LECCIÓN
CONDENSADA

# 13.1 Distribuciones de probabilidad

En esta lección

- Dibujarás la gráfica de la **distribución de probabilidad** de una **variable aleatoria continua**
- Encontrarás unas probabilidades al hallar o aproximar las áreas bajo una curva de distribución de probabilidad
- Extenderás las definiciones de **moda, mediana,** y **media** a las distribuciones de probabilidad

En tiempos electorales, a menudo las estaciones de televisión, los diarios, y las revistas llevan a cabo encuestas. Al encuestar una pequeña **muestra** de votantes, esperan obtener información sobre cómo se siente la **población** completa de votantes sobre cierto candidato o ciertas cuestiones. En capítulos anteriores aprendiste algunas *estadísticas*—por ejemplo, la media, la mediana, y la desviación estándar—que se pueden usar para describir una muestra. Los números correspondientes que describen a la población completa se llaman los **parámetros.** Cuánto más grande sea la muestra, más cerca estarán sus estadísticas a los parámetros.

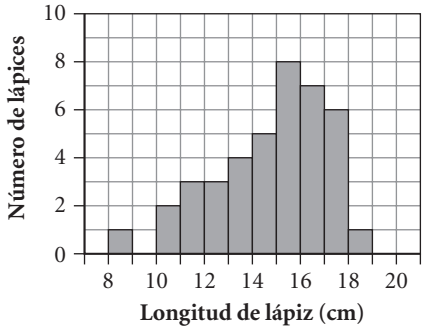
En el capítulo anterior, trabajaste con unas variables aleatorias discretas. Los datos tenían valores enteros—por ejemplo, 10 caras o 3 cruces. En ocasiones, los datos pueden tomar cualquier valor real dentro de un intervalo. Esto se representa mediante una **variable aleatoria continua.** Por ejemplo, la altura de una persona elegida al azar es una variable aleatoria continua. Una persona podría tener una altura de 165 cm o de 166 cm, pero cualquier medida entre tales mediciones enteras también es posible (por ejemplo, 165.25 cm ó 165.67897 cm).

## Investigación: Longitud de lápices

La investigación en tu libro requiere que reúnes unos datos sobre las longitudes de los lápices de todos tus compañeros de aula. Los resultados siguientes se basan en tales datos. (Las longitudes están en centímetros.)

{16.9, 18.7, 11.3, 13.8, 15.2, 17.0, 16.5, 16.6, 11.8, 17.2, 15.5, 15.7, 17.0, 11.4, 16.5, 16.0, 13.4, 15.7, 15.5, 14.1, 12.3, 13.8, 15.5, 15.7, 10.7, 15.6, 12.1, 14.4, 16.5, 17.9, 8.2, 17.8, 17.6, 14.1, 16.7, 14.6, 12.3, 10.0, 13.2, 14.3}

Crea un histograma de estos datos con columnas que representan incrementos de 1 cm. El histograma debe parecerse al siguiente.




Divide el número de lápices de cada columna entre el número total de lápices. Debes obtener los siguientes resultados:

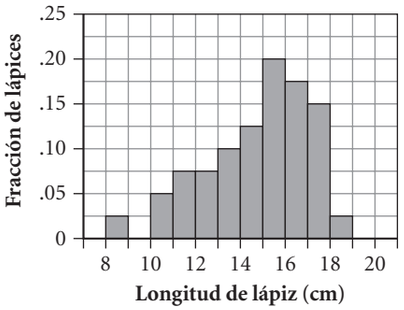
| | | | |
|---|---|---|---|
| 8–9: .025 | 9–10: 0 | 10–11: .05 | 11–12: .075 |
| 12–13: .075 | 13–14: .1 | 14–15: .125 | 15–16: .2 |
| 16–17: .175 | 17–18: .15 | 18–19: .025 | |

Haz un segundo histograma usando estos nuevos valores como los valores *y*. Tu histograma debe parecerse al que presentamos en la próxima página. Esta gráfica tiene la misma forma que la anterior, pero la escala vertical es diferente.

**(continúa)**

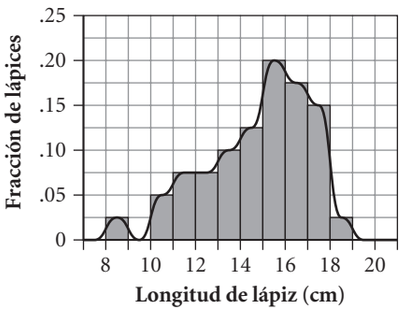
## Lección 13.1 • Distribuciones de probabilidad (continuación)

La altura de cada barra representa la fracción de lápices con longitudes dentro del intervalo dado. Como el ancho de cada barra es 1, el área de cada barra también representa la fracción de lápices dentro del intervalo. Debido a que se tomaron en cuenta todos los lápices, el área total de todas las barras debe ser 1. Puedes verificar esto sumando las áreas. Observa que el área de cada cuadrado de la cuadrícula es .025.




Imagina que reúnas y mides cada vez más lápices y dibujas un histograma usando la fracción de lápices como la altura de la columna. Dibuja un histograma correspondiente a un número infinito de lápices. Asegúrate de poder justificar la forma de tu histograma.

Imagina que haces una encuesta completa y precisa de todos los lápices del mundo. Supón que la distribución de longitudes sea aproximadamente la misma que la de la muestra anterior. Además, supón que utilices un número infinito de columnas delgadas. (Cada ancho de columna representa una fracción infinitamente pequeña de un centímetro.) Para aproximar esta gráfica, dibuja una curva lisa sobre la parte superior de tu histograma. Haz que el área entre la curva y el eje horizontal sea aproximadamente la misma que el área del histograma. Trata de asegurar que el área extra encerrada por la curva por encima del histograma sea igual que el área cortada en las esquinas de las columnas, como en la curva que se muestra aquí.




Usa tu curva para estimar las áreas descritas en el Paso 7 en tu libro. Las siguientes estimaciones se basan en la curva que se muestra aquí.

**a.** Aproximadamente .025

**b.** Aproximadamente 3(.025), ó .075

**c.** Aproximadamente 32.5(.025), ó .8125

**d.** 0

El segundo histograma que hiciste en la investigación, donde se muestra la fracción de lápices en cada columna, se llama un **histograma de frecuencias relativas.** La curva lisa que dibujaste se aproxima a la **distribución de probabilidad** de una variable aleatoria continua para el conjunto infinito de mediciones.

Las áreas que encontraste son las probabilidades de que la longitud de un lápiz escogido al azar satisfaga la condición dada. Si $x$ representa la variable aleatoria continua que da las longitudes de los lápices en centímetros, entonces puedes escribir estas áreas como

$P(x < 10)$ $P(11 < x < 12)$ $P(x > 12.5)$ $P(x = 11)$

En una distribución de probabilidad continua, la probabilidad de tener cualquier resultado sencillo, tal como la probabilidad de que la longitud de un lápiz sea 11 cm, es el área de un segmento de recta, es decir, 0. Teóricamente es posible que un lápiz tenga 11 cm de longitud, pero la probabilidad de escoger un resultado entre un número infinito de ellos es 0. Lee el Ejemplo A en tu libro, donde se ilustra cómo las áreas representan las probabilidades para una variable aleatoria continua.

Después del Ejemplo A, tu libro define la moda, la mediana, y la media de una distribución de probabilidad. Estas definiciones están relacionadas, aunque son un tanto diferentes, con las definiciones que aprendiste anteriormente. Léelas atentamente y asegúrate de que tengan sentido para ti. Después aplica las nuevas definiciones, trabajando el Ejemplo B.

LECCIÓN CONDENSADA

# 13.2 Distribuciones normales

En esta lección

- Descubrirás que la gráfica de una distribución binomial es una curva con forma de campana, que se llama una **curva normal**
- Aprenderás la ecuación de una **distribución normal** con media $\mu$ y desviación estándar $\sigma$
- Usarás las funciones de una calculadora para **graficar una curva normal** y **encontrar áreas** debajo de una porción de la curva

En el Capítulo 12, estudiaste la distribución binomial para variables aleatorias discretas. En general, si un experimento tiene dos resultados, *éxito* y *fracaso,* con una probabilidad de éxito $p$ y una probabilidad de fracaso $q$, entonces la probabilidad de $x$ éxitos en $n$ ensayos es $P(x) = {}_nC_x p^x q^{n-x}$. Observa que $q = 1 - p$, de modo que esto es equivalente a $P(x) = {}_nC_x p^x (1 - p)^{n-x}$. En esta lección, descubrirás algunas propiedades de esta distribución de probabilidad.

## Investigación: La campana

Completa la investigación en tu libro. Después compara tus resultados con los siguientes.

**Paso 1** $p = .5$ y $1 - p = .5$; esto es, la probabilidad de obtener una cara en cualquiera de 5 ensayos es .5, y la probabilidad de no obtener una cara (es decir, obtener una cruz) es .5. Teóricamente, la mitad de los lanzamientos serán caras, de modo que 7 y 8 darán el valor máximo para $P(x)$.

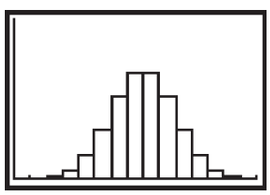
**Paso 2** He aquí la tabla completa:

| **Caras** ($x$) | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| $P(x)$ | .000 | .000 | .003 | .014 | .042 | .092 | .153 | .196 |

| **Caras** ($x$) | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|
| $P(x)$ | .196 | .153 | .092 | .042 | .014 | .003 | .0000 | .000 |

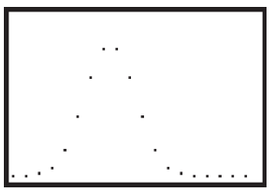
La suma es el total de las probabilidades de todos los resultados posibles de cualquier experimento, 1.

**Paso 3** Este histograma tiene forma de loma y es simétrico. Los valores $y$ van de 0 a aproximadamente .196. El valor máximo se presenta en $x = 7$ y $x = 8$.



[0, 15, 1, 0, 0.25, 0.1]

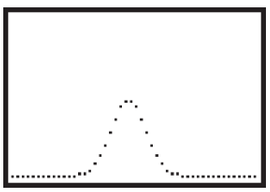
**Paso 4** La función se define para los valores enteros que van de 0 a 15. La gráfica consiste en puntos discretos que adquieren una forma de loma simétrica, con la misma forma del histograma del Paso 3. La media, la mediana, y la moda son todas iguales a 7.5.



[0, 18.8, 1, −0.01, 0.25, 0.1]

**(continúa)**

**Paso 5** La función describe las distribuciones de probabilidad cuando se lanzan 45 monedas al mismo tiempo. La forma es la misma que la del Paso 4, pero más amplia y más corta. El dominio son los valores enteros desde 0 hasta 45. El rango va de 0 hasta aproximadamente .117. La media, la mediana, y la moda son todas iguales a 22.5.



[0, 47, 1, −0.01, 0.25, 0.1]

**Paso 6** $\bar{x} = 22.5$, $s \approx 3.354$

**Paso 7** Si el número de monedas aumentara, el dominio aumentaría y el rango disminuiría. La forma general sería la misma, pero la gráfica sería más amplia, más lisa, y más plana. La media, la mediana, y la moda se moverían a $\frac{n}{2}$, donde $n$ es el número de monedas. La desviación estándar sería más grande debido a la mayor dispersión.

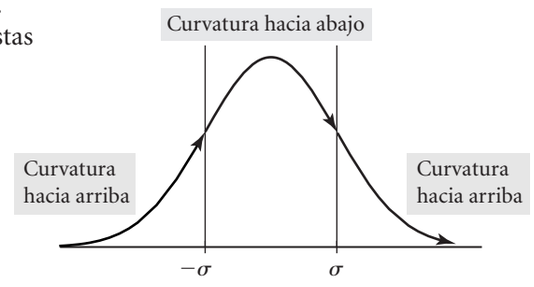
A medida que $n$ se hace cada vez más grande, la distribución binomial se ve cada vez más continua hasta que adquiere la apariencia de la curva en forma de campana que se muestra aquí. Las distribuciones de poblaciones grandes a menudo tienen esta forma. Una curva con esta forma se llama una **curva normal,** y una distribución con esta forma se llama una **distribución normal.**

Observa que utilizas $\bar{x}$ y $s$ para representar la media y la desviación estándar de una muestra, pero usas $\mu$ y $\sigma$ (se pronuncian "myu" y "sigma") para representar la media y la desviación estándar de una población entera.

Empezando en la página 735 en tu libro, se analiza la ecuación de la gráfica de la distribución normal. Lee ese texto y trabaja el Ejemplo A. En el ejemplo, graficas la ecuación general de la curva normal, escribes la ecuación de una distribución normal estándar con media $\mu$ y desviación estándar $\sigma$, y después escribes una ecuación para una curva normal que se ajuste a una distribución binomial dada. La ecuación general de una distribución normal también se da en el recuadro "The Normal Distribution" (la distribución normal) en la página 737.

Puede resultar tedioso introducir la ecuación de la distribución normal en una calculadora. Afortunadamente, la mayoría de las calculadoras proporcionan la ecuación como una función preconfigurada. Sólo tienes que proporcionar la media y la desviación estándar. (Consulta **Calculator Note 13B** para aprender cómo graficar una distribución normal en tu calculadora.)

En este capítulo se usa la notación $n(x,$ *media, desviación estándar*$)$ para representar una distribución normal. Así, por ejemplo, $n(x, 2.6, 1.5)$ representa una distribución normal con media 2.6 y desviación estándar 1.5. La **función distribución normal estándar**—es decir, la función de la distribución normal con media 0 y desviación estándar 1—se denota simplemente como $n(x)$. La notación $N($*inferior, superior, media, desviación estándar*$)$ se usa para representar el área debajo de una parte de la curva normal. Por ejemplo, $N(2, 3, 2.6, 1.5)$ representa el área entre $x = 2$ y $x = 3$ debajo de la curva de la distribución normal con media 2.6 y desviación estándar 1.5. (Consulta **Calculator Note 13C** para aprender cómo hallar estas áreas en tu calculadora.) Lee el Ejemplo B en tu libro.




En los puntos que están a una desviación estándar de la media, la curva normal cambia, de una curvatura hacia abajo a una curvatura hacia arriba. Estos puntos se llaman los **puntos de inflexión.** Puedes estimar la desviación estándar de una distribución normal, ubicando los puntos de inflexión de su gráfica.

LECCIÓN CONDENSADA

# 13.3 Valores *z* e intervalos de confianza

En esta lección

- Descubrirás la **regla 68-95-99.7** para determinar la probabilidad de que un valor de datos esté dentro de una, dos, o tres desviaciones estándares de la media
- Aprenderás cómo transformar los valores $x$ de una distribución normal en **valores** $z$
- Calcularás unos **intervalos de confianza**

Conocer cómo un valor de una muestra se relaciona con el valor medio no te dice si el valor es típico. Por ejemplo, decir que el peso de un perro terrier escocés es de 4 lb por encima de la media no te dice si esta medida es un evento raro o un evento común. Si supieras cuántas desviaciones estándares tiene el peso del perro de la media, podrías evaluar si tal peso es común.

## Investigación: Áreas y distribuciones

Completa la investigación en tu libro, a suponiendo que la medición que hiciste es de 72.4 cm. Supón, también, que todos los compañeros de tu aula obtuvieron respuestas parecidas en el Paso 2. Cuando termines, compara tus resultados con los siguientes.

**Paso 1** Tu gráfica debe ser una curva con forma de campana, simétrica con respecto a $x = 72.4$, con puntos de inflexión en aproximadamente 71.6 y 73.2 (los valores $x$ que están a una desviación estándar de la media).

**Paso 2**

**a.** $N(71.6, 73.2, 72.4, 0.8) \approx .683$, ó aproximadamente 68%

**b.** $N(72.4 - 2(0.8), 72.4 + 2(.8), 72.4, 0.8) = N(70.8, 74, 72.4, 0.8) \approx .954$, ó aproximadamente 95%

**c.** $N(72.4 - 3(0.8), 72.4 + 3(.8), 72.4, 0.8) = N(70, 74.8, 72.4, 0.8) \approx .997$, ó aproximadamente 99.7%

**Paso 3** Para una distribución normal, existe un 68% de probabilidad de que un valor esté dentro de una desviación estándar de la media, un 95% de probabilidad de que un valor esté dentro de dos desviaciones estándares de la media, y un 99.7% de probabilidad de que un valor esté dentro de tres desviaciones estándares de la media.

En una distribución normal, el **valor *z*** de $x$ es el número de desviaciones estándares a las cuales $x$ se encuentra de la media. En la investigación, encontraste que la probabilidad de que una nueva medida tenga un valor $z$ entre $-1$ y 1 es 68%, la probabilidad de que tenga un valor $z$ entre $-2$ y 2 es 95%, y la probabilidad de que tenga un valor $z$ entre $-3$ y 3 es 99.7%.

Puedes pensar en los valores $z$ de $x$ como la imagen de $x$ bajo una transformación que traslada y estira o encoge la distribución normal, para convertirla en la distribución normal estándar $n(x)$ con media $\mu = 0$ y desviación estándar $\sigma = 1$. La transformación de valores $x$ en valores $z$ se llama **estandarización de la variable** y se puede calcular con la ecuación $z = \frac{x - \mu}{\sigma}$, donde $\mu$ y $\sigma$ son la media y la desviación estándar de la distribución normal de $x$. Trabaja el Ejemplo A en tu libro, que ilustra la estandarización de la variable. Observa que, cuando se usa un proceso de tanteos en la parte c, debes probar solamente los intervalos que son simétricos con respecto a la media.

**(continúa)**

## Lección 13.3 • Valores *z* e intervalos de confianza (continuación)

No hay manera de saber con certeza de la cercanía entre la media de la población normalmente distribuida y la media de una muestra. Sin embargo, puedes describir el nivel de confianza que tienes de que la media de la población se encuentre dentro de un intervalo dado, centrada en la media de la muestra. Un **intervalo de confianza** $p$% es un intervalo alrededor de la media de la muestra, $\bar{x}$, en el que puedes tener $p$% de confianza de que la media de la población, $\mu$, se encuentre ahí. Específicamente, si $z$ es el número de desviaciones estándares desde la media, dentro de la cual se encuentra el $p$% de los datos normalmente distribuidos, entonces el intervalo de confianza $p$% de una muestra de tamaño $n$ es

$$\bar{x} - \frac{z\sigma}{\sqrt{n}} < \mu < \bar{x} + \frac{z\sigma}{\sqrt{n}}$$

En muchas situaciones reales, no conocerás la desviación estándar de la población. Sin embargo, si el tamaño de la muestra es lo suficientemente grande, por lo general $n > 30$, puedes usar la desviación estándar de la muestra, $s$, en lugar de $\sigma$.

La regla 68-95-99.7 te dice cuáles valores $z$ debes usar si quieres tener una confianza de 68%, 95%, ó 99.7%. La tabla en la página 748 de tu libro da los valores $z$ para algunos otros intervalos de confianza de uso común. En el Ejemplo B en tu libro, se calculan los intervalos de confianza utilizando los valores de la tabla. Lee ese ejemplo, y después intenta resolver el problema en el ejemplo siguiente.

***EJEMPLO***

El gerente de control de calidad de una empresa que empaca cereal sacó una muestra aleatoria de 30 cajas del cereal *Morning Crunch* de las líneas de producción y pesó el contenido de cada caja. El peso medio de la muestra fue de 9.8 onzas y la desviación estándar fue de 0.42 onzas.

**a.** Encuentra los intervalos de confianza de 68% y 90%.

**b.** La etiqueta de la caja de *Morning Crunch* dice que el peso es de 10 onzas. ¿Crees que el gerente de control de calidad debe informar que hay un problema con el peso de las cajas? Explica tu respuesta.

▸ ***Solución***

**a.** Usando la desviación estándar de la muestra en lugar de $\sigma$, el intervalo de confianza de 68% es

$$\left(9.8 - \frac{1(0.42)}{\sqrt{30}},\ 9.8 + \frac{1(0.42)}{\sqrt{30}}\right), \text{ ó aproximadamente } (9.72,\ 9.88)$$

El gerente de control de calidad tiene una confianza de 68% de que el peso medio está entre 9.72 y 9.88 onzas.

En la tabla de tu libro se indica que el valor $z$ correspondiente a un intervalo de confianza de 90% es 1.645, de modo que el intervalo de confianza de 90% es

$$\left(9.8 - \frac{1.645(0.42)}{\sqrt{30}},\ 9.8 + \frac{1.645(0.42)}{\sqrt{30}}\right), \text{ ó aproximadamente } (9.67,\ 9.93)$$

El gerente de control de calidad tiene una confianza de 90% de que el peso está entre 9.67 y 9.93 onzas.

**b.** De acuerdo con la respuesta a la parte a, el gerente de control de calidad puede tener una certeza del 90% de que la población media (es decir, el peso medio de todas las cajas producidas) se encuentra dentro del intervalo (9.67, 9.93). Este intervalo no incluye la cifra de 10 onzas, que es el peso que se alega en la caja, y por eso el gerente debe informar que hay problemas con el peso.

LECCIÓN CONDENSADA

# 13.4 El Teorema del límite central

En esta lección

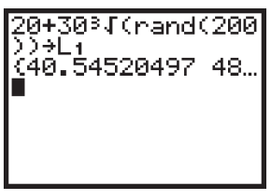
- Descubrirás cómo la media de una muestra se relaciona con la media de la población, en los casos en que la población no está distribuida normalmente
- Aplicarás el **Teorema del límite central** para probar una afirmación con respecto a una media de población
- Usarás el proceso de **inferencia** para resolver un problema de la vida real

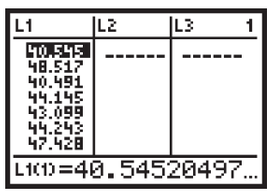
En la Lección 13.3, usaste una muestra estadística para estimar los parámetros de una población distribuida normalmente. En esta lección, explorarás lo que te puede decir una muestra en los casos en que la población no está distribuida normalmente.

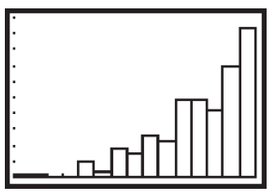
## Investigación: Medias de muestras

El Paso 1 te pide crear una población de un tipo particular: uniforme, normal, sesgada a la izquierda, o sesgada a la derecha. Abajo se muestran los resultados para una población sesgada a la izquierda. Debes seguir los pasos por tu cuenta para al menos una de las otras posibilidades. (Consulta **Calculator Note 13D** para crear una lista de 200 valores, cada uno de ellos entre 20 y 50, para el tipo de población que escoges.)

Los comandos que siguen a la izquierda generan una población sesgada a la izquierda de 200 valores entre 20 y 50. El histograma confirma que los datos están sesgados a la izquierda. (Los valores son aleatorios, así que obtendrás una población diferente y diferentes resultados si generas una población sesgada a la izquierda.)



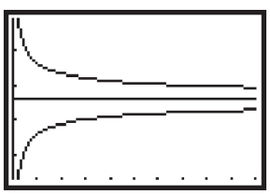



[20, 50, 2, 0, 100, 10]

**Paso 2** La media de la población, $\mu$, es 43.07 y la desviación estándar, $\sigma$, es 5.75.

**Paso 3** El comando L1(randInt(1,200)) elegirá al azar unos valores de datos de la lista L1. La selección de tres valores al azar resulta en la muestra {49.07, 46.93, 47.43}. La media de esta muestra es 47.81. Al sumar dos valores más, elegidos al azar, se obtiene {49.07, 46.93, 47.43, 46.08, 41.72}, cuya media es 46.25. Sumando dos valores más, se obtiene {49.07, 46.93, 47.43, 46.08, 41.72, 42.62, 37.82}, cuya media es 44.52. Las medias de la muestra se acercan bastante a la media de la población, y a medida que agregas más valores, se acercan más a la media de la población.

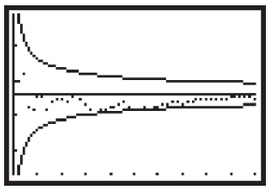
**Paso 4** A continuación se muestran las gráficas de $y = \mu$, $y = \mu - \frac{2\sigma}{\sqrt{x}}$, y $y = \mu + \frac{2\sigma}{\sqrt{x}}$.



[0, 50, 5, 31.57, 54.57, 5]

(continúa)

**Paso 5** La rutina recursiva de la ilustración que sigue agrega valores elegidos al azar, uno por uno, a una muestra y después grafica cada punto (*número muestreado, media*). (En la rutina, *N* es el número de valores muestreados y *T* es el total (suma) de los valores muestreados.) Al ejecutar la rutina 50 veces, se grafican 50 puntos, como se muestra en la gráfica abajo a la derecha.

[0, 50, 5, 31.57, 54.57, 5]

Observa que a medida que aumenta el tamaño de muestra, los puntos (que representan la media de la muestra) se aproximan cada vez más a la recta $y = \mu$ (que representa la media de la población), con las cotas inferior y superior $y = \mu - \frac{2\sigma}{\sqrt{x}}$ y $y = \mu + \frac{2\sigma}{\sqrt{x}}$. En otras palabras, para una muestra de tamaño *n,* la media de la muestra se encuentra dentro de $-\frac{2\sigma}{\sqrt{n}}$ y $\frac{2\sigma}{\sqrt{n}}$ de la media de la población.

**Paso 6** Si repites el Paso 5, obtendrás unos resultados parecidos.

**Paso 7** No importa con qué tipo de distribución de población empieces (uniforme, normal, sesgada a la derecha, etcétera), encontrarás que a medida que aumenta el tamaño de muestra, la media de la muestra se aproxima cada vez más a la media de la población. (¡Asegúrate de verificar esto para al menos un otro tipo de población!)

Acabas de descubrir que la media de una muestra se aproxima a la media de la población, y que la aproximación es mejor para las muestras más grandes. De hecho, las medias de muestra están distribuidas normalmente, aunque esto no sea cierto para la población. Además, puedes usar las medias de muestra para predecir la desviación estándar de la población. Estas observaciones se resumen en el **Teorema del límite central.** Este teorema se encuentra en la página 753 de tu libro. Lee su texto y el párrafo que le sigue.

Trabaja el Ejemplo A, en el cual se aplica el Teorema del límite central para evaluar una afirmación sobre la media de una población. Observa que probar la afirmación de la compañía implica responder la pregunta, "Si la media de la población es realmente 324 mg, ¿cuál es la probabilidad de seleccionar una muestra con una media de 319.96?" En la solución se llega a la conclusión de que podrías esperar una muestra con esta media o una menor solamente el 5.7% del tiempo.

El proceso utilizado en el Ejemplo A se llama **inferencia.** Ésta implica plantear una hipótesis sobre los parámetros de la población (por ejemplo, "La media es 324 mg"), decidir las circunstancias en que la hipótesis sería improbable (por ejemplo, "Una muestra aleatoria de 25 tabletas tomada de la población tiene una media que se presentaría menos del 10% del tiempo"), reunir los datos, y rechazar la hipótesis o aceptarla, basándose en las probabilidades.

El Ejemplo B te lleva por el proceso de inferencia paso a paso. Lee el ejemplo atentamente. Asegúrate de entender el significado de la **hipótesis nula.** Después lee el resto de la lección, que explica el significado de una **muestra aleatoria simple.**

LECCIÓN CONDENSADA

# 13.5 Datos bivariados y correlación

En esta lección

- Determinarás si dos variables están **correlacionadas**
- Usarás el **coeficiente de correlación** para determinar la fuerza de una correlación

Muchos problemas estadísticos de la vida real implican predecir las asociaciones entre dos variables. Por ejemplo, algunos investigadores podrían desear determinar si hay una asociación entre el número de miligramos de vitamina C que una persona consume y el número de resfriados que la persona adquiere. El proceso de recolección de datos sobre dos variables posiblemente relacionadas se llama el **muestreo bivariado.**

En esta lección, aprenderás determinar si existe una asociación lineal entre unas variables y la fuerza de tal asociación. Una asociación lineal entre unas variables se llama una **correlación.** La medida estadística de asociación lineal que más se utiliza es el **coeficiente de correlación.**

## Investigación: Búsqueda de conexiones

**Paso 1** Observa la encuesta que está en el Paso 1 de la investigación en tu libro.

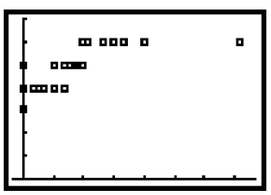
**Paso 2** Enumera al menos dos pares de variables que crees que tendrán una *correlación positiva* (a medida que aumenta una, la otra tiende a aumentar también). Una posibilidad podría ser el número de minutos que se pasa haciendo las tareas (pregunta 1) y el número de clases (pregunta 5).

Ahora, enumera al menos dos pares de variables que crees que tendrán una *correlación negativa* (a medida que aumenta una, la otra tiende a disminuir). Una posibilidad es el número de clases (pregunta 5) y el tiempo que se pasa en platicar con amigos, llamarles por teléfono, mandarles mensajes de e-mail, o escribirles (pregunta 2).

Finalmente, enumera dos pares de variables que crees que tendrán una *correlación débil.* Una posibilidad podría ser el tiempo que se pasa comunicándose con amigos (pregunta 2) y el tiempo que se pasa mirando la televisión o escuchando música (pregunta 3).

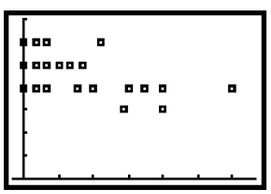
**Paso 3** La tabla que se presenta más adelante muestra los resultados obtenidos por una aula. Introduce los datos en cinco listas de la calculadora. Grafica los puntos correspondientes a cada par de listas y encuentra los coeficientes de correlación de cada par. (Consulta **Calculator Note 13F**.) He aquí los coeficientes de correlación y las gráficas correspondientes a las relaciones mencionadas en el Paso 2. Sin embargo, debes construir gráficas y encontrar los coeficientes de correlación de *todos* los pares posibles.

**1 versus 5**



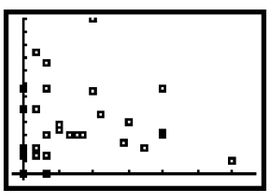
$[-10, 230, 30, 0, 7, 1]$
$r \approx .771$

**2 versus 5**



$[-10, 200, 30, 0, 7, 1]$
$r \approx -.632$

**2 versus 4**



$[-10, 200, 30, -10, 360, 30]$
$r \approx -.059$

**(continúa)**

Debes hacer las siguientes observaciones:

- Las gráficas crecientes tienen coeficientes de correlación positivos y las gráficas decrecientes tienen coeficientes de correlación negativos.
- Cuanto más fuerte sea la correlación, más cerca estará el coeficiente de correlación a $\pm 1$. Las correlaciones débiles tienen coeficientes de correlación cercanos a 0.

| Estudiante | Pregunta 1 | Pregunta 2 | Pregunta 3 | Pregunta 4 | Pregunta 5 |
|---|---|---|---|---|---|
| 1 | 0 | 120 | 60 | 100 | 3 |
| 2 | 20 | 120 | 90 | 200 | 4 |
| 3 | 80 | 65 | 80 | 140 | 6 |
| 4 | 55 | 20 | 220 | 260 | 5 |
| 5 | 10 | 20 | 155 | 200 | 4 |
| 6 | 15 | 0 | 145 | 200 | 4 |
| 7 | 90 | 10 | 80 | 150 | 6 |
| 8 | 215 | 10 | 0 | 60 | 6 |
| 9 | 100 | 0 | 140 | 150 | 6 |
| 10 | 60 | 30 | 120 | 105 | 5 |
| 11 | 65 | 0 | 120 | 150 | 6 |
| 12 | 10 | 60 | 300 | 360 | 4 |
| 13 | 120 | 0 | 0 | 45 | 6 |
| 14 | 30 | 45 | 285 | 90 | 4 |
| 15 | 40 | 60 | 150 | 190 | 4 |
| 16 | 0 | 85 | 150 | 75 | 3 |
| 17 | 0 | 180 | 30 | 30 | 4 |
| 18 | 80 | 0 | 0 | 0 | 6 |
| 19 | 90 | 20 | 0 | 0 | 6 |
| 20 | 45 | 10 | 180 | 285 | 5 |
| 21 | 10 | 120 | 0 | 90 | 4 |
| 22 | 40 | 30 | 100 | 115 | 5 |
| 23 | 0 | 0 | 360 | 60 | 5 |
| 24 | 30 | 50 | 45 | 90 | 5 |
| 25 | 60 | 20 | 30 | 90 | 6 |
| 26 | 45 | 20 | 30 | 45 | 5 |
| 27 | 20 | 105 | 20 | 60 | 4 |
| 28 | 0 | 90 | 0 | 120 | 4 |
| 29 | 50 | 40 | 0 | 90 | 5 |
| 30 | 40 | 10 | 0 | 45 | 4 |

**(continúa)**

**Paso 4** Escribe un párrafo que describe las correlaciones que descubriste. Menciona los pares que no están correlacionados y que tú pensabas que deberían estarlo. He aquí algunas cosas que podrías mencionar: Las correlaciones entre el tiempo que se pasa haciendo las tareas y el número de clases, y entre el tiempo que se pasa mirando televisión y la hora de acostarse, son relativamente fuertes y positivas. Las correlaciones entre el tiempo que se pasa platicando con amigos y el número de clases, y entre el tiempo que se pasa haciendo las tareas y el tiempo que se pasa platicando con amigos, son relativamente fuertes y negativas. Los otros pares de variables parecen no estar correlacionados. Estos datos se obtuvieron de una muestra pequeña que no fue muy aleatoria, de modo es posible que no sean buenos para predecir los resultados de la escuela completa.

El texto situado entre la investigación y el Ejemplo A en tu libro ofrece información sobre el coeficiente de correlación y sobre cómo se derivó su fórmula. Este texto también explica que en la estadística, las variables $x$ y $y$ a menudo se llaman las variables de **explicación** y de **respuesta.** Lee este texto y después trabaja el Ejemplo A, que muestra cómo calcular el coeficiente de correlación usando la fórmula.

Es muy importante no confundir la correlación con la **causalidad.** El hecho de que dos variables estén fuertemente correlacionadas no significa que un cambio en una variable *cause* un cambio en la otra. El Ejemplo B ilustra este punto. Lee dicho ejemplo atentamente.

LECCIÓN CONDENSADA

# 13.6 La recta de mínimos cuadrados

En esta lección

- Aprenderás cómo ajustar la **recta de mínimos cuadrados** a un conjunto de datos
- Descubrirás cómo la recta de mínimos cuadrados adquirió su nombre
- Usarás el **error cuadrático medio** para comparar una recta de mínimos cuadrados con una recta mediana-mediana

En el Capítulo 3, aprendiste cómo ajustar una recta mediana-mediana a los datos. En esta lección, conocerás un tipo diferente de recta de ajuste, que se llama la **recta de mínimos cuadrados.** La ecuación de la recta de mínimos cuadrados es $z_y = rz_x$, donde $r$ es el coeficiente de correlación, y $z_x$ y $z_y$ son los valores $z$ para $x$ y $y$, respectivamente. En la práctica, deseas que la ecuación represente la relación entre $x$ y $y$, no entre sus valores $z$. Usando la definición del valor $z$, puedes reescribir la ecuación como

$$\frac{y - \bar{y}}{s_y} = r\left(\frac{x - \bar{x}}{s_x}\right) \text{ ó } \hat{y} = \bar{y} + r\left(\frac{s_y}{s_x}\right)(x - \bar{x})$$

Para encontrar más detalles sobre la recta de mínimos cuadrados, lee el texto que precede el Ejemplo A en tu libro. Después lee el Ejemplo A, que ilustra cómo encontrar la recta de mínimos cuadrados para un conjunto de datos dado.

## Investigación: Relación de variables

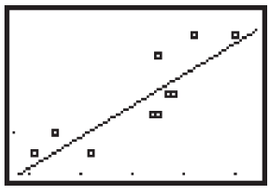
El Paso 1 de la investigación en tu libro te pide obtener mediciones de las partes del cuerpo de tus compañeros de aula. He aquí una muestra de datos correspondientes a 10 estudiantes (con las medidas en centímetros). Usa estos datos para completar los Pasos 2–4, y después compara tus resultados con los siguientes, en donde se busca la relación entre una cuarta (la amplitud de la mano extendida) y la longitud del dedo meñique.

| Estudiante | Cuarta | Largo del pie | Largo del dedo meñique | Estatura | Largo del cúbito | Largo de la pierna inferior | Largo del brazo superior | Ancho de la uña del pulgar |
|---|---|---|---|---|---|---|---|---|
| 1 | 18.5 | 23.5 | 6 | 159 | 41.6 | 47 | 31 | 1.2 |
| 2 | 21.2 | 26 | 6.5 | 170 | 42 | 48 | 36 | 1.3 |
| 3 | 20.8 | 25.8 | 6.2 | 173 | 44.2 | 49.5 | 36.2 | 1.3 |
| 4 | 20.4 | 24.5 | 6.1 | 164 | 42.1 | 48.5 | 34 | 1.3 |
| 5 | 18.1 | 23 | 5.9 | 155 | 39.5 | 46 | 30.2 | 1.2 |
| 6 | 22 | 28.4 | 6.5 | 189 | 47.5 | 52.2 | 40.5 | 1.4 |
| 7 | 20.5 | 29 | 6.4 | 191 | 47.9 | 52.6 | 41 | 1.4 |
| 8 | 20.7 | 26.2 | 6.2 | 175 | 46 | 51 | 34 | 1.3 |
| 9 | 20.5 | 25 | 6.1 | 168 | 44 | 50 | 35.4 | 1.2 |
| 10 | 19.2 | 24.1 | 5.9 | 158 | 43.5 | 48 | 33.6 | 1.1 |

**Paso 2** El coeficiente de correlación para estas variables es aproximadamente .858, de modo que los datos están linealmente relacionados.

**(continúa)**

**Paso 3** Usando una calculadora, $\bar{x} = 20.19$, $\bar{y} = 6.18$, $s_x \approx 1.219$, $s_y \approx 0.225$, y $r = .858$. Sustituyendo estos valores en la ecuación general, se obtiene $\hat{y} = 6.18 + .858\left(\frac{0.225}{1.219}\right)(x - 20.19)$, ó $\hat{y} = 2.981 + 0.158x$. La recta parece ser un buen ajuste.



[17.71, 22.39, 1, 5.798, 6.602, 1]

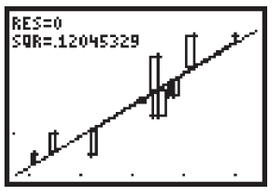
**Paso 4**

**a.** Es la misma recta.

**b.** La suma de los residuos es 0.

**c.** La recta de mínimos cuadrados es la recta para la cual la suma de los cuadrados de los residuos es la más pequeña posible, y la suma de los residuos es 0.

**d.** A medida que el valor de la cuarta aumenta, también lo hace la longitud del dedo meñique. Los datos se modelan bastante bien mediante una recta.




[17.71, 22.39, 1, 5.798, 6.602, 1]

Puedes medir la precisión de una recta de mínimos cuadrados al calcular el error cuadrático medio, del mismo modo en que lo hiciste para la recta mediana-mediana en el Capítulo 3. El Ejemplo B en tu libro ajusta tanto una recta de mínimos cuadrados como una recta mediana-mediana a un conjunto de datos, y determina cuál es el mejor ajuste al calcular el error cuadrático medio. Trabaja ese ejemplo y después intenta el ejemplo siguiente.

***EJEMPLO***

Encuentra la recta de mínimos cuadrados y la recta mediana-mediana para los datos (*largo del cúbito, largo del brazo superior*) de la investigación. Después encuentra el error cuadrático medio de ambos modelos. ¿Cuál recta se ajusta mejor?

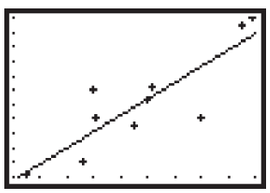
▶ ***Solución***

Usa tu calculadora para hallar ambos modelos y sus errores cuadráticos medios.

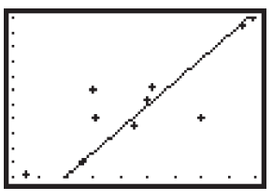
Recta de mínimos cuadrados: $\hat{y} = -13.982 + 1.122x$; error cuadrático medio: 1.923

Recta mediana-mediana: $\hat{y} = -35.904 + 1.610x$; error cuadrático medio: 2.446

Debido a que el error cuadrático medio es menor para la recta de mínimos cuadrados, ésta es un mejor ajuste. Puedes verificar esto de manera visual si graficas las rectas. La gráfica a la izquierda muestra el modelo de mínimos cuadrados. La gráfica a la derecha muestra el modelo mediana-mediana.



[39, 48, 1, 30, 41, 1]



[39, 48, 1, 30, 41, 1]

A menudo la recta de mínimos cuadrados se llama la "recta de mejor ajuste" porque tiene la más pequeña suma de cuadrados de los errores entre los puntos de datos y las predicciones de la recta. Sin embargo, debido a que pone un mismo énfasis en cada punto, la recta de mínimos cuadrados puede verse afectada por los valores externos. Cuando ajustes una recta, siempre resulta una buena idea verificar la recta visualmente. En ocasiones, la recta mediana-mediana u otra recta es un mejor ajuste que la recta de mínimos cuadrados.

LECCIÓN CONDENSADA

# 13.7 Regresión no lineal

En esta lección

- **Linealizarás** unos datos para determinar si podría ajustarse una función de potencias o una exponencial
- Usarás tu calculadora para **ajustar diferentes funciones no lineales** a los datos
- Usarás los **errores cuadráticos medios** y las **gráficas de residuos** para comparar cómo se ajustan las diferentes funciones a los datos.

Has hecho regresión lineal para ajustar una recta a datos que muestran una fuerte tendencia lineal. En esta lección, encontrarás modelos para ajustar a los datos que muestran una clara tendencia no lineal.

En las páginas 780 y 781 de tu libro, se repasan los tipos de funciones que has estudiado en este curso, y se dan la ecuación general y la gráfica de cada tipo. Lee el texto y observa las gráficas.

El Ejemplo A en tu libro muestra que al observar una gráfica de datos, puedes predecir el tipo de función que podría ajustarse (o al menos descartar los tipos que no se ajustarán). Si piensas que un modelo de potencias o uno exponencial, puede ajustarse a tus datos, puedes verificar si son buenos modelos mediante la **linealización** de los datos. Si el ajuste es bueno, entonces puedes usar las técnicas de regresión lineal para encontrar un modelo. Esto se ilustra en el Ejemplo B. Lee Ejemplo A y Ejemplo B atentamente. Después trabaja el ejemplo siguiente.

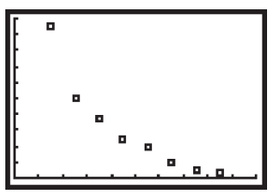
***EJEMPLO***

Encuentra una función que modele estos datos:

| $x$ | 3 | 5 | 7 | 9 | 11 | 13 | 15 | 17 |
|---|---|---|---|---|---|---|---|---|
| $y$ | 19.1 | 10.0 | 7.5 | 4.7 | 4.0 | 1.8 | 1.1 | 0.7 |

► ***Solución***

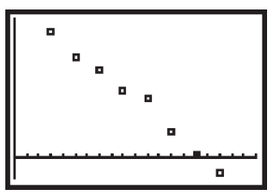
Según la gráfica siguiente, parece que una función exponencial, de potencias, o polinomial podría ajustarse.



[0, 20, 2, 0, 20, 2]

Si los datos son exponenciales, una ecuación de la forma $y = ab^x$ ajustará. Tomando el logaritmo en ambos lados, se obtiene $\log y = \log a + x \log b$. Ésta es una ecuación lineal, en la que las variables son $x$ y $\log y$. Por tanto, si los datos son exponenciales, la gráfica de $(x, \log y)$ será lineal.

La gráfica de $(x, \log y)$, que se muestra aquí, parece ser bastante lineal. (Puedes verificar que una función de potencias *no* es un buen ajuste si graficas $(\log x, \log y)$ y verificas que la gráfica no es lineal.)



[0, 20, 2, −0.2, 1.4, 0.2]

**(continúa)**

Usa la regresión lineal para encontrar que la recta de mínimos cuadrados para estos datos transformados es $\hat{y} = 1.57 - 0.10x$. Ahora sustituye $\hat{y}$ por $(\log y)$ y resuelve para $y$.

$\log y = 1.57 - 0.10x$ Recta de mínimos cuadrados.

$y = 10^{1.57-0.10x}$ Definición de logaritmo.

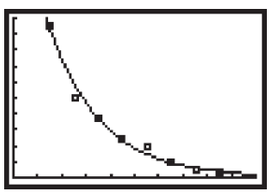
$y = 10^{1.57} \cdot 10^{-0.10x}$ Propiedad multiplicativa de los exponentes.

$y = 37.153\left(10^{-0.10}\right)^x$ Evalúa $10^{1.57}$ y aplica la propiedad de la potencia de los exponentes.

$y = 37.153(0.794)^x$ Evalúa $10^{-0.10.}$

La gráfica muestra que $\hat{y} = 37.153(0.794)^x$ se ajusta bien a los datos.

El error cuadrático medio para esta ecuación es aproximadamente 0.854, que es bastante pequeño.



[0, 20, 2, 0, 20, 2]

Tu calculadora posee comandos para ajustar muchos tipos de funciones, incluyendo las funciones cúbicas, cuadráticas, exponenciales, de potencias, sinusoidales, y logísticas. Usa los comandos de la calculadora para verificar la función del ejemplo anterior. Después lee el Ejemplo C en tu libro, donde se ajusta una función cúbica a los datos.

## Investigación: Experimento de la botella que se sale

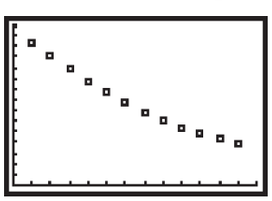
Lee la investigación en tu libro y asegúrate de entender cómo funciona el experimento. Completa los pasos usando esta muestra de datos, y después compara tus resultados con los siguientes.

| Tiempo (s) | Altura del agua (mm) |
|---|---|
| 0 | 150 |
| 10 | 134 |
| 20 | 120 |
| 30 | 109 |
| 40 | 97 |

| Tiempo (s) | Altura del agua (mm) |
|---|---|
| 50 | 86 |
| 60 | 77 |
| 70 | 68 |
| 80 | 60 |

| Tiempo (s) | Altura del agua (mm) |
|---|---|
| 90 | 54 |
| 100 | 48 |
| 110 | 43 |
| 120 | 39 |

**Paso 2** Los datos parecen ser exponenciales, cuadráticos, o cúbicos.



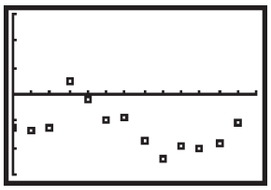
[0, 130, 10, 0, 150, 10]

**(continúa)**

**Paso 3** $y = 151.234(0.989)^x$; $y = 0.005x^2 - 1.500x + 149.110$; $y = -0.000005x^3 + 0.006x^2 - 1.544x + 149.456$

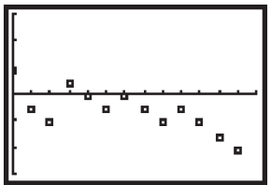
**Paso 4** He aquí las gráficas de los residuos de cada modelo del Paso 3.

Exponencial; error cuadrático medio aproximadamente 1.595



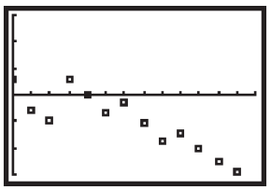
[0, 130, 10, −3, 3, 1]

Cuadrático; error cuadrático medio aproximadamente 1.095



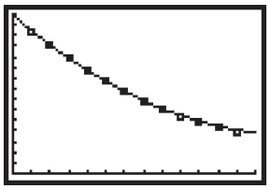
[0, 130, 10, −3, 3, 1]

Cúbico; error cuadrático medio aproximadamente 1.422



[0, 130, 10, −3, 3, 1]

**Paso 5** La curva cuadrática es la mejor. Tiene el menor error cuadrático medio y el patrón menos notable en los residuos.



[0, 130, 10, 0, 150, 10]

**Paso 6** Ninguna de estas ecuaciones da una respuesta razonable para el caso en que la botella queda vacía. De hecho, la curva cuadrática predice que la botella *nunca* estará vacía. Entonces, ninguno de los modelos es bueno para predecir los valores de datos fuera del dominio dado.