana, la moda, el rango, la varianza y la desviación estándar para cada una de las muestras:

a) 4, 7, 2, 2
b) 1, 8, 9, 4, 4
c) 0, 0, 1, 1, 10
d) 3, 3, 3
e) 8, 14, 15, 16, 22

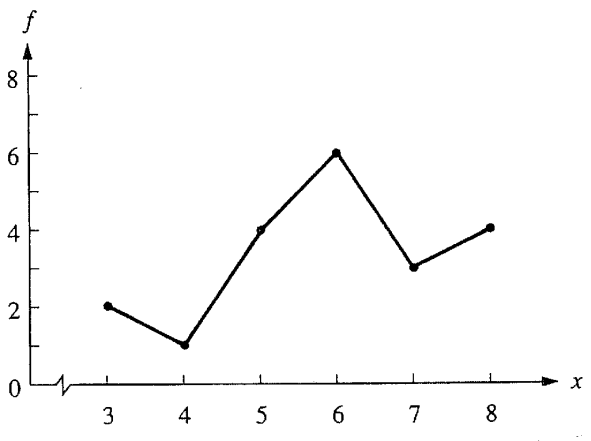
3. Calcule el puntaje z para x en cada una de las situaciones siguientes:

a) $x = 22, \mu = 15, \sigma = 2$
b) $x = -10, \mu = 5, \sigma = 8$
c) $x = 0, \bar{x} = 12, s = 6$
d) $x = 12.5, \bar{x} = 22, s = 0.4$
e) $x = 17, \bar{x} = 15, s^2 = 4$

4. Calcule la media, la mediana, la moda, la varianza y la desviación estándar para la tabla de frecuencia de datos muestrales:

x	f
0	1
1	3
2	2
3	4

5. Encuentre la mediana, la moda, la varianza y la desviación estándar para los datos muestrales ilustrados por la gráfica lineal siguiente:



6. Se recabaron los datos muestrales:

8	8	26	10	8
8	8	18	8	14
20	10	6	14	14

a) Encuentre $\bar{x}$ y s.
b) Si al recabar los datos se cometió un error y la medida original de 26 hubiera sido 20, ¿debería crecer o decrecer s? Explique.
c) Con las hipótesis del inciso anterior pero suponiendo que la medida original hubiera sido 8. ¿Aumentaría o disminuiría s? Explique.

7. Un grupo de cálculo tiene 30 alumnos. Las calificaciones que siguen son las obtenidas en un examen por los alumnos que se sientan en la primera fila: 87, 83, 89, 71 y 95.
a) ¿Esta colección de calificaciones es una muestra o una población?
b) Calcule la media y la desviación estándar de los datos.
c) Encuentre los puntajes estándar para los valores 71 y 95.

8. Para cada uno de los conjuntos siguientes, especifique una medida de tendencia central apropiada y dé su valor. Justifique su elección en cada caso.

a)

Peso en libras
3
2
4
13
4
4

b)

Clasificación	Número
Profesor	25
Profesor asociado	24
Profesor asistente	13
Instructor	10

c)

Partido	Número
Demócrata	200
Republicano	300
Socialista	50
Independiente	17

d)

Calificación	Número
A	2
B	3
C	1

e)

Velocidad	Número
Rápido	25
Lento	75

9. Los siguientes datos representan los cargos mensuales, en dólares, del servicio telefónico en 19 ciudades del mundo: 7.28, 8.54, 15.28, 5.51, 3.17, 6.34, 3.80, 4.59, 5.12, 9.98, 7.04, 10.00, 11.96, 5.48, 2.30, 5.85, 9.39, 8.73, 7.66.[25]

a) Encuentre $\bar{x}$.

b) Determine s.

c) Calcule el puntaje z para el cargo mensual del servicio telefónico en Nueva York ($x = 10.00$ dólares).

d) Construya una gráfica de caja.

10. Se seleccionaron cincuenta domicilios para determinar el número de habitantes hombres. Los datos obtenidos se enlistan aquí:

0	1	2	1	3	0	1	4	0	1
1	1	1	1	1	0	1	3	2	3
1	0	1	2	2	1	1	2	2	1
0	0	0	0	0	0	0	1	1	1
2	1	0	1	1	2	2	0	1	0

a) Encuentre $\bar{x}$.

b) Encuentre s.

c) ¿Cuántas medidas distan menos de una desviación estándar de la media?

11. Los siguientes son promedios EPA de rendimiento en millas por galón, para 15 automóviles compactos y subcompactos modelo 1989: 30, 31, 34, 31, 35, 41, 27, 35, 20, 47, 27, 29, 34, 38 y 32.

a) Ubique $\bar{x}$ y s.

b) Encuentre los cuartiles y el cuarto decil.

c) Construya una gráfica de caja.

12. Los datos siguientes representan las ventas anuales de armas, en billones de dólares, de Estados Unidos a países del tercer mundo, de 1976 a 1983: 8.2, 9.8, 10.1, 9.2, 6.4, 6.8, 7.9 y 9.7. Encuentre:

a) $\bar{x}$

b) s

c) Q_3 y D_7

Aplicaciones de computación

1. Se listan aquí las calificaciones del examen de ingreso de una muestra de 100 principiantes que acuden a una universidad en el medio oeste:

432 257 502 506 425 479 387 394 282 423 606
417 596 395 517 512 501 620 142 556 671 633
340 489 646 394 440 323 367 554 544 347 576
320 505 356 428 797 353 532 294 555 512 433
454 563 299 355 455 452 412 436 562 602 561
630 375 338 244 283 452 412 326 564 350 664
279 284 221 432 446 284 492 348 401 267 372
617 285 195 309 637 314 415 546 577 282 370
353 457 394 485 276 377 170 690 583 273 393
258

Use un programa computacional para:

a) encontrar la media muestral.

b) hallar la desviación estándar muestral.

c) calcular el rango.

d) hallar los puntos cuartiles.

e) ubicar el vigésimo percentil.

f) fijar el cuarto decil.

g) construir un histograma.

h) dibujar un diagrama de tallo y hojas.

i) trazar una gráfica de caja.

2. Los datos siguientes representan los pasos, en centésimos de libra, de una muestra de niños recién nacidos registrados durante el año pasado en el Memorial Hospital.

631 827 631 734 938 604 583 753 554 890 781
750 756 779 758 821 612 780 843 581 743 951
669 682 714 711 930 927 744 727 857 602 875
571 829 759 875 902 808 766 866 590 623 986
793 835 674 770 842 738 838 726 609 717 657
702 916 618 855 770 680 847 679 754 733 787
869 825 808 715 723 728 849 958 760 875 841
917 851 848 768 750 700 793 870 627 641 795
732 582 856 913 809 804 820 602 779 651 773
591

Use un programa computacional para:

a) encontrar la media muestral.

b) hallar la desviación muestral estándar.

c) calcular el rango.

d) hallar los puntos cuartiles.

e) ubicar el sexagésimo percentil.

f) fijar el sexto decil.

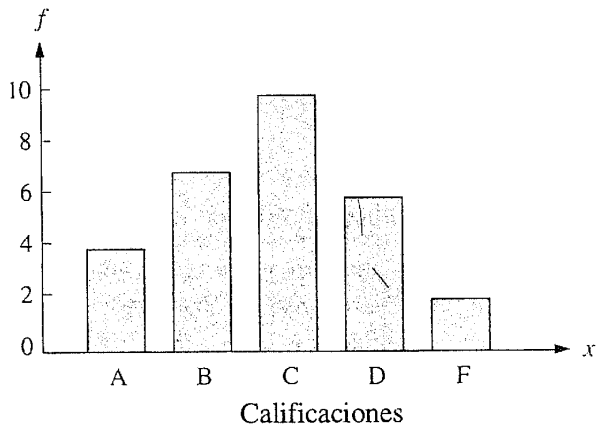
g) construir un histograma.

h) trazar un diagrama de tallo y hojas.

i) dibujar una gráfica de caja.

EXAMEN DE CONOCIMIENTOS DEL CAPÍTULO

1. Las calificaciones finales de una sección del grupo 209 de matemáticas se ilustran en la gráfica de barras adjunta.



a) ¿Qué medida de tendencia central deberá usarse para describir la calificación central? Explique la razón de su respuesta.
b) Utilice su respuesta anterior para encontrar la(s) calificación(es) central(es).
c) ¿Cuántos estudiantes están representados en la gráfica?
d) ¿Qué porcentaje de estudiantes recibieron una calificación de C?
e) ¿Qué porcentaje de estudiantes recibieron una calificación de C o mejor?

2. Considere la muestra 3, 8, 7, 12 y 10 para encontrar:
a) el rango
b) la media
c) la mediana
d) el rango promedio
e) la varianza
f) la desviación estándar
g) el puntaje estándar para la medida 10
h) el IQR.

3. Considere la siguiente tabla de frecuencia para una población:

x	f
4	2
8	3
3	5

a) Encuentre μ.
b) Determine σ.

4. ¿Qué puede decir de x en relación con el resto de los datos si x:
a) ¿tiene un puntaje z de o?
b) ¿posee un puntaje estándar de 2?
c) ¿tiene un puntaje de z de -1?

5. ¿En cuál de las situaciones siguientes es mayor el puntaje original x respecto a su conjunto de datos?
a) $x = 37, \bar{x} = 20, s = 10$
b) $x = 500, \bar{x} = 200, s = 250$
c) $x = 3.0, \bar{x} = 1.0, s = 0.7$

6. Si $\mu = 8$ y $\sigma^2 = 4$, encuentre el puntaje original x para $z = -2$.

7. Construya una gráfica de caja para los datos del problema 2.

8. Suponga que una muestra consiste en cinco medidas, 30, 80, 50, 40 y x. Determine el valor de x tal que la media, la mediana y la moda sean todas iguales.

9. ¿Tiene la mayor parte de la gente una medida de pie mayor que el promedio? Justifique su respuesta.

4 Análisis descriptivos de datos bivariados

DESCRIPCIÓN

4.1 Dependencia lineal y covarianza

4.2 Correlación

4.3 Regresión y predicción

OBJETIVOS DEL CAPITULO

En este capítulo estudiaremos

- Qué es un diagrama de dispersión y cómo se usa.
- Covarianza.
- Correlación.
- Cómo determinar el coeficiente de correlación r.
- El método de los mínimos cuadrados para determinar la ecuación de predicción.
- Cómo determinar la ecuación de mínimos cuadrados, que estima cómo están relacionadas dos variables.
- Cómo usar la ecuación de regresión con propósitos de predicción.
- Cómo se relacionan el coeficiente de correlación y la pendiente de la recta de regresión.
- Qué es la suma de cuadrados para el error y cómo se calcula.

MOTIVADOR 4

El clima parece afectar la ofensiva en beisbol. La tabla adjunta indica una relación entre la temperatura y la ofensiva de 1987 a 1989.[26]

Temperatura	Porcentaje de bateo	Carreras por juego	Jonrones por juego
0°-59°	0.248	8.0	1.40
60°-69°	0.253	8.5	1.65
70°-79°	0.259	8.6	1.69
80°-89°	0.263	9.1	1.85
90° en adelante	0.263	9.1	1.83

Los datos sugieren que cuando la temperatura aumenta, la ofensiva mejora; un estudio sobre la relación entre la temperatura y la ofensiva utiliza regresión y correlación, que son los temas que veremos en este capítulo.

Panorama del capítulo

Los análisis estadísticos utilizan frecuentemente datos cuantitativos de naturaleza *bivariada*; esto es, a cada elemento de la muestra le corresponde un par de medidas. Los siguientes son ejemplos de **datos bivariados:**

- Salarios y edades de maestros del distrito A
- Pulsaciones por minuto y presión sanguínea sistólica de estudiantes del grupo 209 de matemáticas
- Estaturas y pesos de un grupo del club de Scouts

- Precipitación pluvial diaria y temperatura diaria promedio en Frotsburg durante diez días
- Ingreso en la primavera y el verano de 1985 en 20 universidades

Este capítulo tratará gráficas de datos bivariados, midiendo la fuerza de una relación lineal y describiendo relaciones lineales entre dos variables. En todo el capítulo trataremos únicamente relaciones lineales (**línea recta**).

SECCIÓN 4.1 *Dependencia lineal y covarianza*

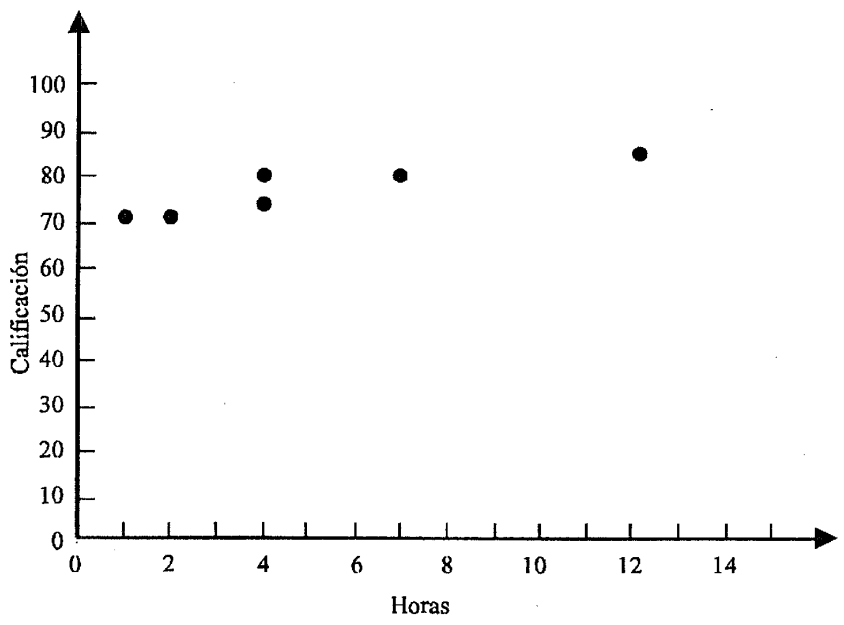
Los datos bivariados pueden verse como una colección de pares ordenados (x, y), donde la medida x en el primer conjunto de datos es la pareja de la medida y en el segundo conjunto; el valor perteneciente al primer conjunto se escribe siempre primero en la pareja. Se acostumbra llamar **variable independiente** a la variable x y **variable dependiente** a la variable y. La aplicación indicará usualmente cuál conjunto de datos se asocia con la variable independiente: estos pares ordenados se pueden dibujar en un sistema coordinado. La gráfica resultante se llama **diagrama de dispersión**.

EJEMPLO 4.1

Considere la colección adjunta de datos pareados; representan el número de horas de estudio (x) y la calificación recibida (y) en un examen para una muestra de seis estudiantes.

Estudiante	A	B	C	D	E	F
x: horas	1	2	4	4	7	12
y: calificación	71	71	74	80	80	86

Un diagrama de dispersión para los datos es como sigue:



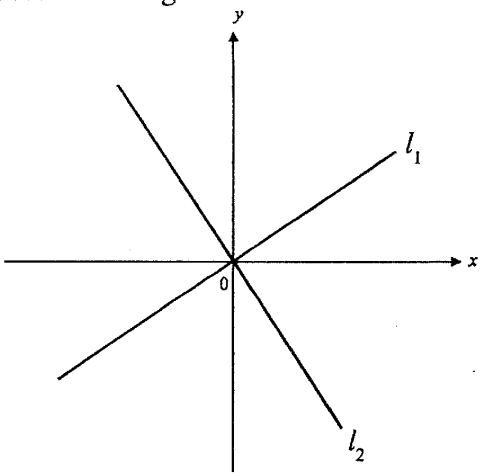
Estamos interesados en determinar cuándo hay una **dependencia lineal** entre las dos variables, es decir, queremos determinar si la variable y tiene una tendencia a crecer o a decrecer, cuando la variable x aumenta. Si examinamos el diagrama de dispersión del ejemplo 4.1, parece que existe una tendencia de y a crecer cuando x lo hace. En este caso, decimos que hay algún grado de dependencia lineal entre x y y. Si la tendencia es que la variable y crezca cuando la variable x crece, la dependencia se llama *positiva*; en cambio, si la tendencia es que y disminuya cuando x crece, la dependencia es *negativa*; si no hay tendencia de y a crecer o decrecer cuando la variable x crece, entonces no hay dependencia lineal.

Si los datos pareados tienen una relación lineal perfecta, la pendiente —inclinación— de la línea recta indica el tipo de relación de dependencia para las variables utilizadas; si la línea tiene una pendiente positiva —si sube de izquierda a derecha—, entonces la dependencia es positiva, mientras que si tiene una pendiente negativa —si baja de izquierda a derecha—, la dependencia es negativa.

Considere las dos líneas de la figura 4.1; ambas pasan a través del origen; la línea etiquetada con l_1 tiene una pendiente positiva y la línea l_2 tiene una pendiente negativa. Note que todos los puntos, excepto el origen de la línea l_1, caen en los cuadrantes I y III; para un punto en estos cuadrantes, ambas coordinadas son positivas o negativas.

FIGURA 4.1

Líneas rectas a través del origen



Es posible asignarle un peso a cualquier punto contenido en una recta que pasa a través del origen; este peso es un número definido por el producto de sus coordenadas. Si el punto está contenido en los cuadrantes I y III, el peso asignado es positivo, mientras que si el punto está contenido en los cuadrantes II y IV, el peso es negativo. En la figura 4.1, cada punto en la línea l_2, excepto el origen, tiene un peso negativo y cada punto en la línea l_1 (excepto el origen) tiene peso positivo; al origen se le asigna un peso de 0.

Para resumir:

1. Cualquier línea a través del origen con pendiente positiva tiene un producto de coordenadas que es no negativo.
2. Cualquier línea a través del origen con pendiente negativa tiene un producto de coordenadas que es no positivo.

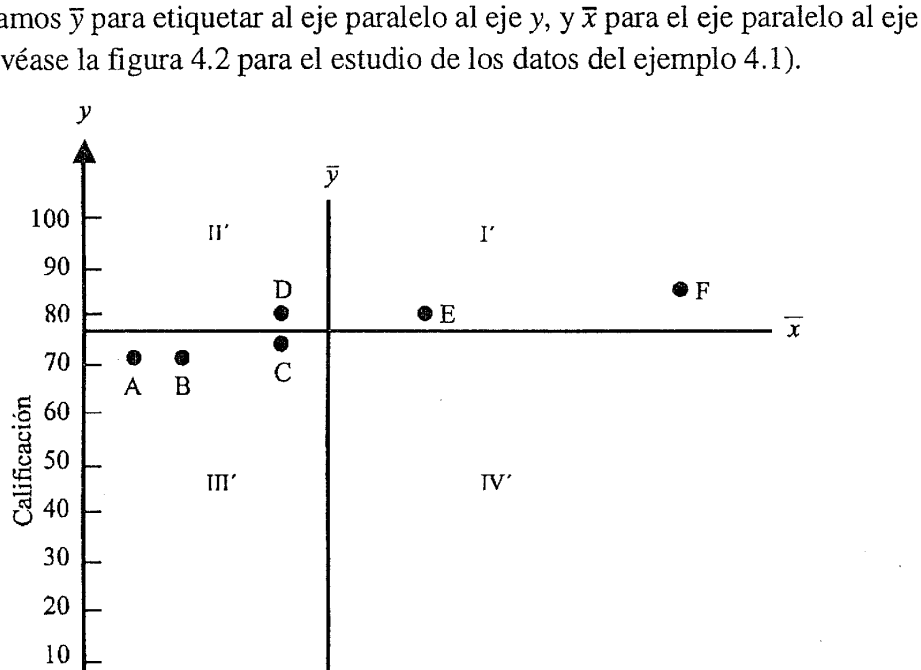
La desviación de valores para x y y puede usarse para crear una fórmula para medir el grado de dependencia lineal. Recuerde los hechos siguientes relativos a desviación de valores $x - \bar{x}$:

1. Una medida está debajo de la media si la desviación de su valor es negativa.
2. Una medida está arriba de la media si la desviación de su valor es positiva.
3. Una medida es igual a la media si la desviación de su valor es cero.

Si transformamos cada medida de una pareja (x, y) a su correspondiente desviación de valor, resulta la pareja $(x - \bar{x}, y - \bar{y})$; entonces el diagrama de dispersión de los pares de desviaciones de los valores tiene una interpretación interesante. El punto $(\bar{x}, \bar{y})$ se llama el **centroide** del diagrama de dispersión y sirve como punto de referencia. Si dibujamos dos líneas a través del centroide, una paralela al eje x y otra al eje y, entonces estas dos líneas pueden servir como líneas de referencia o ejes, para la desviación de los valores; usamos $\bar{y}$ para etiquetar al eje paralelo al eje y, y $\bar{x}$ para el eje paralelo al eje x (véase la figura 4.2 para el estudio de los datos del ejemplo 4.1).

FIGURA 4.2

Diagrama de dispersión de desviaciones de valores de los datos en el ejemplo 4.1



Estas nuevas líneas de referencia establecen cuatro cuadrantes: I', II', III' y IV'. Una pareja de desviaciones de valores se dibujará en el cuadrante I' si la desviación de sus valores x y y es positiva; en el cuadrante II', si la desviación de su valor x es negativa y la de su valor y positiva; en el cuadrante III', si la desviación de sus valores x es negativa y la de su valor y es negativa; y en el cuadrante IV' si la desviación de su valor x es positiva y la de su valor y negativa. Las distancias perpendiculares de los puntos medidos desde los ejes $\bar{x}$ y $\bar{y}$ representan las desviaciones del centroide; los pares de desviaciones de valores $(x - \bar{x}, y - \bar{y})$ se dibujan con respecto a los ejes $\bar{x}$ y $\bar{y}$ de la misma forma en que los pares (x, y) se dibujan con respecto a los ejes x y y.

El producto de las dos desviaciones de valores para una pareja determina un peso; si todos los puntos en un diagrama de dispersión están contenidos en los cuadrantes I' y III', entonces los pesos son todos positivos y si los puntos están contenidos en los cuadrantes II' y IV', todos los pesos son negativos. La suma de los pesos de todos los puntos de un diagrama de dispersión indica la fuerza de la dependencia lineal; si la suma de los pesos es positiva, la dependencia lineal también lo es, pero si esa suma es negativa, la dependencia es negativa; si la suma de los pesos es cero, no hay dependencia lineal entre las variables x y y. La tabla 4.1 contiene información de las medidas pareadas usadas anteriormente. La media de las medidas x es $\bar{x} = 5$ y la media de las y es $\bar{y} = 77$.

TABLA 4.1

Coordenadas, desviaciones de valores, cuadrantes y pesos para los datos en el ejemplo 4.1

Estudiante	Coordenadas x, y	$x-\bar{x}$	$y-\bar{y}$	Coordenadas $(\bar{x}, \bar{y})$	Cuadrante	Peso
A	(1, 71)	–4	–6	(–4, –6)	III'	24
B	(2, 71)	–3	–6	(–3, –6)	III'	18
C	(4, 74)	–1	–3	(–1, –3)	III'	3
D	(4, 80)	–1	3	(–1, 3)	II'	–3
E	(7, 80)	2	3	(2, 3)	I'	6
F	(12, 86)	7	9	(7, 9)	I'	63
		0	0			111

Note las relaciones siguientes de la tabla 4.1:

1. Todos los estudiantes excepto D están identificados con pares que tienen pesos positivos y dibujados en los cuadrantes I' y III'.
2. El estudiante D está identificado con un par cuyo peso es negativo y dibujado en el cuadrante II'.
3. $\Sigma (x - \bar{x}) = 0$ y $\Sigma (y - \bar{y}) = 0$.
4. El diagrama de dispersión está dominado por puntos en los cuadrantes I' y III' con pesos positivos. Esto se constata porque la suma de los pesos es 111.

Covarianza muestral

Cualquier peso asignado a una pareja de desviaciones de puntajes contribuye a la suma de todos los pesos; la suma de los pesos de las desviaciones de los puntajes proporciona una medida total de la dependencia de las variables; representa la tendencia combinada de los puntos que habrán de estar ya sea en los cuadrantes I' o III', o en el II' o el IV'. Si n representa el número de pares y dividimos la suma de los productos de las desviaciones de los valores entre $n - 1$, obtenemos, en algún sentido, una medida promedio de la dependencia lineal, llamada la **covarianza muestral**, denotada por $\text{cov}(x, y)$ o S_{xy}. Así, la covarianza muestral está dada por:

Covarianza muestral

$$\text{cov}(x, y) = S_{xy} = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{n - 1} \tag{4.1}$$

EJEMPLO 4.2

La covarianza muestral para los datos del examen del ejemplo 4.1 es

$$s_{xy} = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{n - 1}$$

$$= \frac{111}{5} = 22.2$$

Este resultado indica una dependencia lineal positiva entre la cantidad de tiempo de estudio y la calificación obtenida.

MINITAB puede usarse para determinar la covarianza muestral en los datos anteriores. La pantalla 4.1 contiene las órdenes usadas y las respuestas correspondientes.

Pantalla 4.1

```
MTB > NAME C1 'HOURS' C2 'GRADE'
MTB > SET C1
DATA > 1 2 4 4 7 12
DATA > END
MTB > SET C2
DATA > 71 71 74 80 80 86
DATA > END
MTB > COVARIANCE C1 C2

              HOURS     GRADE
HOURS       16.0000
GRADE       22.2000    36.000
```

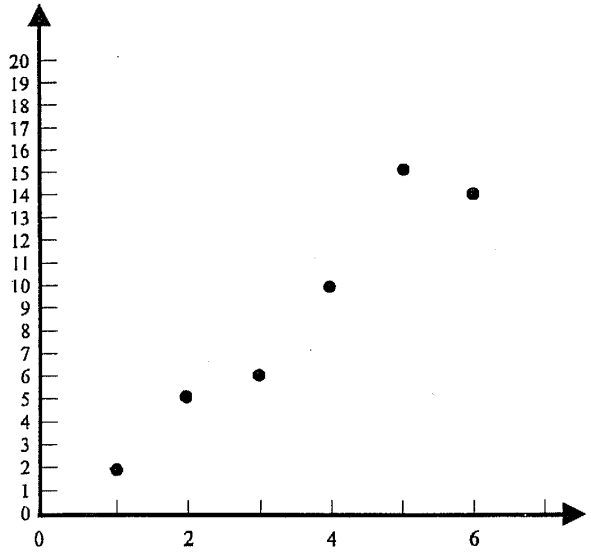
Las entradas 16.000 y 36.000 en las dos últimas líneas de la respuesta representan la varianza muestral de horas y grados, respectivamente.

EJEMPLO 4.3

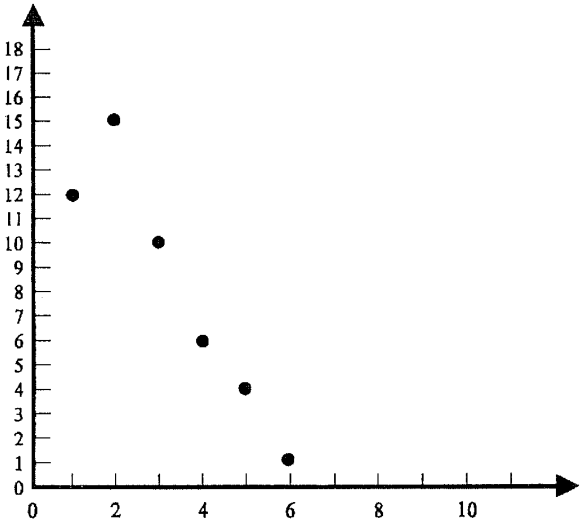
La figura 4.3 muestra diagramas de dispersión que representan una dependencia lineal negativa, una dependencia lineal positiva y una dependencia lineal cero; advierta en cada caso que el promedio de los productos de las desviaciones de los valores determina el tipo de dependencia.

FIGURA 4.3

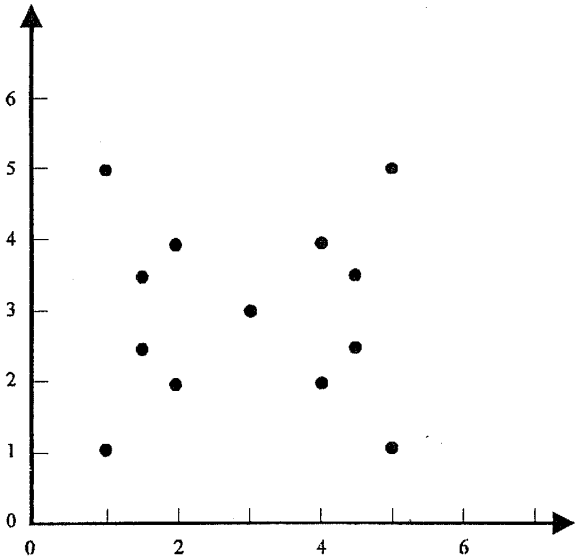
Diagramas de dispersión que ilustran dependencia positiva, negativa y falta de dependencia



Dependencia lineal positiva



Dependencia lineal negativa



Ninguna dependencia lineal

La covarianza muestral es similar a la varianza muestral en el sentido de que para la varianza muestral, la suma de cuadrados SS se divide entre $n - 1$, y para la covarianza muestral la suma de los productos de las desviaciones de los valores se divide entre $n - 1$.

$$s^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1} \qquad s_{xy} = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{n - 1}$$

↑ Varianza muestral ↑ Covarianza muestral

Mientras mayor sea el valor de la varianza muestral s^2, mayor variación hay en los datos; en general, mientras mayor sea la magnitud de la covarianza muestral s_{xy}, más fuerte es la dependencia lineal, pero ambas medidas pueden ser afectadas severamente por la presencia de observaciones aberrantes y que éstas tienen una tendencia general a inflarlas. La covarianza muestral tiene la seria desventaja de que dos diagramas de dispersión pueden tener la misma dependencia lineal, aunque posean covarianzas muestrales diferentes.

EJEMPLO 4.4

Para ver que dos diagramas de dispersión diferentes pueden tener la misma dependencia lineal, usemos los dos conjuntos siguientes de datos pareados, ambos con una dependencia lineal perfecta. La relación usada para generar los pares es $y = x$.

Conjunto de datos A

x	y
1	1
3	3

Las medias para el conjunto de datos A son: $\bar{x} = 2$ $\bar{y} = 2$

Conjunto de datos B

x	y
2	2
6	6

Las medias para el conjunto de datos B son: $\bar{x} = 4$ $\bar{y} = 4$

Las desviaciones de valores y los productos de las desviaciones de valores correspondientes se ilustran en la tabla 4.2.

TABLA 4.2

Suma de productos de desviación de valores para los conjuntos de datos A y B

Conjunto de datos A					Conjunto de datos B				
x	y	$x-2$	$y-2$	$(x-2)(y-2)$	x	y	$x-4$	$y-4$	$(x-4)(y-4)$
1	1	−1	−1	1	2	2	−2	−2	4
3	3	1	1	1	6	6	2	2	4
		0	0	2			0	0	8

La covarianza para el conjunto de datos A es:

$$s_{xy} = \frac{\Sigma (x - 2)(y - 2)}{n - 1} = \frac{2}{1} = 2$$

La covarianza para el conjunto de datos B es:

$$s_{xy} = \frac{\Sigma (x - 4)(y - 4)}{n - 1} = \frac{8}{1} = 8$$

Dos puntos determinan una recta. La dependencia lineal en cada caso no puede ser más fuerte, es perfecta; pero la covarianza para el conjunto de datos B es cuatro veces la covarianza para A. Un resultado de esta observación es que un cambio en las unidades de medida para x o y afectará el valor de la covarianza. En la sección 4.2 usaremos otra medida de dependencia lineal a la que no afecta la unidad de medición. La medida usa la suma de los productos de las desviaciones de los valores x y y.

GRUPO DE EJERCICIOS 4.1

Habilidades básicas

1. Considere el conjunto de datos bivariados:

x	0	1	4	−5	2	6
y	−2	−1	2	−7	0	4

a) Dibuje un diagrama de dispersión.
b) Determine si la dependencia es positiva o negativa usando el diagrama de dispersión.
c) Calcule el valor de la covarianza muestral. ¿Qué tipo de relación de dependencia indica?

2. Considere los datos bivariados:

x	0	4	8	1	3	−1
y	2	6	−14	0	−4	4

a) Dibuje un diagrama de dispersión.
b) Resuelva si la dependencia es positiva o negativa usando el diagrama de dispersión.
c) Calcule el valor de $\text{cov}(x, y)$. ¿Qué tipo de relación de dependencia indica?

3. Determine la covarianza muestral para los datos pareados:

x	1	2	3	7	8	9
y	2	1	4	18	10	19

4. Determine la covarianza muestral para los datos pareados adjuntos.

x	1	2	3	7	8	9
y	16	9	4	4	9	16

Más aplicaciones

5. Las calificaciones de ocho estudiantes del grupo 101 en matemáticas (x) e inglés (y) son como sigue:

x	77	81	94	50	72	63	88	95
y	82	47	85	66	65	72	89	95

a) Dibuje un diagrama de dispersión.
b) Determine si la dependencia es positiva o negativa usando el diagrama de dispersión.
c) Calcule el valor de s_{xy}. ¿Qué tipo de relación de dependencia indica?

6. Los datos siguientes representan los puntajes (x) en el SAT, matemáticas y (y) en el GPA para un grupo de 10 estudiantes:

x	450	376	514	678	501	734	325	400	398	681
y	3.5	2.5	2.1	3.6	2.7	3.8	1.8	2.4	2.0	1.9

a) Dibuje un diagrama de dispersión.
b) Resuelva si la dependencia es positiva o negativa usando el diagrama de dispersión.
c) Calcule la suma de los productos de las desviaciones de los puntajes. ¿Qué tipo de relación de dependencia indica?

7. Los datos que siguen representan los tamaños de los motores en pulgadas cúbicas y la estimación de millas por galón, para siete automóviles subcompactos.

Coche	Tamaño del motor	Millas/galón
Chevette	98	31
Sentra	98	35
Colt	86	41
Isuzu I-Mark	111	27
Mercedes 190D	134	35
Firebird	173	20
VW Rabbit	97	47

a) Trace un diagrama de dispersión.
b) Fije la covarianza muestral.

8. Las razones de precio-ganancia (PG) y el porcentaje rendido para siete tipos de acciones son:

Razón PG	2.4	2.4	3.4	2.9	4.0	3.8	2.7
Porcentaje rendido	4.2	0.7	10	4.6	6.2	6.3	8.4

a) Dibuje un diagrama de dispersión.
b) Resuelva la covarianza muestral.

9. El número de bebidas alcohólicas consumidas y la concentración de alcohol en la sangre para una muestra de seis sujetos con pesos corporales semejantes, utilizados en un experimento son:

Número de bebidas	2	3	4	5	6	7
Concentración de alcohol en la sangre	0.05	0.09	0.11	0.13	0.17	0.20

a) Dibuje un diagrama de dispersión para los datos.
b) Determine sxy.
c) ¿Qué tipo de relación de dependencia existe entre el número de bebidas consumidas y el nivel de alcohol en la sangre?

10. Considere el conjunto siguiente de datos bivariados:

x	1	2	3	4	5	6	7
y	12	7	4	3	4	7	12

a) Dibuje un diagrama de dispersión.
b) Calcule el valor de la covarianza muestral.
c) ¿Qué tipo de relación de dependencia existe entre x y y? Analice los resultados de los incisos a) y b) de este ejercicio.

11. Considere el conjunto de datos bivariados:

x	1	2	3	4	5
y	3	5	7	9	11

a) Dibuje un diagrama de dispersión.
b) Calcule el valor de la covarianza muestral.

12. Considere los datos bivariados:

x	1	2	3	4	5
y	−1	−3	−5	−7	−9

a) Dibuje un diagrama de dispersión.
b) Calcule el valor de la covarianza muestral.

Un paso más allá

13. Demuestre que:

$$s_{xy} = \frac{n\,\Sigma\, xy - (\Sigma\, x)(\Sigma\, y)}{n(n-1)}.$$

SECCIÓN 4.2 *Correlación*

Uno de los objetivos principales en estadística es la posibilidad de estimar o predecir el valor de una variable que depende de otra. El **análisis de regresión** es un método usado para estudiar la relación entre dos o más variables y para predecir valores de una de ellas; en muchas aplicaciones existe una relación entre las variables que pueden usarse con propósitos de predicción. El **análisis de correlación** es un método usado por los estadísticos para determinar la fuerza de la relación o dependencia lineal existente entre las variables; si la fuerza de la dependencia lineal es pequeña, entonces no será fructífero usar el análisis de regresión para encontrar la relación lineal y usarla con propósitos de predicción.

En la sección 4.1 aprendimos a construir diagramas de dispersión; ellos representan un medio gráfico de determinar si existe una relación lineal entre dos variables; si todos los puntos caen exactamente en una línea recta, entonces decimos que las dos variables tienen una *correlación lineal perfecta*; si los puntos están cercanos a una línea recta, se dice que las dos variables tienen una *correlación lineal fuerte*; si la línea recta tiene una pendiente positiva, decimos que las dos variables tienen *correlación lineal positiva*; y si la línea tiene pendiente negativa, decimos que las variables tienen *correlación lineal negativa*. Y si la recta tiene una pendiente de cero, decimos que *no hay correlación lineal* entre las dos variables.

La primera marca mundial para el recorrido de la milla final fue un tiempo de 4:56, registrado en 1864; desde entonces, el recorrido de la milla ha sido mejorado a 3:47.3 y en 1945 fue el último año en que la marca de la milla estuvo arriba de los 4 minutos. La tabla 4.3 muestra la evolución en el tiempo de la marca mundial para el recorrido de la milla, desde 1945 hasta 1985. En la figura 4.4 se muestra un diagrama de dispersión para los datos de las marcas.

TABLA 4.3

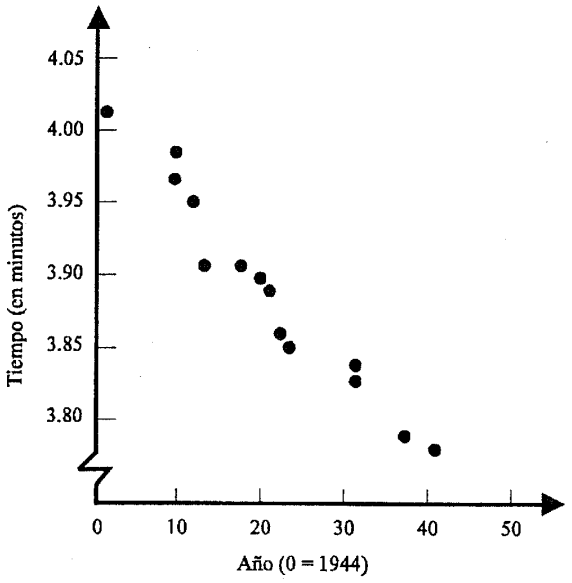
Marcas mundiales para el recorrido de la milla de 1945 a 1985

Año	País	Tiempo
1945	Suecia	4:01.4
1954	Estados Unidos	3:59.4
1954	Austria	3:58.0
1957	Gran Bretaña	3:57.2
1958	Australia	3:54.5
1962	Nueva Zelandia	3:54.4
1964	Nueva Zelandia	3:54.1
1965	Francia	3:53.6
1966	Estados Unidos	3:51.3
1967	Estados Unidos	3:51.1
1975	Tanzania	3:50.0
1975	Nueva Zelandia	3:49.4
1981	Gran Bretaña	3:47.3
1985	Gran Bretaña	3:46.3

Una ojeada al diagrama de dispersión de la figura 4.4, sugiere que existe una correlación negativa para los datos y que sería razonable una aproximación lineal para ellos; la correlación es negativa, pues los puntos del diagrama de dispersión parecen estar cercanos a una recta de pendiente negativa. Obtener la aproximación lineal a los puntos del diagrama de dispersión requiere análisis de regresión, que exploraremos en la sección 4.3. Los entusiastas del deporte han hecho muchas especulaciones sobre el año en que se dé un tiempo de 3:40 para la milla, y algunos expertos en el campo e investigadores han hecho predicciones para el año 2 000 usando análisis de regresión.

FIGURA 4.4

Marcas mundiales para los tiempos de recorrido de la milla de1945 a 1985



Considere las calificaciones del SAT (matemáticas) y los puntajes promedio de calificaciones al ingresar, GPA, de cada estudiante de segundo año inscrito este semestre en una universidad. Pudiera gustarnos obtener respuestas a las dos preguntas siguientes:

1. ¿Hay una relación lineal entre las calificaciones del SAT y los GPA?
2. Si la hay, ¿cuál es?

La respuesta para la primera pregunta requiere correlación y para la segunda pregunta se utiliza regresión. Considere los datos en la tabla 4.4, que son las calificaciones del SAT (matemáticas) y los GPA de ingreso, para una muestra de diez estudiantes de segundo año inscritos en una universidad estatal.

TABLA 4.4

Calificaciones del SAT (matemáticas) y GPA

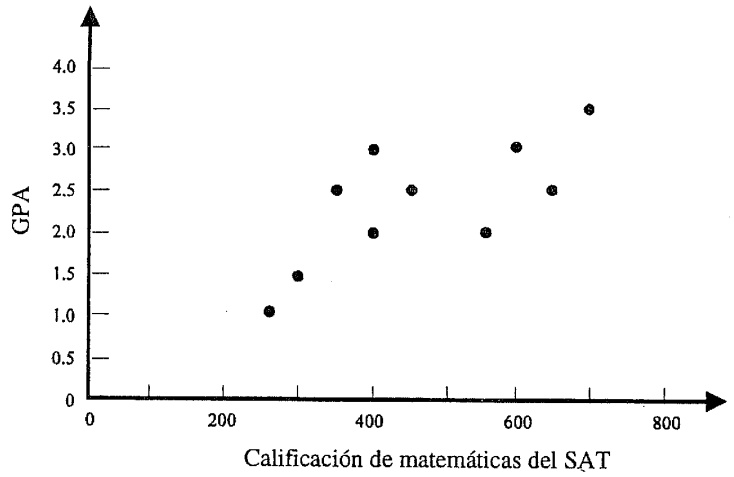
Número de estudiante	Calificación del SAT (matemáticas)	GPA
1	450	2.5
2	600	3.0
3	550	2.0
4	400	3.0
5	350	2.5
6	650	2.5
7	300	1.5
8	400	2.0
9	700	3.5
10	250	1.0

Los datos pueden exhibirse también en un diagrama de dispersión, como el mostrado en la figura 4.5, donde las calificaciones del SAT (matemáticas) se muestran en el eje horizontal y los GPA en el eje vertical; note que la correlación entre GPA y SAT parece ser positiva, pues los puntos parecen caer cerca de una recta de pendiente positiva.

En la sección 4.1 se usó la covarianza muestral para determinar la dependencia lineal entre dos variables. Si los puntos de un diagrama de dispersión se distribuyen desde abajo a la izquierda hasta arriba a la derecha, entonces hay una dependencia positiva (véase la figura 4.3), y si los puntos se distribuyen de arriba a la izquierda a abajo a la derecha, entonces hay una dependencia negativa (figura 4.4). En otras palabras, si los valores de y tienden a crecer cuando los valores de x lo hacen, entonces decimos que la correlación es positiva, mientras que si los valores de y tienden a disminuir cuando los de x crecen, decimos que es negativa; si los puntos de un diagrama de dispersión no caen en una línea recta, es imposible establecer la magnitud de la correlación, a menos que se use una fórmula que tome en cuenta las variaciones de un diagrama de dispersión lineal.

La correlación lineal o fuerza de la dependencia lineal para los diagramas de dispersión mostrados en las figuras 4.4 y 4.5, puede medirse por la covarianza muestral estudiada en la sección 4.1; pero hay dos problemas al usarla para medir la fuerza de la relación o dependencia lineal entre dos variables; primero, la covarianza depende de las unidades de medida. Si cambiamos las unidades para x y y, entonces la covarianza cambia. Segundo, no hay cotas en los valores de la covarianza. Por fortuna podemos remediar estos dos problemas.

FIGURA 4.5
Diagrama de dispersión para SAT y GPA



Lo que necesitamos para medir la fuerza de la relación lineal es un índice que posea las cuatro propiedades siguientes:

1. No estar ligado a las unidades de medida; sus valores no dependen de las unidades de medida de cada variable.
2. Su valor es igual a 1 si los puntos están en una línea recta con pendiente positiva.
3. Su valor es igual a –1 si los puntos están en una línea recta con pendiente negativa.
4. Su valor es cero si no hay relación lineal entre las variables.

Recuerde de la sección 3.4 que la media y la desviación estándar de una colección de medidas tienen la misma unidad de medida que las medidas de la colección; en consecuencia, las desviaciones de los valores para un conjunto de medidas tienen la misma unidad de medida que las medidas individuales. Si dividimos las desviaciones de los valores entre la desviación estándar, tendremos un conjunto de números que no tienen unidad de medida, son libres de unidades. Recuerde también de la sección 3.4 que estos cocientes se refieren a *puntajes z*; si una desviación de un valor es positiva o negativa, su puntaje z será también positivo o negativo.

Si en lugar de usar la suma de los productos de las desviaciones de los valores para x y y en el numerador de la fórmula de la covarianza muestral, para obtener un índice de la dependencia, usáramos la suma de los productos de los puntajes z, $\Sigma z_x z_y$, donde z_x representa un puntaje z para x, y z_y el puntaje z para y, obtendríamos un índice que cumpliría con las cuatro propiedades ya mencionadas. Este nuevo indicador se llama el coeficiente de Pearson y se denota por r:

Coeficiente de correlación de Pearson

$$r = \frac{\Sigma z_x z_y}{n-1} \tag{4.2}$$

donde n es el número de parejas usadas en la muestra. Si el valor de r es igual a 1 o a –1, entonces existe una correlación lineal o relación lineal perfecta

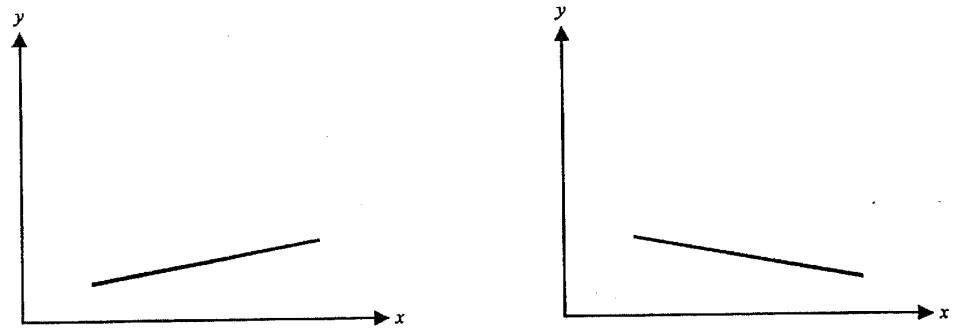
entre las variables, mientras que si $r = 0$ no hay correlación o relación lineal; esto significa que cuando x tiende a crecer, no hay una tendencia definida de los valores de y a crecer o a decrecer. Note que un valor de $r = 0$ no necesariamente significa la falta de una relación entre x y y. Puede existir una relación no lineal [véase la figura 4.7 (c)].

EJEMPLO 4.5

El diagrama de dispersión en la figura 4.6 exhibe dependencias o relaciones lineales perfectas entre x y y.

FIGURA 4.6

Diagramas de dispersión que muestran relaciones lineales perfectas



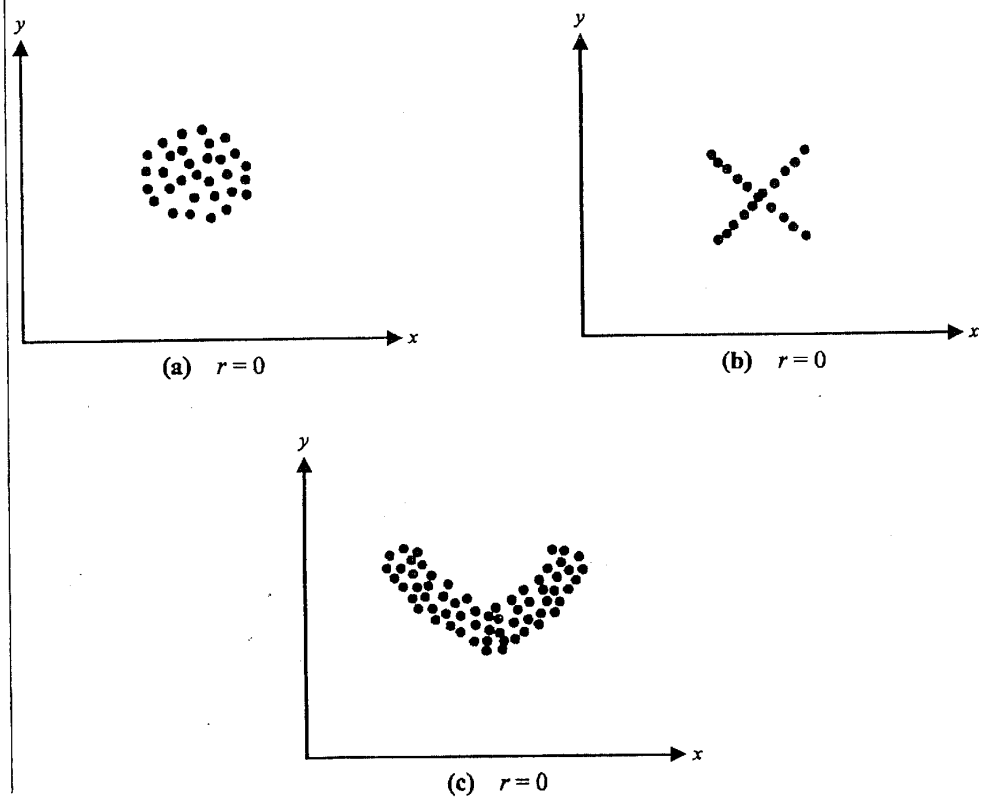
(a) $r = 1$ **(b)** $r = -1$

EJEMPLO 4.6

Los diagramas de dispersión en la figura 4.7 muestran que no hay relación lineal.

FIGURA 4.7

Diagramas de dispersión que muestran falta de relación lineal



(a) $r = 0$ **(b)** $r = 0$

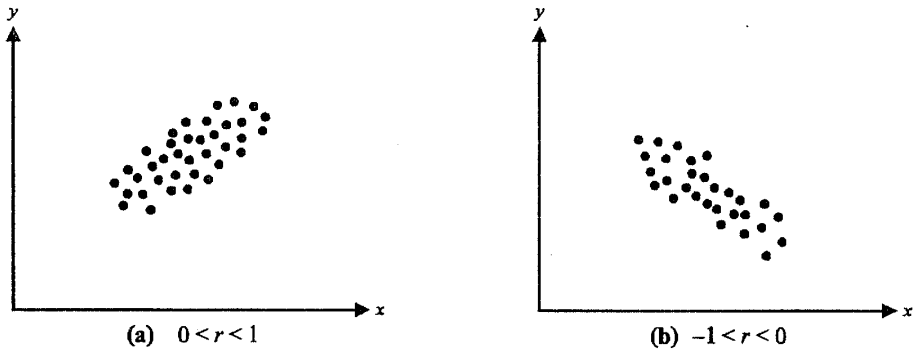
(c) $r = 0$

EJEMPLO 4.7

Los diagramas de dispersión en la figura 4.8 exhiben alguna relación lineal.

FIGURA 4.8

Diagramas de dispersión que muestran algunas relaciones lineales



(a) $0 < r < 1$ (b) $-1 < r < 0$

APLICACIÓN 4.1

Calcule el valor del coeficiente de correlación de Pearson (r) para la muestra de datos pareados de la sección 4.1, que representan el número de horas (x) de estudio para un examen y la calificación recibida (y) para una muestra de seis estudiantes.

Estudiante	A	B	C	D	E	F
x: horas	1	2	4	4	7	12
y: calificación	71	71	74	80	80	86

Solución:

1. La media y la desviación estándar para cada grupo son como sigue:
   x: $\bar{x} = 5$ y $s = 4$
   y: $\bar{y} = 77$ y $s = 6$
2. El puntaje z para x está dado por $z_x = (x - 5)/4$, y el puntaje z para y está dado por $z_y = (y - 77)/6$.
3. Los cálculos están organizados en esta tabla:

x	y	$x-5$	$y-77$	z_x	z_y	$z_x z_y$
1	71	−4	−6	−1	−1	1
2	71	−3	−6	−0.75	−1	0.75
4	74	−1	−3	−0.25	−0.5	0.125
4	80	−1	3	−0.35	0.5	−0.125
7	80	2	3	0.5	0.5	0.25
12	86	7	9	1.75	1.5	2.625
						4.625

4. El valor del coeficiente de correlación de Pearson es, usando la fórmula 4.2:

$$r = \frac{\Sigma z_x z_y}{n - 1}$$

$$= \frac{4.625}{6 - 1} = \frac{4.625}{5} = 0.925 \quad \blacksquare$$

Hay una fórmula para calcular el valor del coeficiente de correlación de Pearson (r), que utiliza el concepto de suma de cuadrados presentado en la sección 3.2.

$$SS_x = \Sigma x^2 - \frac{(\Sigma x)^2}{n} \tag{4.3}$$

Como estamos tratando con datos bivariados (x, y), SS_x significaría la suma de cuadrados para x, mientras que SS_y denotará la suma de cuadrados para y. Así, tenemos:

$$SS_y = \Sigma y^2 - \frac{(\Sigma y)^2}{n} \tag{4.4}$$

Si reescribimos el lado derecho de la fórmula (4.3), remplazando uno de los factores en que aparece x por un factor en que figure y, obtenemos la expresión siguiente:

$$\Sigma xx - \frac{(\Sigma x)(\Sigma x)}{n} \rightarrow \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}$$

La expresión resultante se llama suma de productos cruzados y se denota por SS_{xy}.

Suma de productos cruzados

$$SS_{xy} = \Sigma_{xy} - \frac{(\Sigma x)\ (\Sigma y)}{n} \tag{4.5}$$

Los valores de SS_x, SS_y, SS_{xy} se usan para calcular los valores del coeficiente de correlación de Pearson. La fórmula para calcular el valor r es:

Fórmula para calcular r

$$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} \tag{4.6}$$

APLICACIÓN 4.2

Consulte la aplicación 4.1. Calcule el valor de r usando la fórmula 4.6 para los datos pareados.

Solución: Primero calculamos los valores de SS_x, SS_y y SS_{xy}. La tabla 4.5 organiza los cálculos necesarios para determinar las sumas usadas en las fórmulas.

TABLA 4.5
Cálculos para obtener SS_x, SS_y y SS_{xy}

x	y	x^2	y^2	xy
1	71	1	5,041	71
2	71	4	5,041	142
4	74	16	5,476	296
4	80	16	6,400	320
7	80	49	6,400	560
12	86	144	7,396	1032
30	462	230	35,754	2421

Determine SS_x,

$$SS_x = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

$$= 230 - \frac{(30)^2}{6} = 80$$

Calcule SS_y

$$SS_y = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$$

$$= 35{,}754 - \frac{(462)^2}{6} = 180$$

Obtenga SS_{xy},

$$SS_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}$$

$$= 2421 - \frac{(30)(462)}{6} = 111$$

Use la fórmula (4.6) para determinar r, el coeficiente de correlación de Pearson

$$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

$$= \frac{111}{\sqrt{(80)(180)}} = 0.925$$

que concuerda con el resultado encontrado en la aplicación 4.1. ■

Pantalla 4.2

La pantalla 4.2 ilustra el uso de MINITAB para obtener el coeficiente de correlación para los datos de la aplicación 4.1.

```
MTB > READ C1 C2
DATA > 1 71
DATA > 2 71
DATA > 4 71
DATA > 4 80
DATA > 7 80
DATA > 12 86
DATA > END
     6 ROWS READ
MTB > NAME C1 'HOURS' NAME C2 'GRADE'
MTX > CORRELATION C1 C2

CORRELATION OF HOURS AND GRADES = 0.925
```

Codificación para simplificar los cálculos de r

Con frecuencia los valores de x y y hacen muy pesado calcular SS_x, SS_y y SS_{xy}; para simplificar el proceso tanto como sea posible, a menudo se usa la codificación. La **codificación** requiere aplicar transformaciones lineales a los datos. Las transformaciones lineales son del tipo siguiente:

$$U = ax + b$$

$$V = cy + d$$

donde $a \geq 0$ y $c \geq 0$. Entonces el coeficiente de correlación de Pearson entre U y V es igual al de x y y. La aplicación 4.3 muestra el proceso.

APLICACIÓN 4.3

Use la codificación para calcular el valor de r para los datos bivariados exhibidos en la tabla siguiente:

x	168	169	170	171
y	0.6	0.9	0.9	0.5

Solución: Sea $U = x - 167$ ($a = 1$, $b = -167$) y $V = 10y$ ($c = 10$, $d = 0$). Para encontrar los valores de U, sustituimos los valores de x en la ecuación $U = x - 167$, y para encontrar los valores de V sustituimos los valores de y en la ecuación $V = 10y$. Los datos transformados son:

U	1	2	3	4
V	6	9	2	5

Si usamos la fórmula (4.6), el valor de r para los datos bivariados (U, V) es $r = -0.447$. Por lo tanto, el coeficiente de correlación de Pearson para x y y es $r = -0.447$. ■

APLICACIÓN 4.4

Encuentre el coeficiente de correlación de Pearson r para los datos de la tabla 4.3.

Solución: Usaremos la transformación x = año – 1944 para codificar los datos a fin de simplificar los cálculos; el año 1945 será codificado como 1, el año 1946 como 2 y así sucesivamente; además, los tiempo se expresarán en minutos. En la tabla 4.6, x representa el año codificado y y el tiempo en minutos.

TABLA 4.6

Datos codificados para la aplicación 4.3

Año	Tiempo	x	y	x^2	y^2	xy
1945	4:10.4	1	4.023	1	16.1845	4.023
1954	3:59.4	10	3.990	100	15.9201	39.900
1954	3:58.0	10	3.967	100	15.7371	3 9.670
1957	3:57.2	13	3.953	169	15.6262	51.389
1958	3:54.5	14	3.908	196	15.2725	54.712
1962	3:54.4	18	3.907	324	14.2646	7 0.326
1964	3:54.1	20	3.902	400	15.2256	78.040
1965	3:53.6	21	3.893	441	15.1554	81.753
1966	3:51.3	22	3.855	484	14.8610	84.810
1967	3:51.1	23	3.852	529	14.8379	88.596
1975	3:50.0	31	3.833	961	14.6919	118.823
1975	3:49.4	31	3.823	961	14.6153	118.512
1981	3:47.3	37	3.788	1369	14.3489	140.156
1985	3:46.3	41	3.772	1681	14.2280	154.652
		292	54.466	7716	211.9690	1125.363

Para calcular el valor de r se siguen los pasos:

1. Calcule el valor de SS_x:

$$SS_x = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$
$$= 7716 - \frac{(292)^2}{14} = 1625.7143$$

2. Obtenga el valor de SS_y:

$$SS_y = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$$
$$= 211.9690 - \frac{(54.466)^2}{14} = 0.07291743$$

3. Determine el valor de SS_{xy}:

$$SS_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}$$
$$= 1125.363 - \frac{(292)(54.466)}{14} = -10.6421429$$

4. Establezca el valor de r.

$$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$
$$= \frac{-10.6421429}{\sqrt{(1625.7143)(0.07291743)}} = -0.9774$$

El valor obtenido de r indica una relación lineal fuerte entre el año y el tiempo. Advierta que el valor negativo de r significa que cuando el número del año aumenta, la marca mundial del tiempo de la milla decrece. ■

El coeficiente de correlación de Pearson (r) deberá interpretarse sólo como una medida matemática de la fuerza de la relación lineal entre dos variables. Nota:

> Un valor alto de r no necesariamente deberá entenderse como la existencia de una relación causa-efecto entre las variables, porque ambas variables pueden haber sido influidas por otras. Recuerde que un valor alto de r significa que las dos variables tienden de manera simultánea a variar en la misma dirección; la tendencia es un fenómeno matemático y no siempre implica una relación directa entre las variables.

EJEMPLO 4.8

Si el número de reuniones religiosas y el de crímenes violentos se registraran cada mes en un grupo de ciudades cuya población es muy variada, los datos indicarían probablemente una correlación positiva alta; pero sería ridículo concluir que el número de crímenes violentos y el de reuniones religiosas están relacionados directamente; una tercera variable, por ejemplo la población, está causando que los

crímenes y las reuniones religiosas varíen en el mismo sentido. La correlación entre crimen y reuniones religiosas es un ejemplo de **correlación espuria**, causada por una tercera variable; como consecuencia, deberemos ser muy cuidadosos respecto a deducir una relación causal de una correlación observada, pues la correlación puede resultar espuria.

GRUPO DE EJERCICIOS 4.2

Habilidades básicas

1. Para los datos bivariados exhibidos en la tabla adjunta, encuentre:
   a) SS_x
   b) SS_y
   c) SS_{xy}

x	1	5	2	4	8	9
y	3	7	2	6	7	4

2. De la tabla adjunta encuentre:
   a) SS_x
   b) SS_y
   c) SS_{xy}

x	2	7	8	1	5	9	3
y	8	7	1	4	7	4	5

3. Para cada uno de los conjuntos siguientes de datos bivariados, ¿esperaría usted una correlación positiva, una correlación negativa o una falta de correlación?
   a) Medidas de zapatos y medidas de sombreros.
   b) Consumo promedio de cerveza en los adultos para los 23 condados de Maryland y el número de nacimientos en esos mismo condados, durante el año pasado.
   c) Salarios promedio de los maestros y puntajes promedio en matemáticas en el SAT, para los sistemas escolares públicos de Pennsylvania.
   d) Pesos de coches y rendimiento de gasolina.

4. De los conjuntos siguientes de datos bivariados, ¿esperaría usted una correlación positiva, una correlación negativa o una falta de correlación?
   a) Pesos y estaturas de niños de 6 años de edad.
   b) Presión sanguínea y ritmo cardiaco para mujeres de 30 años edad.
   c) Promedio de precipitación pluvial en pulgadas y de producción en fanegas de los árboles de durazno en Clark County, Georgia, durante los últimos 10 años.
   d) Diámetros y áreas de círculos.

5. Considere los datos pareados siguientes:

x	1	5	11	17
y	1	7	19	19

   Use la fórmula (4.2) para determinar el valor del coeficiente de correlación de Pearson (r).

6. Calcule r para los datos adjuntos mediante la fórmula (4.2).

x	1	1	4	10	10	16
y	1	2	2	10	13	14

7. Considere el conjunto de datos bivariados:

x	0	1	4	−5	2	6
y	−2	−1	2	−7	0	4

   a) Dibuje un diagrama de dispersión.
   b) Use ese diagrama para determinar si la correlación es positiva o negativa.
   c) Calcule el valor de r.

8. Considere el conjunto:

x	0	4	8	1	3	−1
y	2	6	−14	0	−4	4

   a) Dibuje un diagrama de dispersión.
   b) Use el diagrama de dispersión para determinar si la correlación es positiva o negativa.
   c) Calcule el valor de r.

9. Considere el siguiente conjunto de datos bivariados:

x	1	2	3	4	5	6	7
y	12	7	4	3	4	7	12

   a) Dibuje un diagrama de dispersión.
   b) Calcule el valor del coeficiente de correlación de Pearson r.
   c) Analice los resultados de las partes a) y b).

Mas aplicaciones

10. Las calificaciones de ocho estudiantes inscritos tanto en el grupo 101 de matemáticas (x) como en el grupo 101 de inglés (y), se muestran en la tabla adjunta.

x	77	81	94	50	72	63	88	95
y	82	47	85	66	65	72	89	95

a) Dibuje un diagrama de dispersión.
b) Use ese diagrama para determinar si la correlación es positiva o negativa.
c) Calcule el valor de r.

11. Los datos de la tabla representan los puntajes de matemáticas en el SAT (x) y en la GPA (y), de un grupo de diez estudiantes.

x	450	375	514	678	501	734	325	400	398	618
y	3.5	2.5	2.1	3.6	2.7	3.8	1.8	2.4	2.0	1.9

a) Dibuje un diagrama de dispersión.
b) Use ese diagrama de dispersión para determinar si la correlación es positiva o negativa.
c) Calcule el valor de r.

12. Los datos adjuntos representan los tamaños de motor en pulgadas cúbicas, y las millas por galón estimadas para siete automóviles subcompactos.

Coche	Tamaño del motor	Millas por galón
Chevette	98	31
Sentra	98	35
Colt	86	41
Isuzu I-Mark	111	27
Mercedes 190D	134	35
Firebird	173	20
VW Rabbit	97	47

Determine el coeficiente de correlación de Pearson r.

13. Los datos en la tabla adjunta representan las razones precio-ganancia (PG) y el porcentaje de rendimiento en siete tipos de acciones. Calcule el coeficiente de correlación r.

Razón PG	2.4	2.4	3.4	2.9	4.0	3.8	2.7
Porcentaje de rendimiento	4.2	0.7	10	4.6	6.2	6.3	8.4

14. Los datos siguientes indican los montos en billones de dólares de las importaciones y exportaciones agrícolas en Estados Unidos.[27]

Año	Costos de importación	Precios de exportación
1974	21.9	10.2
1975	21.9	9.3
1976	23.0	11.0
1977	23.6	13.4
1978	29.4	14.8
1979	34.7	16.7
1980	41.2	17.4
1981	43.3	16.8
1982	39.1	15.4

Encuentre el coeficiente de correlación de Pearson para costos de importación y precios de exportación; verifique su resultado codificando los costos de importación como $U = 10$ (costos de importación) $- 219$, y los precios de exportación como $V = 10$ (precio de exportación) $- 93$ y determine r para las variables U y V.

15. La tabla adjunta informa que el número de escuelas estadounidenses (K-12) con equipos de video (VCR) y de cómputo (PC), creció significativamente en la década pasada.[28]

Año	1982-1983	1983-1984	1984-1985	1985-1986	1986-1987	1987-1988	1988-1989
Núm. de escuelas	83,648	82,952	81,971	81,461	81,408	80,999	82,089
Núm. de PC	30,859	55,175	70,255	74,379	76,242	76,899	78,784
Núm. de VCR	25,663	36,545	56,151	64,744	70,037	73,495	80,776

a) Calcule el coeficiente de correlación de Pearson para el número de escuelas y de PC.
b) Calcule r para el número de PC y el de VCR.

16. Consulte el ejercicio 15. Calcule el valor de r para el número de escuelas y el número de VCR; verifique su resultado codificando el número de escuelas como $U =$ (número de escuelas) $-$ 80,999 y el número de VCR como $V =$ (número VCR) $- 25{,}663$ y determine el valor de r para U y V.

17. La tabla siguiente de las presiones sanguíneas (en mm Hg) en un grupo de pacientes hipertensos, junto con el nivel de dosificación (en mg), de un fármaco contra la hipertensión.

Dosis	Presión
1	275
2	235
3	193
4	128
5	105

Determine el coeficiente de correlación de Pearson.

18. En un experimento para investigar el efecto del incremento en la dosis de un cierto barbitúrico, en μM/kg, sobre el tiempo de sueño (en horas), se hicieron las siguientes lecturas en cada uno de los tres niveles de dosificación.

Dosis	Tiempo	Dosis	Tiempo
3	4	10	7
3	6	15	13
3	5	15	11
10	9	15	9
10	8		

Determine el coeficiente de correlación de Pearson entre las dosis y el tiempo de sueño.

19. Para los datos en la aplicación 4.4, verifique que el coeficiente de correlación entre x y y es igual al coeficiente de correlación de Pearson entre U y V.

Un paso más allá

20. Los niveles de ingreso de las personas de raza negra siguen estando muy por debajo de los de la gente blanca. La tabla adjunta lista los ingresos per cápita, en dólares, de 1979 a 1988 para los dos tipos.[29]

Año	Blancos	Negros
1979	12,342	7,241
1980	11,820	6,897
1981	11,686	6,675
1982	11,679	6,571
1983	12,026	6,836
1984	12,455	7,147
1985	12,832	7,520
1986	13,332	7,779
1987	13,687	7,961
1988	13,896	8,271

a) Calcule el valor de la r de Pearson para los ingresos per cápita de los dos grupos.
b) Codifique los años 1979 = 0, 1980 = 1, etc., para determinar el valor de r para el año y el ingreso de los blancos.
c) Codifique los años 1979 = 0, 1980 = 1 y así sucesivamente, para encontrar el valor de r para el año y el ingreso de los negros.

21. Para los datos en el ejercicio 5, demuestre que:
$$SS_{xy} = \Sigma(x - \bar{x})(y - \bar{y}).$$

22. Verifique:
$$SS_{xy} = \Sigma(x - \bar{x})(y - \bar{y}).$$

23. Compruebe que:
$$SS_{xy} = \Sigma(x - \bar{x})y.$$

24. Demuestre que $r = \text{cov}(x, y)/s_x s_y$, donde s_x es la desviación estándar de los puntajes x, y s_y de los puntajes y.

25. Sean $U = ax + b$ y $V = cy + d$ con a y c mayores que cero, compruebe que el coeficiente de correlación de Pearson entre U y V es idénticamente igual al que existe entre x y y.

26. Verifique que $\text{cov}(z_x, z_y) = \dfrac{\Sigma z_x z_y}{n-1} = r$.

SECCIÓN 4.3 *Regresión y predicción*

En la sección 4.2 aprendimos cómo determinar la fuerza de la relación lineal entre dos variables usando diagramas de dispersión y el coeficiente de correlación r. Si la fuerza de la relación lineal se determina usando el coeficiente de correlación r y ésta resulta ser alta, puede ser deseable describir la relación en términos de una ecuación; determinar la relación

lineal requiere el estudio de regresión, como veremos después; una ecuación de regresión puede usarse con propósitos de predicción.

Una relación lineal entre dos variables x y y puede definirse por la **ecuación lineal** $y = b + mx$; la constante m representa la **pendiente** de la línea recta y la b representa la **intercepción** con el eje y. La relación entre temperaturas Farenheit y Celsius es una relación lineal. Su ecuación es

$$F^\circ = 32 + \frac{9}{5} C^\circ$$

Dada una temperatura en grados centígrados, digamos 25, podemos determinarla en grados Farenheit:

$$\begin{aligned} F^\circ &= 32 + \frac{9}{5} C^\circ \\ &= 32 + \frac{9}{5}(25) \\ &= 32 + 45 = 77^\circ \end{aligned}$$

Suponga que estamos interesados en estudiar la relación entre calificaciones del SAT (matemáticas) y los GPA de los estudiantes de primer año; más aún, después de construir un diagrama de dispersión para los datos bivariados de la clase de primer año del año pasado y determinar el coeficiente de correlación r, decidimos encontrar la relación lineal, llamada **ecuación de regresión.** Usando el método apropiado (presentado a continuación), hemos determinado que la ecuación es $\hat{y} = -1.33 + 0.007x$, donde x representa la calificación en matemáticas en el SAT y $\hat{y}$ (léase "y testada") representa el GPA predicho para el fin del primer año; esta ecuación puede usarse con propósitos predictivos. Supongamos también que María es estudiante regular de preparatoria y que ha solicitado su admisión a la universidad, presenta el examen y recibe una calificación de 480 en matemáticas; gracias a la ecuación de regresión, $\hat{y} = -1.33 + 0.007x$, podemos predecir su éxito durante el primer año en la universidad; para determinar el GPA predicho, sustituimos $x = 480$ en la ecuación de regresión $\hat{y} = -1.33 + 0.007x$ y despejamos $\hat{y}$.

$$\begin{aligned} \hat{y} &= -1.33 + 0.007x \\ &= -1.33 + (0.007)(480) = 2.03 \end{aligned}$$

Nuestra predicción para el GPA de María al final del primer año sería 2.03.

La universidad desearía adoptar una política de admisión según la cual ningún estudiante debe ser admitido si no tiene una predicción de GPA al menos de 1.25 para su primer año. ¿Cuál sería la calificación de "rechazo" en el SAT, para la universidad, en la que el estudiante tendría una predicción en el GPA de al menos 1.25? Esto puede encontrarse despejando x de la ecuación siguiente cuando $\hat{y} = 1.25$:

$$\hat{y} = -1.33 + 0.007x$$

Si sustituimos 1.25 para $\hat{y}$ tenemos:

$$1.25 = -1.33 + 0.007x$$

Y si despejamos x resulta:

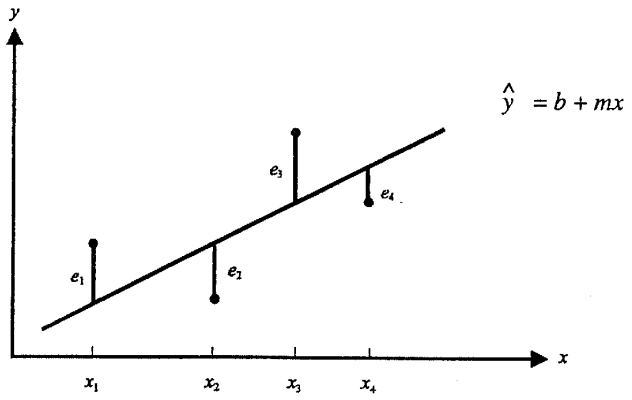
$$2.58 = 0.007x$$
$$x = 369$$

Así, la calificación de rechazo en el SAT sería 369 y a cualquier estudiante con una calificación debajo de 369 se le negaría la admisión a causa del pobre desempeño mostrado; desde luego, esta situación de predecir el éxito no es tan simple y deben ponerse en juego muchas otras variables.

En muchas aplicaciones prácticas que utilizan datos bivariados, los puntos del diagrama de dispersión no caen en una línea recta; entonces el problema es identificar una línea recta cercana a todos los puntos del diagrama de dispersión, donde la "cercanía" se juzga mediante los cuadrados de las distancias verticales de los puntos a la línea recta. Esta recta se presenta por la ecuación $\hat{y} = b + mx$ y se llama la **recta de mejor ajuste** o **recta de regresión;** el símbolo m representa la pendiente de la recta y el b, la intercepción con el eje y. Para esta recta, la suma de los cuadrados de las distancias verticales es lo más pequeña posible; el procedimiento para determinar la recta de mejor ajuste se llama **método de los mínimos cuadrados**; este método identificará, entre todas las rectas que pueden dibujarse en un diagrama de dispersión, a la que produce la suma mínima de los cuadrados de las desviaciones de los puntos del diagrama respecto a la recta.

Con el propósito de ilustrar, suponga que nuestro diagrama de dispersión tiene sólo cuatro puntos. Si e_1 expresa la distancia del primer punto a alguna recta representada por $\hat{y} = b + mx$, entonces la distancia vertical corresponde al error de usar la recta para predecir y_1 usando x_1 (véase la figura 4.9). Así, $e_i = y_i - \hat{y}_i$ expresa el error al predecir el i-ésimo valor de y usando el i-ésimo valor de x.

FIGURA 4.9
Distancias verticales de puntos a la recta de regresión



Si la recta de mejor ajuste es $\hat{y} = b + mx$, entonces por el principio del método de los mínimos cuadrados Σe_i^2 es un mínimo. Esto es,

$$\Sigma e_i^2 = e_1^2 + e_2^2 + e_3^2 + e_4^2$$
$$= (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + (y_3 - \hat{y}_3)^2 + (y_4 - \hat{y}_4)^2$$

es un mínimo. En general, la suma Σe_i^2 se denomina la **suma de cuadrados de los errores** y se denota por SSE. Por tanto, tenemos

Suma de cuadrados de los errores

$$\text{SSE} = \Sigma e_i^2 = \Sigma(y - \hat{y})^2 \tag{4.7}$$

Deberá notarse que, puesto que $e_i = y_i - \hat{y}_i$, los puntos arriba de la recta tendrán errores positivos, los ubicados en la recta errores cero y los puntos bajo la recta tendrán errores negativos; si se suman los errores de todos los puntos, siempre resultará una suma de cero. Esta es la razón por la que los errores deben elevarse al cuadrado antes de sumarlos.

EJEMPLO 4.9

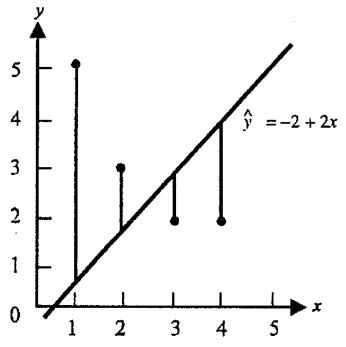
Considere los datos siguientes.

x	3	2	4	1
y	2	3	2	5

Calculemos SSE para alguna recta, digamos $\hat{y} = -2 + 2x$ dibujada en el diagrama de dispersión; la recta representada por la ecuación $\hat{y} = -2 + 2x$ está dibujada en el diagrama de dispersión ilustrado en la figura 4.10. La tabla siguiente se usará para organizar los cálculos:

x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
3	2	4	−2	4
2	3	2	1	1
4	2	6	−4	16
1	5	0	5	25
			0	

FIGURA 4.10
Diagrama de dispersión que muestra la recta $\hat{y} = -2 + 2x$



La primera entrada en la columna $\hat{y}$ se encontró sustituyendo $x = 3$ en la ecuación $\hat{y} = -2 + 2x$ despejando $\hat{y}$:

$$\hat{y} = 2 + 2x$$
$$= -2 + (2)(3) = 4$$

Note que la suma de los errores es:

$$\Sigma(y - \hat{y}) = -2 + 1 - 4 + 5 = 0$$

Y la suma de los cuadrados de los errores:

$$\text{SSE} = \Sigma (y - \hat{y})^2$$
$$= 4 + 1 + 16 + 25 = 46$$

Por lo tanto, para la recta representada por $\hat{y} = 2 + 2x$, SSE = 46.

Si ninguna otra recta que se dibuje en el diagrama de dispersión del ejemplo 4.9 produce un valor de SSE menor que 46, entonces la recta representada por la ecuación $\hat{y} = -2 + 2x$ es *la recta de regresión* o *la recta de mejor ajuste*. Desde luego, no sería productivo un método de ensayo y

error para seleccionar la mejor recta según el criterio de los mínimos cuadrados; por suerte, la determinación de m y b en la ecuación $\hat{y} = b + mx$ que minimice SSE puede realizarse usando álgebra o derivadas parciales de cálculo, y los detalles pueden omitirse. Las fórmulas de mínimos cuadrados para encontrar m y b son como sigue:

Constantes para la recta regresión

$$m = \frac{SS_{xy}}{SS_x} \tag{4.8}$$

$$b = \bar{y} - m\bar{x} \tag{4.9}$$

APLICACIÓN 4.5

Para los datos en el ejemplo 4.9, encuentre la ecuación de regresión y calcule SSE.

Solución: Los cálculos se organizan usando la tabla siguiente:

	x	y	x^2	y^2	xy
	3	2	9	4	6
	2	3	4	9	6
	4	2	16	4	8
	1	5	1	25	5
Sumas	10	12	30	42	25

La suma de cuadrados para x es:

$$SS_x = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 30 - \frac{(10)^2}{4} = 5$$

La suma de productos cruzados es:

$$SS_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} = 25 - \frac{(10)(12)}{4} = -5$$

Por la fórmula (4.8), la pendiente de la recta de regresión será:

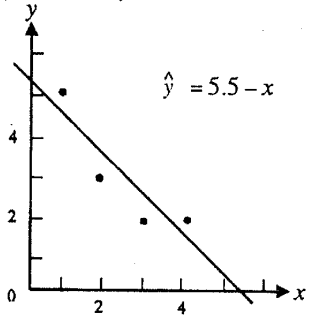
$$m = \frac{SS_{xy}}{SS_x} = \frac{-5}{5} = -1$$

Por la fórmula (4.9), la intercepción de la recta de regresión con el eje y será:

$$b = \bar{y} - m\bar{x} = \frac{12}{4} - (-1)\frac{10}{4} = 3 + 2.5 = 5.5$$

FIGURA 4.11

Diagrama de dispersión para la aplicación 4.5



De esta manera, la ecuación de regresión es $\hat{y} = 5.5 - x$. Su gráfica aparece en el diagrama de dispersión en la figura 4.11.

Para encontrar SSE, organizamos nuestros cálculos en la tabla siguiente:

x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
3	2	2.5	−0.5	0.25
2	3	3.5	−0.5	0.25
4	2	1.5	0.5	0.25
1	5	4.5	0.5	0.25
			0	1

De la última columna de la tabla, tenemos:

$$\begin{aligned} SSE &= \Sigma(y - \hat{y})^2 \\ &= 0.25 + 0.25 + 0.25 + 0.25 \\ &= 1 \end{aligned}$$ ■

Para los datos usados en el ejemplo 4.9, de todas las rectas que se pueden dibujar en el diagrama de dispersión, la recta de regresión produce la suma mínima de cuadrados de errores, que es 1; cualquier otra recta produce una suma de cuadrados de errores mayor que 1; la recta cuya ecuación es $y = -2 + 2x$ produce una suma de cuadrados de errores de 46. Es interesante observar que el punto $(\bar{x}, \bar{y})$ está siempre en la recta de regresión.

EJEMPLO 4.10

Para los datos en el ejemplo 4.9, $\bar{x} = 2.5$ y $\bar{y} = 3$ y el punto (2.5, 3) satisface la ecuación $\hat{y} = 5.5 - x$, porque $3 = 5.5 - 2.5$. Así, el punto (2.5, 3), en consecuencia está en la recta.

Para un gran conjunto de datos bivariados, sería demasiado tardado seguir el método anterior para encontrar SSE; en su lugar, se puede usar la fórmula siguiente para obtener SSE.

Fórmula para calcular SSE

$$SSE = SS_y - m\,SS_{xy} \tag{4.10}$$

APLICACIÓN 4.6

Para los datos en la aplicación 4.5 encuentre SSE usando la fórmula (4.10).

Solución: Calculamos primero SS_y usando la tabla en la aplicación 4.5:

$$\begin{aligned} SS_y &= \Sigma y^2 - \frac{(\Sigma y)^2}{n} \\ &= 42 - \frac{(12)^2}{4} = 6 \end{aligned}$$

Mediante la fórmula (4.10) obtenemos:

$$\begin{aligned} SSE &= SS_y - m\ SS_x \\ &= 6 - (-1)(-5) \\ &= 6 - 5 = 1 \end{aligned}$$

En consecuencia, SSE = 1, como se calculó en la aplicación 4.5. ■

APLICACIÓN 4.7

Para los datos de la marca mundial de la milla en la tabla 4.3, encuentre la ecuación de regresión y úsela para determinar el año para el cual el tiempo predicho para correr la milla será 3:40.

Solución: De la aplicación 4.4, tenemos:

$$\begin{aligned} SS_x &= 1625.7143 \\ SS_{xy} &= -10.6421429 \end{aligned}$$

También, la media de x es:

$$\bar{x} = \frac{\Sigma x}{n} = \frac{292}{14} = 20.85714286$$

y la media de y:

$$\bar{y} = \frac{\Sigma y}{n} = \frac{54.466}{14} = 3.890428571$$

Obtenemos la pendiente de la recta de regresión aplicando la fórmula (4.8):

$$m = \frac{SS_{xy}}{SS_x} = \frac{-10.6421429}{1625.7143} = -0.006546133$$

Y por la fórmula (4.9), la intercepción de la recta de regresión es:

$$b = \bar{y} - m\bar{x} = 3.890428571 - (-0.006546133)(20.85714286) = 4.026962213$$

Por lo tanto, la ecuación de regresión será:

$$\begin{aligned} \hat{y} &= b + mx \\ &= 4.026962213 - 0.006546133x \end{aligned}$$

Si $y = 3{:}40 = 3.667$ minutos y despejamos x resulta $x = 55.0$. Como los años se codificaron mediante la transformación x = año – 1944, el año en que ocurrirá el 3:40 en la milla será:

$$\begin{aligned} \text{Año} &= x + 1944 \\ &= 55 + 1944 \\ &= 1999 \end{aligned}$$

Así, el año en que se predice un tiempo de 3:40 para correr la milla es 1999; note que el año 1999 no se obtuvo usando la predicción en un sentido normal, donde x se usa para predecir y; para hacer eso, necesitaríamos determinar la ecuación de regresión usando la aproximación de los mínimos cuadrados para predecir y de x (véase el ejercicio 26 al final de esta sección). ■

La ecuación de regresión en la aplicación 4.7 da lugar a preguntas interesantes. ¿Qué ocurre con los tiempos de 3 o 2 minutos para la milla? Encontramos difícil creer que algún ser humano llegue a correr la milla en 2 minutos; incluso, la ecuación de regresión producirá un valor de $x = 310$ si $y = 2$; si decodificamos el valor de x encontramos que la ecuación de regresión estima que los 2 minutos para la milla se alcanzarán en el año 2254. Deberemos ser siempre cuidadosos al hacer predicciones alejadas de los valores de la variable x contenidos en los datos muestrales; en nuestro ejemplo, los valores registrados para la variable x representan años de 1945 a 1985; con propósitos predictivos, sólo deben usarse valores de x iguales o cercanos a estos valores.