Ejemplo 6.3: Dada una distribución normal estándar, calcule el valor de k tal que

a) $P(Z > k) = 0.3015$, y

b) $P(k < Z < -0.18) = 0.4197$.

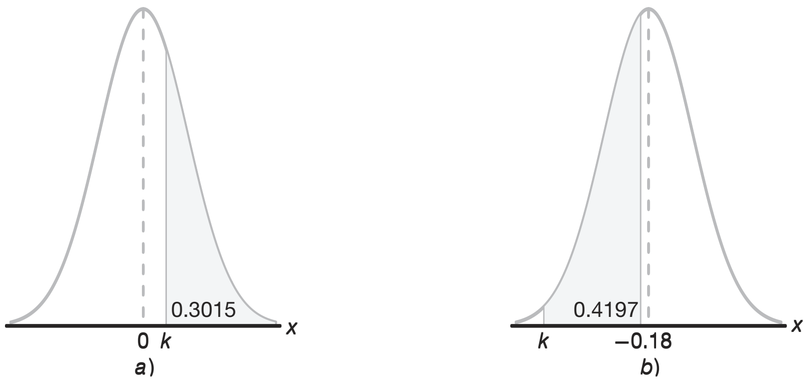

Figura 6.10: Áreas para el ejemplo 6.3.

***Solución*:** La distribución y las áreas deseadas se muestran en la figura 6.10.

a) En la figura 6.10*a* vemos que el valor k que deja un área de 0.3015 a la derecha debe dejar entonces un área de 0.6985 a la izquierda. De la tabla A.3 se sigue que $k = 0.52$.

b) En la tabla A.3 observamos el área total a la izquierda de -0.18 es igual a 0.4286. En la figura 6.10*b* vemos que el área entre k y -0.18 es 0.4197, de manera que el área a la izquierda de k debe ser $0.4286 - 0.4197 = 0.0089$. Por lo tanto, a partir de la tabla A.3 tenemos $k = -2.37$.

Ejemplo 6.4: Dada una variable aleatoria X que tiene una distribución normal con $\mu = 50$ y $\sigma = 10$, calcule la probabilidad de que X tome un valor entre 45 y 62.

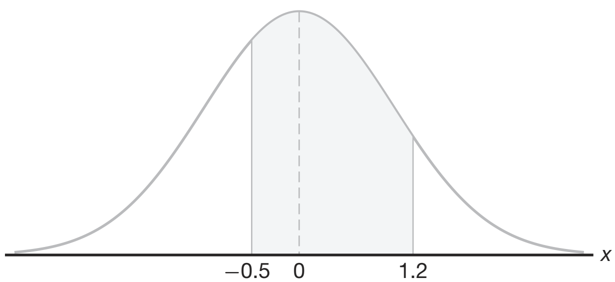

Figura 6.11: Área para el ejemplo 6.4.

***Solución*:** Los valores z que corresponden a $x_1 = 45$ y $x_2 = 62$ son

$$z_1 = \frac{45 - 50}{10} = -0.5 \text{ y } z_2 = \frac{62 - 50}{10} = 1.2.$$

Por lo tanto,

$$P(45 < X < 62) = P(-0.5 < Z < 1.2).$$

$P(-0.5 < Z < 1.2)$ se muestra mediante el área de la región sombreada de la figura 6.11. Esta área se puede calcular restando el área a la izquierda de la ordenada $z = -0.5$ de toda el área a la izquierda de $z = 1.2$. Si usamos la tabla A.3, tenemos

$$\begin{aligned} P(45 < X < 62) &= P(-0.5 < Z < 1.2) = P(Z < 1.2) - P(Z < -0.5) \\ &= 0.8849 - 0.3085 = 0.5764. \end{aligned}$$

Ejemplo 6.5: Dado que X tiene una distribución normal con $\mu = 300$ y $\sigma = 50$, calcule la probabilidad de que X tome un valor mayor que 362.

***Solución*:** La distribución de probabilidad normal que muestra el área sombreada que se desea se presenta en la figura 6.12. Para calcular $P(X > 362)$ necesitamos evaluar el área bajo la curva normal a la derecha de $x = 362$. Esto se puede realizar transformando $x = 362$ al valor z correspondiente, obteniendo el área a la izquierda de z de la tabla A.3 y después restando esta área de 1. Encontramos que

$$z = \frac{362 - 300}{50} = 1.24.$$

De ahí,

$$P(X > 362) = P(Z > 1.24) = 1 - P(Z < 1.24) = 1 - 0.8925 = 0.1075.$$

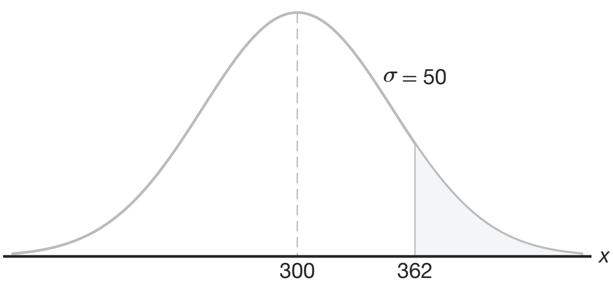

Figura 6.12: Área para el ejemplo 6.5.

De acuerdo con el teorema de Chebyshev en la página 137, la probabilidad de que una variable aleatoria tome un valor dentro de 2 desviaciones estándar de la media es de por lo menos 3/4. Si la variable aleatoria tiene una distribución normal, los valores z que corresponden a $x_1 = \mu - 2\sigma$ y $x_2 = \mu + 2\sigma$ se calculan fácilmente y son

$$z_1 = \frac{(\mu - 2\sigma) - \mu}{\sigma} = -2 \text{ y } z_2 = \frac{(\mu + 2\sigma) - \mu}{\sigma} = 2.$$

De ahí,

$$\begin{aligned} P(\mu - 2\sigma < X < \mu + 2\sigma) &= P(-2 < Z < 2) = P(Z < 2) - P(Z < -2) \\ &= 0.9772 - 0.0228 = 0.9544, \end{aligned}$$

que es una afirmación mucho más firme que la que se establece mediante el teorema de Chebyshev.

Uso de la curva normal a la inversa

En ocasiones se nos pide calcular el valor de z que corresponde a una probabilidad específica que cae entre los valores que se listan en la tabla A.3 (véase el ejemplo 6.6). Por conveniencia, siempre elegiremos el valor z que corresponde a la probabilidad tabular que está más cerca de la probabilidad que se especifica.

Los dos ejemplos anteriores se resolvieron al ir primero de un valor de x a un valor z y después calcular el área que se desea. En el ejemplo 6.6 invertimos el proceso y comenzamos con un área o probabilidad conocida, calculamos el valor z y después determinamos x reacomodando la fórmula

$$z = \frac{x - \mu}{\sigma} \text{ para obtener } x = \sigma z + \mu.$$

Ejemplo 6.6: Dada una distribución normal con $\mu = 40$ y $\sigma = 6$, calcule el valor de x que tiene

a) 45% del área a la izquierda, y

b) 14% del área a la derecha.

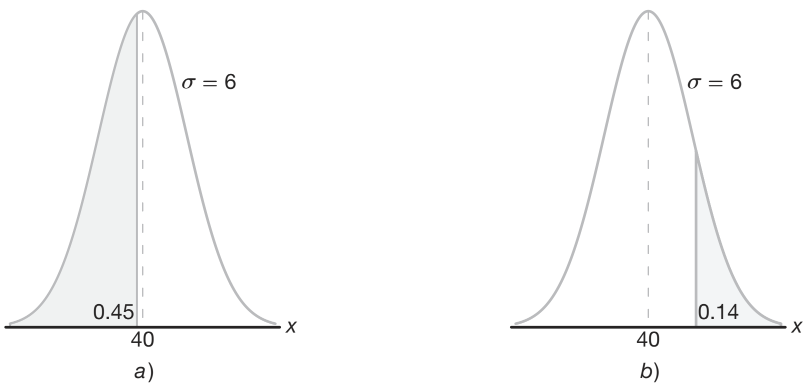

Figura 6.13: Áreas para el ejemplo 6.6.

***Solución*:** *a*) En la figura 6.13*a* se sombrea un área de 0.45 a la izquierda del valor x deseado. Necesitamos un valor z que deje un área de 0.45 a la izquierda. En la tabla A.3 encontramos $P(Z < -0.13) = 0.45$, es decir, que el valor z que se desea es -0.13. Por lo tanto,

$$x = (6)(-0.13) + 40 = 39.22.$$

b) En la figura 6.13*b* sombreamos un área igual a 0.14 a la derecha del valor x deseado. Esta vez necesitamos un valor z que deje 0.14 del área a la derecha y, por lo tanto, un área de 0.86 a la izquierda. De nuevo, a partir de la tabla A.3 encontramos $P(Z < 1.08) = 0.86$, así que el valor z deseado es 1.08 y

$$x = (6)(1.08) + 40 = 46.48.$$

6.4 Aplicaciones de la distribución normal

En los siguientes ejemplos se abordan algunos de los muchos problemas en los que se puede aplicar la distribución normal. El uso de la curva normal para aproximar probabilidades binomiales se estudia en la sección 6.5.

Ejemplo 6.7: Cierto tipo de batería de almacenamiento dura, en promedio, 3.0 años, con una desviación estándar de 0.5 años. Suponga que la duración de la batería se distribuye normalmente y calcule la probabilidad de que una batería determinada dure menos de 2.3 años.

***Solución*:** Empiece construyendo un diagrama como el de la figura 6.14, que muestra la distribución dada de la duración de las baterías y el área deseada. Para calcular la $P(X < 2.3)$ necesitamos evaluar el área bajo la curva normal a la izquierda de 2.3. Esto se logra calculando el área a la izquierda del valor z correspondiente. De donde encontramos que

$$z = \frac{2.3 - 3}{0.5} = -1.4,$$

y entonces, usando la tabla A.3, tenemos

$$P(X < 2.3) = P(Z < -1.4) = 0.0808.$$

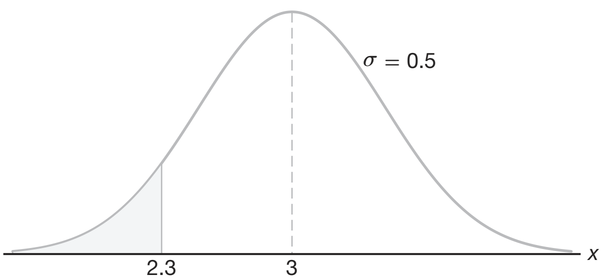

Figura 6.14: Área para el ejemplo 6.7.

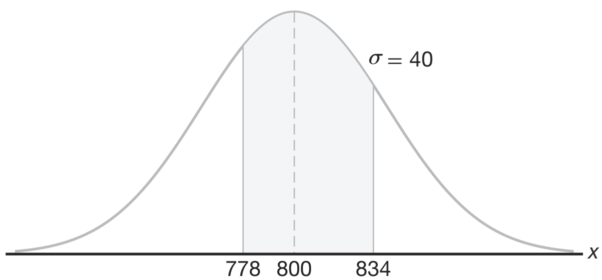

Figura 6.15: Área para el ejemplo 6.8.

Ejemplo 6.8 Una empresa de material eléctrico fabrica bombillas de luz cuya duración, antes de quemarse, se distribuye normalmente con una media igual a 800 horas y una desviación estándar de 40 horas. Calcule la probabilidad de que una bombilla se queme entre 778 y 834 horas.

***Solución*:** La distribución de vida de las bombillas se ilustra en la figura 6.15. Los valores z que corresponden a $x_1 = 778$ y $x_2 = 834$ son

$$z_1 = \frac{778 - 800}{40} = -0.55 \text{ y } z_2 = \frac{834 - 800}{40} = 0.85.$$

Por lo tanto,

$$\begin{aligned} P(778 < X < 834) &= P(-0.55 < Z < 0.85) = P(Z < 0.85) - P(Z < -0.55) \\ &= 0.8023 - 0.2912 = 0.5111. \end{aligned}$$

Ejemplo 6.9: En un proceso industrial el diámetro de un cojinete de bolas es una medida importante. El comprador establece que las especificaciones en el diámetro sean 3.0 ± 0.01 cm. Esto implica que no se aceptará ninguna parte que no cumpla estas especificaciones. Se sabe que en el proceso el diámetro de un cojinete tiene una distribución normal con media $\mu = 3.0$ y una desviación estándar $\sigma = 0.005$. En promedio, ¿cuántos de los cojinetes fabricados se descartarán?

***Solución*:** La distribución de los diámetros se ilustra en la figura 6.16. Los valores que corresponden a los límites especificados son $x_1 = 2.99$ y $x_2 = 3.01$. Los valores z correspondientes son

$$z_1 = \frac{2.99 - 3.0}{0.005} = -2.0 \text{ y } z_2 = \frac{3.01 - 3.0}{0.005} = +2.0.$$

Por lo tanto,

$$P(2.99 < X < 3.01) = P(-2.0 < Z < 2.0).$$

A partir de la tabla A.3, $P(Z < -2.0) = 0.0228$. Debido a la simetría de la distribución normal, encontramos que

$$P(Z < -2.0) + P(Z > 2.0) = 2(0.0228) = 0.0456.$$

Como resultado se anticipa que, en promedio, se descartarán 4.56% de los cojinetes fabricados.

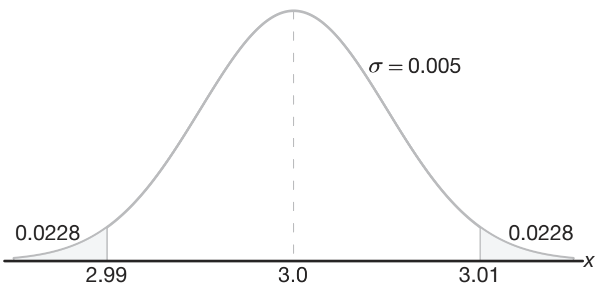

Figura 6.16: Área para el ejemplo 6.9.

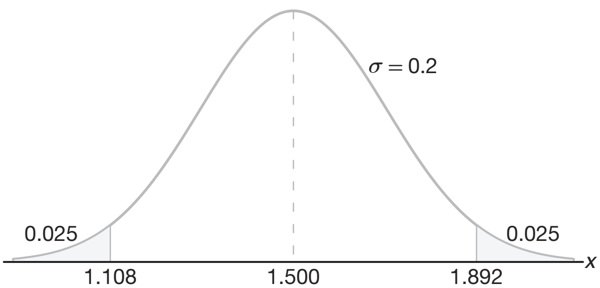

Figura 6.17: Especificaciones para el ejemplo 6.10.

Ejemplo 6.10: Se utilizan medidores para rechazar todos los componentes en los que cierta dimensión no esté dentro de la especificación $1.50 \pm d$. Se sabe que esta medida se distribuye normalmente con una media de 1.50 y una desviación estándar de 0.2. Determine el valor d tal que las especificaciones "cubran" 95% de las mediciones.

***Solución*:** A partir de la tabla A.3 sabemos que

$$P(-1.96 < Z < 1.96) = 0.95.$$

Por lo tanto,

$$1.96 = \frac{(1.50 + d) - 1.50}{0.2},$$

de la que obtenemos

$$d = (0.2)(1.96) = 0.392.$$

En la figura 6.17 se muestra una ilustración de las especificaciones.

Ejemplo 6.11: Cierta máquina fabrica resistencias eléctricas que tienen una resistencia media de 40 ohms y una desviación estándar de 2 ohms. Si se supone que la resistencia sigue una distribución normal y que se puede medir con cualquier grado de precisión, ¿qué porcentaje de resistencias tendrán una resistencia que exceda 43 ohms?

***Solución*:** Se obtiene un porcentaje multiplicando la frecuencia relativa por 100%. Como la frecuencia relativa para un intervalo es igual a la probabilidad de caer en el intervalo, debemos calcular el área a la derecha de $x = 43$ en la figura 6.18. Esto se puede hacer transformando $x = 43$ al valor z correspondiente, con lo cual se obtiene el área a la izquierda de z de la tabla A.3, y después se resta esta área de 1. Encontramos que

$$z = \frac{43 - 40}{2} = 1.5.$$

Por lo tanto,

$$P(X > 43) = P(Z > 1.5) = 1 - P(Z < 1.5) = 1 - 0.9332 = 0.0668.$$

Así, 6.68% de las resistencias tendrán una resistencia que exceda 43 ohms.

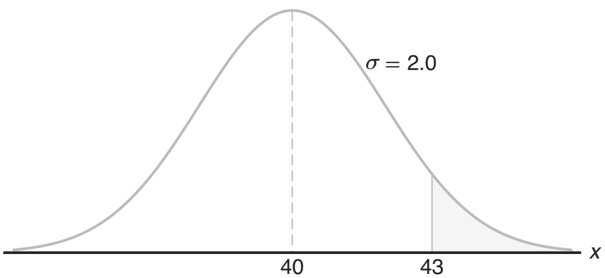

Figura 6.18: Área para el ejemplo 6.11.

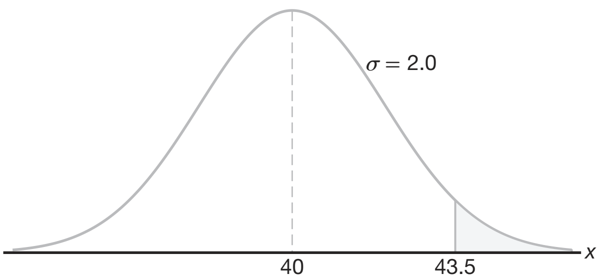

Figura 6.19: Área para el ejemplo 6.12.

Ejemplo 6.12: Calcule el porcentaje de resistencias que excedan 43 ohms para el ejemplo 6.11 si la resistencia se mide al ohm más cercano.

***Solución*:** Este problema difiere del ejemplo 6.11 en que ahora asignamos una medida de 43 ohms a todos los resistores cuyas resistencias sean mayores que 42.5 y menores que 43.5. Lo que estamos haciendo realmente es aproximar una distribución discreta por medio de una distribución continua normal. El área que se requiere es la región sombreada a la derecha de 43.5 en la figura 6.19. Encontramos ahora que

$$z = \frac{43.5 - 40}{2} = 1.75.$$

En consecuencia,

$$P(X > 43.5) = P(Z > 1.75) = 1 - P(Z < 1.75) = 1 - 0.9599 = 0.0401.$$

Por lo tanto, 4.01% de las resistencias exceden 43 ohms cuando se miden al ohm más cercano. La diferencia 6.68% – 4.01% = 2.67% entre esta respuesta y la del ejemplo 6.11 representa todos los valores de resistencias mayores que 43 y menores que 43.5, que ahora se registran como de 43 ohms.

Ejemplo 6.13: La calificación promedio para un examen es 74 y la desviación estándar es 7. Si 12% del grupo obtiene *A* y las calificaciones siguen una curva que tiene una distribución normal, ¿cuál es la *A* más baja posible y la *B* más alta posible?

***Solución*:** En este ejemplo comenzamos con un área de probabilidad conocida, calculamos el valor z y después determinamos x con la fórmula $x = \sigma z + \mu$. Un área de 0.12, que corresponde a la fracción de estudiantes que reciben *A,* está sombreada en la figura 6.20. Necesitamos un valor z que deje 0.12 del área a la derecha y, por lo tanto, un área de 0.88 a la izquierda. A partir de la tabla A.3, $P(Z < 1.18)$ tiene el valor más cercano a 0.88, de manera que el valor z que se desea es 1.18. En consecuencia,

$$x = (7)(1.18) + 74 = 82.26.$$

Por lo tanto, la *A* más baja es 83 y la *B* más alta es 82.

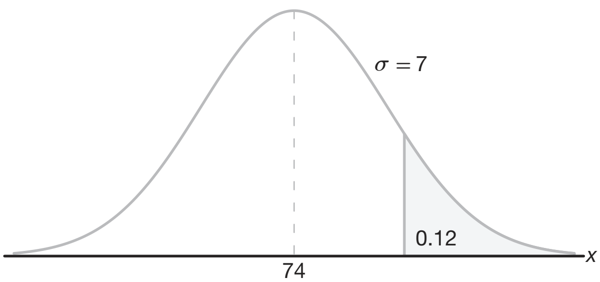

Figura 6.20: Área para el ejemplo 6.13.

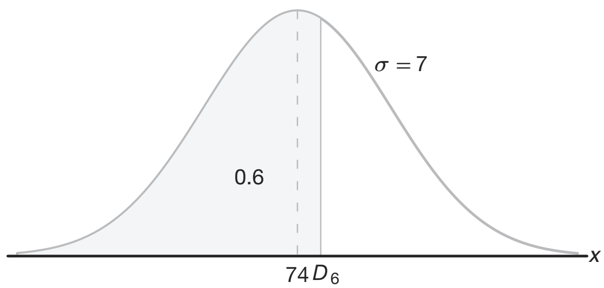

Figura 6.21: Área para el ejemplo 6.14.

Ejemplo 6.14: Remítase al ejemplo 6.13 y calcule el sexto decil.

***Solución*:** El sexto decil, escrito como D_6, es el valor x que deja 60% del área a la izquierda, como se muestra en la figura 6.21. En la tabla A.3 encontramos que $P(Z < 0.25) \approx 0.6$, de manera que el valor z deseado es 0.25. Ahora, $x = (7)(0.25) + 74 = 75.75$. Por lo tanto, $D_6 = 75.75$. Es decir, 60% de las calificaciones son 75 o menos.

Ejercicios

6.1 Dada una distribución continua uniforme, demuestre que

a) $\mu = \frac{A+B}{2}$, y

b) $\sigma^2 = \frac{(B-A)^2}{12}$.

6.2 Suponga que X tiene una distribución continua uniforme de 1 a 5. Determine la probabilidad condicional $P(X > 2.5 \mid X \le 4)$.

6.3 La cantidad de café diaria, en litros, que sirve una máquina que se localiza en el vestíbulo de un aeropuerto es una variable aleatoria X que tiene una distribución continua uniforme con $A = 7$ y $B = 10$. Calcule la probabilidad de que en un día determinado la cantidad de café que sirve esta máquina sea

a) a lo sumo 8.8 litros;

b) más de 7.4 litros, pero menos de 9.5 litros;

c) al menos 8.5 litros.

6.4 Un autobús llega cada 10 minutos a una parada. Se supone que el tiempo de espera para un individuo en particular es una variable aleatoria con distribución continua uniforme.

a) ¿Cuál es la probabilidad de que el individuo espere más de 7 minutos?
b) ¿Cuál es la probabilidad de que el individuo espere entre 2 y 7 minutos?

6.5 Dada una distribución normal estándar, calcule el área bajo la curva que está
a) a la izquierda de $z = -1.39$;
b) a la derecha de $z = 1.96$;
c) entre $z = -2.16$ y $z = -0.65$;
d) a la izquierda de $z = 1.43$;
e) a la derecha de $z = -0.89$;
f) entre $z = -0.48$ y $z = 1.74$.

6.6 Calcule el valor de z si el área bajo una curva normal estándar
a) a la derecha de z es 0.3622;
b) a la izquierda de z es 0.1131;
c) entre 0 y z, con $z > 0$, es 0.4838;
d) entre $-z$ y z, con $z > 0$, es 0.9500.

6.7 Dada una distribución normal estándar, calcule el valor de k tal que
a) $P(Z > k) = 0.2946$;
b) $P(Z < k) = 0.0427$;
c) $P(-0.93 < Z < k) = 0.7235$.

6.8 Dada una distribución normal con $\mu = 30$ y $\sigma = 6$, calcule
a) el área de la curva normal a la derecha de $x = 17$;
b) el área de la curva normal a la izquierda de $x = 22$;
c) el área de la curva normal entre $x = 32$ y $x = 41$;
d) el valor de x que tiene 80% del área de la curva normal a la izquierda;
e) los dos valores de x que contienen 75% central del área de la curva normal.

6.9 Dada la variable X normalmente distribuida con una media de 18 y una desviación estándar de 2.5, calcule
a) $P(X < 15)$;
b) el valor de k tal que $P(X < k) = 0.2236$;
c) el valor de k tal que $P(X > k) = 0.1814$;
d) $P(17 < X < 21)$.

6.10 De acuerdo con el teorema de Chebyshev, la probabilidad de que cualquier variable aleatoria tome un valor dentro de 3 desviaciones estándar de la media es de al menos $8/9$. Si se sabe que la distribución de probabilidad de una variable aleatoria X es normal con media μ y varianza σ^2, ¿cuál es el valor exacto de $P(\mu - 3\sigma < X < \mu + 3\sigma)$?

6.11 Una máquina expendedora de bebidas gaseosas se regula para que sirva un promedio de 200 mililitros por vaso. Si la cantidad de bebida se distribuye normalmente con una desviación estándar igual a 15 mililitros,
a) ¿qué fracción de los vasos contendrá más de 224 mililitros?
b) ¿cuál es la probabilidad de que un vaso contenga entre 191 y 209 mililitros?
c) ¿cuántos vasos probablemente se derramarán si se utilizan vasos de 230 mililitros para las siguientes 1000 bebidas?
d) ¿por debajo de qué valor obtendremos el 25% más bajo en el llenado de las bebidas?

6.12 Las barras de pan de centeno que cierta panadería distribuye a las tiendas locales tienen una longitud promedio de 30 centímetros y una desviación estándar de 2 centímetros. Si se supone que las longitudes están distribuidas normalmente, ¿qué porcentaje de las barras son
a) más largas que 31.7 centímetros?
b) de entre 29.3 y 33.5 centímetros de longitud?
c) más cortas que 25.5 centímetros?

6.13 Un investigador informa que unos ratones a los que primero se les restringen drásticamente sus dietas y después se les enriquecen con vitaminas y proteínas vivirán un promedio de 40 meses. Si suponemos que la vida de tales ratones se distribuye normalmente, con una desviación estándar de 6.3 meses, calcule la probabilidad de que un ratón determinado viva
a) más de 32 meses;
b) menos de 28 meses;
c) entre 37 y 49 meses.

6.14 El diámetro interior del anillo de un pistón terminado se distribuye normalmente con una media de 10 centímetros y una desviación estándar de 0.03 centímetros.
a) ¿Qué proporción de anillos tendrá diámetros interiores que excedan 10.075 centímetros?
b) ¿Cuál es la probabilidad de que el anillo de un pistón tenga un diámetro interior de entre 9.97 y 10.03 centímetros?
c) ¿Por debajo de qué valor del diámetro interior caerá el 15% de los anillos de pistón?

6.15 Un abogado viaja todos los días de su casa en los suburbios a su oficina en el centro de la ciudad. El tiempo promedio para un viaje sólo de ida es de 24 minutos, con una desviación estándar de 3.8 minutos. Si se supone que la distribución de los tiempos de viaje está distribuida normalmente.
a) ¿Cuál es la probabilidad de que un viaje tome al menos $1/2$ hora?
b) Si la oficina abre a las 9:00 A.M. y él sale diario de su casa a las 8:45 A.M., ¿qué porcentaje de las veces llegará tarde al trabajo?

c) Si sale de su casa a las 8:35 A.M. y el café se sirve en la oficina de 8:50 A.M. a 9:00 A.M., ¿cuál es la probabilidad de que se pierda el café?

d) Calcule la duración mayor en la que se encuentra el 15% de los viajes más lentos.

e) Calcule la probabilidad de que 2 de los siguientes 3 viajes tomen al menos 1/2 hora.

6.16 En el ejemplar de noviembre de 1990 de *Chemical Engineering Progress*, un estudio analiza el porcentaje de pureza del oxígeno de cierto proveedor. Suponga que la media fue de 99.61, con una desviación estándar de 0.08. Suponga que la distribución del porcentaje de pureza fue aproximadamente normal.

a) ¿Qué porcentaje de los valores de pureza esperaría que estuvieran entre 99.5 y 99.7?

b) ¿Qué valor de pureza esperaría que excediera exactamente 5% de la población?

6.17 La vida promedio de cierto tipo de motor pequeño es de 10 años, con una desviación estándar de 2 años. El fabricante reemplaza gratis todos los motores que fallen dentro del periodo de garantía. Si estuviera dispuesto a reemplazar sólo 3% de los motores que fallan, ¿cuánto tiempo de garantía debería ofrecer? Suponga que la duración de un motor sigue una distribución normal.

6.18 La estatura de 1000 estudiantes se distribuye normalmente con una media de 174.5 centímetros y una desviación estándar de 6.9 centímetros. Si se supone que las estaturas se redondean al medio centímetro más cercano, ¿cuántos de estos estudiantes esperaría que tuvieran una estatura

a) menor que 160.0 centímetros?

b) de entre 171.5 y 182.0 centímetros inclusive?

c) igual a 175.0 centímetros?

d) mayor o igual que 188.0 centímetros?

6.19 Una empresa paga a sus empleados un salario promedio de $15.90 por hora, con una desviación estándar de $1.50. Si los salarios se distribuyen aproximadamente de forma normal y se redondean al centavo más cercano,

a) ¿qué porcentaje de los trabajadores recibe salarios de entre $13.75 y $16.22 por hora?

b) ¿el 5% de los salarios más altos por hora de los empleados es mayor a qué cantidad?

6.20 Los pesos de un gran número de *poodle* miniatura se distribuyen aproximadamente de forma normal con una media de 8 kilogramos y una desviación estándar de 0.9 kilogramos. Si las mediciones se redondean al décimo de kilogramo más cercano, calcule la fracción de estos *poodle* con pesos

a) por arriba de 9.5 kilogramos;

b) a lo sumo 8.6 kilogramos;

c) entre 7.3 y 9.1 kilogramos.

6.21 La resistencia a la tensión de cierto componente de metal se distribuye normalmente con una media de 10,000 kilogramos por centímetro cuadrado y una desviación estándar de 100 kilogramos por centímetro cuadrado. Las mediciones se redondean a los 50 kilogramos por centímetro cuadrado más cercanos.

a) ¿Qué proporción de estos componentes excede a 10,150 kilogramos por centímetro cuadrado de resistencia a la tensión?

b) Si las especificaciones requieren que todos los componentes tengan una resistencia a la tensión de entre 9800 y 10,200 kilogramos por centímetro cuadrado, ¿qué proporción de piezas esperaría que se descartara?

6.22 Si un conjunto de observaciones se distribuye de manera normal, ¿qué porcentaje de éstas difieren de la media en

a) más de 1.3σ?

b) menos de 0.52σ?

6.23 El coeficiente intelectual (CI) de 600 aspirantes a cierta universidad se distribuye aproximadamente de forma normal con una media de 115 y una desviación estándar de 12. Si la universidad requiere un CI de al menos 95, ¿cuántos de estos estudiantes serán rechazados con base en éste sin importar sus otras calificaciones? Tome en cuenta que el CI de los aspirantes se redondea al entero más cercano.

6.5 Aproximación normal a la binomial

Las probabilidades asociadas con experimentos binomiales se obtienen fácilmente a partir de la fórmula $b(x; n, p)$ de la distribución binomial o de la tabla A.1 cuando n es pequeña. Además, las probabilidades binomiales están disponibles en muchos paquetes de software. Sin embargo, resulta aleccionador conocer la relación entre la distribución binomial y la normal. En la sección 5.5 explicamos cómo se puede utilizar la distribución de Poisson para aproximar probabilidades binomiales cuando n es muy grande y p se acerca mucho a 0 o a 1. Tanto la distribución binomial como la de Poisson son discretas. La primera aplicación de una distribución continua de probabilidad para aproximar probabilidades sobre un espacio muestral discreto se demostró en el ejemplo 6.12, donde se utilizó la curva normal. La distribución normal a menudo es una buena aproximación a una distribución discreta cuando la última adquiere una forma de campana simétrica. Desde un punto de vista teórico, algunas distribuciones convergen a la normal a medida que sus parámetros se aproximan a ciertos límites. La distribución normal es una distribución de aproximación conveniente, ya que la función de distribución acumulativa se tabula con mucha facilidad. La distribución binomial se aproxima bien por medio de la normal en problemas prácticos cuando se trabaja con la función de distribución acumulativa. Ahora plantearemos un teorema que nos permitirá utilizar áreas bajo la curva normal para aproximar propiedades binomiales cuando n es suficientemente grande.

Teorema 6.3: Si X es una variable aleatoria binomial con media $\mu = np$ y varianza $\sigma^2 = npq$, entonces la forma limitante de la distribución de

$$Z = \frac{X - np}{\sqrt{npq}},$$

conforme $n \to \infty$, es la distribución normal estándar $n(z; 0, 1)$.

Resulta que la distribución normal con $\mu = np$ y $\sigma^2 = np(1 - p)$ no sólo ofrece una aproximación muy precisa a la distribución binomial cuando n es grande y p no está extremadamente cerca de 0 o de 1, sino que también brinda una aproximación bastante buena aun cuando n es pequeña y p está razonablemente cerca de $1/2$.

Para ilustrar la aproximación normal a la distribución binomial primero dibujamos el histograma para $b(x; 15, 0.4)$ y después superponemos la curva normal particular con la misma media y varianza que la variable binomial X. En consecuencia, dibujamos una curva normal con

$$\mu = np = (15)(0.4) = 6 \text{ y } \sigma^2 = npq = (15)(0.4)(0.6) = 3.6.$$

El histograma de $b(x; 15, 0.4)$ y la curva normal superpuesta correspondiente, que está determinada por completo por su media y su varianza, se ilustran en la figura 6.22.

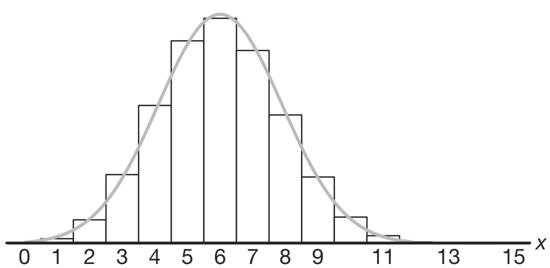

Figura 6.22: Aproximación normal de $b(x; 15, 0.4)$.

La probabilidad exacta de que la variable aleatoria binomial X tome un valor determinado x es igual al área de la barra cuya base se centra en x. Por ejemplo, la probabilidad exacta de que X tome el valor 4 es igual al área del rectángulo con base centrada en $x = 4$. Si usamos la tabla A.1, encontramos que esta área es

$$P(X = 4) = b(4; 15, 0.4) = 0.1268,$$

que es aproximadamente igual al área de la región sombreada bajo la curva normal entre las dos ordenadas $x_1 = 3.5$ y $x_2 = 4.5$ en la figura 6.23. Al convertir a valores z, tenemos

$$z_1 = \frac{3.5 - 6}{1.897} = -1.32 \quad \text{y} \quad z_2 = \frac{4.5 - 6}{1.897} = -0.79.$$

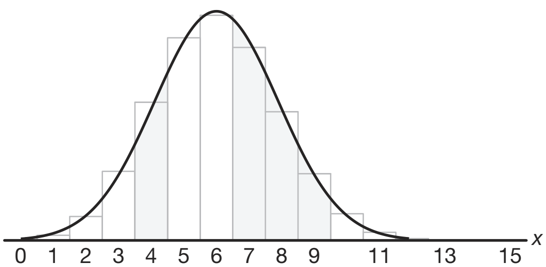

Figura 6.23: Aproximación normal de $b(x; 15, 0.4)$ y $\sum_{x=7}^{9} b(x; 15, 0.4)$.

Si X es una variable aleatoria binomial y Z una variable normal estándar, entonces,

$$\begin{aligned} P(X = 4) &= b(4; 15, 0.4) \approx P(-1.32 < Z < -0.79) \\ &= P(Z < -0.79) - P(Z < -1.32) = 0.2148 - 0.0934 = 0.1214. \end{aligned}$$

Esto se aproxima bastante al valor exacto de 0.1268.

La aproximación normal es más útil en el cálculo de sumatorias binomiales para valores grandes de n. Si nos remitimos a la figura 6.23, nos podríamos interesar en la probabilidad de que X tome un valor de 7 a 9. La probabilidad exacta es dada por

$$\begin{aligned} P(7 \leq X \leq 9) &= \sum_{x=0}^{9} b(x; 15, 0.4) - \sum_{x=0}^{6} b(x; 15, 0.4) \\ &= 0.9662 - 0.6098 = 0.3564, \end{aligned}$$

que es igual a la sumatoria de las áreas de los rectángulos cuyas bases están centradas en $x = 7$, 8 y 9. Para la aproximación normal calculamos el área de la región sombreada bajo la curva entre las ordenadas $x_1 = 6.5$ y $x_2 = 9.5$ de la figura 6.23. Los valores z correspondientes son

$$z_1 = \frac{6.5 - 6}{1.897} = 0.26 \quad \text{y} \quad z_2 = \frac{9.5 - 6}{1.897} = 1.85.$$

Ahora,

$$P(7 \le X \le 9) \approx P(0.26 < Z < 1.85) = P(Z < 1.85) - P(Z < 0.26)$$
$$= 0.9678 - 0.6026 = 0.3652.$$

Una vez más, la aproximación de la curva normal ofrece un valor que se acerca al valor exacto de 0.3564. El grado de exactitud, que depende de qué tan bien se ajuste la curva al histograma, se incrementa a medida que aumenta n. Esto es particularmente cierto cuando p no está muy cerca de $1/2$ y el histograma ya no es simétrico. Las figuras 6.24 y 6.25 muestran los histogramas para $b(x; 6, 0.2)$ y $b(x; 15, 0.2)$, respectivamente. Es evidente que una curva normal se ajustará mucho mejor al histograma cuando $n = 15$ que cuando $n = 6$.

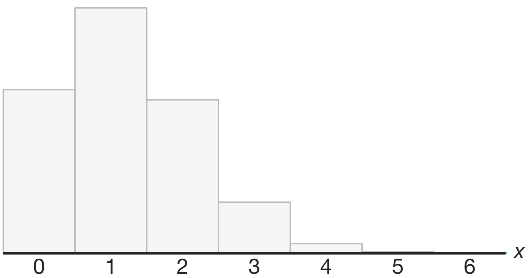

Figura 6.24: Histograma para $b(x; 6, 0.2)$.

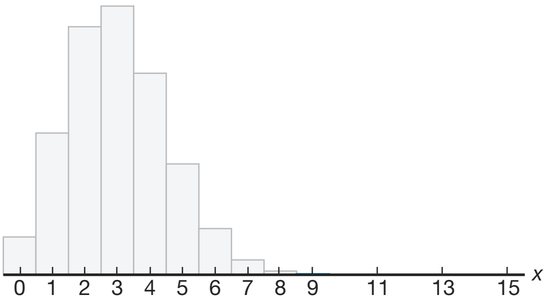

Figura 6.25: Histograma para $b(x; 15, 0.2)$.

En las ilustraciones de la aproximación normal a la binomial se hizo evidente que si buscamos el área bajo la curva normal hacia la izquierda de, digamos x, es más preciso utilizar $x + 0.5$. Esto es una corrección para dar cabida al hecho de que una distribución discreta se aproxima mediante una distribución continua. La corrección **+0.5** se llama **corrección de continuidad**. La explicación anterior conduce a la siguiente aproximación normal formal a la binomial.

Aproximación normal a la distribución binomial

Sea X una variable aleatoria binomial con parámetros n y p. Para una n grande, X tiene aproximadamente una distribución normal con $\mu = np$ y $\sigma^2 = npq = np(1 - p)$ y

$$\begin{aligned} P(X \le x) &= \sum_{k=0}^{x} b(k; n, p) \\ &\approx \text{área bajo la curva normal a la izquierda de } x + 0.5 \\ &= P\left(Z \le \frac{x + 0.5 - np}{\sqrt{npq}}\right), \end{aligned}$$

y la aproximación será buena si np y $n(1 - p)$ son mayores que o iguales a 5.

Como indicamos antes, la calidad de la aproximación es muy buena para n grande. Si p está cerca de $1/2$, un tamaño de la muestra moderado o pequeño será suficiente para una aproximación razonable. Ofrecemos la tabla 6.1 como una indicación de la calidad de la aproximación. Se presentan tanto la aproximación normal como las probabilidades binomiales acumulativas reales. Observe que en $p = 0.05$ y $p = 0.10$ la aproximación es muy burda para $n = 10$. Sin embargo, incluso para $n = 10$, observe la mejoría para $p = 0.50$. Por otro lado, cuando p es fija en $p = 0.05$, observe cómo mejora la aproximación conforme vamos de $n = 20$ a $n = 100$.

Tabla 6.1: Aproximación normal y probabilidades binomiales acumulativas reales

	$p = 0.05, n = 10$		$p = 0.10, n = 10$		$p = 0.50, n = 10$	
r	**Binomial**	**Normal**	**Binomial**	**Normal**	**Binomial**	**Normal**
0	0.5987	0.5000	0.3487	0.2981	0.0010	0.0022
1	0.9139	0.9265	0.7361	0.7019	0.0107	0.0136
2	0.9885	0.9981	0.9298	0.9429	0.0547	0.0571
3	0.9990	1.0000	0.9872	0.9959	0.1719	0.1711
4	1.0000	1.0000	0.9984	0.9999	0.3770	0.3745
5			1.0000	1.0000	0.6230	0.6255
6					0.8281	0.8289
7					0.9453	0.9429
8					0.9893	0.9864
9					0.9990	0.9978
10					1.0000	0.9997

	$p = 0.05$					
	$n = 20$		$n = 50$		$n = 100$	
r	**Binomial**	**Normal**	**Binomial**	**Normal**	**Binomial**	**Normal**
0	0.3585	0.3015	0.0769	0.0968	0.0059	0.0197
1	0.7358	0.6985	0.2794	0.2578	0.0371	0.0537
2	0.9245	0.9382	0.5405	0.5000	0.1183	0.1251
3	0.9841	0.9948	0.7604	0.7422	0.2578	0.2451
4	0.9974	0.9998	0.8964	0.9032	0.4360	0.4090
5	0.9997	1.0000	0.9622	0.9744	0.6160	0.5910
6	1.0000	1.0000	0.9882	0.9953	0.7660	0.7549
7			0.9968	0.9994	0.8720	0.8749
8			0.9992	0.9999	0.9369	0.9463
9			0.9998	1.0000	0.9718	0.9803
10			1.0000	1.0000	0.9885	0.9941

Ejemplo 6.15: Un paciente que padece una rara enfermedad de la sangre tiene 0.4 de probabilidad de recuperarse. Si se sabe que 100 personas contrajeron esta enfermedad, ¿cuál es la probabilidad de que sobrevivan menos de 30?

***Solución*:** Representemos con la variable binomial X el número de pacientes que sobreviven. Como $n = 100$, deberíamos obtener resultados muy precisos usando la aproximación de la curva normal con

$$\mu = np = (100)(0.4) = 40 \text{ y } \sigma = \sqrt{npq} = \sqrt{(100)(0.4)(0.6)} = 4.899.$$

Para obtener la probabilidad que se desea, tenemos que calcular el área a la izquierda de $x = 29.5$.

El valor z que corresponde a 29.5 es

$$z = \frac{29.5 - 40}{4.899} = -2.14,$$

y la probabilidad de que menos de 30 de los 100 pacientes sobrevivan está dada por la región sombreada en la figura 6.26. Por lo tanto,

$$P(X < 30) \approx P(Z < -2.14) = 0.0162.$$

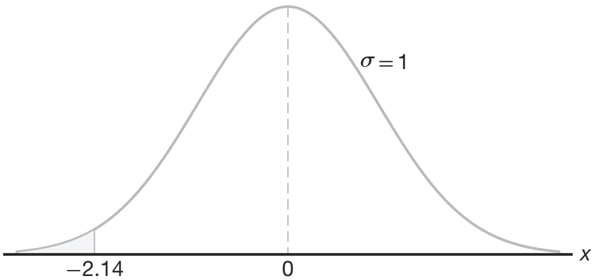

Figura 6.26: Área para el ejemplo 6.15.

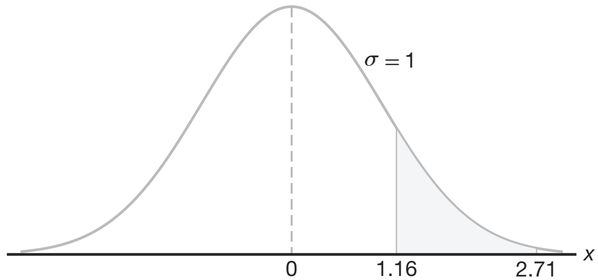

Figura 6.27: Área para el ejemplo 6.16.

Ejemplo 6.16: Un examen de opción múltiple tiene 200 preguntas, cada una con 4 respuestas posibles, de las que sólo una es la correcta. ¿Cuál es la probabilidad de que solamente adivinando se obtengan de 25 a 30 respuestas correctas para 80 de los 200 problemas sobre los que el estudiante no tiene conocimientos?

***Solución*:** La probabilidad de adivinar una respuesta correcta para cada una de las 80 preguntas es $p = 1/4$. Si X representa el número de respuestas correctas sólo porque se adivinaron, entonces,

$$P(25 \le X \le 30) = \sum_{x=25}^{30} b(x; 80, 1/4).$$

Al usar la aproximación de la curva normal con

$$\mu = np = (80)\left(\frac{1}{4}\right) = 20$$

y

$$\sigma = \sqrt{npq} = \sqrt{(80)(1/4)(3/4)} = 3.873,$$

necesitamos el área entre $x_1 = 24.5$ y $x_2 = 30.5$. Los valores z correspondientes son

$$z_1 = \frac{24.5 - 20}{3.873} = 1.16 \text{ y } z_2 = \frac{30.5 - 20}{3.873} = 2.71.$$

La probabilidad de adivinar correctamente de 25 a 30 preguntas es dada por la región sombreada de la figura 6.27. En la tabla A.3 encontramos que

$$\begin{aligned} P(25 \le X \le 30) &= \sum_{x=25}^{30} b(x; 80, 0.25) \approx P(1.16 < Z < 2.71) \\ &= P(Z < 2.71) - P(Z < 1.16) = 0.9966 - 0.8770 = 0.1196. \end{aligned}$$

Ejercicios

6.24 Se lanza una moneda 400 veces. Utilice la aproximación a la curva normal para calcular la probabilidad de obtener

a) entre 185 y 210 caras;
b) exactamente 205 caras;
c) menos de 176 o más de 227 caras.

6.25 En un proceso para fabricar un componente electrónico, 1% de los artículos resultan defectuosos. Un plan de control de calidad consiste en seleccionar 100 artículos de un proceso de producción y detenerlo o continuar con él si ninguno está defectuoso. Use la aproximación normal a la binomial para calcular

a) la probabilidad de que el proceso continúe con el plan de muestreo descrito;
b) la probabilidad de que el proceso continúe aun si éste va mal (es decir, si la frecuencia de componentes defectuosos cambió a 5.0% de defectuosos).

6.26 Un proceso produce 10% de artículos defectuosos. Si se seleccionan al azar 100 artículos del proceso, ¿cuál es la probabilidad de que el número de defectuosos

a) exceda los 13?
b) sea menor que 8?

6.27 Un paciente tiene 0.9 de probabilidad de recuperarse de una operación de corazón delicada. De los siguientes 100 pacientes que se someten a esta operación, ¿cuál es la probabilidad de que

a) sobrevivan entre 84 y 95 inclusive?
b) sobrevivan menos de 86?

6.28 Investigadores de la Universidad George Washington y del Instituto Nacional de Salud informan que aproximadamente 75% de las personas cree que "los tranquilizantes funcionan muy bien para lograr que una persona esté más tranquila y relajada". De las siguientes 80 personas entrevistadas, ¿cuál es la probabilidad de que

a) al menos 50 tengan esta opinión?
b) a lo sumo 56 tengan esta opinión?

6.29 Si 20% de los residentes de una ciudad de Estados Unidos prefieren un teléfono blanco sobre cualquier otro color disponible, ¿cuál es la probabilidad de que, de los siguientes 1000 teléfonos que se instalen en esa ciudad,

a) entre 170 y 185 sean blancos?
b) al menos 210 pero no más de 225 sean blancos?

6.30 Un fabricante de medicamentos sostiene que cierto medicamento cura una enfermedad de la sangre, en promedio, 80% de las veces. Para verificar la aseveración, inspectores gubernamentales utilizan el medicamento en una muestra de 100 individuos y deciden aceptar la afirmación si se curan 75 o más.

a) ¿Cuál es la probabilidad de que los inspectores gubernamentales rechacen la aseveración si la probabilidad de curación es, de hecho, de 0.8?
b) ¿Cuál es la probabilidad de que el gobierno acepte la afirmación si la probabilidad de curación resulta tan baja como 0.7?

6.31 Una sexta parte de los estudiantes de primer año que entran a una escuela estatal grande provienen de otros estados. Si son asignados al azar a los 180 dormitorios de un edificio, ¿cuál es la probabilidad de que en un determinado dormitorio al menos una quinta parte de los estudiantes provenga de otro estado?

6.32 Una empresa farmacéutica sabe que aproximadamente 5% de sus píldoras anticonceptivas no contiene la cantidad suficiente de un ingrediente, lo que las vuelve ineficaces. ¿Cuál es la probabilidad de que menos de 10 píldoras en una muestra de 200 sean ineficaces?

6.33 Estadísticas publicadas por la National Highway Traffic Safety Administration y el National Safety Council revelan que en una noche promedio de fin de semana, uno de cada 10 conductores está ebrio. Si la siguiente noche de sábado se revisan 400 conductores al azar, ¿cuál es la probabilidad de que el número de conductores ebrios sea

a) menor que 32?
b) mayor que 49?
c) al menos 35 pero menos que 47?

6.34 Un par de dados se lanza 180 veces. ¿Cuál es la probabilidad de que ocurra un total de 7

a) al menos 25 veces?
b) entre 33 y 41 veces?
c) exactamente 30 veces?

6.35 Una empresa produce partes componentes para un motor. Las especificaciones de las partes sugieren que sólo 95% de los artículos las cumplen. Las partes para los clientes se embarcan en lotes de 100.

a) ¿Cuál es la probabilidad de que más de 2 artículos estén defectuosos en un lote determinado?
b) ¿Cuál es la probabilidad de que más de 10 artículos de un lote estén defectuosos?

6.36 Una práctica común por parte de las aerolíneas consiste en vender más boletos que el número real de asientos para un vuelo específico porque los clientes que compran boletos no siempre se presentan a abordar el avión. Suponga que el porcentaje de pasajeros que no se presentan a la hora del vuelo es de 2%. Para un vuelo particular con 197 asientos, se vendieron un total de 200 boletos. ¿Cuál es la probabilidad de que la aerolínea haya sobrevendido el vuelo?

6.37 El nivel X de colesterol en la sangre en muchachos de 14 años tiene aproximadamente una distribución normal, con una media de 170 y una desviación estándar de 30.

a) Determine la probabilidad de que el nivel de colesterol en la sangre de un muchacho de 14 años elegido al azar exceda 230.

b) En una escuela secundaria hay 300 muchachos de 14 años. Determine la probabilidad de que por lo menos 8 de ellos tengan un nivel de colesterol superior a 230.

6.38 Una empresa de telemarketing tiene una máquina especial para abrir cartas que abre y extrae el contenido de los sobres. Si un sobre se colocara de forma incorrecta en la máquina, no se podría extraer su contenido, o incluso se podría dañar. En este caso se dice que "falló" la máquina.

a) Si la probabilidad de que falle la máquina es de 0.01, ¿cuál es la probabilidad de que ocurra más de una falla en un lote de 20 sobres?

b) Si la probabilidad de que falle la máquina es de 0.01 y se abrirá un lote de 500 sobres, ¿cuál es la probabilidad de que ocurran más de 8 fallas?

6.6 Distribución gamma y distribución exponencial

Aunque la distribución normal se puede utilizar para resolver muchos problemas de ingeniería y ciencias, aún hay numerosas situaciones que requieren diferentes tipos de funciones de densidad. En esta sección se estudiarán dos de estas funciones de densidad, la **distribución gamma** y la **distribución exponencial**.

Resulta que la distribución exponencial es un caso especial de la distribución gamma, y ambas tienen un gran número de aplicaciones. La distribución exponencial y la distribución gamma desempeñan un papel importante en la teoría de colas y en problemas de confiabilidad. Los tiempos entre llegadas en instalaciones de servicio y los tiempos de operación antes de que partes componentes y sistemas eléctricos empiecen a fallar a menudo se representan bien mediante la distribución exponencial. La relación entre la distribución gamma y la exponencial permite que la gamma se utilice en problemas similares. En la siguiente sección se presentarán más detalles y ejemplos.

La distribución gamma deriva su nombre de la bien conocida **función gamma**, que se estudia en muchas áreas de las matemáticas. Antes de estudiar la distribución gamma repasaremos esta función y algunas de sus propiedades importantes.

Definición 6.2: La **función gamma** se define como

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} \, dx, \quad \text{para} \quad \alpha > 0.$$

Las siguientes son algunas propiedades sencillas de la función gamma.

a) $\Gamma(n) = (n-1)(n-2)\cdots(1)\,\Gamma(1)$ para una integral positiva n.

Para ver la demostración, al integrar por partes con $u = x^{\alpha-1}$ y $dv = e^{-x}\,dx$, obtenemos

$$\Gamma(\alpha) = -e^{-x}\, x^{\alpha-1}\Big|_0^\infty + \int_0^\infty e^{-x}(\alpha-1)x^{\alpha-2}\,dx = (\alpha-1)\int_0^\infty x^{\alpha-2}e^{-x}\,dx,$$

para $\alpha > 1$, que produce la fórmula recursiva

$$\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1).$$

El resultado proviene de la aplicación repetida de la fórmula recursiva. Si utilizamos este resultado, podemos demostrar con facilidad las siguientes dos propiedades.

b) $\Gamma(n) = (n-1)!$ para una integral positiva n.

c) $\Gamma(1) = 1$.

Asimismo, tenemos la siguiente propiedad de $\Gamma(\alpha)$, que el lector deberá verificar (véase el ejercicio 6.39 de la página 206).

(véase el ejercicio 6.39 de la página 206)

d) $\Gamma(1/2) = \sqrt{\pi}$.

A continuación se define la **distribución gamma**.

Distribución gamma

La variable aleatoria continua X tiene una **distribución gamma**, con parámetros α y β, si su función de densidad está dada por

$$f(x;\alpha,\beta) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, & x > 0, \\ 0, & \text{en otro caso,} \end{cases}$$

donde $\alpha > 0$ y $\beta > 0$.

En la figura 6.28 se muestran gráficas de varias distribuciones gamma para ciertos valores específicos de los parámetros α y β. La distribución gamma especial para la que $\alpha = 1$ se llama **distribución exponencial**.

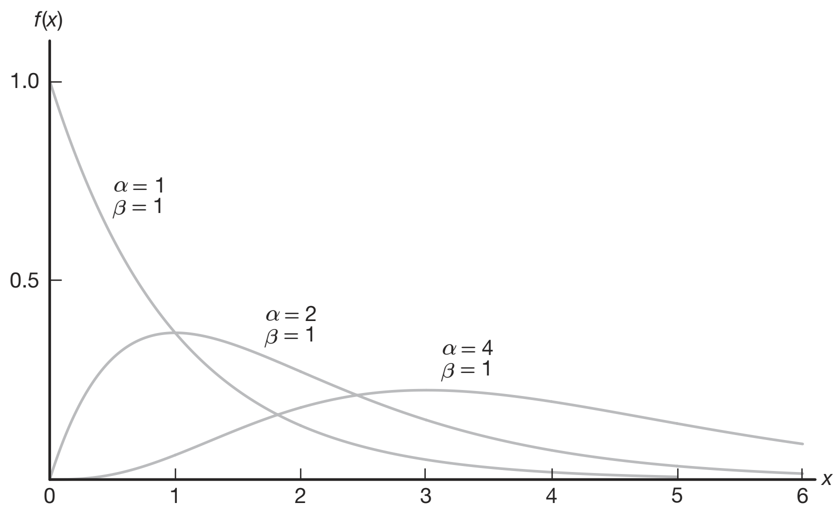

Figura 6.28: Distribuciones gamma.

Distribución exponencial

La variable aleatoria continua X tiene una **distribución exponencial**, con parámetro β, si su función de densidad es dada por

$$f(x;\beta) = \begin{cases} \frac{1}{\beta} e^{-x/\beta}, & x > 0, \\ 0, & \text{en otro caso,} \end{cases}$$

donde $\beta > 0$.

El siguiente teorema y corolario proporcionan la media y la varianza de la distribución gamma y la exponencial.

Teorema 6.4: La media y la varianza de la distribución gamma son

$$\mu = \alpha\beta \text{ y } \sigma^2 = \alpha\beta^2.$$

La demostración de este teorema se encuentra en el apéndice A.26.

Corolario 6.1: La media y la varianza de la distribución exponencial son

$$\mu = \beta \text{ y } \sigma^2 = \beta^2.$$

Relación con el proceso de Poisson

Continuaremos con las aplicaciones de la distribución exponencial y después regresaremos a la distribución gamma. Las aplicaciones más importantes de la distribución exponencial son situaciones donde se aplica el proceso de Poisson (véase la sección 5.5). El lector debería recordar que el proceso de Poisson permite utilizar la distribución discreta llamada distribución de Poisson. Recuerde que la distribución de Poisson se utiliza para calcular la probabilidad de números específicos de "eventos" durante un *periodo o espacio* particulares. En muchas aplicaciones la variable aleatoria es el tiempo o la cantidad de espacio. Por ejemplo, un ingeniero industrial se podría interesar en un modelo de tiempo T entre las llegadas en una intersección congestionada durante las horas de mayor afluencia en una ciudad grande. Una llegada representa el evento de Poisson.

La relación entre la distribución exponencial (a menudo denominada exponencial negativa) y el proceso de Poisson es muy simple. En el capítulo 5 la distribución de Poisson se desarrolló como una distribución de un solo parámetro con parámetro λ, donde λ se interpreta como el número medio de eventos por *unidad de "tiempo"*. Considere ahora la *variable aleatoria* descrita por el tiempo que se requiere para que ocurra el primer evento. Si utilizamos la distribución de Poisson, vemos que la probabilidad de que no ocurra algún evento, en el periodo hasta el tiempo t, es dada por

$$p(0; \lambda t) = \frac{e^{-\lambda t}(\lambda t)^0}{0!} = e^{-\lambda t}.$$

Ahora podemos utilizar lo anterior y hacer que X sea el tiempo para el primer evento de Poisson. La probabilidad de que la duración del tiempo hasta el primer evento exceda x es la misma que la probabilidad de que no ocurra algún evento de Poisson en x. Esto último, por supuesto, es dado por $e^{-\lambda x}$. Como resultado,

$$P(X > x) = e^{-\lambda x}.$$

Así, la función de distribución acumulativa para X es dada por

$$P(0 \leq X \leq x) = 1 - e^{-\lambda x}.$$

Ahora, para poder reconocer la presencia de la distribución exponencial, podemos diferenciar la función de distribución acumulativa anterior con el fin de obtener la función de densidad

$$f(x) = \lambda e^{-\lambda x},$$

que es la función de densidad de la distribución exponencial con $\lambda = 1/\beta$.

Aplicaciones de la distribución exponencial y la distribución gamma

En la explicación anterior establecimos las bases para la aplicación de la distribución exponencial en el "tiempo de llegada" o tiempo para problemas con eventos de Poisson. Aquí ilustraremos algunas aplicaciones de modelado y después procederemos a analizar el papel que la distribución gamma desempeña en ellas. Observe que la media de la distribución exponencial es el parámetro β, el recíproco del parámetro en la distribución de Poisson. El lector debería recordar que con frecuencia se dice que la distribución de Poisson no tiene memoria, lo cual implica que las ocurrencias en periodos sucesivos son independientes. El importante parámetro β es el tiempo promedio entre eventos. En la teoría de confiabilidad, donde la falla de equipo con frecuencia se ajusta a este proceso de Poisson, β se denomina **tiempo medio entre fallas**. Muchas descomposturas de equipo siguen el proceso de Poisson y por ello se aplica la distribución exponencial. Otras aplicaciones incluyen tiempos de supervivencia en experimentos biomédicos y tiempo de respuesta de computadoras.

En el siguiente ejemplo mostramos una aplicación simple de la distribución exponencial a un problema de confiabilidad. La distribución binomial también desempeña un papel en la solución.

Ejemplo 6.17: Suponga que un sistema contiene cierto tipo de componente cuyo tiempo de operación antes de fallar, en años, está dado por T. La variable aleatoria T se modela bien mediante la distribución exponencial con tiempo medio de operación antes de fallar $\beta = 5$. Si se instalan 5 de estos componentes en diferentes sistemas, ¿cuál es la probabilidad de que al final de 8 años al menos dos aún funcionen?

***Solución*:** La probabilidad de que un componente determinado siga funcionando después de 8 años es dada por

$$P(T > 8) = \frac{1}{5}\int_8^{\infty} e^{-t/5}\, dt = e^{-8/5} \approx 0.2.$$

Representemos con X el número de componentes que todavía funcionan después de 8 años. Entonces, utilizando la distribución binomial tenemos

$$P(X \geq 2) = \sum_{x=2}^{5} b(x; 5, 0.2) = 1 - \sum_{x=0}^{1} b(x; 5, 0.2) = 1 - 0.7373 = 0.2627.$$

En el capítulo 3 se incluyen ejercicios y ejemplos en los que el lector ya se enfrentó a la distribución exponencial. Otros que implican problemas de tiempo de espera y de confiabilidad se pueden encontrar en el ejemplo 6.24 y en los ejercicios y ejercicios de repaso al final de este capítulo.

La propiedad de falta de memoria y su efecto en la distribución exponencial

En los tipos de aplicación de la distribución exponencial en los problemas de confiabilidad y de tiempo de vida de una máquina o de un componente influye la **propiedad de**

falta de memoria de la distribución exponencial. Por ejemplo, en el caso de, digamos, un componente electrónico, en el que la distribución del tiempo de vida es exponencial, la probabilidad de que el componente dure, por ejemplo, t horas, es decir, $P(X \geq t)$, es igual que la probabilidad condicional

$$P(X \geq t_0 + t \mid X \geq t_0).$$

Entonces, si el componente "alcanza" las t_0 horas, la probabilidad de que dure otras t horas es igual que la probabilidad de que dure t horas. No hay "castigo" a través del desgaste como resultado de durar las primeras t_0 horas. Por lo tanto, cuando la propiedad de falta de memoria es justificada es más adecuada la distribución exponencial. Pero si la falla del componente es resultado del desgaste lento o gradual (como en el caso del desgaste mecánico), entonces la distribución exponencial no es aplicable y serían más adecuadas la distribución gamma o la de Weibull (sección 6.10).

La importancia de la distribución gamma radica en el hecho de que define una familia en la cual otras distribuciones son casos especiales. Pero la propia distribución gamma tiene aplicaciones importantes en tiempo de espera y teoría de confiabilidad. Mientras que la distribución exponencial describe el tiempo que transcurre hasta la ocurrencia de un evento de Poisson (o el tiempo entre eventos de Poisson), el tiempo (o espacio) que transcurre hasta que *ocurre un número específico de eventos de Poisson* es una variable aleatoria, cuya función de densidad es descrita por la distribución gamma. Este número específico de eventos es el parámetro α en la función de densidad gamma. De esta manera se facilita comprender que cuando $\alpha = 1$, ocurre el caso especial de la distribución exponencial. La densidad gamma se puede desarrollar a partir de su relación con el proceso de Poisson de la misma manera en que lo hicimos con la densidad exponencial. Los detalles se dejan al lector. El siguiente es un ejemplo numérico de cómo se utiliza la distribución gamma en una aplicación de tiempo de espera.

Ejemplo 6.18: Suponga que las llamadas telefónicas que llegan a un conmutador particular siguen un proceso de Poisson con un promedio de 5 llamadas entrantes por minuto. ¿Cuál es la probabilidad de que transcurra hasta un minuto en el momento en que han entrado 2 llamadas al conmutador?

***Solución*:** Se aplica el proceso de Poisson, con un lapso de tiempo hasta que ocurren 2 eventos de Poisson que sigue una distribución gamma con $\beta = 1/5$ y $\alpha = 2$. Denote con X el tiempo en minutos que transcurre antes de que lleguen 2 llamadas. La probabilidad que se requiere está dada por

$$P(X \leq 1) = \int_0^1 \frac{1}{\beta^2} x e^{-x/\beta}\, dx = 25 \int_0^1 x e^{-5x}\, dx = 1 - e^{-5}(1+5) = 0.96.$$

Mientras el origen de la distribución gamma trata con el tiempo (o espacio) hasta la ocurrencia de α eventos de Poisson, hay muchos ejemplos donde una distribución gamma funciona muy bien aunque no exista una estructura de Poisson clara. Esto es particularmente cierto para problemas de **tiempo de supervivencia** en aplicaciones de ingeniería y biomédicas.

Ejemplo 6.19: En un estudio biomédico con ratas se utiliza una investigación de respuesta a la dosis para determinar el efecto de la dosis de un tóxico en su tiempo de supervivencia. El tóxico es producido por el combustible que utilizan los aviones y, en consecuencia, descargan con frecuencia a la atmósfera. Para cierta dosis del tóxico, el estudio determina que el tiempo de supervivencia de las ratas, en semanas, tiene una distribución gamma con $\alpha = 5$ y $\beta = 10$. ¿Cuál es la probabilidad de que una rata no sobreviva más de 60 semanas?

Solución: Sea la variable aleatoria X el tiempo de supervivencia (tiempo hasta la muerte). La probabilidad que se requiere es

$$P(X \leq 60) = \frac{1}{\beta^5} \int_0^{60} \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(5)}\, dx.$$

La integral anterior se puede resolver mediante la **función gamma incompleta**, que se convierte en la función de distribución acumulativa para la distribución gamma. Esta función se escribe como

$$F(x;\alpha) = \int_0^{x} \frac{y^{\alpha-1} e^{-y}}{\Gamma(\alpha)}\, dy.$$

Si permitimos que $y = x/\beta$, de modo que $x = \beta y$, tenemos

$$P(X \leq 60) = \int_0^{6} \frac{y^4 e^{-y}}{\Gamma(5)}\, dy,$$

que se denota como $F(6; 5)$ en la tabla de la función gamma incompleta del apéndice A.23. Observe que esto permite un cálculo rápido de las probabilidades para la distribución gamma. De hecho, para este problema la probabilidad de que la rata no sobreviva más de 60 días es dada por

$$P(X \leq 60) = F(6;5) = 0.715.$$

Ejemplo 6.20: A partir de datos previos se sabe que la longitud de tiempo, en meses, entre las quejas de los clientes sobre cierto producto es una distribución gamma con $\alpha = 2$ y $\beta = 4$. Se realizaron cambios para hacer más estrictos los requerimientos del control de calidad después de los cuales pasaron 20 meses antes de la primera queja. ¿Parecería que los cambios realizados en el control de calidad resultaron eficaces?

Solución: Sea X el tiempo para que se presente la primera queja, el cual, en las condiciones anteriores a los cambios, seguía una distribución gamma con $\alpha = 2$ y $\beta = 4$. La pregunta se centra alrededor de qué tan raro es $X \geq 20$ dado que α y β permanecen con los valores 2 y 4, repectivamente. En otras palabras, en las condiciones anteriores ¿es razonable un "tiempo para la queja" tan grande como 20 meses? Por consiguiente, si seguimos la solución del ejemplo 6.19,

$$P(X \geq 20) = 1 - \frac{1}{\beta^\alpha} \int_0^{20} \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha)} dx.$$

De nuevo, usando $y = x/\beta$ tenemos

$$P(X \geq 20) = 1 - \int_0^{5} \frac{y e^{-y}}{\Gamma(2)}\, dy = 1 - F(5;2) = 1 - 0.96 = 0.04,$$

donde $F(5; 2) = 0.96$ se obtiene de la tabla A.23.

Como resultado, podríamos concluir que las condiciones de la distribución gamma con $\alpha = 2$ y $\beta = 4$ no son sustentadas por los datos de que un tiempo observado para la queja sea tan extenso como 20 meses. Entonces, es razonable concluir que el trabajo de control de calidad resultó eficaz.

Ejemplo 6.21: Considere el ejercicio 3.31 de la página 94. Con base en abundantes pruebas se determinó que el tiempo Y en años antes de que se requiera una reparación mayor para cierta lavadora se caracteriza por la función de densidad

$$f(y) = \begin{cases} \frac{1}{4} e^{-y/4}, & y \geq 0, \\ 0, & \text{en otro caso.} \end{cases}$$

Observe que Y es una variable aleatoria exponencial con $\mu = 4$ años. Se considera que la lavadora es una ganga si no hay probabilidades de que requiera una reparación mayor antes de cumplir 6 años de haber sido comprada. ¿Cuál es la probabilidad de $P(Y > 6)$? ¿Cuál es la probabilidad de que la lavadora requiera una reparación mayor durante el primer año?

Solución: Considere la función de distribución acumulativa $F(y)$ para la distribución exponencial,

$$F(y) = \frac{1}{\beta}\int_0^y e^{-t/\beta}\,dt = 1 - e^{-y/\beta}.$$

De manera que

$$P(Y > 6) = 1 - F(6) = e^{-3/2} = 0.2231.$$

Por lo tanto, la probabilidad de que la lavadora requiera una reparación mayor después de seis años es de 0.223. Desde luego, la probabilidad de que requiera reparación antes del sexto año es de 0.777. Así, se podría concluir que la lavadora no es realmente una ganga. La probabilidad de que se requiera una reparación mayor durante el primer año es

$$P(Y < 1) = 1 - e^{-1/4} = 1 - 0.779 = 0.221.$$

6.7 Distribución chi cuadrada

Otro caso especial muy importante de la distribución gamma se obtiene al permitir que $\alpha = v/2$ y $\beta = 2$, donde v es un entero positivo. Este resultado se conoce como **distribución chi cuadrada**. La distribución tiene un solo parámetro, v, denominado **grados de libertad**.

Distribución chi cuadrada

La variable aleatoria continua X tiene una **distribución chi cuadrada**, con v **grados de libertad**, si su función de densidad es dada por

$$f(x; v) = \begin{cases} \frac{1}{2^{v/2}\Gamma(v/2)} x^{v/2-1} e^{-x/2}, & x > 0, \\ 0, & \text{en otro caso,} \end{cases}$$

donde v es un entero positivo.

La distribución chi cuadrada desempeña un papel fundamental en la inferencia estadística. Tiene una aplicación considerable tanto en la metodología como en la teoría. Aunque no estudiaremos con detalle sus aplicaciones en este capítulo, es importante tener en cuenta que los capítulos 8, 9 y 16 contienen aplicaciones importantes. La distribución chi cuadrada es un componente importante de la prueba estadística de hipótesis y de la estimación estadística.

Los temas en los que se trata con distribuciones de muestreo, análisis de varianza y estadística no paramétrica implican el uso extenso de la distribución chi cuadrada.

Teorema 6.5: La media y la varianza de la distribución chi cuadrada son

$$\mu = v \text{ y } \sigma^2 = 2v.$$

6.8 Distribución beta

Una extensión de la distribución uniforme es la distribución beta. Primero definiremos una **función beta**.

Definición 6.3: Una **función beta** es definida por

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \text{ para } \alpha, \beta > 0,$$

donde $\Gamma(\alpha)$ es la función gamma.

Distribución beta

La variable aleatoria continua X tiene una **distribución beta** con los parámetros $\alpha > 0$ y $\beta > 0$, si su función de densidad es dada por

$$f(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & 0 < x < 1, \\ 0, & \text{en otro caso.} \end{cases}$$

Observe que la distribución uniforme sobre (0, 1) es una distribución beta con los parámetros $\alpha = 1$ y $\beta = 1$.

Teorema 6.6: La media y la varianza de una distribución beta en la que los parámetros α y β son

$$\mu = \frac{\alpha}{\alpha+\beta} \text{ y } \sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)},$$

respectivamente.

Para la distribución uniforme sobre (0, 1), la media y la varianza son

$$\mu = \frac{1}{1+1} = \frac{1}{2} \text{ y } \sigma^2 = \frac{(1)(1)}{(1+1)^2(1+1+1)} = \frac{1}{12},$$

respectivamente.

6.9 Distribución logarítmica normal

La distribución logarítmica normal se utiliza en una amplia variedad de aplicaciones. La distribución se aplica en casos donde una transformación logarítmica natural tiene como resultado una distribución normal.

Distribución logarítmica normal

La variable aleatoria continua X tiene una **distribución logarítmica normal** si la variable aleatoria $Y = \ln(X)$ tiene una distribución normal con media μ y desviación estándar σ. La función de densidad de X que resulta es

$$f(x; \mu, \sigma) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{1}{2\sigma^2}[\ln(x)-\mu]^2}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

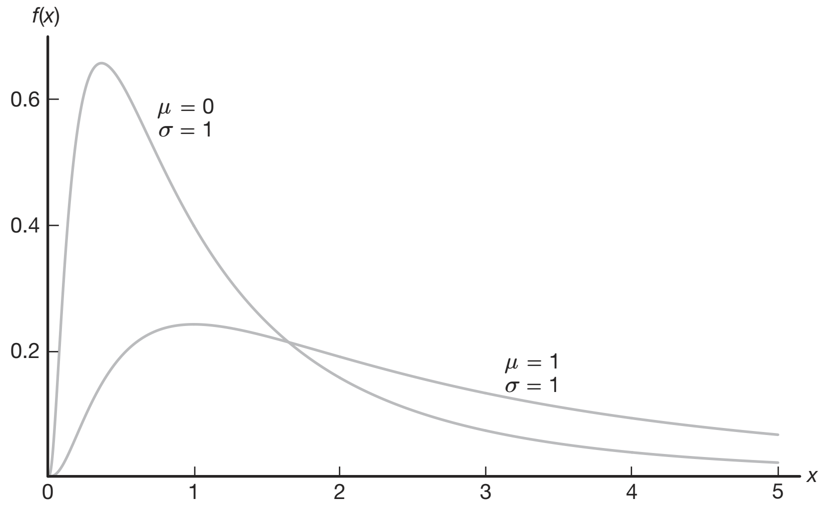

Figura 6.29: Distribuciones logarítmicas normales.

Las gráficas de las distribuciones logarítmicas normales se ilustran en la figura 6.29.

Teorema 6.7: La media y la varianza de la distribución logarítmica normal son

$$\mu = e^{\mu+\sigma^2/2} \text{ y } \sigma^2 = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1).$$

La función de distribución acumulativa es muy simple debido a su relación con la distribución normal. El uso de la función de distribución se ilustra con el siguiente ejemplo.

Ejemplo 6.22: Se sabe que históricamente la concentración de contaminantes producidos por plantas químicas exhiben un comportamiento que se parece a una distribución logarítmica normal. Esto es importante cuando se consideran cuestiones relacionadas con el cumplimiento de las regulaciones gubernamentales. Suponga que la concentración de cierto contaminante, en partes por millón, tiene una distribución logarítmica normal con los parámetros $\mu = 3.2$ y $\sigma = 1$. ¿Cuál es la probabilidad de que la concentración exceda 8 partes por millón?

***Solución*:** Sea la variable aleatoria X la concentración de contaminantes. Entonces

$$P(X > 8) = 1 - P(X \leq 8).$$

Como ln(X) tiene una distribución normal con media $\mu = 3.2$ y desviación estándar $\sigma = 1$,

$$P(X \leq 8) = \Phi\left[\frac{\ln(8) - 3.2}{1}\right] = \Phi(-1.12) = 0.1314.$$

Aquí, utilizamos el símbolo Φ para denotar la función de distribución acumulativa de la distribución normal estándar. Como resultado, la probabilidad de que la concentración del contaminante exceda 8 partes por millón es 0.1314.

Ejemplo 6.23: La vida, en miles de millas, de un cierto tipo de control electrónico para locomotoras tiene una distribución aproximadamente logarítmica normal con $\mu = 5.149$ y $\sigma = 0.737$. Calcule el quinto percentil de la vida de un control electrónico como éste.

***Solución*:** A partir de la tabla A.3 sabemos que $P(Z < -1.645) = 0.05$. Denote como X la vida del control electrónico. Puesto que $\ln(X)$ tiene una distribución normal con media $\mu = 5.149$ y $\sigma = 0.737$, el quinto percentil de X se calcula como

$$\ln(x) = 5.149 + (0.737)(-1.645) = 3.937.$$

Por lo tanto, $x = 51.265$. Esto significa que sólo 5% de los controles tendrán un tiempo de vida menor que 51,265 millas.

6.10 Distribución de Weibull (opcional)

La tecnología actual permite que los ingenieros diseñen muchos sistemas complicados cuya operación y seguridad dependen de la confiabilidad de los diversos componentes que conforman los sistemas. Por ejemplo, un fusible se puede quemar, una columna de acero se puede torcer o un dispositivo sensor de calor puede fallar. Componentes idénticos, sujetos a idénticas condiciones ambientales, fallarán en momentos diferentes e impredecibles. Ya examinamos el papel que desempeñan las distribuciones gamma y exponencial en estos tipos de problemas. Otra distribución que se ha utilizado ampliamente en años recientes para tratar con tales problemas es la **distribución de Weibull**, introducida por el físico sueco Waloddi Weibull en 1939.

Distribución de Weibull

La variable aleatoria continua X tiene una **distribución de Weibull**, con parámetros α y β, si su función de densidad es dada por

$$f(x; \alpha, \beta) = \begin{cases} \alpha\beta x^{\beta-1} e^{-\alpha x^{\beta}}, & x > 0, \\ 0, & \text{en otro caso,} \end{cases}$$

donde $\alpha > 0$ y $\beta > 0$.

En la figura 6.30 se ilustran las gráficas de la distribución de Weibull para $\alpha = 1$ y diversos valores del parámetro β. Vemos que las curvas cambian de manera considerable para diferentes valores del parámetro β. Si permitimos que $\beta = 1$, la distribución de Weibull se reduce a la distribución exponencial. Para valores de $\beta > 1$ las curvas adoptan ligeramente la forma de campana y se asemejan a las curvas normales, pero muestran algo de asimetría.

La media y la varianza de la distribución de Weibull se establecen en el siguiente teorema. Se solicita al lector que haga la demostración en el ejercicio 6.52 de la página 206.

Teorema 6.8: La media y la varianza de la distribución de Weibull son

$$\mu = \alpha^{-1/\beta}\Gamma\left(1 + \frac{1}{\beta}\right) \text{ y } \sigma^2 = \alpha^{-2/\beta}\left\{\Gamma\left(1 + \frac{2}{\beta}\right) - \left[\Gamma\left(1 + \frac{1}{\beta}\right)\right]^2\right\}.$$

Al igual que la distribución gamma y la exponencial, la distribución de Weibull se aplica a problemas de confiabilidad y de prueba de vida como los de **tiempo de operación**

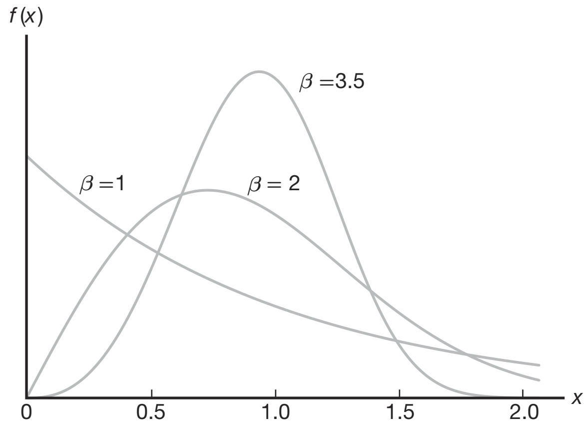

Figura 6.30: Distribuciones de Weibull ($\alpha = 1$).

antes de la falla o la **duración de la vida** de un componente, que se miden desde algún tiempo específico hasta que falla. Representemos este tiempo de operación antes de la falla mediante la variable aleatoria continua T, con función de densidad de probabilidad $f(t)$, donde $f(t)$ es la distribución de Weibull. Ésta tiene la flexibilidad inherente de no requerir la propiedad de falta de memoria de la distribución exponencial. La función de distribución acumulativa (fda) para la distribución de Weibull se puede escribir en forma cerrada y realmente es muy útil para calcular probabilidades.

Fda para la distribución de Weibull

La **función de distribución acumulativa para la distribución de Weibull** es dada por

$$F(x) = 1 - e^{-\alpha x^{\beta}}, \quad \text{para} \quad x \geq 0,$$

para $\alpha > 0$ y $\beta > 0$.

Ejemplo 6.24: El tiempo de vida X, en horas, de un artículo en el taller mecánico tiene una distribución de Weibull con $\alpha = 0.01$ y $\beta = 2$. ¿Cuál es la probabilidad de que falle antes de 8 horas de uso?

***Solución*:** $P(X < 8) = F(8) = 1 - e^{-(0.01)8^2} = 1 - 0.527 = 0.473.$

La tasa de fallas para la distribución de Weibull

Cuando se aplica la distribución de Weibull, con frecuencia es útil determinar la **tasa de fallas** (algunas veces denominada tasa de riesgo) para tener conocimiento del desgaste o deterioro del componente. Comencemos por definir la confiabilidad de un componente o producto como la *probabilidad de que funcione adecuadamente por al menos un tiempo específico en condiciones experimentales específicas*. Por lo tanto, si $R(t)$ se define como la confiabilidad del componente dado en el tiempo t, escribimos

$$R(t) = P(T > t) = \int_t^{\infty} f(t)\, dt = 1 - F(t),$$

donde $F(t)$ es la función de distribución acumulativa de T. La probabilidad condicional de que un componente fallará en el intervalo de $T = t$ a $T = t + \Delta t$, dado que sobrevive hasta el tiempo t, es

$$\frac{F(t + \Delta t) - F(t)}{R(t)}.$$

Al dividir esta proporción entre Δt y tomar el límite como $\Delta t \to 0$, obtenemos la **tasa de fallas**, denotada por $Z(t)$. De aquí,

$$Z(t) = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \frac{1}{R(t)} = \frac{F'(t)}{R(t)} = \frac{f(t)}{R(t)} = \frac{f(t)}{1 - F(t)},$$

que expresa la tasa de fallas en términos de la distribución del tiempo de operación antes de la falla.

Como $Z(t) = f(t)/[1 - F(t)]$, entonces la tasa de falla es dada como sigue:

Tasa de fallas para la distribución de Weibull

La **tasa de fallas** en el tiempo t para la distribución de Weibull es dada por

$$Z(t) = \alpha\beta t^{\beta-1}, \qquad t > 0.$$

Interpretación de la tasa de fallas

La cantidad $Z(t)$ es bien llamada tasa de fallas porque realmente cuantifica la tasa de cambio con el tiempo de la probabilidad condicional de que el componente dure una Δt adicional *dado que ha durado el tiempo t*. La tasa de disminución (o crecimiento) con el tiempo también es importante. Los siguientes puntos son fundamentales.

a) Si $\beta = 1$, la tasa de fallas $= \alpha$, es decir, una constante. Esto, como se indicó anteriormente, es el caso especial de la distribución exponencial en que predomina la falta de memoria.

b) Si $\beta > 1$, $Z(t)$ es una función creciente del tiempo t que indica que el componente se desgasta con el tiempo.

c) Si $\beta < 1$, $Z(t)$ es una función decreciente del tiempo t y, por lo tanto, el componente se fortalece o endurece con el paso del tiempo.

Por ejemplo, el artículo en el taller mecánico del ejemplo 6.24 tiene $\beta = 2$ y, por consiguiente, se desgasta con el tiempo. De hecho, la función de la tasa de fallas es dada por $Z(t) = .02t$. Por otro lado, suponga un parámetro donde $\beta = 3/4$ y $\alpha = 2$. En ese caso, $Z(t) = 1.5/t^{1/4}$ y, por lo tanto, el componente se hace más fuerte con el tiempo.

Ejercicios

6.39 Utilice la función gamma con $y = \sqrt{2x}$ para demostrar que $\Gamma(1/2) = \sqrt{\pi}$.

6.40 En cierta ciudad, el consumo diario de agua (en millones de litros) sigue aproximadamente una distribución gamma con $\alpha = 2$ y $\beta = 3$. Si la capacidad diaria de dicha ciudad es de 9 millones de litros de agua, ¿cuál es la probabilidad de que en cualquier día dado el suministro de agua sea inadecuado?

6.41 Si una variable aleatoria X tiene una distribución gamma con $\alpha = 2$ y $\beta = 1$, calcule $P(1.8 < X < 2.4)$.

6.42 Suponga que el tiempo, en horas, necesario para reparar una bomba de calor es una variable aleatoria X que tiene una distribución gamma con los parámetros $\alpha = 2$ y $\beta = 1/2$. ¿Cuál es la probabilidad de que la siguiente llamada de servicio requiera

a) a lo sumo una hora para reparar la bomba de calor?
b) al menos dos horas para reparar la bomba de calor?

6.43 *a*) Calcule la media y la varianza del consumo diario de agua del ejercicio 6.40.

b) De acuerdo con el teorema de Chebyshev, ¿hay por lo menos 3/4 de probabilidad de que el consumo de agua en cualquier día determinado caiga dentro de cuál intervalo?

6.44 En cierta ciudad el consumo diario de energía eléctrica, en millones de kilowatts-hora, es una variable aleatoria X que tiene una distribución gamma con media $\mu = 6$ y varianza $\sigma^2 = 12$.

a) Calcule los valores de α y β.
b) Calcule la probabilidad de que en cualquier día dado el consumo diario de energía exceda los 12 millones de kilowatts-hora.

6.45 El tiempo necesario para que un individuo sea atendido en una cafetería es una variable aleatoria que tiene una distribución exponencial con una media de 4 minutos. ¿Cuál es la probabilidad de que una persona sea atendida en menos de 3 minutos en al menos 4 de los siguientes 6 días?

6.46 La vida, en años, de cierto interruptor eléctrico tiene una distribución exponencial con una vida promedio de $\beta = 2$. Si 100 de estos interruptores se instalan en diferentes sistemas, ¿cuál es la probabilidad de que, a lo sumo, fallen 30 durante el primer año?

6.47 Suponga que la vida de servicio de la batería de un auxiliar auditivo, en años, es una variable aleatoria que tiene una distribución de Weibull con $\alpha = 1/2$ y $\beta = 2$.

a) ¿Cuánto tiempo se puede esperar que dure tal batería?
b) ¿Cuál es la probabilidad de que tal batería esté funcionando después de 2 años?

6.48 Derive la media y la varianza de la distribución beta.

6.49 Suponga que la variable aleatoria X tiene una distribución beta con $\alpha = 1$ y $\beta = 3$.

a) Determine la media y la mediana de X.
b) Determine la varianza de X.
c) Calcule la probabilidad de que $X > 1/3$.

6.50 Si la proporción de una marca de televisores que requiere servicio durante el primer año de operación es una variable aleatoria que tiene una distribución beta con $\alpha = 3$ y $\beta = 2$, ¿cuál es la probabilidad de que al menos 80% de los nuevos modelos de esta marca que se vendieron este año requieran servicio durante su primer año de operación?

6.51 Las vidas de ciertos sellos para automóvil tienen la distribución de Weibull con tasa de fallas $Z(t) = 1/\sqrt{t}$. Calcule la probabilidad de que tal sello aún esté intacto después de 4 años.

6.52 Derive la media y la varianza de la distribución de Weibull.

6.53 En una investigación biomédica se determinó que el tiempo de supervivencia, en semanas, de un animal cuando se le somete a cierta exposición de radiación gamma tiene una distribución gamma con $\alpha = 5$ y $\beta = 10$.

a) ¿Cuál es el tiempo medio de supervivencia de un animal seleccionado al azar del tipo que se utilizó en el experimento?
b) ¿Cuál es la desviación estándar del tiempo de supervivencia?
c) ¿Cuál es la probabilidad de que un animal sobreviva más de 30 semanas?

6.54 Se sabe que la vida, en semanas, de cierto tipo de transistor tiene una distribución gamma con una media de 10 semanas y una desviación estándar de $\sqrt{50}$ semanas.

a) ¿Cuál es la probabilidad de que un transistor de este tipo dure a lo sumo 50 semanas?
b) ¿Cuál es la probabilidad de que un transistor de este tipo no sobreviva las primeras 10 semanas?

6.55 El tiempo de respuesta de una computadora es una aplicación importante de las distribuciones gamma y exponencial. Suponga que un estudio de cierto sistema de cómputo revela que el tiempo de respuesta, en segundos, tiene una distribución exponencial con una media de 3 segundos.

a) ¿Cuál es la probabilidad de que el tiempo de respuesta exceda 5 segundos?
b) ¿Cuál es la probabilidad de que el tiempo de respuesta exceda 10 segundos?

6.56 Los datos de frecuencia a menudo tienen una distribución logarítmica normal. Se estudia el uso promedio de potencia (dB por hora) para una empresa específica y se sabe que tiene una distribución logarítmica normal con parámetros $\mu = 4$ y $\sigma = 2$. ¿Cuál es la probabilidad de que la empresa utilice más de 270 dB durante cualquier hora particular?

6.57 Para el ejercicio 6.56, ¿cuál es el uso de la potencia media (dB promedio por hora)? ¿Cuál es la varianza?

6.58 El número de automóviles que llegan a cierta intersección por minuto tiene una distribución de Poisson con una media de 5. Existe interés por el tiempo que transcurre antes de que 10 automóviles aparezcan en la intersección.

a) ¿Cuál es la probabilidad de que más de 10 automóviles aparezcan en la intersección durante cualquier minuto determinado?
b) ¿Cuál es la probabilidad de que transcurran más de 2 minutos antes de que lleguen 10 autos?

6.59 Considere la información del ejercicio 6.58.
a) ¿Cuál es la probabilidad de que transcurra más de 1 minuto entre llegadas?
b) ¿Cuál es el número medio de minutos que transcurre entre las llegadas?

6.60 Demuestre que la función de la tasa de fallas es dada por

$$Z(t) = \alpha\beta t^{\beta-1}, \quad t > 0,$$

si y sólo si la distribución del tiempo que transcurre antes de la falla es la distribución de Weibull

$$f(t) = \alpha\beta t^{\beta-1} e^{-\alpha t^{\beta}}, \quad t > 0.$$

Ejercicios de repaso

6.61 Según un estudio publicado por un grupo de sociólogos de la Universidad de Massachusetts, aproximadamente 49% de los consumidores de Valium en el estado de Massachusetts son empleados de oficina. ¿Cuál es la probabilidad de que entre 482 y 510 de los siguientes 1000 consumidores de Valium seleccionados al azar de dicho estado sean empleados de oficina?

6.62 La distribución exponencial se aplica con frecuencia a los tiempos de espera entre éxitos en un proceso de Poisson. Si el número de llamadas que se reciben por hora en un servicio de respuesta telefónica es una variable aleatoria de Poisson con el parámetro $\lambda = 6$, sabemos que el tiempo, en horas, entre llamadas sucesivas tiene una distribución exponencial con el parámetro $\beta = 1/6$. ¿Cuál es la probabilidad de esperar más de 15 minutos entre cualesquiera 2 llamadas sucesivas?

6.63 Cuando α es un entero positivo n, la distribución gamma también se conoce como **distribución de Erlang**. Al establecer que $\alpha = n$ en la distribución gamma de la página 195, la distribución de Erlang es

$$f(x) = \begin{cases} \frac{x^{n-1} e^{-x/\beta}}{\beta^n (n-1)!}, & x > 0, \\ 0, & \text{en otro caso.} \end{cases}$$

Se puede demostrar que si los tiempos entre eventos sucesivos son independientes, y cada uno tiene una distribución exponencial con el parámetro β, entonces el tiempo de espera total X transcurrido hasta que ocurran n eventos tiene la distribución de Erlang. Con referencia al ejercicio de repaso 6.62, ¿cuál es la probabilidad de que las siguientes 3 llamadas se reciban dentro de los siguientes 30 minutos?

6.64 Un fabricante de cierto tipo de máquina grande desea comprar remaches de uno de dos fabricantes. Es importante que la resistencia a la rotura de cada remache exceda 10,000 psi. Dos fabricantes (*A* y *B*) ofrecen este tipo de remache y ambos tienen remaches cuya resistencia a la rotura está distribuida de forma normal. Las resistencias promedio a la rotura para los fabricantes *A* y *B* son 14,000 psi y 13,000 psi, respectivamente. Las desviaciones estándar son 2000 psi y 1000 psi, respectivamente. ¿Cuál fabricante producirá, en promedio, el menor número de remaches defectuosos?

6.65 De acuerdo con un censo reciente, casi 65% de los hogares en Estados Unidos se componen de una o dos personas. Si se supone que este porcentaje sigue siendo válido en la actualidad, ¿cuál es la probabilidad de que entre 590 y 625 de los siguientes 1000 hogares seleccionados al azar en Estados Unidos consten de una o dos personas?

6.66 Cierto tipo de dispositivo tiene una tasa de fallas anunciada de 0.01 por hora. La tasa de fallas es constante y se aplica la distribución exponencial.
a) ¿Cuál es el tiempo promedio que transcurre antes de la falla?
b) ¿Cuál es la probabilidad de que pasen 200 horas antes de que se observe una falla?

6.67 En una planta de procesamiento químico es importante que el rendimiento de cierto tipo de producto de un lote se mantenga por arriba de 80%. Si permanece por debajo de 80% durante un tiempo prolongado, la empresa pierde dinero. Los lotes producidos ocasionalmente con defectos son de poco interés, pero si varios lotes por día resultan defectuosos, la planta se detiene y se realizan ajustes. Se sabe que el rendimiento se distribuye normalmente con una desviación estándar de 4%.

a) ¿Cuál es la probabilidad de una “falsa alarma” (rendimiento por debajo de 80%) cuando el rendimiento promedio es en realidad de 85%?

b) ¿Cuál es la probabilidad de que un lote tenga un rendimiento que exceda el 80% cuando en realidad el rendimiento promedio es de 79%?

6.68 Para un componente eléctrico que tiene una tasa de fallas de una vez cada 5 horas es importante considerar el tiempo que transcurre para que fallen 2 componentes.

a) Suponiendo que se aplica la distribución gamma, ¿cuál es el tiempo promedio que transcurre para que fallen 2 componentes?

b) ¿Cuál es la probabilidad de que transcurran 12 horas antes de que fallen 2 componentes?

6.69 Se establece que la elongación de una barra de acero bajo una carga particular se distribuye normalmente con una media de 0.05 pulgadas y $\sigma = 0.01$ pulgadas. Calcule la probabilidad de que el alargamiento esté

a) por arriba de 0.1 pulgadas;

b) por abajo de 0.04 pulgadas;

c) entre 0.025 y 0.065 pulgadas.

6.70 Se sabe que un satélite controlado tiene un error (distancia del objetivo) que se distribuye normalmente con una media 0 y una desviación estándar de 4 pies. El fabricante del satélite define un éxito como un disparo en el cual el satélite llega a 10 pies del objetivo. Calcule la probabilidad de que el satélite falle.

6.71 Un técnico planea probar cierto tipo de resina desarrollada en el laboratorio para determinar la naturaleza del tiempo que transcurre antes de que se logre el pegado. Se sabe que el tiempo promedio para el pegado es de 3 horas y que la desviación estándar es de 0.5 horas. Un producto se considerará indeseable si el tiempo de pegado es menor de una hora o mayor de 4 horas. Comente sobre la utilidad de la resina. ¿Con qué frecuencia su desempeño se considera indeseable? Suponga que el tiempo para la unión se distribuye normalmente.

6.72 Considere la información del ejercicio de repaso 6.66. ¿Cuál es la probabilidad de que transcurran menos de 200 horas antes de que ocurran 2 fallas?

6.73 Para el ejercicio de repaso 6.72, ¿cuál es la media y la varianza del tiempo que transcurre antes de que ocurran 2 fallas?

6.74 Se sabe que la tasa promedio de uso de agua (en miles de galones por hora) en cierta comunidad implica la distribución logarítmica normal con los parámetros $\mu = 5$ y $\sigma = 2$. Para propósitos de planeación es importante tener información sobre los periodos de alto consumo. ¿Cuál es la probabilidad de que, para cualquier hora determinada, se usen 50,000 galones de agua?

6.75 Para el ejercicio de repaso 6.74, ¿cuál es la media del uso de agua por hora promedio en miles de galones?

6.76 En el ejercicio 6.54 de la página 206 se supone que la vida de un transistor tiene una distribución gamma con una media de 10 semanas y una desviación estándar de $\sqrt{50}$ semanas. Suponga que la distribución gamma es incorrecta y que se trata de una distribución normal.

a) ¿Cuál es la probabilidad de que el transistor dure a lo sumo 50 semanas?

b) ¿Cuál es la probabilidad de que el transistor no sobreviva las primeras 10 semanas?

c) Comente acerca de la diferencia entre los resultados que obtuvo aquí y los que se obtuvieron en el ejercicio 6.54 de la página 206.

6.77 La distribución beta tiene muchas aplicaciones en problemas de confiabilidad, donde la variable aleatoria básica es una proporción, como sucede en el contexto práctico que se ilustra en el ejercicio 6.50 de la página 206. En este apartado considere el ejercicio de repaso 3.73 de la página 108. Las impurezas en el lote del producto de un proceso químico reflejan un problema grave. Se sabe que la proporción de impurezas Y en un lote tiene la siguiente función de densidad

$$f(y) = \begin{cases} 10(1-y)^9, & 0 \le y \le 1, \\ 0, & \text{en otro caso.} \end{cases}$$

a) Verifique que la anterior sea una función de densidad válida.

b) ¿Cuál es la probabilidad de que un lote se considere no aceptable (es decir, $Y > 0.6$)?

c) ¿Cuáles son los parámetros α y β de la distribución beta que se ilustra aquí?

d) La media de la distribución beta es $\frac{\alpha}{\alpha+\beta}$ ¿Cuál es la proporción media de impurezas en el lote?

e) La varianza de una variable aleatoria beta distribuida es

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

¿Cuál es la varianza de Y en este problema?

6.78 Considere ahora el ejercicio de repaso 3.74 de la página 108. La función de densidad del tiempo Z entre las llamadas, en minutos, a una empresa de suministro eléctrico es dada por

$$f(z) = \begin{cases} \frac{1}{10}e^{-z/10}, & 0 < z < \infty, \\ 0, & \text{en otro caso.} \end{cases}$$

a) ¿Cuál es el tiempo medio entre llamadas?
b) ¿Cuál es la varianza en el tiempo entre llamadas?
c) ¿Cuál es la probabilidad de que el tiempo entre llamadas supere la media?

6.79 Considere el ejercicio de repaso 6.78. Dada la suposición de la distribución exponencial, ¿cuál es el número medio de llamadas por hora? ¿Cuál es la varianza en el número de llamadas por hora?

6.80 En un proyecto experimental sobre el factor humano se determinó que el tiempo de reacción de un piloto ante un estímulo visual es distribuido normalmente con una media de $1/2$ segundo y una desviación estándar de $2/5$ de segundo.
a) ¿Cuál es la probabilidad de que una reacción del piloto tome más de 0.3 segundos?
b) ¿Qué tiempo de reacción se excede el 95% de las veces?

6.81 El tiempo que transcurre entre las fallas de una pieza esencial de equipo es importante en la decisión del uso de equipo auxiliar. Un ingeniero cree que el mejor modelo para el tiempo entre las fallas de un generador es la distribución exponencial con una media de 15 días.
a) Si el generador acaba de fallar, ¿cuál es la probabilidad de que falle en los siguientes 21 días?
b) ¿Cuál es la probabilidad de que el generador funcione durante 30 días sin fallar?

6.82 El periodo de vida de una broca en una operación mecánica, en horas, tiene una distribución de Weibull con $\alpha = 2$ y $\beta = 50$. Calcule la probabilidad de que la broca falle antes de 10 horas de uso.

6.83 Calcule la fda para la distribución de Weibull. [*Sugerencia*: En la definición de una fda haga la transformación $z = y^{\beta}$].

6.84 Explique por qué la naturaleza del escenario en el ejercicio de repaso 6.82 probablemente no se preste a la distribución exponencial.

6.85 A partir de la relación entre la variable aleatoria chi cuadrada y la variable aleatoria gamma, demuestre que la media de la variable aleatoria chi cuadrada es v y que la varianza es $2v$.

6.86 El tiempo que le toma a un usuario de computadora leer su correo electrónico, en segundos, se distribuye como una variable aleatoria logarítmica normal con $\mu = 1.8$ y $\sigma^2 = 4.0$.
a) ¿Cuál es la probabilidad de que el usuario lea el correo durante más de 20 segundos? ¿Y por más de un minuto?
b) ¿Cuál es la probabilidad de que el usuario lea el correo durante un tiempo que sea igual a la media de la distribución logarítmica normal subyacente?

6.87 **Proyecto de grupo:** Pida a grupos de estudiantes que observen durante 2 semanas el número de personas que entra a una cafetería o restaurante de comida rápida específico en el transcurso de una hora, empezando a la misma hora cada día. La hora deberá ser la de mayor tránsito en la cafetería o restaurante. Los datos reunidos corresponderán al número de clientes que entran al lugar durante cada lapso de media hora. De esta manera, cada día se recolectarán 2 datos. Supongamos que la variable aleatoria X, el número de personas que entra cada media hora, tiene una distribución de Poisson. Los estudiantes deberán calcular la media y la varianza muestrales de X utilizando los 28 datos obtenidos.
a) ¿Qué evidencia hay de que la distribución de Poisson es o no correcta?
b) Dado que X es una variable de Poisson, ¿cuál es la distribución de T, el tiempo entre la llegada de las personas al lugar durante un lapso de media hora? Proporcione un estimado numérico del parámetro de esa distribución.
c) Proporcione un estimado de la probabilidad de que el lapso de tiempo entre las 2 llegadas sea menor de 15 minutos.
d) ¿Cuál es la probabilidad estimada de que el lapso entre las 2 llegadas sea mayor de 10 minutos?
e) ¿Cuál es la probabilidad estimada de que 20 minutos después de iniciar la recolección de datos ningún cliente haya llegado?

6.11 Posibles riesgos y errores conceptuales; relación con el material de otros capítulos

Muchos de los riesgos en el uso del material de este capítulo son muy similares a los del capítulo 5. Uno de los peores abusos de la estadística consiste en suponer que se trata de una distribución normal haciendo algún tipo de inferencia estadística, cuando en realidad no es normal. En los capítulos 10 al 15 el lector estudiará las pruebas de hipótesis, en las que se asume normalidad. Además, se le recordará al lector que hay **pruebas de la bondad de ajuste**, además de las rutinas gráficas que se examinan en los capítulos 8 y 10, que permiten verificar los datos para determinar si es razonable la suposición de normalidad.

Debemos hacer advertencias similares con respecto a las suposiciones que a menudo se hacen sobre otras distribuciones, además de la curva normal. En este libro se han presentado ejemplos en los que es necesario calcular las probabilidades de falla de ciertos productos o la probabilidad de recibir una queja durante cierto periodo. Se suelen hacer suposiciones con respecto a cierto tipo de distribución, así como a los valores de los parámetros de la distribución. Observe que los problemas de ejemplo incluyen los valores de los parámetros (por ejemplo, el valor de β para la distribución exponencial). No obstante, en los problemas de la vida real los valores de los parámetros deben ser estimaciones de experiencias o datos reales. Observe el énfasis que se pone en la estimación en los proyectos que aparecen en los capítulos 1, 5 y 6, así como la referencia que se hace en el capítulo 5 a las estimación de parámetros, tema que se analizará ampliamente a partir del capítulo 9.

Capítulo 7

Funciones de variables aleatorias (opcional)

7.1 Introducción

Este capítulo contiene un amplio espectro de material. Los capítulos 5 y 6 tratan tipos específicos de distribuciones, tanto discretas como continuas. Éstas son distribuciones que suelen aplicarse en muchos campos, por ejemplo en el de la confiabilidad, el de control de calidad y el de muestreo de aceptación. En este capítulo comenzamos a estudiar un tema más general: el de la distribución de funciones de variables aleatorias. Se presentan las técnicas generales y se ilustran con ejemplos. Las presentaciones van seguidas por un concepto relacionado, el de *funciones generadoras de momentos*, que pueden ser útiles para el aprendizaje de distribuciones de funciones lineales de variables aleatorias.

En los métodos estadísticos estándar, el resultado de la prueba de hipótesis estadísticas, la estimación, o incluso las gráficas estadísticas, no involucra a una sola variable aleatoria sino a *funciones de una o más variables aleatorias*. Como resultado, la inferencia estadística requiere la distribución de tales funciones. Por ejemplo, es común que se utilicen **promedios de variables aleatorias**. Además, las sumatorias y las combinaciones lineales más generales son importantes. Con frecuencia nos interesa la distribución de las sumas de cuadrados de variables aleatorias, en particular la manera en que se utilizan las técnicas del análisis de varianza, las cuales se estudiarán en los capítulos 11 a 14.

7.2 Transformaciones de variables

Con frecuencia, en la estadística se enfrenta la necesidad de derivar la distribución de probabilidad de una función de una o más variables aleatorias. Por ejemplo, suponga que X es una variable aleatoria discreta con distribución de probabilidad $f(x)$, suponga también que $Y = u(X)$ define una transformación uno a uno entre los valores de X y Y. Queremos encontrar la distribución de probabilidad de Y. Es importante notar que la transformación uno a uno implica que cada valor x está relacionado con un, y sólo un, valor $y = u(x)$, y que cada valor y está relacionado con un, y sólo un, valor $x = w(y)$, donde $w(y)$ se obtiene al resolver $y = u(x)$ para x en términos de y.

A partir de lo expuesto respecto a las distribuciones de probabilidad discreta en el capítulo 3, nos quedó claro que la variable aleatoria Y toma el valor y cuando X toma el valor $w(y)$. En consecuencia, la distribución de probabilidad de Y es dada por

$$g(y) = P(Y = y) = P[X = w(y)] = f[w(y)].$$

Teorema 7.1: Suponga que X es una variable aleatoria **discreta** con distribución de probabilidad $f(x)$. Definamos con $Y = u(X)$ una transformación uno a uno entre los valores de X y Y, de manera que la ecuación $y = u(x)$ se resuelva exclusivamente para x en términos de y, digamos, $x = w(y)$. Entonces, la distribución de probabilidad de Y es

$$g(y) = f[w(y)].$$

Ejemplo 7.1: Sea X una variable aleatoria geométrica con la siguiente distribución de probabilidad

$$f(x) = \frac{3}{4}\left(\frac{1}{4}\right)^{x-1}, \qquad x = 1, 2, 3, \ldots.$$

Calcule la distribución de probabilidad de la variable aleatoria $Y = X^2$.

***Solución*:** Como todos los valores de X son positivos, la transformación define una correspondencia uno a uno entre los valores x y y, $y = x^2$ y $x = \sqrt{y}$. Por lo tanto,

$$g(y) = \begin{cases} f(\sqrt{y}) = \frac{3}{4}\left(\frac{1}{4}\right)^{\sqrt{y}-1}, & y = 1, 4, 9, \ldots, \\ 0, & \text{en cualquier caso.} \end{cases}$$

De manera similar, para una transformación de dos dimensiones, tenemos el resultado en el teorema 7.2.

Teorema 7.2: Suponga que X_1 y X_2 son variables aleatorias **discretas**, con distribución de probabilidad conjunta $f(x_1, x_2)$. Definamos con $Y_1 = u_1(X_1, X_2)$ y $Y_2 = u_2(X_1, X_2)$ una transformación uno a uno entre los puntos (x_1, x_2) y (y_1, y_2), de manera que las ecuaciones

$$y_1 = u_1(x_1, x_2) \qquad \text{y} \qquad y_2 = u_2(x_1, x_2)$$

se pueden resolver exclusivamente para x_1 y x_2 en términos de y_1 y y_2, digamos $x_1 = w_1(y_1, y_2)$ y $x_2 = w_2(y_1, y_2)$. Entonces, la distribución de probabilidad conjunta de Y_1 y Y_2 es

$$g(y_1, y_2) = f[w_1(y_1, y_2), w_2(y_1, y_2)].$$

El teorema 7.2 es muy útil para encontrar la distribución de alguna variable aleatoria $Y_1 = u_1(X_1, X_2)$, donde X_1 y X_2 son variables aleatorias discretas con distribución de probabilidad conjunta $f(x_1, x_2)$. Definimos simplemente una segunda función, digamos $Y_2 = u_2(X_1, X_2)$, manteniendo una correspondencia uno a uno entre los puntos (x_1, x_2) y (y_1, y_2), y obtenemos la distribución de probabilidad conjunta $g(y_1, y_2)$. La distribución de Y_1 es precisamente la distribución marginal de $g(y_1, y_2)$ que se encuentra sumando los valores y_2. Si denotamos la distribución de Y_1 con $h(y_1)$, podemos escribir

$$h(y_1) = \sum_{y_2} g(y_1, y_2).$$

Ejemplo 7.2: Sean X_1 y X_2 dos variables aleatorias independientes que tienen distribuciones de Poisson con los parámetros μ_1 y μ_2, respectivamente. Calcule la distribución de la variable aleatoria $Y_1 = X_1 + X_2$.

***Solución*:** Como X_1 y X_2 son independientes, podemos escribir

$$f(x_1, x_2) = f(x_1)f(x_2) = \frac{e^{-\mu_1}\mu_1^{x_1}}{x_1!}\frac{e^{-\mu_2}\mu_2^{x_2}}{x_2!} = \frac{e^{-(\mu_1+\mu_2)}\mu_1^{x_1}\mu_2^{x_2}}{x_1!x_2!},$$

donde $x_1 = 0, 1, 2,...$ y $x_2 = 0, 1, 2,....$ Definamos ahora una segunda variable aleatoria, digamos $Y_2 = X_2$. Las funciones inversas son dadas por $x_1 = y_1 - y_2$ y $x_2 = y_2$. Si usamos el teorema 7.2, encontramos que la distribución de probabilidad conjunta de Y_1 y Y_2 es

$$g(y_1, y_2) = \frac{e^{-(\mu_1+\mu_2)}\mu_1^{y_1-y_2}\mu_2^{y_2}}{(y_1 - y_2)!y_2!},$$

donde $y_1 = 0, 1, 2,...$ y $y_2 = 0, 1, 2,..., y_1$. Advierta que, como $x_1 > 0$, la transformación $x_1 = y_1 - x_2$ implica que y_2 y, por lo tanto, x_2 siempre deben ser menores o iguales que y_1. En consecuencia, la distribución de probabilidad marginal de Y_1 es

$$\begin{aligned} h(y_1) &= \sum_{y_2=0}^{y_1} g(y_1, y_2) = e^{-(\mu_1+\mu_2)} \sum_{y_2=0}^{y_1} \frac{\mu_1^{y_1-y_2}\mu_2^{y_2}}{(y_1 - y_2)!y_2!} \\ &= \frac{e^{-(\mu_1+\mu_2)}}{y_1!} \sum_{y_2=0}^{y_1} \frac{y_1!}{y_2!(y_1 - y_2)!}\mu_1^{y_1-y_2}\mu_2^{y_2} \\ &= \frac{e^{-(\mu_1+\mu_2)}}{y_1!} \sum_{y_2=0}^{y_1} \binom{y_1}{y_2}\mu_1^{y_1-y_2}\mu_2^{y_2}. \end{aligned}$$

Al reconocer esta suma como la expansión binomial de $(\mu_1 + \mu_2)^{y_1}$, obtenemos

$$h(y_1) = \frac{e^{-(\mu_1+\mu_2)}(\mu_1 + \mu_2)^{y_1}}{y_1!}, \qquad y_1 = 0, 1, 2, \ldots,$$

a partir de lo cual concluimos que la suma de las dos variables aleatorias independientes que tienen distribuciones de Poisson, con los parámetros μ_1 y μ_2, tiene una distribución de Poisson con el parámetro $\mu_1 + \mu_2$. ▟

Para calcular la distribución de probabilidad de la variable aleatoria $Y = u(X)$, cuando X es una variable aleatoria continua y la transformación es uno a uno, necesitaremos el teorema 7.3. La demostración de este teorema se deja al lector.

Teorema 7.3: Suponga que X es una variable aleatoria **continua** con distribución de probabilidad $f(x)$. Definamos con $Y = u(X)$ una correspondencia uno a uno entre los valores de X y Y, de manera que la ecuación $y = u(x)$ se resuelva exclusivamente para x en términos de y, digamos $x = w(y)$. Entonces, la distribución de probabilidad de Y es

$$g(y) = f[w(y)]|J|,$$

donde $J = w'(y)$ y se llama **jacobiano** de la transformación.

Ejemplo 7.3: Sea X una variable aleatoria continua con la siguiente distribución de probabilidad

$$f(x) = \begin{cases} \frac{x}{12}, & 1 < x < 5, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Calcule la distribución de probabilidad de la variable aleatoria $Y = 2X - 3$.

***Solución*:** La solución inversa de $y = 2x - 3$ produce $x = (y + 3)/2$, de la que obtenemos $J = w'(y) = dx/dy = 1/2$. Por lo tanto, usando el teorema 7.3 encontramos que la función de densidad de Y es

$$g(y) = \begin{cases} \frac{(y+3)/2}{12}\left(\frac{1}{2}\right) = \frac{y+3}{48}, & -1 < y < 7, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Para calcular la distribución de probabilidad conjunta de las variables aleatorias $Y_1 = u_1(X_1, X_2)$ y $Y_2 = u_2(X_1, X_2)$, cuando X_1 y X_2 son continuas y la transformación es uno a uno, necesitamos un teorema adicional análogo al teorema 7.2, el cual establecemos sin demostración.

Teorema 7.4: Suponga que X_1 y X_2 son variables aleatorias **continuas** con distribución de probabilidad conjunta $f(x_1, x_2)$. Definamos con $Y_1 = u_1(X_1, X_2)$ y $Y_2 = u_2(X_1, X_2)$ una transformación uno a uno entre los puntos (x_1, x_2) y (y_1, y_2), de manera que las ecuaciones $y_1 = u_1(x_1, x_2)$ y $y_2 = u_2(x_1, x_2)$ se resuelven exclusivamente para x_1 y x_2 en términos de y_1 y y_2, digamos $x_1 = w_1(y_1, y_2)$ y $x_2 = w_2(y_1, y_2)$. Entonces, la distribución de probabilidad conjunta de Y_1 y Y_2 es

$$g(y_1, y_2) = f[w_1(y_1, y_2), w_2(y_1, y_2)]|J|,$$

donde el jacobiano es el determinante 2×2

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix}$$

y $\frac{\partial x_1}{\partial y_1}$ es simplemente la derivada de $x_1 = w_1(y_1, y_2)$ respecto a y_1, con y_2 constante, que en cálculo se denomina derivada parcial de x_1 respecto a y_1. Las otras derivadas parciales se definen de manera similar.

Ejemplo 7.4: Sean X_1 y X_2 dos variables aleatorias continuas con la siguiente distribución de probabilidad conjunta

$$f(x_1, x_2) = \begin{cases} 4x_1x_2, & 0 < x_1 < 1,\ 0 < x_2 < 1, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Calcule la distribución de probabilidad conjunta de $Y_1 = X_1^2$ y $Y_2 = X_1X_2$.

***Solución*:** Las soluciones inversas de $y_1 = x_1^2$ y $y_2 = x_1x_2$ son $x_1 = \sqrt{y_1}$ y $x_2 = y_2/\sqrt{y_1}$, de las que obtenemos

$$J = \begin{vmatrix} 1/(2\sqrt{y_1}) & 0 \\ -y_2/2y_1^{3/2} & 1/\sqrt{y_1} \end{vmatrix} = \frac{1}{2y_1}.$$

Para determinar el conjunto B de puntos en el plano y_1y_2 en el que se traza el conjunto A de puntos en el plano x_1x_2 escribimos

$$x_1 = \sqrt{y_1} \qquad \text{y} \qquad x_2 = y_2/\sqrt{y_1}.$$

Luego, al establecer $x_1 = 0$, $x_2 = 0$, $x_1 = 1$ y $x_2 = 1$, las fronteras del conjunto A se transforman en $y_1 = 0$, $y_2 = 0$, $y_1 = 1$ y $y_2 = \sqrt{y_1}$ o $y_2^2 = y_1$. Las dos regiones se ilustran en la figura 7.1. Al trazar el conjunto $A = \{(x_1, x_2) \mid 0 < x_1 < 1, 0 < x_2 < 1\}$ en el conjunto $B = \{(y_1, y_2) \mid y_2^2 < y_1 < 1,\ 0 < y_2 < 1\}$, se vuelve evidente que la transformación es uno a uno. Del teorema 7.4, la distribución de probabilidad conjunta de Y_1 y Y_2 es

$$g(y_1, y_2) = 4(\sqrt{y_1})\frac{y_2}{\sqrt{y_1}}\frac{1}{2y_1} = \begin{cases} \frac{2y_2}{y_1}, & y_2^2 < y_1 < 1,\ 0 < y_2 < 1, \\ 0, & \text{en cualquier caso.} \end{cases}$$

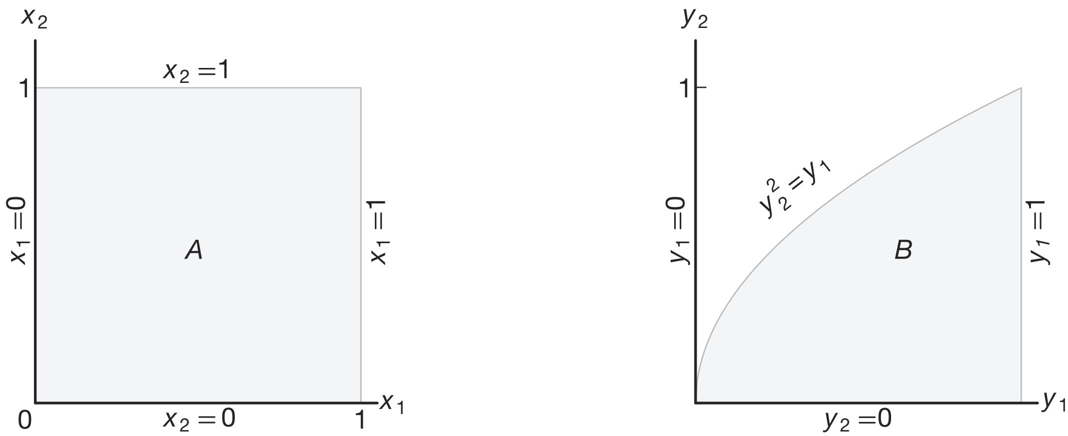

Figura 7.1: Gráfica del conjunto A en el conjunto B.

A menudo surgen problemas cuando deseamos encontrar la distribución de probabilidad de la variable aleatoria $Y = u(X)$ y X es una variable aleatoria continua y la transformación no es uno a uno. Es decir, a cada valor x le corresponde exactamente un valor y; pero a cada valor y le corresponde más de un valor x. Por ejemplo, suponga que $f(x)$ es positiva en el intervalo $-1 < x < 2$ y cero en cualquier caso. Considere la transformación $y = x^2$. En este caso, $x = \pm\sqrt{y}$ para $0 < y < 1$ y $x = \sqrt{y}$ para $1 < y < 4$. Para el intervalo $1 < y < 4$, la distribución de probabilidad de Y se calcula como antes, con el teorema 7.3. Es decir,

$$g(y) = f[w(y)]|J| = \frac{f(\sqrt{y})}{2\sqrt{y}}, \qquad 1 < y < 4.$$

Sin embargo, cuando $0 < y < 1$, podemos dividir el intervalo $-1 < x < 1$ para obtener las dos funciones inversas

$$x = -\sqrt{y}, \qquad -1 < x < 0, \qquad \text{y} \qquad x = \sqrt{y}, \qquad 0 < x < 1.$$

Entonces, a todo valor y le corresponde un solo valor x para cada partición. En la figura 7.2 vemos que

$$P(a < Y < b) = P(-\sqrt{b} < X < -\sqrt{a}) + P(\sqrt{a} < X < \sqrt{b})$$
$$= \int_{-\sqrt{b}}^{-\sqrt{a}} f(x)\ dx + \int_{\sqrt{a}}^{\sqrt{b}} f(x)\ dx.$$

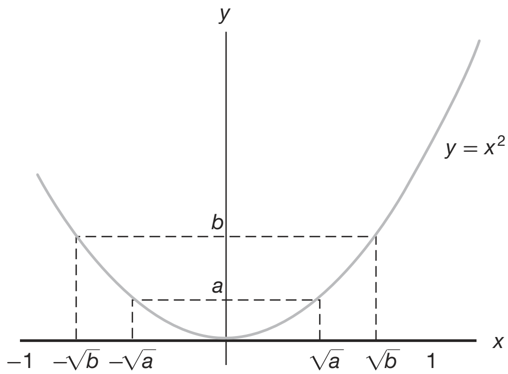

Figura 7.2: Función decreciente y creciente.

Al cambiar la variable de integración de x a y, obtenemos

$$P(a < Y < b) = \int_{b}^{a} f(-\sqrt{y})J_1\ dy + \int_{a}^{b} f(\sqrt{y})J_2\ dy$$
$$= -\int_{a}^{b} f(-\sqrt{y})J_1\ dy + \int_{a}^{b} f(\sqrt{y})J_2\ dy,$$

donde

$$J_1 = \frac{d(-\sqrt{y})}{dy} = \frac{-1}{2\sqrt{y}} = -|J_1|$$

y

$$J_2 = \frac{d(\sqrt{y})}{dy} = \frac{1}{2\sqrt{y}} = |J_2|.$$

Por lo tanto, podemos escribir

$$P(a < Y < b) = \int_{a}^{b} [f(-\sqrt{y})|J_1| + f(\sqrt{y})|J_2|]\ dy,$$

y entonces

$$g(y) = f(-\sqrt{y})|J_1| + f(\sqrt{y})|J_2| = \frac{f(-\sqrt{y}) + f(\sqrt{y})}{2\sqrt{y}}, \qquad 0 < y < 1.$$

La distribución de probabilidad de Y para $0 < y < 4$ se puede escribir ahora como

$$g(y) = \begin{cases} \frac{f(-\sqrt{y})+f(\sqrt{y})}{2\sqrt{y}}, & 0 < y < 1, \\ \frac{f(\sqrt{y})}{2\sqrt{y}}, & 1 < y < 4, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Este procedimiento para calcular $g(y)$ cuando $0 < y < 1$ se generaliza en el teorema 7.5 para k funciones inversas. Para transformaciones de funciones de diversas variables que no son uno a uno se recomienda al lector *Introduction to Mathematical Statistics* de Hogg, McKean y Craig (2005; véase la bibliografía).

Teorema 7.5: Suponga que X es una variable aleatoria **continua** con distribución de probabilidad $f(x)$. Definamos con $Y = u(X)$ una transformación entre los valores de X y Y que no es uno a uno. Si el intervalo sobre el que se define X se puede dividir en k conjuntos mutuamente disjuntos de manera que cada una de las funciones inversas

$$x_1 = w_1(y), \quad x_2 = w_2(y), \quad \ldots, \quad x_k = w_k(y)$$

de $y = u(x)$ defina una correspondencia uno a uno, entonces la distribución de probabilidad de Y es

$$g(y) = \sum_{i=1}^{k} f[w_i(y)]|J_i|,$$

donde $J_i = w_i'(y),\ i = 1, 2, \ldots, k$.

Ejemplo 7.5: Demuestre que $Y = (X - \mu)^2/\sigma^2$ tiene una distribución chi cuadrada con 1 grado de libertad cuando X tiene una distribución normal con media μ y varianza σ^2.

***Solución*:** Sea $Z = (X - \mu)/\sigma$, donde la variable aleatoria Z tiene la distribución normal estándar

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \qquad -\infty < z < \infty.$$

Ahora debemos calcular la distribución de la variable aleatoria $Y = Z^2$. Las soluciones inversas de $y = z^2$ son $z = \pm\sqrt{y}$. Si designamos $z_1 = -\sqrt{y}$ y $z_2 = \sqrt{y}$, entonces $J_1 = -1/2\sqrt{y}$ y $J_2 = -1/2\sqrt{y}$. Entonces, por el teorema 7.5, tenemos

$$g(y) = \frac{1}{\sqrt{2\pi}} e^{-y/2} \left| \frac{-1}{2\sqrt{y}} \right| + \frac{1}{\sqrt{2\pi}} e^{-y/2} \left| \frac{1}{2\sqrt{y}} \right| = \frac{1}{\sqrt{2\pi}} y^{1/2-1} e^{-y/2}, \qquad y > 0.$$

Como $g(y)$ es una función de densidad, se deduce que

$$1 = \frac{1}{\sqrt{2\pi}} \int_0^\infty y^{1/2-1} e^{-y/2}\, dy = \frac{\Gamma(1/2)}{\sqrt{\pi}} \int_0^\infty \frac{y^{1/2-1} e^{-y/2}}{\sqrt{2}\Gamma(1/2)}\, dy = \frac{\Gamma(1/2)}{\sqrt{\pi}},$$

la integral es el área bajo una curva de probabilidad gamma con los parámetros $\alpha = 1/2$ y $\beta = 2$. Por lo tanto, $\sqrt{\pi} = \Gamma(1/2)$ y la densidad de Y es dada por

$$g(y) = \begin{cases} \frac{1}{\sqrt{2}\Gamma(1/2)} y^{1/2-1} e^{-y/2}, & y > 0, \\ 0, & \text{en cualquier caso.} \end{cases}$$

que se considera una distribución chi cuadrada con 1 grado de libertad.

7.3 Momentos y funciones generadoras de momentos

En esta sección nos concentramos en aplicaciones de las funciones generadoras de momentos. El propósito evidente de la función generadora de momentos es la determinación de los momentos de variables aleatorias. Sin embargo, la contribución más importante consiste en establecer distribuciones de funciones de variables aleatorias.

Si $g(X) = X^r$ para $r = 0, 1, 2, 3,...$, la definición 7.1 proporciona un valor esperado que se denomina r-ésimo **momento alrededor del origen** de la variable aleatoria X, que denotamos con μ'_r.

Definición 7.1: El r-ésimo **momento alrededor del origen** de la variable aleatoria X es dado por

$$\mu'_r = E(X^r) = \begin{cases} \sum_x x^r f(x), & \text{si } X \text{ es discreta,} \\ \int_{-\infty}^{\infty} x^r f(x)\, dx, & \text{si } X \text{ es continua.} \end{cases}$$

Como el primer y segundo momentos alrededor del origen son dados por $\mu'_1 = E(X)$ y $\mu'_2 = E(X^2)$, podemos escribir la media y la varianza de una variable aleatoria como

$$\mu = \mu'_1 \qquad \text{y} \qquad \sigma^2 = \mu'_2 - \mu^2.$$

Aunque los momentos de una variable aleatoria se pueden determinar directamente a partir de la definición 7.1, existe un procedimiento alternativo, el cual requiere que utilicemos una **función generadora de momentos**.

Definición 7.2: La **función generadora de momentos** de la variable aleatoria X es dada por $E(e^{tX})$, y se denota con $M_X(t)$. Por lo tanto,

$$M_X(t) = E(e^{tX}) = \begin{cases} \sum_x e^{tx} f(x), & \text{si } X \text{ es discreta,} \\ \int_{-\infty}^{\infty} e^{tx} f(x)\, dx, & \text{si } X \text{ es continua.} \end{cases}$$

Las funciones generadoras de momentos existirán sólo si la sumatoria o integral de la definición 7.2 converge. Si existe una función generadora de momentos de una variable aleatoria X, se puede utilizar para generar todos los momentos de dicha variable. El método se describe en el teorema 7.6 sin demostración.

Teorema 7.6: Sea X una variable aleatoria con función generadora de momentos $M_X(t)$. Entonces,

$$\left.\frac{d^r M_X(t)}{dt^r}\right|_{t=0} = \mu'_r.$$

Ejemplo 7.6: Calcule la función generadora de momentos de la variable aleatoria binomial X y después utilícela para verificar que $\mu = np$ y $\sigma^2 = npq$.

***Solución*:** A partir de la definición 7.2 tenemos

$$M_X(t) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x q^{n-x} = \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x q^{n-x}.$$

Al reconocer a esta última sumatoria como la expansión binomial de $(pe^t + q)^n$ obtenemos

$$M_X(t) = (pe^t + q)^n.$$

Así,

$$\frac{dM_X(t)}{dt} = n(pe^t + q)^{n-1}pe^t$$

y

$$\frac{d^2M_X(t)}{dt^2} = np[e^t(n-1)(pe^t + q)^{n-2}pe^t + (pe^t + q)^{n-1}e^t].$$

Al establecer $t = 0$ obtenemos

$$\mu_1' = np \text{ y } \mu_2' = np[(n-1)p + 1].$$

Por consiguiente,

$$\mu = \mu_1' = np \text{ y } \sigma^2 = \mu_2' - \mu^2 = np(1-p) = npq,$$

que coincide con los resultados que se obtuvieron en el capítulo 5. ▄

Ejemplo 7.7: Demuestre que la función generadora de momentos de la variable aleatoria X, la cual tiene una distribución de probabilidad normal con media μ y varianza σ^2, es dada por

$$M_X(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right).$$

***Solución*:** A partir de la definición 7.2, la función generadora de momentos de la variable aleatoria normal X es

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{x^2 - 2(\mu + t\sigma^2)x + \mu^2}{2\sigma^2}\right] dx. \end{aligned}$$

Si completamos el cuadrado en el exponente, podemos escribir

$$x^2 - 2(\mu + t\sigma^2)x + \mu^2 = [x - (\mu + t\sigma^2)]^2 - 2\mu t\sigma^2 - t^2\sigma^4$$

y, entonces,

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{[x - (\mu + t\sigma^2)]^2 - 2\mu t\sigma^2 - t^2\sigma^4}{2\sigma^2}\right\} dx \\ &= \exp\left(\frac{2\mu t + \sigma^2 t^2}{2}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{[x - (\mu + t\sigma^2)]^2}{2\sigma^2}\right\} dx. \end{aligned}$$

Sea $w = [x - (\mu + t\sigma^2)]/\sigma$; entonces $dx = \sigma\, dw$ y

$$M_X(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-w^2/2}\, dw = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right),$$

ya que la última integral representa el área bajo una curva de densidad normal estándar y, en consecuencia, es igual a 1. ∎

Aunque el método de transformación de variables brinda una forma eficaz para determinar la distribución de una función de múltiples variables, existe un procedimiento alternativo, y que a menudo se prefiere cuando la función a analizar es una combinación lineal de variables aleatorias independientes. Este procedimiento utiliza las propiedades de las funciones generadoras de momentos que se estudian en los siguientes cuatro teoremas. Para no rebasar el alcance matemático de este libro, establecemos el teorema 7.7 sin demostración.

Teorema 7.7: (**Teorema de unicidad**) Sean X y Y dos variables aleatorias con funciones generadoras de momentos $M_X(t)$ y $M_Y(t)$, respectivamente. Si $M_X(t) = M_Y(t)$ para todos los valores de t, entonces X y Y tienen la misma distribución de probabilidad.

Teorema 7.8: $M_{X+a}(t) = e^{at} M_X(t)$.

Prueba: $M_{X+a}(t) = E[e^{t(X+a)}] = e^{at} E(e^{tX}) = e^{at} M_X(t)$. ∎

Teorema 7.9: $M_{aX}(t) = M_X(at)$.

Prueba: $M_{aX}(t) = E[e^{t(aX)}] = E[e^{(at)X}] = M_X(at)$. ∎

Teorema 7.10: Si $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes con funciones generadoras de momentos $M_{X_1}(t), M_{X_2}(t), \ldots, M_{X_n}(t)$, respectivamente, y $Y = X_1 + X_2 + \ldots + X_n$, entonces,

$$M_Y(t) = M_{X_1}(t) M_{X_2}(t) \cdots M_{X_n}(t).$$

La demostración del teorema 7.10 se deja al lector.

Los teoremas 7.7 a 7.10 son fundamentales para entender las funciones generadoras de momentos. A continuación se presenta un ejemplo como ilustración. Hay muchas situaciones en que necesitamos conocer la distribución de la suma de las variables aleatorias. Podemos utilizar los teoremas 7.7 y 7.10, así como el resultado del ejercicio 7.19 de la página 224, para calcular la distribución de una suma de dos variables aleatorias independientes de Poisson, con funciones generadoras de momentos dadas por

$$M_{X_1}(t) = e^{\mu_1(e^t-1)} \text{ y } M_{X_2}(t) = e^{\mu_2(e^t-1)},$$

respectivamente. De acuerdo con el teorema 7.10, la función generadora de momentos de la variable aleatoria $Y_1 = X_1 + X_2$ es

$$M_{Y_1}(t) = M_{X_1}(t) M_{X_2}(t) = e^{\mu_1(e^t-1)} e^{\mu_2(e^t-1)} = e^{(\mu_1+\mu_2)(e^t-1)},$$

que de inmediato identificamos como la función generadora de momentos de una variable aleatoria que tiene una distribución de Poisson con el parámetro $\mu_1 + \mu_2$. Por lo tanto, de acuerdo con el teorema 7.7, de nuevo concluimos que la suma de dos variables aleatorias independientes, que tienen distribuciones de Poisson con los parámetros μ_1 y μ_2, tiene una distribución de Poisson con el parámetro $\mu_1 + \mu_2$.

Combinaciones lineales de variables aleatorias

En estadística aplicada a menudo se necesita conocer la distribución de probabilidad de una combinación lineal de variables aleatorias normales independientes. Obtengamos la distribución de la variable aleatoria $Y = a_1X_1 + a_2 X_2$ cuando X_1 es una variable normal con media μ_1 y varianza σ_1^2 y X_2 también es una variable normal, pero independiente de X_1, con media μ_2 y varianza σ_2^2. Primero, por medio del teorema 7.10, obtenemos

$$M_Y(t) = M_{a_1X_1}(t)M_{a_2X_2}(t),$$

y después, usando el teorema 7.9, obtenemos

$$M_Y(t) = M_{X_1}(a_1t)M_{X_2}(a_2t).$$

Si sustituimos a_1t por t, y despues a_2t por t, en una función generadora de momentos de la distribución normal derivada en el ejemplo 7.7, tenemos

$$\begin{aligned} M_Y(t) &= \exp(a_1\mu_1 t + a_1^2\sigma_1^2 t^2/2 + a_2\mu_2 t + a_2^2\sigma_2^2 t^2/2) \\ &= \exp[(a_1\mu_1 + a_2\mu_2)t + (a_1^2\sigma_1^2 + a_2^2\sigma_2^2)t^2/2], \end{aligned}$$

que reconocemos como la función generadora de momentos de una distribución que es normal, con media $a_1\mu_1 + a_2\mu_2$ y varianza $a_1^2\sigma_1^2 + a_2^2\sigma_2^2$.

Al generalizar para el caso de n variables normales independientes, establecemos el siguiente resultado.

Teorema 7.11: Si $X_1, X_2,..., X_n$ son variables aleatorias independientes que tienen distribuciones normales con medias $\mu_1, \mu_2,... \mu_n$ y varianzas $\sigma_1^2, \sigma_2^2,..., \sigma_n^2$, respectivamente, entonces la variable aleatoria

$$Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n$$

tiene una distribución normal con media

$$\mu_Y = a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n$$

y varianza

$$\sigma_Y^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2.$$

Ahora es evidente que la distribución de Poisson y la distribución normal tienen una propiedad reproductiva, en el sentido de que la suma de variables aleatorias independientes que tengan cualquiera de estas distribuciones es una variable aleatoria que también tiene el mismo tipo de distribución. La distribución chi cuadrada también posee esta propiedad reproductiva.

Teorema 7.12: Si $X_1, X_2,..., X_n$ son variables aleatorias mutuamente independientes, que tienen distribuciones chi cuadrada con $v_1, v_2,..., v_n$ grados de libertad, respectivamente, entonces la variable aleatoria

$$Y = X_1 + X_2 + \cdots + X_n$$

tiene una distribución chi cuadrada con $v = v_1 + v_2 +...+ v_n$ grados de libertad.

***Prueba*:** Por medio del teorema 7.10 y el ejercicio 7.21,

$$M_Y(t) = M_{X_1}(t)M_{X_2}(t)\cdots M_{X_n}(t) \text{ y } M_{X_i}(t) = (1-2t)^{-v_i/2}, \; i = 1, 2, \ldots, n.$$

Por lo tanto,

$$M_Y(t) = (1-2t)^{-v_1/2}(1-2t)^{-v_2/2} \dots (1-2t)^{-v_n/2} = (1-2t)^{-(v_1+v_2+\cdots+v_n)/2},$$

que reconocemos como la función generadora de momentos de una distribución chi cuadrada con $v = v_1 + v_2 + \ldots + v_n$ grados de libertad.

Corolario 7.1: Si $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes que tienen distribuciones normales idénticas, con media μ y varianza σ^2, entonces la variable aleatoria

$$Y = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2$$

tiene una distribución chi cuadrada con $v = n$ grados de libertad.

Este corolario es una consecuencia inmediata del ejemplo 7.5, y establece una relación entre la muy importante distribución chi cuadrada y la distribución normal. También debe brindar al lector una idea muy clara de lo que significa el parámetro llamado grados de libertad. En futuros capítulos el concepto de grados de libertad desempeñará un papel cada vez más relevante.

Corolario 7.2: Si $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes y X_i tiene una distribución normal con media μ_i y varianza σ_i^2 para $i = 1, 2, \ldots, n$, entonces la variable aleatoria

$$Y = \sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2$$

tiene una distribución chi cuadrada con $v = n$ grados de libertad.

Ejercicios

7.1 Sea X una variable aleatoria que tiene la siguiente probabilidad

$$f(x) = \begin{cases} \frac{1}{3}, & x = 1, 2, 3, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Calcule la distribución de probabilidad de la variable aleatoria $Y = 2X - 1$.

7.2 Sea X una variable aleatoria binomial con la siguiente distribución de probabilidad

$$f(x) = \begin{cases} \binom{3}{x} \left(\frac{2}{5}\right)^x \left(\frac{3}{5}\right)^{3-x}, & x = 0, 1, 2, 3, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Calcule la distribución de probabilidad de la variable aleatoria $Y = X^2$.

7.3 Sean X_1 y X_2 variables aleatorias discretas con la siguiente distribución multinomial conjunta

$$f(x_1, x_2) = \binom{2}{x_1, x_2, 2-x_1-x_2} \left(\frac{1}{4}\right)^{x_1} \left(\frac{1}{3}\right)^{x_2} \left(\frac{5}{12}\right)^{2-x_1-x_2}$$

para $x_1 = 0, 1, 2$; $x_2 = 0, 1, 2$; $x_1 + x_2 \le 2$; y cero en cualquier caso. Calcule la distribución de probabilidad conjunta de $Y_1 = X_1 + X_2$ y $Y_2 = X_1 - X_2$.

7.4 Sean X_1 y X_2 variables aleatorias discretas con la siguiente distribución de probabilidad conjunta

$$f(x_1, x_2) = \begin{cases} \frac{x_1 x_2}{18}, & x_1 = 1, 2;\ x_2 = 1, 2, 3, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Calcule la distribución de probabilidad de la variable aleatoria $Y = X_1 X_2$.

7.5 Si X tiene la siguiente distribución de probabilidad

$$f(x) = \begin{cases} 1, & 0 < x < 1, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Demuestre que la variable aleatoria $Y = -2\ln X$ tiene una distribución chi cuadrada con 2 grados de libertad.

7.6 Dada la variable aleatoria X con la siguiente distribución de probabilidad

$$f(x) = \begin{cases} 2x, & 0 < x < 1, \\ 0, & \text{en cualquier caso,} \end{cases}$$

calcule la distribución de probabilidad de $Y = 8X^3$.

7.7 La velocidad de una molécula en un gas uniforme en equilibrio es una variable aleatoria V, cuya distribución de probabilidad es dada por

$$f(v) = \begin{cases} kv^2 e^{-bv^2}, & v > 0, \\ 0, & \text{en cualquier caso,} \end{cases}$$

donde k es una constante adecuada y b depende de la temperatura absoluta y de la masa de la molécula. Calcule la distribución de probabilidad de la energía cinética de la molécula W, donde $W = mV^2/2$.

7.8 La utilidad de un distribuidor, en unidades de \$5000, sobre un automóvil nuevo, es dada por $Y = X^2$, donde X es una variable aleatoria que tiene la siguiente función de densidad

$$f(x) = \begin{cases} 2(1-x), & 0 < x < 1, \\ 0, & \text{en cualquier caso.} \end{cases}$$

a) Calcule la función de densidad de probabilidad de la variable aleatoria Y.
b) Utilice la función de densidad de Y para calcular la probabilidad de que la utilidad sobre el siguiente automóvil nuevo que venda este distribuidor sea menor que \$500.

7.9 El periodo hospitalario, en días, para pacientes que siguen un tratamiento para cierto tipo de enfermedad del riñón es una variable aleatoria $Y = X + 4$, donde X tiene la siguiente función de densidad

$$f(x) = \begin{cases} \frac{32}{(x+4)^3}, & x > 0, \\ 0, & \text{en cualquier caso.} \end{cases}$$

a) Calcule la función de densidad de probabilidad de la variable aleatoria Y.
b) Utilice la función de densidad de Y para calcular la probabilidad de que el periodo hospitalario para un paciente que sigue este tratamiento exceda los 8 días.

7.10 Las variables aleatorias X y Y, que representan los pesos de cremas y chiclosos, respectivamente, en cajas de un kilogramo de chocolates que contienen una combinación de cremas, chiclosos y envinados, tienen la siguiente función de densidad conjunta

$$f(x, y) = \begin{cases} 24xy, & 0 \le x \le 1,\ 0 \le y \le 1, x + y \le 1, \\ 0, & \text{en cualquier caso.} \end{cases}$$

a) Calcule la función de densidad de probabilidad de la variable aleatoria $Z = X + Y$.
b) Utilice la función de densidad de Z para calcular la probabilidad de que, en una determinada caja, la suma de los pesos de las cremas y los chiclosos sea por lo menos 1/2 del peso total, pero menos de 3/4.

7.11 La cantidad de queroseno en un tanque al inicio de cualquier día, en miles de litros, es una cantidad aleatoria Y, de la cual una cantidad aleatoria X se vende durante ese día. Suponga que la función de densidad conjunta de estas variables es dada por

$$f(x, y) = \begin{cases} 2, & 0 < x < y,\ 0 < y < 1, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Calcule la función de densidad de probabilidad para la cantidad de queroseno que queda en el tanque al final del día.

7.12 Sean X_1 y X_2 variables aleatorias independientes que tienen cada una la siguiente distribución de probabilidad

$$f(x) = \begin{cases} e^{-x}, & x > 0, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Demuestre que las variables aleatorias Y_1 y Y_2 son independientes cuando $Y_1 = X_1 + X_2$ y $Y_2 = X_1/(X_1 + X_2)$.

7.13 Una corriente de I amperios que fluye a través de una resistencia de R ohms varía de acuerdo con la siguiente distribución de probabilidad

$$f(i) = \begin{cases} 6i(1-i), & 0 < i < 1, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Si la resistencia varía independientemente de la corriente de acuerdo con la siguiente distribución de probabilidad

$$g(r) = \begin{cases} 2r, & 0 < r < 1, \\ 0, & \text{en cualquier caso,} \end{cases}$$

calcule la distribución de probabilidad para la potencia $W = I^2R$ watts.

7.14 Sea X una variable aleatoria con la siguiente distribución de probabilidad

$$f(x) = \begin{cases} \frac{1+x}{2}, & -1 < x < 1, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Calcule la distribución de probabilidad de la variable aleatoria $Y = X^2$.

7.15 Si X tiene la siguiente distribución de probabilidad

$$f(x) = \begin{cases} \frac{2(x+1)}{9}, & -1 < x < 2, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Calcule la distribución de probabilidad de la variable aleatoria $Y = X^2$.

7.16 Demuestre que el r-ésimo momento respecto al origen de la distribución gamma es

$$\mu_r' = \frac{\beta^r \Gamma(\alpha + r)}{\Gamma(\alpha)}.$$

[*Sugerencia*: Sustituya $y = x/\beta$ en la integral que define μ_r' y después utilice la función gamma para evaluar la integral].

7.17 Una variable aleatoria X tiene la siguiente distribución uniforme discreta

$$f(x;k) = \begin{cases} \frac{1}{k}, & x = 1, 2, \ldots, k, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Demuestre que la función generadora de momentos de X es

$$M_X(t) = \frac{e^t(1 - e^{kt})}{k(1 - e^t)}.$$

7.18 Una variable aleatoria X tiene la distribución geométrica $g(x; p) = pq^{x-1}$ para $x = 1, 2, 3, \ldots$. Demuestre que la función generadora de momentos de X es

$$M_X(t) = \frac{pe^t}{1 - qe^t}, \qquad t < \ln q,$$

y después use $M_X(t)$ para calcular la media y la varianza de la distribución geométrica.

7.19 Una variable aleatoria X tiene la distribución de Poisson $p(x; \mu) = e^{-\mu}\mu^x/x!$ para $x = 0, 1, 2, \ldots$. Demuestre que la función generadora de momentos de X es

$$M_X(t) = e^{\mu(e^t - 1)}.$$

Utilice $M_X(t)$ para calcular la media y la varianza de la distribución de Poisson.

7.20 La función generadora de momentos de cierta variable aleatoria de Poisson X es dada por

$$M_X(t) = e^{4(e^t - 1)}.$$

Calcule $P(\mu - 2\sigma < X < \mu + 2\sigma)$.

7.21 Demuestre que la función generadora de momentos de la variable aleatoria X, que tiene una distribución chi cuadrada con v grados de libertad, es

$$M_X(t) = (1 - 2t)^{-v/2}.$$

7.22 Con la función generadora de momentos del ejemplo 7.21 demuestre que la media y la varianza de la distribución chi cuadrada con v grados de libertad son, respectivamente, v y $2v$.

7.23 Si tanto X como Y, distribuidas de manera independiente, siguen distribuciones exponenciales con parámetro medio 1, calcule las distribuciones de

a) $U = X + Y$;
b) $V = X/(X + Y)$.

7.24 Mediante la expansión de e^{tx} en una serie de Maclaurin y la integración término por término, demuestre que

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} e^{tx} f(x)\, dx \\ &= 1 + \mu t + \mu_2' \frac{t^2}{2!} + \ldots + \mu_r' \frac{t^r}{r!} + \ldots \end{aligned}$$

Capítulo 8

Distribuciones de muestreo fundamentales y descripciones de datos

8.1 Muestreo aleatorio

El resultado de un experimento estadístico se puede registrar como un valor numérico o como una representación descriptiva. Cuando se lanza un par de dados y lo que nos interesa es el resultado total, registramos un valor numérico. Sin embargo, si a los estudiantes de cierta escuela se les hacen pruebas de sangre para averiguar cuál es su tipo, podría ser más útil una representación descriptiva. La sangre de una persona se puede clasificar de 8 maneras. Puede ser AB, A, B u O, cada una con un signo de más o de menos, lo cual depende de la presencia o ausencia del antígeno Rh.

En este capítulo nos enfocamos en el muestreo de distribuciones o poblaciones, y estudiamos cantidades tan importantes como la *media de la muestra* y la *varianza de la muestra*, que serán de importancia fundamental en los capítulos siguientes. Además, en los próximos capítulos intentamos introducir al lector al papel que desempeñarán la media y la varianza de la muestra en la inferencia estadística. El uso de las computadoras modernas de alta velocidad permite a los científicos e ingenieros incrementar enormemente su uso de la inferencia estadística formal con técnicas gráficas. La mayoría de las veces la inferencia formal parece muy árida y quizás incluso abstracta para el profesional o el gerente que desea que el análisis estadístico sea una guía para la toma de decisiones.

Poblaciones y muestras

Comenzamos esta sección presentando los conceptos de *poblaciones* y *muestras*. Ambas se mencionan de forma extensa en el capítulo 1; sin embargo, aquí será necesario estudiarlas más ampliamente, en particular en el contexto del concepto de variables aleatorias. La totalidad de observaciones que nos interesan, ya sean de número finito o infinito, constituye lo que llamamos **población**. En alguna época el término *población* se refería a observaciones que se obtenían de estudios estadísticos aplicados a personas. En la actualidad el estadístico utiliza la palabra para referirse a observaciones sobre cualquier cuestión de interés, ya sea de grupos de personas, de animales o de todos los resultados posibles de algún complicado sistema biológico o de ingeniería.

Definición 8.1: Una **población** consta de la totalidad de las observaciones en las que estamos interesados.

El número de observaciones en la población se define como el tamaño de la población. Si en la escuela hay 600 estudiantes que clasificamos de acuerdo con su tipo de sangre, decimos que tenemos una población de tamaño 600. Los números en las cartas de una baraja, las estaturas de los residentes de cierta ciudad y las longitudes de los peces en un lago específico son ejemplos de poblaciones de tamaño finito. En cada caso el número total de observaciones es un número finito. Las observaciones que se obtienen al medir diariamente la presión atmosférica desde el pasado hasta el futuro, o todas las mediciones de la profundidad de un lago desde cualquier posición concebible son ejemplos de poblaciones cuyos tamaños son infinitos. Algunas poblaciones finitas son tan grandes que en teoría las supondríamos infinitas, lo cual es cierto si se considera la población de la vida útil de cierto tipo de batería de almacenamiento que se está fabricando para distribuirla en forma masiva en todo el país.

Cada observación en una población es un valor de una variable aleatoria X que tiene alguna distribución de probabilidad $f(x)$. Si se inspeccionan artículos que salen de una línea de ensamble para buscar defectos, entonces cada observación en la población podría ser un valor 0 o 1 de la variable aleatoria X de Bernoulli, con una distribución de probabilidad

$$b(x; 1, p) = p^x q^{1-x}, \qquad x = 0, 1$$

donde 0 indica un artículo sin defecto y 1 indica un artículo defectuoso. De hecho, se supone que p, la probabilidad de que cualquier artículo esté defectuoso, permanece constante de una prueba a otra. En el experimento del tipo de sangre la variable aleatoria X representa el tipo de sangre y se supone que toma un valor del 1 al 8. A cada estudiante se le asigna uno de los valores de la variable aleatoria discreta. Las duraciones de las baterías de almacenamiento son valores que toma una variable aleatoria continua que quizá tiene una distribución normal. De ahora en adelante, cuando nos refiramos a una "población binomial", a una "población normal" o, en general, a la "población $f(x)$", aludiremos a una población cuyas observaciones son valores de una variable aleatoria que tiene una distribución binomial, una distribución normal o la distribución de probabilidad $f(x)$. Por ello, a la media y a la varianza de una variable aleatoria o distribución de probabilidad también se les denomina la media y la varianza de la población correspondiente.

En el campo de la inferencia estadística, el estadístico se interesa en llegar a conclusiones respecto a una población, cuando es imposible o poco práctico conocer todo el conjunto de observaciones que la constituyen. Por ejemplo, al intentar determinar la longitud de la vida promedio de cierta marca de bombilla, sería imposible probarlas todas si tenemos que dejar algunas para venderlas. Los costos desmesurados que implicaría estudiar a toda la población también constituirían un factor que impediría hacerlo. Por lo tanto, debemos depender de un subconjunto de observaciones de la población que nos ayude a realizar inferencias respecto a ella. Esto nos lleva a considerar el concepto de muestreo.

Definición 8.2: Una **muestra** es un subconjunto de una población.

Para que las inferencias que hacemos sobre la población a partir de la muestra sean válidas, debemos obtener muestras que sean representativas de ella. Con mucha frecuencia nos sentimos tentados a elegir una muestra seleccionando a los miembros más convenientes de la población. Tal procedimiento podría conducir a inferencias erróneas respecto a la población. Se dice que cualquier procedimiento de muestreo que produzca inferencias que sobreestimen o subestimen de forma consistente alguna característica de la población está **sesgado**. Para eliminar cualquier posibilidad de sesgo en el procedimiento de muestreo es deseable elegir una **muestra aleatoria**, lo cual significa que las observaciones se realicen de forma independiente y al azar.

Para seleccionar una muestra aleatoria de tamaño n de una población $f(x)$ definimos la variable aleatoria X_i, $i = 1, 2,..., n$, que representa la i-ésima medición o valor de la muestra que observamos. Si las mediciones se obtienen repitiendo el experimento n veces independientes en, esencialmente, las mismas condiciones, las variables aleatorias $X_1, X_2,..., X_n$ constituirán entonces una muestra aleatoria de la población $f(x)$ con valores numéricos $x_1, x_2,..., x_n$. Debido a las condiciones idénticas en las que se seleccionan los elementos de la muestra, es razonable suponer que las n variables aleatorias $X_1, X_2,..., X_n$ son independientes y que cada una tiene la misma distribución de probabilidad $f(x)$. Es decir, las distribuciones de probabilidad de $X_1, X_2,..., X_n$son, respectivamente, $f(x_1), f(x_2),..., f(x_n)$, y su distribución de probabilidad conjunta es $f(x_1, x_2, \ldots, x_n) = f(x_1) f(x_2) \cdots f(x_n)$. El concepto de muestra aleatoria se describe de manera formal en la siguiente definición.

Definición 8.3: Sean $X_1, X_2,..., X_n$ variables aleatorias independientes n, cada una con la misma distribución de probabilidad $f(x)$. Definimos $X_1, X_2,..., X_n$ como una **muestra aleatoria** de tamaño n de la población $f(x)$ y escribimos su distribución de probabilidad conjunta como

$$f(x_1, x_2, \ldots, x_n) = f(x_1) f(x_2) \cdots f(x_n).$$

Si se realiza una selección aleatoria de $n = 8$ baterías de almacenamiento de un proceso de fabricación que mantiene las mismas especificaciones, y al registrar la duración de cada batería se encuentra que la primera medición x_1 es un valor de X_1, la segunda medición x_2 es un valor de X_2, y así sucesivamente, entonces $x_1, x_2,..., x_8$ son los valores de la muestra aleatoria $X_1, X_2,..., X_8$. Si suponemos que la población de vidas útiles de las baterías es normal, los valores posibles de cualquier X_i, $i = 1, 2,..., 8$ serán exactamente los mismos que los de la población original, por consiguiente, X_i tiene una distribución normal idéntica a la de X.

8.2 Algunos estadísticos importantes

Nuestro principal propósito al seleccionar muestras aleatorias consiste en obtener información acerca de los parámetros desconocidos de la población. Suponga, por ejemplo, que deseamos concluir algo respecto a la proporción de consumidores de café en Estados Unidos que prefieren cierta marca de café. Sería imposible interrogar a cada consumidor estadounidense de café para calcular el valor del parámetro p que representa la proporción de la población. En vez de esto se selecciona una muestra aleatoria grande y se calcula la proporción $\hat{p}$ de personas en esta muestra que prefieren la marca de café en cuestión. El valor $\hat{p}$ se utiliza ahora para hacer una inferencia respecto a la proporción p verdadera.

Ahora, $\hat{p}$ es una función de los valores observados en la muestra aleatoria; ya que es posible tomar muchas muestras aleatorias de la misma población, esperaríamos que $\hat{p}$ variara un poco de una a otra muestra. Es decir, $\hat{p}$ es un valor de una variable aleatoria que representamos con P. Tal variable aleatoria se llama **estadístico**.

Definición 8.4: Cualquier función de las variables aleatorias que forman una muestra aleatoria se llama **estadístico**.

Medidas de localización de una muestra: la media, la mediana y la moda muestrales

En el capítulo 4 presentamos los parámetros μ y σ^2, que miden el centro y la variabilidad de una distribución de probabilidad. Éstos son parámetros de población constantes y de ninguna manera se ven afectados o influidos por las observaciones de una muestra aleatoria. Definiremos, sin embargo, algunos estadísticos importantes que describen las medidas correspondientes de una muestra aleatoria. Los estadísticos que más se utilizan para medir el centro de un conjunto de datos, acomodados en orden de magnitud, son la **media**, la **mediana** y la **moda**. Aunque los primeros dos estadísticos se expusieron en el capítulo 1, repetiremos las definiciones. Sean $X_1, X_2,..., X_n$ representaciones de n variables aleatorias.

a) Media muestral:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Observe que el estadístico $\bar{X}$ toma el valor $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ cuando X_1 toma el valor x_1, X_2 toma el valor x_2 y así sucesivamente. El término *media muestral* se aplica tanto al estadístico $\bar{X}$ como a su valor calculado $\bar{x}$.

b) Mediana muestral:

$$\tilde{x} = \begin{cases} x_{(n+1)/2}, & \text{si } n \text{ es impar,} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}), & \text{si } n \text{ es par.} \end{cases}$$

La mediana muestral también es una medida de localización que indica el valor central de la muestra. En la sección 1.3 se presentan ejemplos de la media muestral y de la mediana muestral. La moda muestral se define de la siguiente manera:

c) La moda muestral es el valor que ocurre con mayor frecuencia en la muestra.

Ejemplo 8.1: Suponga que un conjunto de datos consta de las siguientes observaciones:

0.32 0.53 0.28 0.37 0.47 0.43 0.36 0.42 0.38 0.43

La moda de la muestra es 0.43, ya que este valor aparece con más frecuencia que los demás. ■

Como se expuso en el capítulo 1, una medida de localización o tendencia central en una muestra no da por sí misma una indicación clara de la naturaleza de ésta, de manera que también debe considerarse una medida de variabilidad en la muestra.

Las medidas de variabilidad de una muestra: la varianza, la desviación estándar y el rango de la muestra

La variabilidad en la muestra refleja cómo se dispersan las observaciones a partir del promedio. Se remite al lector al capítulo 1 para un análisis más amplio. Es posible tener dos conjuntos de observaciones con las mismas media o mediana que difieran de manera considerable en la variabilidad de sus mediciones sobre el promedio.

Considere las siguientes mediciones, en litros, para dos muestras de jugo de naranja envasado por las empresas A y B:

Muestra A	0.97	1.00	0.94	1.03	1.06
Muestra B	1.06	1.01	0.88	0.91	1.14

Ambas muestras tienen la misma media, 1.00 litros. Es muy evidente que la empresa A envasa el jugo de naranja con un contenido más uniforme que la B. Decimos que la **variabilidad** o la **dispersión** de las observaciones a partir del promedio es menor para la muestra A que para la muestra B. Por lo tanto, al comprar jugo de naranja, tendríamos más confianza en que el envase que seleccionemos se acerque al promedio anunciado si se lo compramos a la empresa A.

En el capítulo 1 presentamos varias medidas de la variabilidad de una muestra, como la **varianza muestral**, la **desviación estándar muestral** y el **rango de la muestra**. En este capítulo nos enfocaremos sobre todo en la varianza de la muestra. Nuevamente, sea que $X_1, X_2,..., X_n$ representan n variables aleatorias.

a) La varianza muestral:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2. \tag{8.2.1}$$

El valor calculado de S^2 para una muestra dada se denota con s^2. Observe que S^2 se define esencialmente como el promedio de los cuadrados de las desviaciones de las observaciones a partir de su media. La razón para utilizar $n - 1$ como divisor, en vez de la elección más obvia n, quedará más clara en el capítulo 9.

Ejemplo 8.2: Una comparación de los precios de café en 4 tiendas de abarrotes de San Diego, seleccionadas al azar, mostró aumentos en comparación con el mes anterior de 12, 15, 17 y 20 centavos por bolsa de una libra. Calcule la varianza de esta muestra aleatoria de aumentos de precio.

***Solución*:** Si calculamos la media de la muestra, obtenemos

$$\bar{x} = \frac{12+15+17+20}{4} = 16 \text{ centavos.}$$

Por lo tanto,

$$s^2 = \frac{1}{3}\sum_{i=1}^{4}(x_i - 16)^2 = \frac{(12-16)^2 + (15-16)^2 + (17-16)^2 + (20-16)^2}{3}$$

$$= \frac{(-4)^2 + (-1)^2 + (1)^2 + (4)^2}{3} = \frac{34}{3}.$$

Mientras que la expresión para la varianza de la muestra de la definición 8.6 ilustra mejor que S^2 es una medida de variabilidad, una expresión alternativa tiene cierto mérito, de manera que el lector debería conocerla. El siguiente teorema contiene tal expresión.

Teorema 8.1: Si S^2 es la varianza de una muestra aleatoria de tamaño n, podemos escribir

$$S^2 = \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right].$$

***Prueba*:** Por definición,

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i^2 - 2\bar{X}X_i + \bar{X}^2)$$
$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum_{i=1}^{n} X_i + n\bar{X}^2\right].$$

Como en el capítulo 1, a continuación se definen la **desviación estándar muestral** y el **rango muestral**:

b) Desviación estándar muestral:

$$S = \sqrt{S^2},$$

donde S^2 es la varianza muestral.

Permitamos que $X_{\text{máx}}$ denote el más grande de los valores X_i y $X_{\text{mín}}$ el más pequeño.

c) Rango muestral:

$$R = X_{\text{máx}} - X_{\text{mín}}.$$

Ejemplo 8.3: Calcule la varianza de los datos 3, 4, 5, 6, 6 y 7, que representan el número de truchas atrapadas por una muestra aleatoria de 6 pescadores, el 19 de junio de 1996, en el lago Muskoka.

***Solución*:** Encontramos que $\sum_{i=1}^{6} x_i^2 = 171$, $\sum_{i=1}^{6} x_i = 31$ y $n = 6$. De aquí,

$$s^2 = \frac{1}{(6)(5)}[(6)(171) - (31)^2] = \frac{13}{6}.$$

Por consiguiente, la desviación estándar de la muestra $s = \sqrt{13/6} = 1.47$ y el rango muestral es $7 - 3 = 4$.

Ejercicios

8.1 Defina las poblaciones adecuadas a partir de las cuales se seleccionaron las siguientes muestras:

a) Se llamó por teléfono a personas de 200 casas en la ciudad de Richmond y se les pidió nombrar al candidato por el que votarían en la elección del presidente de la mesa directiva de la escuela.

b) Se lanzó 100 veces una moneda y se registraron 34 cruces.

c) Se probaron 200 pares de un nuevo tipo de calzado deportivo en un torneo de tenis profesional para determinar su duración y se encontró que, en promedio, duraron 4 meses.

d) En cinco ocasiones diferentes a una abogada le tomó 21, 26, 24, 22 y 21 minutos conducir desde su casa en los suburbios hasta su oficina en el centro de la ciudad.

8.2 El tiempo, en minutos, que 10 pacientes esperan en un consultorio médico antes de recibir tratamiento se registraron como sigue: 5, 11, 9, 5, 10, 15, 6, 10, 5 y 10. Trate los datos como una muestra aleatoria y calcule

a) la media;
b) la mediana;
c) la moda.

8.3 Los tiempos que los 9 individuos de una muestra aleatoria tardan en reaccionar ante un estimulante se registraron como 2.5, 3.6, 3.1, 4.3, 2.9, 2.3, 2.6, 4.1 y 3.4 segundos. Calcule

a) la media;
b) la mediana.

8.4 El número de multas emitidas por infracciones de tránsito por 8 oficiales estatales durante el fin de semana del día en Conmemoración de los Caídos es 5, 4, 7, 7, 6, 3, 8 y 6.

a) Si estos valores representan el número de multas emitidas por una muestra aleatoria de 8 oficiales estatales del condado de Montgomery, en Virginia, defina una población adecuada.
b) Si los valores representan el número de multas emitidas por una muestra aleatoria de 8 oficiales estatales de Carolina del Sur, defina una población adecuada.

8.5 El número de respuestas incorrectas en un examen de competencia de verdadero-falso para una muestra aleatoria de 15 estudiantes se registraron de la siguiente manera: 2, 1, 3, 0, 1, 3, 6, 0, 3, 3, 5, 2, 1, 4 y 2. Calcule

a) la media;
b) la mediana;
c) la moda.

8.6 Calcule la media, la mediana y la moda para la muestra, cuyas observaciones, 15, 7, 8, 95, 19, 12, 8, 22 y 14 representan el número de días de incapacidad médica reportados en 9 solicitudes de devolución de impuestos. ¿Qué valor parece ser la mejor medida del centro de esos datos? Explique las razones de su preferencia.

8.7 Una muestra aleatoria de empleados de una fábrica local prometieron los siguientes donativos, en dólares, al United Fund: 100, 40, 75, 15, 20, 100, 75, 50, 30, 10, 55, 75, 25, 50, 90, 80, 15, 25, 45 y 100. Calcule

a) la media;
b) la moda.

8.8 De acuerdo con la escritora ecologista Jacqueline Killeen, los fosfatos que contienen los detergentes de uso casero pasan directamente a nuestros sistemas de desagüe, ocasionando que los lagos se conviertan en pantanos, los cuales a la larga se volverán desiertos. Los siguientes datos muestran la cantidad de fosfatos por carga de lavado, en gramos, para una muestra aleatoria de diversos tipos de detergentes que se usan de acuerdo con las instrucciones prescritas:

Detergente para ropa	**Fosfatos por carga (gramos)**
A & P Blue Sail	48
Dash	47
Concentrated All	42
Cold Water All	42
Breeze	41
Oxydol	34
Ajax	31
Sears	30
Fab	29
Cold Power	29
Bold	29
Rinso	26

Para los datos de fosfato dados, calcule

a) la media;
b) la mediana;
c) la moda.

8.9 Considere los datos del ejercicio 8.2 y calcule

a) el rango;
b) la desviación estándar.

8.10 Para la muestra de tiempos de reacción del ejercicio 8.3 calcule

a) el rango;
b) la varianza, utilizando la fórmula de la forma (8.2.1).

8.11 Para los datos del ejercicio 8.5 calcule la varianza utilizando la fórmula

a) de la forma (8.2.1);
b) del teorema 8.1.

8.12 El contenido de alquitrán de 8 marcas de cigarrillos que se seleccionan al azar de la lista más reciente publicada por la Comisión Federal de Comercio es el siguiente: 7.3, 8.6, 10.4, 16.1, 12.2, 15.1, 14.5 y 9.3 miligramos. Calcule

a) la media;
b) la varianza.

8.13 Los promedios de calificaciones de 20 estudiantes universitarios del último año, seleccionados al azar de una clase que se va a graduar, son los siguientes:

3.2	1.9	2.7	2.4	2.8
2.9	3.8	3.0	2.5	3.3
1.8	2.5	3.7	2.8	2.0
3.2	2.3	2.1	2.5	1.9

Calcule la desviación estándar.

8.14 *a*) Demuestre que la varianza de la muestra permanece sin cambio si a cada valor de la muestra se le suma o se le resta una constante c.

b) Demuestre que la varianza de la muestra se vuelve c^2 veces su valor original si cada observación de la muestra se multiplica por c.

8.15 Verifique que la varianza de la muestra 4, 9, 3, 6, 4 y 7 es 5.1, y utilice este hecho, junto con los resultados del ejercicio 8.14, para calcular

a) la varianza de la muestra 12, 27, 9, 18, 12 y 21;

b) la varianza de la muestra 9, 14, 8, 11, 9 y 12.

8.16 En la temporada 2004-2005 el equipo de futbol americano de la Universidad del Sur de California tuvo las siguientes diferencias de puntuación en los 13 partidos que jugó.

11 49 32 3 6 38 38 30 8 4 31 5 36

Calcule

a) la media de la diferencia de puntos;

b) la mediana de las diferencias de puntos.

8.3 Distribuciones muestrales

El campo de la inferencia estadística trata básicamente con generalizaciones y predicciones. Por ejemplo, con base en las opiniones de varias personas entrevistadas en la calle, los estadounidenses podrían afirmar que en una próxima elección 60% de los votantes de la ciudad de Detroit favorecerían a cierto candidato. En este caso tratamos con una muestra aleatoria de opiniones de una población finita muy grande. Por otro lado, con base en las estimaciones de 3 contratistas seleccionados al azar, de los 30 que laboran actualmente en esta ciudad, podríamos afirmar que el costo promedio de construir una residencia en Charleston, Carolina del Sur, está entre \$330,000 y \$335,000. La población que se va a muestrear aquí también es finita, pero muy pequeña. Finalmente, consideremos una máquina despachadora de bebida gaseosa que está diseñada para servir en promedio 240 mililitros de bebida. Un ejecutivo de la empresa calcula la media de 40 bebidas servidas y obtiene $\bar{x} = 236$ mililitros y, con base en este valor, decide que la máquina está sirviendo bebidas con un contenido promedio de $\mu = 240$ mililitros. Las 40 bebidas servidas representan una muestra de la población infinita de posibles bebidas que despachará esta máquina.

Inferencias sobre la población a partir de información de la muestra

En cada uno de los ejemplos anteriores calculamos un estadístico de una muestra que se selecciona de la población, y con base en tales estadísticos hicimos varias afirmaciones respecto a los valores de los parámetros de la población, que pueden ser o no ciertas. El ejecutivo de la empresa decide que la máquina despachadora está sirviendo bebidas con un contenido promedio de 240 mililitros, aunque la media de la muestra fue de 236 mililitros, porque conoce la teoría del muestreo según la cual, si $\mu = 240$ mililitros, tal valor de la muestra podría ocurrir fácilmente. De hecho, si realiza pruebas similares, cada hora por ejemplo, esperaría que los valores del estadístico $\bar{x}$ fluctuaran por arriba y por abajo de $\mu = 240$ mililitros. Sólo cuando el valor de $\bar{x}$ difiera considerablemente de 240 mililitros el ejecutivo de la empresa tomará medidas para ajustar la máquina.

Como un estadístico es una variable aleatoria que depende sólo de la muestra observada, debe tener una distribución de probabilidad.

Definición 8.5: La distribución de probabilidad de un estadístico se denomina **distribución muestral**.

La distribución muestral de un estadístico depende de la distribución de la población, del tamaño de las muestras y del método de selección de las muestras. En lo que resta de este capítulo estudiaremos varias de las distribuciones muestrales más importantes de los estadísticos que se utilizan con frecuencia. Las aplicaciones de tales distribuciones muestrales a problemas de inferencia estadística se consideran en la mayoría de los capítulos posteriores. La distribución de probabilidad de $\bar{X}$ se llama **distribución muestral de la media**.

¿Qué es la distribución muestral de $\bar{X}$?

Se deberían considerar las distribuciones muestrales de $\bar{X}$ y S^2 como los mecanismos a partir de los cuales se puede hacer inferencias acerca de los parámetros μ y σ^2. La distribución muestral de $\bar{X}$ con tamaño muestral n es la distribución que resulta cuando un **experimento se lleva a cabo una y otra vez** (siempre con una muestra de tamaño n) **y resultan los diversos valores de** $\bar{X}$. Por lo tanto, esta distribución muestral describe la variabilidad de los promedios muestrales alrededor de la media de la población μ. En el caso de la máquina despachadora de bebidas, el conocer la distribución muestral de $\bar{X}$ le permite al analista encontrar una discrepancia "típica" entre un valor $\bar{x}$ observado y el verdadero valor de μ. Se aplica el mismo principio en el caso de la distribución de S^2. La distribución muestral produce información acerca de la variabilidad de los valores de s^2 alrededor de σ^2 en experimentos que se repiten.

8.4 Distribución muestral de medias y el teorema del límite central

La primera distribución muestral importante a considerar es la de la media $\bar{X}$. Suponga que de una población normal con media μ y varianza σ^2 se toma una muestra aleatoria de n observaciones. Cada observación X_i, $i = 1, 2,..., n$, de la muestra aleatoria tendrá entonces la misma distribución normal que la población de donde se tomó. Así, por la propiedad reproductiva de la distribución normal que se estableció en el teorema 7.11, concluimos que

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

tiene una distribución normal con media

$$\mu_{\bar{X}} = \frac{1}{n}(\underbrace{\mu + \mu + \cdots + \mu}_{n \text{ términos}}) = \mu \text{ y varianza } \sigma^2_{\bar{X}} = \frac{1}{n^2}(\underbrace{\sigma^2 + \sigma^2 + \cdots + \sigma^2}_{n \text{ términos}}) = \frac{\sigma^2}{n}.$$

Si tomamos muestras de una población con distribución desconocida, ya sea finita o infinita, la distribución muestral de $\bar{X}$ aún será aproximadamente normal con media μ y varianza σ^2/n, siempre que el tamaño de la muestra sea grande. Este asombroso resultado es una consecuencia inmediata del siguiente teorema, que se conoce como teorema del límite central.

El teorema del límite central

Teorema 8.2: **Teorema del límite central:** Si $\bar{X}$ es la media de una muestra aleatoria de tamaño n, tomada de una población con media μ y varianza finita σ^2, entonces la forma límite de la distribución de

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}},$$

a medida que $n \to \infty$, es la distribución normal estándar $n(z; 0, 1)$.

La aproximación normal para $\bar{X}$ por lo general será buena si $n \geq 30$, siempre y cuando la distribución de la población no sea muy asimétrica. Si $n < 30$, la aproximación será buena sólo si la población no es muy diferente de una distribución normal y, como antes se estableció, si se sabe que la población es normal, la distribución muestral de $\bar{X}$ seguirá siendo una distribución normal exacta, sin importar qué tan pequeño sea el tamaño de las muestras.

El tamaño de la muestra $n = 30$ es un lineamiento para el teorema del límite central. Sin embargo, como indica el planteamiento del teorema, la suposición de normalidad en la distribución de $\bar{X}$ se vuelve más precisa a medida que n se hace más grande. De hecho, la figura 8.1 ilustra cómo funciona el teorema. La figura indica cómo la distribución de $\bar{X}$ se acerca más a la normalidad a medida que aumenta n, empezando con la distribución claramente asimétrica de una observación individual ($n = 1$). También ilustra que la media de $\bar{X}$ sigue siendo μ para cualquier tamaño de la muestra y que la varianza de $\bar{X}$ se vuelve más pequeña a medida que aumenta n.

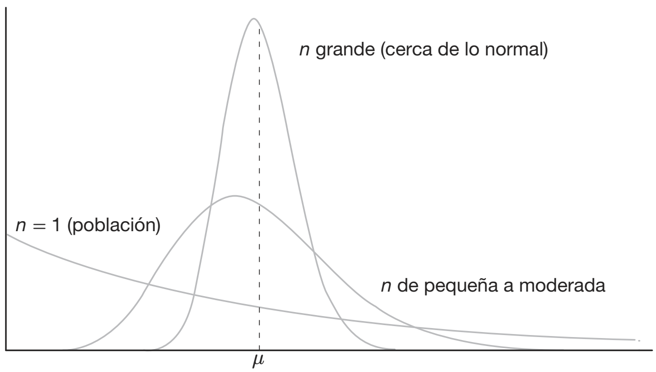

Figura 8.1: Ejemplo del teorema del límite central (distribución de $\bar{X}$ para $n = 1$, n moderada y n grande).

Ejemplo 8.4: Una empresa de material eléctrico fabrica bombillas que tienen una duración que se distribuye aproximadamente en forma normal, con media de 800 horas y desviación estándar de 40 horas. Calcule la probabilidad de que una muestra aleatoria de 16 bombillas tenga una vida promedio de menos de 775 horas.

Solución: La distribución muestral de $\bar{X}$ será aproximadamente normal, con $\mu_{\bar{X}} = 800$ y $\sigma_{\bar{X}} = 40/\sqrt{16} = 10$. La probabilidad que se desea es determinada por el área de la región sombreada de la figura 8.2.

En lo que corresponde a $\bar{x} = 775$, obtenemos que

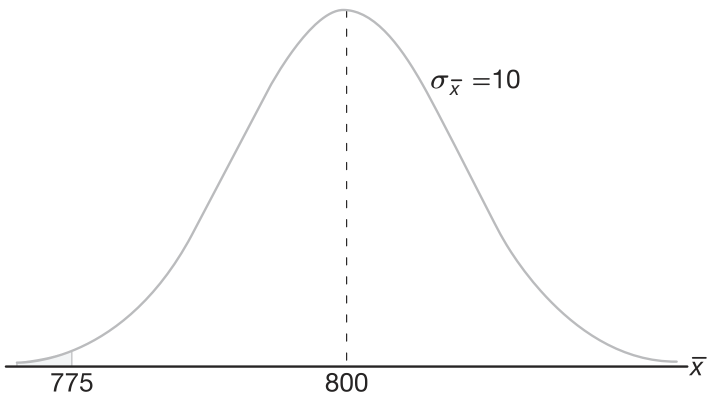

Figura 8.2: Área para el ejemplo 8.4.

$$z = \frac{775 - 800}{10} = -2.5,$$

y, por lo tanto,

$$P(\bar{X} < 775) = P(Z < -2.5) = 0.0062.$$

Inferencias sobre la media de la población

Una aplicación muy importante del teorema del límite central consiste en determinar valores razonables de la media de la población μ. Temas como prueba de hipótesis, estimación, control de calidad y muchos otros utilizan el teorema del límite central. El siguiente ejemplo ilustra cómo se utiliza el teorema del límite central con respecto a su relación con μ, la media poblacional, aunque la aplicación formal de los temas precedentes se deja para capítulos posteriores.

En el siguiente estudio de caso proporcionamos un ejemplo en el que se hace una inferencia utilizando la distribución muestral de $\bar{X}$. En este ejemplo sencillo se conocen μ y σ. El teorema del límite central y el concepto general de las distribuciones muestrales a menudo se utilizan para proporcionar evidencias acerca de algún aspecto importante de una distribución, por ejemplo uno de sus parámetros. En el caso del teorema del límite central el parámetro que nos interesa es la media μ. La inferencia que se hace acerca de μ puede adoptar una de varias formas. Con frecuencia el analista desea que los datos (en la forma de $\bar{x}$) respalden (o no) alguna conjetura predeterminada respecto al valor de μ. El uso de lo que sabemos sobre la distribución de muestreo puede contribuir a responder este tipo de pregunta. En el siguiente estudio de caso el concepto de prueba de hipótesis conduce a un objetivo formal que destacaremos en capítulos posteriores.

Estudio de caso 8.1: **Partes para automóviles.** Un importante proceso de fabricación produce partes de componentes cilíndricos para la industria automotriz. Es importante que el proceso produzca partes que tengan un diámetro medio de 5.0 milímetros. El ingeniero implicado asume que la media de la población es de 5.0 milímetros. Se lleva a cabo un experimento donde se seleccionan al azar 100 partes elaboradas por el proceso y se mide el diámetro de cada una de ellas. Se sabe que la desviación estándar de la población es $\sigma = 0.1$ milímetros. El experimento indica un diámetro promedio muestral de $\bar{x} = 5.027$ milímetros. ¿Esta información de la muestra parece apoyar o refutar la suposición del ingeniero?

***Solución*:** Este ejemplo refleja el tipo de problemas que a menudo se presentan y que se resuelven con las herramientas de pruebas de hipótesis que se presentan en los siguientes capítulos. No utilizaremos aquí el formalismo asociado con la prueba de hipótesis, pero ilustraremos los principios y la lógica que se utilizan.

El hecho de que los datos apoyen o refuten la suposición depende de la probabilidad de que datos similares a los que se obtuvieron en este experimento ($\bar{x} = 5.027$) pueden ocurrir con facilidad cuando de hecho $\mu = 5.0$ (figura 8.3). En otras palabras, ¿qué tan probable es que se pueda obtener $\bar{x} \geq 5.027$ con $n = 100$, si la media de la población es $\mu = 5.0$? Si esta probabilidad sugiere que $\bar{x} = 5.027$ no es poco razonable, no se refuta la suposición. Si la probabilidad es muy baja, se puede argumentar con certidumbre que los datos no apoyan la suposición de que $\mu = 5.0$. La probabilidad que elegimos para el cálculo es dada por $P(|\bar{X} - 5| \geq 0.027)$.

En otras palabras, si la media μ es 5, ¿cuál es la probabilidad de que $\bar{X}$ se desvíe

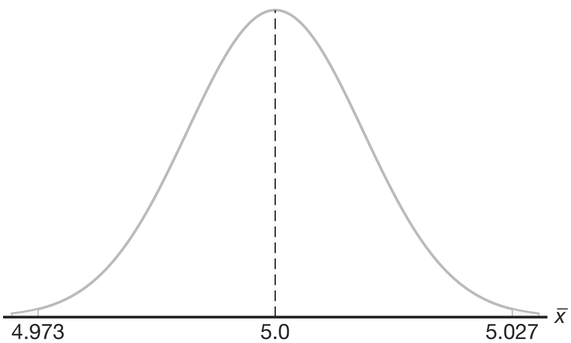

Figura 8.3: Área para el estudio de caso 8.1.

cuando mucho hasta 0.027 milímetros?

$$P(|\bar{X} - 5| \geq 0.027) = P(\bar{X} - 5 \geq 0.027) + P(\bar{X} - 5 \leq -0.027)$$
$$= 2P\left(\frac{\bar{X} - 5}{0.1/\sqrt{100}} \geq 2.7\right).$$

Aquí simplemente estandarizamos $\bar{X}$ de acuerdo con el teorema del límite central. Si la suposición $\mu = 5.0$ es cierta, $\frac{\bar{X} - 5}{0.1/\sqrt{100}}$ debería ser $N(0, 1)$. Por consiguiente,

$$2P\left(\frac{\bar{X} - 5}{0.1/\sqrt{100}} \geq 2.7\right) = 2P(Z \geq 2.7) = 2(0.0035) = 0.007.$$

Por lo tanto, se experimentaría por casualidad que una $\bar{x}$ estaría a 0.027 milímetros de la media en tan sólo 7 de 1000 experimentos. Como resultado, este experimento con $\bar{x} = 5.027$ ciertamente no ofrece evidencia que apoye la suposición de que $\mu = 5.0$. De hecho, ¡la refuta consistentemente!

Ejemplo 8.5: El viaje en un autobús especial para ir de un campus de una universidad al campus de otra en una ciudad toma, en promedio, 28 minutos, con una desviación estándar de 5 minutos. En cierta semana un autobús hizo el viaje 40 veces. ¿Cuál es la probabilidad de que el tiempo promedio del viaje sea mayor a 30 minutos? Suponga que el tiempo promedio se redondea al entero más cercano.

***Solución*:** En este caso $\mu = 28$ y $\sigma = 3$. Necesitamos calcular la probabilidad $P(\bar{X} > 30)$ con $n = 40$. Como el tiempo se mide en una escala continua redondeada al minuto más cercano, una $\bar{x}$ mayor que 30 sería equivalente a $\bar{x} \geq 30.5$. Por lo tanto,

$$P(\bar{X} > 30) = P\left(\frac{\bar{X} - 28}{5/\sqrt{40}} \geq \frac{30.5 - 28}{5/\sqrt{40}}\right) = P(Z \geq 3.16) = 0.0008.$$

Hay sólo una ligera probabilidad de que el tiempo promedio de un viaje del autobús exceda 30 minutos. En la figura 8.4 se presenta una gráfica ilustrativa.

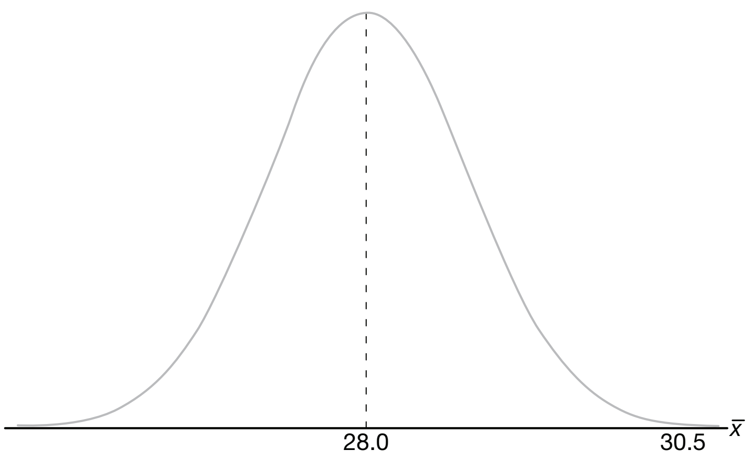

Figura 8.4: Área para el ejemplo 8.5.

Distribución muestral de la diferencia entre dos medias

La ilustración del estudio de caso 8.1 se refiere a conceptos de inferencia estadística sobre una sola media μ. El ingeniero estaba interesado en respaldar una suposición con respecto a una sola media de población. Una aplicación mucho más importante incluye dos poblaciones. Un científico o ingeniero se podrían interesar en un experimento donde se comparan dos métodos de producción: el 1 y el 2. La base para tal comparación es $\mu_1 - \mu_2$, la diferencia entre las medias de población.

Suponga que tenemos dos poblaciones, la primera con media μ_1 y varianza σ_1^2, y la segunda con media μ_2 y varianza σ_2^2. Representemos con el estadístico $\bar{X}_1$ la media de una muestra aleatoria de tamaño n_1, seleccionada de la primera población, y con el estadístico $\bar{X}_2$ la media de una muestra aleatoria de tamaño n_2 seleccionada de la segunda población, independiente de la muestra de la primera población. ¿Qué podríamos decir acerca de la distribución muestral de la diferencia $\bar{X}_1 - \bar{X}_2$ para muestras repetidas de tamaños n_1 y n_2? De acuerdo con el teorema 8.2, tanto la variable $\bar{X}_1$ como la variable $\bar{X}_2$ están distribuidas más o menos de forma normal con medias μ_1 y μ_2 y varianzas σ_1^2/n_1 y σ_2^2/n_2, respectivamente. Esta aproximación mejora a medida que aumentan n_1 y n_2. Al elegir muestras independientes de las dos poblaciones nos aseguramos de que las variables $\bar{X}_1$ y $\bar{X}_2$ sean independientes y, usando el teorema 7.11, con $a_1 = 1$ y $a_2 = -1$, concluimos que $\bar{X}_1 - \bar{X}_2$ se distribuye aproximadamente de forma normal con media

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2$$

y varianza

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

El teorema del límite central se puede ampliar fácilmente al caso de dos muestras y dos poblaciones.

Teorema 8.3: Si se extraen al azar muestras independientes de tamaños n_1 y n_2 de dos poblaciones, discretas o continuas, con medias μ_1 y μ_2 y varianzas σ_1^2 y σ_2^2, respectivamente, entonces la distribución muestral de las diferencias de las medias, $\bar{X}_1 - \bar{X}_2$, tiene una distribución aproximadamente normal, con media y varianza dadas por

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2 \text{ y } \sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

De aquí,

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$$

es aproximadamente una variable normal estándar.

Si tanto n_1 como n_2 son mayores o iguales que 30, la aproximación normal para la distribución de $\bar{X}_1 - \bar{X}_2$ es muy buena cuando las distribuciones subyacentes no están tan alejadas de la normal. Sin embargo, aun cuando n_1 y n_2 sean menores que 30, la aproximación normal es hasta cierto punto buena, excepto cuando las poblaciones no son definitivamente normales. Por supuesto, si ambas poblaciones son normales, entonces $\bar{X}_1 - \bar{X}_2$ tiene una distribución normal sin importar de qué tamaño sean n_1 y n_2.

La utilidad de la distribución muestral de la diferencia entre los dos promedios muestrales es muy similar a la que se describe en el estudio de caso 8.1 en la página 235 para el caso de una sola media. Ahora presentaremos el estudio de caso 8.2, que se enfoca en el uso de la diferencia entre dos medias muestrales para respaldar (o no) la suposición de que dos medias de población son iguales.

Estudio de caso 8.2: **Tiempo de secado de pinturas.** Se llevan a cabo dos experimentos independientes en los que se comparan dos tipos diferentes de pintura, el *A* y el *B*. Con la pintura tipo *A* se pintan 18 especímenes y se registra el tiempo (en horas) que cada uno tarda en secar. Lo mismo se hace con la pintura tipo *B*. Se sabe que la desviación estándar de población de ambas es 1.0.

Si se supone que los especímenes pintados se secan en el mismo tiempo medio con los dos tipos de pintura, calcule $P(\bar{X}_A - \bar{X}_B > 1.0)$, donde $\bar{X}_A$ y $\bar{X}_B$ son los tiempos promedio de secado para muestras de tamaño $n_A = n_B = 18$.

***Solución*:** A partir de la distribución de muestreo de $\bar{X}_A - \bar{X}_B$ sabemos que la distribución es aproximadamente normal con media

$$\mu_{\bar{X}_A - \bar{X}_B} = \mu_A - \mu_B = 0$$

y varianza

$$\sigma^2_{\bar{X}_A - \bar{X}_B} = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B} = \frac{1}{18} + \frac{1}{18} = \frac{1}{9}.$$

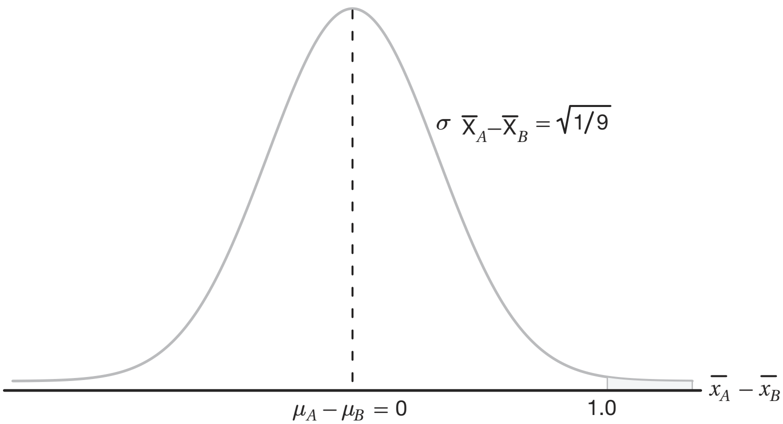

Figura 8.5: Área para el estudio de caso 8.2.

La probabilidad que se desea es dada por la región sombreada en la figura 8.5. En correspondencia con el valor $\bar{X}_A - \bar{X}_B = 1.0$, tenemos

$$z = \frac{1 - (\mu_A - \mu_B)}{\sqrt{1/9}} = \frac{1 - 0}{\sqrt{1/9}} = 3.0;$$

de modo que

$$P(Z > 3.0) = 1 - P(Z < 3.0) = 1 - 0.9987 = 0.0013.$$

¿Qué aprendemos del estudio de caso 8.2?

La mecánica en el cálculo se basa en la suposición de que $\mu_A = \mu_B$. Suponga, sin embargo, que el experimento realmente se lleva a cabo con el fin de hacer una inferencia respecto a la igualdad de μ_A y μ_B, los tiempos medios de secado de las dos poblaciones. Si se encontrara que los dos promedios difieren por una hora (o más), este resultado sería una evidencia que nos llevaría a concluir que el tiempo medio de secado de la población no es igual para los dos tipos de pintura. Por otro lado, suponga que la diferencia en los dos promedios muestrales es tan pequeña como, digamos, 15 minutos. Si $\mu_A = \mu_B$,

$$P[(\bar{X}_A - \bar{X}_B) > 0.25 \text{ horas}] = P\left(\frac{\bar{X}_A - \bar{X}_B - 0}{\sqrt{1/9}} > \frac{3}{4}\right)$$
$$= P\left(Z > \frac{3}{4}\right) = 1 - P(Z < 0.75) = 1 - 0.7734 = 0.2266.$$

Como esta probabilidad no es baja, se concluiría que una diferencia de 15 minutos en las medias de las muestras puede ocurrir por azar, es decir, sucede con frecuencia aunque $\mu_A = \mu_B$. Por lo tanto, este tipo de diferencia en el tiempo promedio de secado ciertamente *no es una señal clara* de que $\mu_A \neq \mu_B$.

Como indicamos al principio, en los capítulos siguientes se observará un formalismo más detallado con respecto a éste y a otros tipos de inferencia estadística, por ejemplo, la prueba de hipótesis. El teorema del límite central y las distribuciones de muestreo que se presentan en las siguientes tres secciones también desempeñarán un papel fundamental.

Ejemplo 8.6: Los cinescopios para televisor del fabricante A tienen una duración media de 6.5 años y una desviación estándar de 0.9 años; mientras que los del fabricante B tienen una duración media de 6.0 años y una desviación estándar de 0.8 años. ¿Cuál es la probabilidad de que una muestra aleatoria de 36 cinescopios del fabricante A tenga por lo menos 1 año más de vida media que una muestra de 49 cinescopios del fabricante B?

***Solución*:** Tenemos la siguiente información:

Población 1	**Población 2**
$\mu_1 = 6.5$	$\mu_2 = 6.0$
$\sigma_1 = 0.9$	$\sigma_2 = 0.8$
$n_1 = 36$	$n_2 = 49$

Si utilizamos el teorema 8.3, la distribución muestral de $\bar{X}_1 - \bar{X}_2$ será aproximadamente normal y tendrá una media y una desviación estándar de

$$\mu_{\bar{X}_1 - \bar{X}_2} = 6.5 - 6.0 = 0.5 \quad \text{y} \quad \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{0.81}{36} + \frac{0.64}{49}} = 0.189.$$

La probabilidad de que 36 cinescopios del fabricante A tengan por lo menos 1 año más de vida media que 49 cinescopios del fabricante B es dada por el área de la región sombreada de la figura 8.6. Con respecto al valor $\bar{x}_1 - \bar{x}_2 = 1.0$, encontramos que

$$z = \frac{1.0 - 0.5}{0.189} = 2.65,$$

y de aquí

$$P(\bar{X}_1 - \bar{X}_2 \geq 1.0) = P(Z > 2.65) = 1 - P(Z < 2.65)$$
$$= 1 - 0.9960 = 0.0040.$$

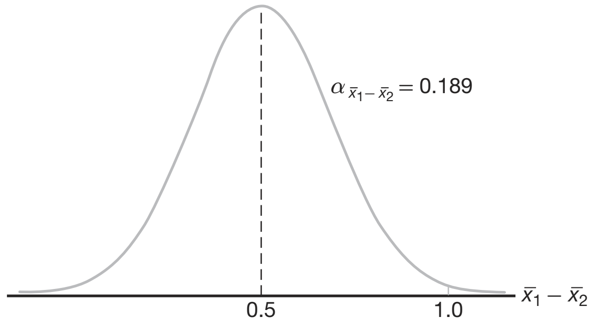

Figura 8.6: Área para el ejemplo 8.6.

Más sobre la distribución muestral de medias. Aproximación normal a la distribución binomial

En la sección 6.5 analizamos a fondo la aproximación normal a la distribución binomial. Estaban dadas las condiciones sobre los parámetros n y p, para los cuales la distribución de una variable aleatoria binomial se puede aproximar mediante la distribución normal. Los ejemplos y los ejercicios reflejaron la importancia del concepto de "aproximación normal". Resulta que el teorema del límite central da más idea de cómo y por qué funciona esta aproximación. Sabemos con certeza que una variable aleatoria binomial es el número X de éxitos en n pruebas independientes, donde el resultado de cada prueba es binario. En el capítulo 1 también vimos que la proporción calculada en un experimento así es un promedio de un conjunto de ceros y unos. De hecho, mientras que la proporción X/n es un promedio, X es la suma de este conjunto de ceros y unos, y tanto X como X/n son casi normales si n es suficientemente grande. Desde luego, a partir de lo que aprendimos en el capítulo 6, sabemos que hay condiciones de n y p que afectan la calidad de la aproximación; a saber, $np \geq 5$ y $nq \geq 5$.

Ejercicios

8.17 Si se extraen todas las muestras posibles de tamaño 16 de una población normal con media igual a 50 y desviación estándar igual a 5, ¿cuál es la probabilidad de que una media muestral $\bar{X}$ caiga en el intervalo que va de $\mu_{\bar{X}} - 1.9\sigma_{\bar{X}}$ a $\mu_{\bar{X}} - 0.4\sigma_{\bar{X}}$? Suponga que las medias muestrales se pueden medir con cualquier grado de precisión.

8.18 Si la desviación estándar de la media para la distribución muestral de muestras aleatorias de tamaño 36 de una población grande o infinita es 2, ¿qué tan grande debe ser el tamaño de la muestra si la desviación estándar se reduce a 1.2?

8.19 Se fabrica cierto tipo de hilo con una resistencia a la tensión media de 78.3 kilogramos y una desviación estándar de 5.6 kilogramos. ¿Cómo cambia la varianza de la media muestral cuando el tamaño de la muestra

a) aumenta de 64 a 196?
b) disminuye de 784 a 49?

8.20 Dada la población uniforme discreta

$$f(x) = \begin{cases} \frac{1}{3}, & x = 2, 4, 6, \\ 0, & \text{en otro caso,} \end{cases}$$

calcule la probabilidad de que una muestra aleatoria de tamaño 54, seleccionada con reemplazo, produzca una media muestral mayor que 4.1 pero menor que 4.4. Suponga que las medias se miden al décimo más cercano.

8.21 Una máquina de bebidas gaseosas se ajusta de manera que la cantidad de bebida que sirve promedie 240 mililitros con una desviación estándar de 15 mililitros. La máquina se verifica periódicamente tomando una muestra de 40 bebidas y calculando el contenido promedio. Si la media de las 40 bebidas es un valor dentro del intervalo $\mu_{\bar{X}} \pm 2\sigma_{\bar{X}}$, se piensa que la máquina opera satisfactoriamente; de lo contrario, se ajusta. En la sección 8.3 el ejecutivo de la empresa encontró que la media de 40 bebidas era $\bar{x} = 236$ mililitros y concluyó que la máquina no necesitaba un ajuste. ¿Fue ésta una decisión razonable?

8.22 Las estaturas de 1000 estudiantes se distribuyen aproximadamente de forma normal con una media de 174.5 centímetros y una desviación estándar de 6.9 centímetros. Si se extraen 200 muestras aleatorias de tamaño 25 de esta población y las medias se registran al décimo de centímetro más cercano, determine

a) la media y la desviación estándar de la distribución muestral de $\bar{X}$;
b) el número de las medias muestrales que caen entre 172.5 y 175.8 centímetros;
c) el número de medias muestrales que caen por debajo de 172.0 centímetros.

8.23 La variable aleatoria X, que representa el número de cerezas en un tarta, tiene la siguiente distribución de probabilidad:

x	4	5	6	7
$P(X = x)$	0.2	0.4	0.3	0.1

a) Calcule la media μ y la varianza σ^2 de X.
b) Calcule la media $\mu_{\bar{X}}$ y la varianza $\sigma^2_{\bar{X}}$ de la media $\bar{X}$ para muestras aleatorias de 36 tartas de cereza.
c) Calcule la probabilidad de que el número promedio de cerezas en 36 tartas sea menor que 5.5.

8.24 Si cierta máquina fabrica resistencias eléctricas que tienen una resistencia media de 40 ohms y una desviación estándar de 2 ohms, ¿cuál es la probabilidad de que una muestra aleatoria de 36 de estas resistencias tenga una resistencia combinada de más de 1458 ohms?

8.25 La vida media de una máquina para elaborar pan es de 7 años, con una desviación estándar de 1 año. Suponga que la vida de estas máquinas sigue aproximadamente una distribución normal y calcule

a) la probabilidad de que la vida media de una muestra aleatoria de 9 de estas máquinas caiga entre 6.4 y 7.2 años;
b) el valor de x a la derecha del cual caería 15% de las medias calculadas de muestras aleatorias de tamaño 9.

8.26 La cantidad de tiempo que le toma al cajero de un banco con servicio en el automóvil atender a un cliente es una variable aleatoria con una media $\mu = 3.2$ minutos y una desviación estándar $\sigma = 1.6$ minutos. Si se observa una muestra aleatoria de 64 clientes, calcule la probabilidad de que el tiempo medio que el cliente pasa en la ventanilla del cajero sea

a) a lo sumo 2.7 minutos;
b) más de 3.5 minutos;
c) al menos 3.2 minutos pero menos de 3.4 minutos.

8.27 En un proceso químico la cantidad de cierto tipo de impureza en el producto es difícil de controlar y por ello es una variable aleatoria. Se especula que la cantidad media de la población de impurezas es 0.20 gramos por gramo del producto. Se sabe que la desviación estándar es 0.1 gramos por gramo. Se realiza un experimento para entender mejor la especulación de que $\mu = 0.2$. El proceso se lleva a cabo 50 veces en un laboratorio y el promedio de la muestra $\bar{x}$ resulta ser 0.23 gramos por gramo. Comente sobre la especulación de que la cantidad media de impurezas es 0.20 gramos por gramo. Utilice el teorema del límite central en su respuesta.

8.28 Se toma una muestra aleatoria de tamaño 25 de una población normal que tiene una media de 80 y una desviación estándar de 5. Una segunda muestra aleatoria de tamaño 36 se toma de una población normal diferente que tiene una media de 75 y una desviación estándar de 3. Calcule la probabilidad de que la media muestral calculada de las 25 mediciones exceda la media muestral calculada de las 36 mediciones por lo menos 3.4 pero menos de 5.9. Suponga que las diferencias de las medias se miden al décimo más cercano.

8.29 La distribución de alturas de cierta raza de perros *terrier* tiene una media de 72 centímetros y una desviación estándar de 10 centímetros; en tanto que la distribución de alturas de cierta raza de *poodles* tiene una media de 28 centímetros con una desviación estándar de 5 centímetros. Suponga que las medias muestrales se pueden medir con cualquier grado de precisión y calcule la probabilidad de que la media muestral de una muestra aleatoria de alturas de 64 *terriers* exceda la media muestral para una muestra aleatoria de alturas de 100 *poodles* a lo sumo 44.2 centímetros.

8.30 La calificación promedio de los estudiantes de primer año en un examen de aptitudes en cierta universidad es 540, con una desviación estándar de 50. Suponga que las medias se miden con cualquier grado de precisión. ¿Cuál es la probabilidad de que dos grupos seleccionados al azar, que constan de 32 y 50 estudiantes, respectivamente, difieran en sus calificaciones promedio por

a) más de 20 puntos?
b) una cantidad entre 5 y 10 puntos?

8.31 Considere el estudio de caso 8.2 de la página 238. Suponga que en un experimento se utilizaron 18 especímenes para cada tipo de pintura y que $\bar{x}_A - \bar{x}_B$, la diferencia real en el tiempo medio de secado, resultó ser 1.0.

a) ¿Parecería ser un resultado razonable si los dos tiempos promedio de secado de las dos poblaciones realmente son iguales? Utilice el resultado que se obtuvo en el estudio de caso 8.2.

b) Si alguien hiciera el experimento 10,000 veces bajo la condición de que $\mu_A = \mu_B$, ¿en cuántos de esos 10,000 experimentos habría una diferencia $\bar{x}_A - \bar{x}_B$ tan grande como 1.0 (o más grande)?

8.32 Dos máquinas diferentes de llenado de cajas se utilizan para llenar cajas de cereal en una línea de ensamble. La medición fundamental en la que influyen estas máquinas es el peso del producto en las cajas. Los ingenieros están seguros de que la varianza en el peso del producto es $\sigma^2 = 1$ onza. Se realizan experimentos usando ambas máquinas con tamaños muestrales de 36 cada una. Los promedios muestrales para las máquinas A y B son $\bar{x}_A = 4.5$ onzas y $\bar{x}_B = 4.7$ onzas. Los ingenieros se sorprenden de que los dos promedios maestrales para las máquinas de llenado sean tan diferentes.

a) Utilice el teorema del límite central para determinar

$$P(\bar{X}_B - \bar{X}_A \geq 0.2)$$

bajo la condición de que $\mu_A = \mu_B$.

b) ¿Los experimentos mencionados parecen, de cualquier forma, apoyar consistentemente la suposición de que las medias de población de las dos máquinas son diferentes? Explique utilizando la respuesta que encontró en el inciso *a*.

8.33 El benceno es una sustancia química altamente tóxica para los seres humanos. Sin embargo, se utiliza en la fabricación de medicamentos, de tintes y de recubrimientos, así como en la peletería. Las regulaciones del gobierno establecen que el contenido de benceno en el agua que resulte de cualquier proceso de producción en el que participe esta sustancia no debe exceder 7950 partes por millón (ppm). Para un proceso particular de interés, un fabricante recolectó una muestra de agua 25 veces de manera aleatoria y el promedio muestral $\bar{x}$ fue de 7960 ppm. A partir de los datos históricos, se sabe que la desviación estándar σ es 100 ppm.

a) ¿Cuál es la probabilidad de que el promedio muestral en este experimento exceda el límite establecido por el gobierno, si la media de la población es igual al límite? Utilice el teorema del límite central.

b) ¿La $\bar{x} = 7960$ observada en este experimento es firme evidencia de que la media de la población en este proceso excede el límite impuesto por el gobierno? Responda calculando

$$P(\bar{X} \geq 7960 \mid \mu = 7950).$$

Suponga que la distribución de la concentración de benceno es normal.

8.34 En la fabricación de cierto producto de acero se están utilizando dos aleaciones, la A y la B. Se necesita diseñar un experimento para comparar las dos aleaciones en términos de su capacidad de carga máxima en toneladas, es decir, la cantidad máxima de carga que pueden soportar sin romperse. Se sabe que las dos desviaciones estándar de la capacidad de carga son iguales a 5 toneladas cada una. Se realiza un experimento en el que se prueban 30 especímenes de cada aleación (A y B) y se obtienen los siguientes resultados:

$$\bar{x}_A = 49.5, \quad \bar{x}_B = 45.5; \qquad \bar{x}_A - \bar{x}_B = 4.$$

Los fabricantes de la aleación A están convencidos de que esta evidencia demuestra de forma concluyente que $\mu_A > \mu_B$ y, por lo tanto, que su aleación es mejor. Los fabricantes de la aleación B afirman que el experimento fácilmente podría haber resultado $\bar{x}_A - \bar{x}_B = 4$, *incluso si* las dos medias de población fueran iguales. En otras palabras, "¡los resultados no son concluyentes!".

a) Encuentre un argumento que ponga en evidencia el error de los fabricantes de la aleación B. Para ello calcule

$$P(\bar{X}_A - \bar{X}_B > 4 \mid \mu_A = \mu_B).$$

b) ¿Considera que estos datos apoyan fuertemente a la aleación A?

8.35 Considere la situación del ejemplo 8.4 de la página 234. ¿Los resultados que se obtuvieron allí lo llevan a cuestionar la premisa de que $\mu = 800$ horas? Proporcione un resultado probabilístico que indique qué tan *raro* es el evento $\bar{X} \leq 775$ cuando $\mu = 800$. Por otro lado, ¿qué tan raro sería si μ fuera, verdaderamente, digamos, $\neq 760$ horas?

8.36 Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución que sólo puede adoptar valores positivos. Utilice el teorema del límite central para argumentar que si n es tan grande como se requiere, entonces $Y = X_1 X_2 \ldots X_n$ tiene aproximadamente una distribución logarítmica normal.

8.5 Distribución muestral de S^2

En la sección anterior aprendimos acerca de la distribución muestral de $\bar{X}$. El teorema del límite central nos permitió utilizar el hecho de que

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

tiende a $N(0, 1)$ a medida que crece el tamaño de la muestra. *Las distribuciones muestrales de estadísticos importantes* nos permiten conocer información sobre los parámetros. Por lo general, los parámetros son las contrapartes del estadístico en cuestión. Por ejemplo, si un ingeniero se interesa en la resistencia media de la población de cierto tipo de resistencia, sacará provecho de la distribución muestral de $\bar{X}$ una vez que reúna la información de la muestra. Por otro lado, si está estudiando la variabilidad en la resistencia, evidentemente utilizará la distribución muestral de S^2 para conocer la contraparte paramétrica, la varianza de la población σ^2.

Si se extrae una muestra aleatoria de tamaño n de una población normal con media μ y varianza σ^2, y se calcula la varianza muestral, se obtiene un valor del estadístico S^2. Procederemos a considerar la distribución del estadístico $(n-1)S^2/\sigma^2$.

Mediante la suma y la resta de la media muestral $\bar{X}$ es fácil ver que

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\mu)^2 &= \sum_{i=1}^{n}[(X_i-\bar{X})+(\bar{X}-\mu)]^2\\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2+\sum_{i=1}^{n}(\bar{X}-\mu)^2+2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\bar{X})\\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2+n(\bar{X}-\mu)^2.\end{aligned}$$

Al dividir cada término de la igualdad entre σ^2 y sustituir $(n-1)S^2$ por $\sum_{i=1}^{n}(X_i-\bar{X})^2$, obtenemos

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2=\frac{(n-1)S^2}{\sigma^2}+\frac{(\bar{X}-\mu)^2}{\sigma^2/n}.$$

Ahora, de acuerdo con el corolario 7.1 de la página 222, sabemos que

$$\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^2}$$

es una variable aleatoria chi cuadrada con n grados de libertad. Tenemos una variable aleatoria chi cuadrada con n grados de libertad dividida en dos componentes. Observe que en la sección 6.7 demostramos que una distribución chi cuadrada es un caso especial de la distribución gamma. El segundo término del lado derecho es Z^2, que es una variable aleatoria chi cuadrada con 1 grado de libertad, y resulta que $(n-1)S^2/\sigma^2$ es una variable aleatoria chi cuadrada con $n-1$ grados de libertad. Formalizamos esto en el siguiente teorema.

Teorema 8.4: Si S^2 es la varianza de una muestra aleatoria de tamaño n que se toma de una población normal que tiene la varianza σ^2, entonces el estadístico

$$\chi^2=\frac{(n-1)S^2}{\sigma^2}=\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{\sigma^2}$$

tiene una distribución chi cuadrada con $v=n-1$ grados de libertad.

Los valores de la variable aleatoria χ^2 se calculan de cada muestra mediante la fórmula

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}.$$

La probabilidad de que una muestra aleatoria produzca un valor χ^2 mayor que algún valor específico es igual al área bajo la curva a la derecha de este valor. El valor χ^2 por arriba del cual se encuentra un área de α por lo general se representa con χ^2_α. Esto se ilustra mediante la región sombreada de la figura 8.7.

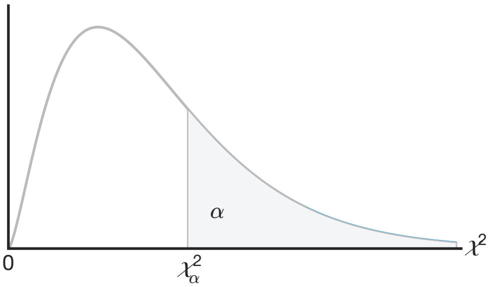

Figura 8.7: La distribución chi cuadrada.

La tabla A.5 da los valores de χ^2_α para diversos valores de α y v. Las áreas, α, son los encabezados de las columnas; los grados de libertad, v, se dan en la columna izquierda, y las entradas de la tabla son los valores χ^2. En consecuencia, el valor χ^2 con 7 grados de libertad, que deja un área de 0.05 a la derecha, es $\chi^2_{0.05} = 14.067$. Debido a la falta de simetría, para encontrar $\chi^2_{0.95} = 2.167$ para $v = 7$ también debemos usar las tablas.

Exactamente 95% de una distribución chi cuadrada cae entre $\chi^2_{0.975}$ y $\chi^2_{0.025}$. Un valor χ^2 que cae a la derecha de $\chi^2_{0.025}$ no tiene probabilidades de ocurrir, a menos que el valor de σ^2 que supusimos sea demasiado pequeño. Lo mismo sucede con un valor χ^2 que cae a la izquierda de $\chi^2_{0.975}$, el cual tampoco es probable que ocurra, a menos que el valor de σ^2 que supusimos sea demasiado grande. En otras palabras, es posible tener un valor χ^2 a la izquierda de $\chi^2_{0.975}$ o a la derecha de $\chi^2_{0.025}$ cuando el valor de σ^2 es correcto; pero si esto sucediera, lo más probable es que el valor de σ^2 que se supuso sea un error.

Ejemplo 8.7: Un fabricante de baterías para automóvil garantiza que su producto durará, en promedio, 3 años con una desviación estándar de 1 año. Si cinco de estas baterías tienen duraciones de 1.9, 2.4, 3.0, 3.5 y 4.2 años, ¿el fabricante continuará convencido de que sus baterías tienen una desviación estándar de 1 año? Suponga que las duraciones de las baterías siguen una distribución normal.

***Solución*:** Primero se calcula la varianza de la muestra usando el teorema 8.1,

$$s^2 = \frac{(5)(48.26) - (15)^2}{(5)(4)} = 0.815.$$

Entonces,

$$\chi^2 = \frac{(4)(0.815)}{1} = 3.26$$

es un valor de una distribución chi cuadrada con 4 grados de libertad. Como 95% de los valores χ^2 con 4 grados de libertad cae entre 0.484 y 11.143, el valor calculado con $\sigma^2 = 1$ es razonable y, por lo tanto, el fabricante no tiene razones para sospechar que la desviación estándar no sea igual a 1 año. ■

Grados de libertad como una medición de la información muestral

Del corolario 7.1 expuesto en la sección 7.3 recuerde que

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2}$$

tiene una distribución χ^2 con *n grados de libertad*. Observe también el teorema 8.4, el cual indica que la variable aleatoria

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2}$$

tiene una distribución χ^2 con $n - 1$ *grados de libertad*. El lector debe también recordar que el término *grados de libertad*, que se utiliza en este contexto idéntico, se estudió en el capítulo 1.

Como antes indicamos, el teorema 8.4 no se demostrará; sin embargo, el lector puede verlo como una indicación de que cuando no se conoce μ y se considera la distribución de

$$\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2},$$

hay **1 grado menos de libertad**, o se pierde un grado de libertad al estimar μ (es decir, cuando μ se reemplaza por $\bar{x}$). En otras palabras, en la muestra aleatoria de la distribución normal hay n grados de libertad o *partes de información* independientes. Cuando los datos (los valores en la muestra) se utilizan para calcular la media, hay un grado menos de libertad en la información que se utiliza para estimar σ^2.

8.6 Distribución *t*

En la sección 8.4 se analizó la utilidad del teorema del límite central. Sus aplicaciones giran en torno a las inferencias sobre una media de la población o a la diferencia entre dos medias de población. En este contexto es evidente la utilidad de utilizar el teorema del límite central y la distribución normal. Sin embargo, se supuso que se conoce la desviación estándar de la población. Esta suposición quizá sea razonable en situaciones en las que el ingeniero está muy familiarizado con el sistema o proceso. Sin embargo, en muchos escenarios experimentales el conocimiento de σ no es ciertamente más razonable que el conocimiento de la media de la población μ. A menudo, de hecho, una estimación de σ debe ser proporcionada por la misma información muestral que produce el promedio muestral $\bar{x}$. Como resultado, un estadístico natural a considerar para tratar con las inferencias sobre μ es

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}},$$

dado que S es el análogo de la muestra para σ. Si el tamaño de la muestra es pequeño, los valores de S^2 fluctúan de forma considerable de una muestra a otra (véase el ejercicio 8.43 de la página 259) y la distribución de T se desvía de forma apreciable de la de una distribución normal estándar.

Si el tamaño de la muestra es suficientemente grande, digamos $n \geq 30$, la distribución de T no difiere mucho de la normal estándar. Sin embargo, para $n < 30$ es útil tratar con la distribución exacta de T. Para desarrollar la distribución muestral de T, supondremos que nuestra muestra aleatoria se seleccionó de una población normal. Podemos escribir, entonces,

$$T = \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}} = \frac{Z}{\sqrt{V/(n-1)}},$$

donde

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

tiene una distribución normal estándar y

$$V = \frac{(n-1)S^2}{\sigma^2}$$

tiene una distribución chi cuadrada con $v = n - 1$ grados de libertad. Al obtener muestras de poblaciones normales se puede demostrar que $\bar{X}$ y S^2 son independientes y, en consecuencia, también lo son Z y V. El siguiente teorema proporciona la definición de una variable aleatoria T como una función de Z (normal estándar) y χ^2. Para completar se proporciona la función de densidad de la distribución t.

Teorema 8.5: Sea Z una variable aleatoria normal estándar y V una variable aleatoria chi cuadrada con v grados de libertad. Si Z y V son independientes, entonces la distribución de la variable aleatoria T, donde

$$T = \frac{Z}{\sqrt{V/v}},$$

es dada por la función de densidad

$$h(t) = \frac{\Gamma[(v+1)/2]}{\Gamma(v/2)\sqrt{\pi v}} \left(1 + \frac{t^2}{v}\right)^{-(v+1)/2}, \quad -\infty < t < \infty.$$

Ésta se conoce como la **distribución *t*** con v grados de libertad.

A partir de lo antes expuesto, y del teorema anterior, se deriva el siguiente corolario.

Corolario 8.1: Sean $X_1, X_2, \ldots, X_n$ variables aleatorias independientes normales con media μ y desviación estándar σ. Sea

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{y} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Entonces la variable aleatoria $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ tiene una distribución t con $v = n - 1$ grados de libertad.

La distribución de probabilidad de T se publicó por primera vez en 1908 en un artículo de W. S. Gosset. En esa época, Gosset trabajaba para una cervecería irlandesa que prohibía a sus empleados que publicaran los resultados de sus investigaciones. Para evadir la prohibición Gosset publicó su trabajo en secreto bajo el seudónimo de "Student". Es por esto que a la distribución de T se le suele llamar distribución t de Student o simplemente distribución t. Para derivar la ecuación de esta distribución Gosset supuso que las muestras se seleccionaban de una población normal. Aunque ésta parecería una suposición muy restrictiva, se puede demostrar que las poblaciones que no son normales y que poseen distribuciones en forma casi de campana aún proporcionan valores de T que se aproximan muy de cerca a la distribución t.

¿Qué apariencia tiene la distribución *t*?

La distribución de T se parece a la distribución de Z en que ambas son simétricas alrededor de una media de cero. Ambas distribuciones tienen forma de campana, pero la distribución t es más variable debido al hecho de que los valores T dependen de las fluctuaciones de dos cantidades, $\bar{X}$ y S^2; mientras que los valores Z dependen sólo de los cambios en $\bar{X}$ de una muestra a otra. La distribución de T difiere de la de Z en que la varianza de T depende del tamaño de la muestra n y siempre es mayor que 1. Sólo cuando el tamaño de la muestra $n \to \infty$ las dos distribuciones serán iguales. En la figura 8.8 se presenta la relación entre una distribución normal estándar ($v = \infty$) y las distribuciones t con 2 y 5 grados de libertad. Los puntos porcentuales de la distribución t se dan en la tabla A.4.

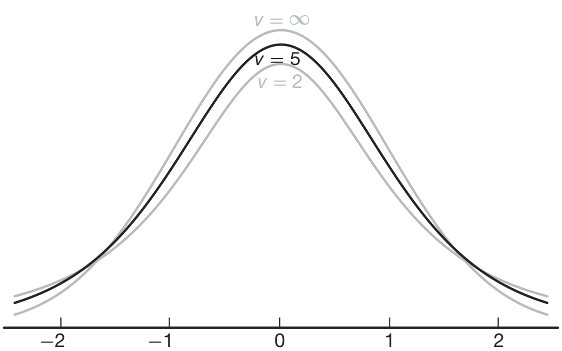

Figura 8.8: Curvas de la distribución t para $v = 2, 5$ y ∞.

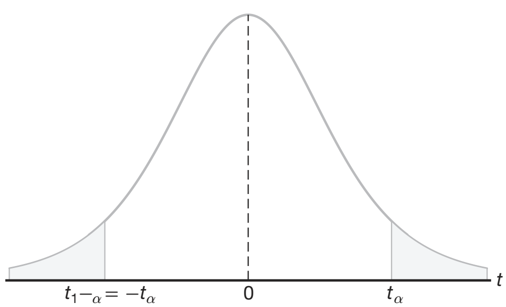

Figura 8.9: Propiedad de simetría (alrededor de 0) de la distribución t.

El valor t por arriba del cual se encuentra un área igual a α por lo general se representa con t_α. Por consiguiente, el valor t con 10 grados de libertad que deja una área de 0.025 a la derecha es $t = 2.228$. Como la distribución t es simétrica alrededor de una media de cero, tenemos $t_{1-\alpha} = -t_\alpha$; es decir, el valor t que deja una área de $1 - \alpha$ a la derecha y, por lo tanto, una área de α a la izquierda es igual al valor t negativo que deja una área de α en la cola derecha de la distribución (véase la figura 8.9). Esto es, $t_{0.95} = -t_{0.05}$, $t_{0.99} = -t_{0.01}$, etcétera.

Ejemplo 8.8: El valor t con $v = 14$ grados de libertad que deja una área de 0.025 a la izquierda y, por lo tanto, una área de 0.975 a la derecha, es

$$t_{0.975} = -t_{0.025} = -2.145.$$

Ejemplo 8.9: Calcule $P(-t_{0.025} < T < t_{0.05})$.

***Solución*:** Como $t_{0.05}$ deja una área de 0.05 a la derecha y $-t_{0.025}$ deja una área de 0.025 a la izquierda, obtenemos una área total de

$$1 - 0.05 - 0.025 = 0.925$$

entre $-t_{0.025}$ y $t_{0.05}$. En consecuencia,

$$P(-t_{0.025} < T < t_{0.05}) = 0.925.$$

Ejemplo 8.10: Calcule k tal que $P(k < T < -1.761) = 0.045$ para una muestra aleatoria de tamaño 15 que se selecciona de una distribución normal y $\frac{\bar{X}-\mu}{s/\sqrt{n}}$.

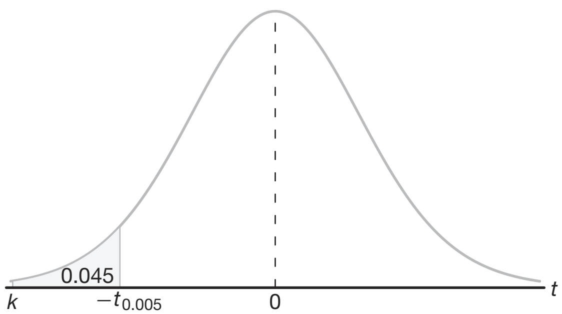

Figura 8.10: Valores t para el ejemplo 8.10.

***Solución*:** A partir de la tabla A.4 advertimos que 1.761 corresponde a $t_{0.05}$ cuando $v = 14$. Por lo tanto, $-t_{0.05} = -1.761$. Puesto que en el enunciado de probabilidad original k está a la izquierda de $-t_{0.05} = -1.761$, tenemos que $k = -t_\alpha$. Entonces, a partir de la figura 8.10, tenemos

$$0.045 = 0.05 - \alpha, \text{ o } \alpha = 0.005.$$

Así, de la tabla A.4 con $v = 14$,

$$k = -t_{0.005} = -2.977 \text{ y } P(-2.977 < T < -1.761) = 0.045.$$

Exactamente 95% de los valores de una distribución t con $v = n - 1$ grados de libertad caen entre $-t_{0.025}$ y $t_{0.025}$. Por supuesto, hay otros valores t que contienen 95% de la distribución, como $-t_{0.02}$ y $t_{0.03}$, pero estos valores no aparecen en la tabla A.4 y, además, el intervalo más corto posible se obtiene eligiendo valores t que dejen exactamente la misma área en las dos colas de nuestra distribución. Un valor t que caiga por debajo de $-t_{0.025}$ o por arriba de $t_{0.025}$ tendería a hacernos creer que ha ocurrido un evento muy raro, o que quizá nuestra suposición acerca de μ es un error. Si esto ocurriera, tendríamos que tomar la decisión de que el valor de μ que supusimos es erróneo. De hecho, un valor t que cae por debajo de $-t_{0.01}$ o por arriba de $t_{0.01}$ proporcionaría incluso evidencia más sólida de que el valor de μ que supusimos es muy improbable. En el capítulo 10 se tratarán procedimientos generales para probar aseveraciones respecto al valor del parámetro μ. El siguiente ejemplo ilustra una vista preliminar del fundamento de tales procedimientos.

Ejemplo 8.11: Un ingeniero químico afirma que el rendimiento medio de la población de un cierto proceso de lotes es 500 gramos por mililitro de materia prima. Para verificar dicha afirmación muestrea 25 lotes cada mes. Si el valor t calculado cae entre $-t_{0.05}$ y $t_{0.05}$, queda satisfecho con su afirmación. ¿Qué conclusión debería sacar de una muestra que tiene una media $\bar{x} = 518$ gramos por mililitro y una desviación estándar muestral $s = 40$ gramos? Suponga que la distribución de rendimientos es aproximadamente normal.

***Solución*:** En la tabla A.4 encontramos que $t_{0.05} = 1.711$ para 24 grados de libertad. Por lo tanto, el ingeniero quedará satisfecho con esta afirmación si una muestra de 25 lotes rinde un valor t entre -1.711 y 1.711. Si $\mu = 500$, entonces,

$$t = \frac{518 - 500}{40/\sqrt{25}} = 2.25,$$

un valor muy superior a 1.711. La probabilidad de obtener un valor t, con $v = 24$, igual o mayor que 2.25, es aproximadamente 0.02. Si $\mu > 500$, el valor de t calculado de la muestra sería más razonable. Por lo tanto, es probable que el ingeniero concluya que el proceso produce un mejor producto del que pensaba.

¿Para qué se utiliza la distribución *t*?

La distribución t se usa ampliamente en problemas relacionados con inferencias acerca de la media de la población (como se ilustra en el ejemplo 8.11) o en problemas que implican muestras comparativas (es decir, en casos donde se trata de determinar si las medias de dos muestras son muy diferentes). El uso de la distribución se ampliará en los capítulos 9, 10, 11 y 12. El lector debería notar que el uso de la distribución t para el estadístico

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

requiere que $X_1, X_2, \ldots, X_n$ sean normales. El uso de la distribución t y la consideración del tamaño de la muestra no se relacionan con el teorema del límite central. El uso de la distribución normal estándar en vez de T para $n \geq 30$ sólo implica, en este caso, que S es un estimador suficientemente bueno de σ. En los siguientes capítulos la distribución t se usa con amplitud.

8.7 Distribución *F*

Recomendamos la distribución t en parte por su aplicación a problemas en los que hay muestreo comparativo, es decir, a problemas en que se tienen que comparar dos medias muestrales. Por ejemplo, algunos de los ejemplos que daremos en los siguientes capítulos adoptarán un método aún más formal; un ingeniero químico reúne datos de dos catalizadores, un biólogo recoge datos sobre dos medios de crecimiento o un químico reúne datos sobre dos métodos de recubrimiento de material para prevenir la corrosión. Si bien es importante que la información muestral aclare lo relacionado con dos medias de población, a menudo éste es el caso en el que comparar la variabilidad es igual de importante, si no es que más. La distribución F tiene una amplia aplicación en la comparación de varianzas muestrales y también es aplicable en problemas que implican dos o más muestras.

El estadístico F se define como el cociente de dos variables aleatorias chi cuadrada independientes, dividida cada una entre su número de grados de libertad. En consecuencia, podemos escribir

$$F = \frac{U/v_1}{V/v_2},$$

donde U y V son variables aleatorias independientes que tienen distribuciones chi cuadrada con v_1 y v_2 grados de libertad, respectivamente. Estableceremos ahora la distribución muestral de F.

Teorema 8.6: Sean U y V dos variables aleatorias independientes que tienen distribuciones chi cuadrada con v_1 y v_2 grados de libertad, respectivamente. Entonces, la distribución de la variable aleatoria $F = \frac{U/v_1}{V/v_2}$ es dada por la función de densidad

$$h(f) = \begin{cases} \frac{\Gamma[(v_1+v_2)/2](v_1/v_2)^{v_1/2}}{\Gamma(v_1/2)\Gamma(v_2/2)} \frac{f^{(v_1/2)-1}}{(1+v_1 f/v_2)^{(v_1+v_2)/2}}, & f > 0, \\ 0, & f \leq 0. \end{cases}$$

Ésta se conoce como la **distribución *F*** con v_1 y v_2 grados de libertad (g.l.).

En capítulos posteriores utilizaremos ampliamente la variable aleatoria F. Sin embargo, no emplearemos la función de densidad, la cual sólo se dará como complemento. La curva de la distribución F no sólo depende de los dos parámetros v_1 y v_2 sino también del orden en el que se establecen. Una vez que tenemos estos dos valores, podemos identificar la curva. En la figura 8.11 se presentan distribuciones F típicas.

Sea f_α el valor f por arriba del cual encontramos un área igual a α. Esto se ilustra mediante la región sombreada de la figura 8.12. La tabla A.6 proporciona valores de f_α sólo para $\alpha = 0.05$ y $\alpha = 0.01$ para varias combinaciones de los grados de libertad v_1 y v_2. Por lo tanto, el valor f con 6 y 10 grados de libertad, que deja un área de 0.05 a la derecha, es $f_{0.05} = 3.22$. Por medio del siguiente teorema, la tabla A.6 también se puede utilizar para encontrar valores de $f_{0.95}$ y $f_{0.99}$. La demostración se deja al lector.

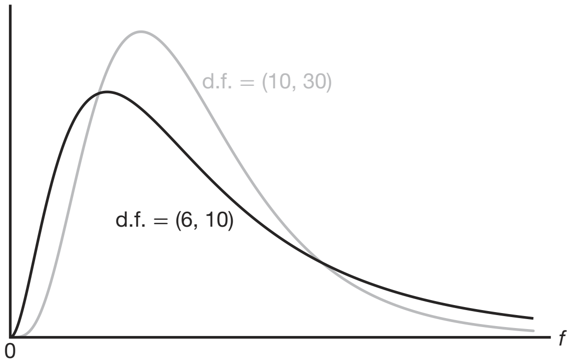

Figura 8.11: Distribuciones F típicas.

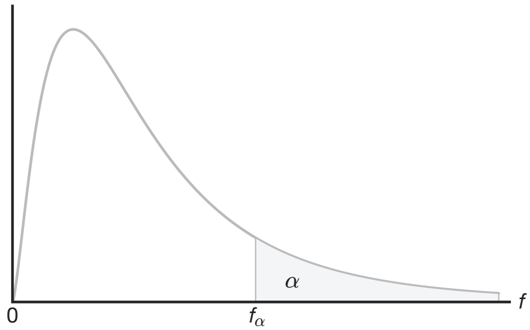

Figura 8.12: Ilustración de la f_α para la distribución F.

Teorema 8.7: Al escribir $f_\alpha(v_1, v_2)$ para f_α con v_1 y v_2 grados de libertad, obtenemos

$$f_{1-\alpha}(v_1, v_2) = \frac{1}{f_\alpha(v_2, v_1)}.$$

Por consiguiente, el valor f con 6 y 10 grados de libertad, que deja una área de 0.95 a la derecha, es

$$f_{0.95}(6, 10) = \frac{1}{f_{0.05}(10, 6)} = \frac{1}{4.06} = 0.246.$$

La distribución *F* con dos varianzas muestrales

Suponga que las muestras aleatorias de tamaños n_1 y n_2 se seleccionan de dos poblaciones normales con varianzas σ_1^2 y σ_2^2, respectivamente. Del teorema 8.4, sabemos que

$$\chi_1^2 = \frac{(n_1 - 1)S_1^2}{\sigma_1^2} \text{ y } \chi_2^2 = \frac{(n_2 - 1)S_2^2}{\sigma_2^2}$$

son variables aleatorias que tienen distribuciones chi cuadrada con $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$ grados de libertad. Además, como las muestras se seleccionan al azar, tratamos con variables aleatorias independientes. Entonces, usando el teorema 8.6 con $\chi_1^2 = U$ y $\chi_2^2 = V$, obtenemos el siguiente resultado.

Teorema 8.8: Si S_1^2 y S_2^2 son las varianzas de muestras aleatorias independientes de tamaño n_1 y n_2 tomadas de poblaciones normales con varianzas σ_1^2 y σ_2^2, respectivamente, entonces,

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$$

tiene una distribución F con $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$ grados de libertad.

¿Para qué se utiliza la distribución F?

Al inicio de esta sección contestamos esta pregunta parcialmente. La distribución F se usa en situaciones de dos muestras para hacer inferencias acerca de las varianzas de población, lo cual implica aplicar el teorema 8.8. Sin embargo, la distribución F también se puede aplicar a muchos otros tipos de problemas que involucren varianzas muestrales. De hecho, la distribución F se llama *distribución de razón de varianzas*. Como ejemplo, considere el estudio de caso 8.2 en el que se compararon las dos pinturas, A y B, en relación con el tiempo medio que tardan en secar, en donde la distribución normal se aplica muy bien (suponiendo que se conocen σ_A y σ_B). Sin embargo, suponga que necesitamos comparar tres tipos de pinturas, digamos A, B y C, y que queremos determinar si las medias de población son equivalentes. Suponga que un resumen de la información importante del experimento es el siguiente:

Pintura	Media muestral	Varianza muestral	Tamaño muestral
A	$\bar{X}_A = 4.5$	$s_A^2 = 0.20$	10
B	$\bar{X}_B = 5.5$	$s_B^2 = 0.14$	10
C	$\bar{X}_C = 6.5$	$s_C^2 = 0.11$	10

El problema se centra alrededor de si los promedios muestrales ($\bar{x}_A$, $\bar{x}_B$, $\bar{x}_C$) están o no suficientemente alejados. La implicación de "suficientemente alejados" resulta muy importante. Parecería razonable que si la variabilidad entre los promedios muestrales es mayor que lo que se esperaría por casualidad, los datos no apoyan la conclusión de que $\mu_A = \mu_B = \mu_C$. Si estos promedios muestrales pudieran ocurrir por casualidad depende de la *variabilidad dentro de las muestras*, cuando se cuantifican por medio de s_A^2, s_B^2 y s_C^2. La idea de los componentes importantes de la variabilidad se observa mejor utilizando algunas gráficas sencillas. Considere la gráfica de los datos brutos de las muestras A, B y C que se presenta en la figura 8.13. Estos datos podrían generar con facilidad la información antes resumida.

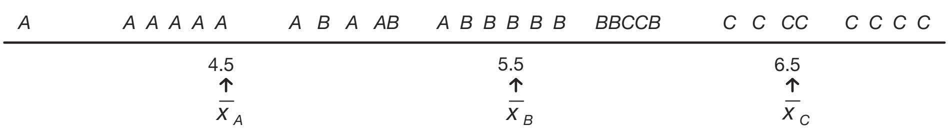

Figura 8.13: Datos de tres muestras diferentes.

Parece evidente que los datos provienen de distribuciones con diferentes medias de población, aunque hay cierto traslape entre las muestras. Un análisis que incluya todos los datos intentaría determinar si la variabilidad entre los promedios muestrales *y* la variabilidad dentro de las muestras podría haber ocurrido conjuntamente *si, de hecho, las poblaciones tienen una media común*. Observe que la clave para este análisis se centra alrededor de las dos siguientes fuentes de variabilidad.

1. Variabilidad dentro de las muestras (entre observaciones en muestras distintas).
2. Variabilidad entre muestras (entre promedios muestrales).

Es evidente que si la variabilidad en 1) es considerablemente mayor que en 2), entonces habrá un traslape considerable en los datos muestrales, una señal de que los datos podrían provenir de una distribución común. En el conjunto de datos que se presenta en la figura 8.14 se encuentra un ejemplo. Por otro lado, es muy improbable que los datos de una distribución con una media común puedan tener una variabilidad entre promedios muestrales que sea considerablemente mayor que la variabilidad dentro de las muestras.

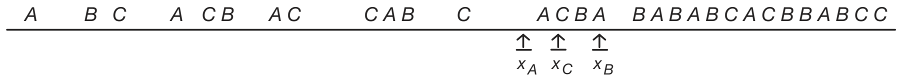

Figura 8.14: Datos que con facilidad podrían provenir de la misma población.

Las fuentes de variabilidad en 1) y 2) generan importantes cocientes de *varianzas muestrales* y los cocientes se utilizan junto con la distribución F. El procedimiento general implicado se llama **análisis de varianza**. Es interesante que en el ejemplo de la pintura aquí descrito tratamos con inferencias sobre tres medias de población pero utilizamos dos fuentes de variabilidad. No proporcionaremos detalles aquí, pero en los capítulos 13, 14 y 15 utilizaremos ampliamente el análisis de varianza en donde, por supuesto, la distribución F desempeña un papel importante.

8.8 Gráficas de cuantiles y de probabilidad

En el capítulo 1 presentamos al lector las distribuciones empíricas. El objetivo es utilizar presentaciones creativas para extraer información acerca de las propiedades de un conjunto de datos. Por ejemplo, los diagramas de tallo y hojas brindan al observador una imagen de la simetría y de otras propiedades de los datos. En este capítulo tratamos con muestras que, por supuesto, son conjuntos de datos experimentales de los que sacamos conclusiones sobre las poblaciones. A menudo, la apariencia de la muestra proporciona información sobre la distribución de la que se tomaron los datos. Por ejemplo, en el capítulo 1 ilustramos la naturaleza general de pares de muestras con gráficas de puntos que presentan una comparación relativa entre la tendencia central y la variabilidad de dos muestras.

En los capítulos siguientes con frecuencia supondremos que una distribución es normal. La información gráfica respecto a la validez de esta suposición se puede obtener a partir de presentaciones como los diagramas de tallo y hojas y los histogramas de frecuencias. Además, en esta sección presentaremos los conceptos de *gráficas de probabilidad normal* y *gráficas de cuantiles*. Estas gráficas se utilizan en estudios con diversos grados de complejidad con el principal objetivo de que las gráficas proporcionen una verificación diagnóstica sobre la suposición de que los datos provienen de una distribución normal.

Podemos caracterizar el análisis estadístico como el proceso de sacar conclusiones acerca de los sistemas en presencia de la variabilidad del sistema. Por ejemplo, el intento de un ingeniero por aprender acerca de un proceso químico a menudo es obstaculizado por la *variabilidad del proceso*. Un estudio que implica el número de artículos defectuosos en un proceso de producción con frecuencia se dificulta por la variabilidad en el método con el que se fabrican. En las secciones anteriores aprendimos acerca de las muestras y los estadísticos que expresan el centro de localización y la variabilidad en la muestra. Tales estadísticos ofrecen medidas simples, en tanto que una presentación gráfica brinda información adicional por medio de una imagen.

Un tipo de gráfica que puede ser especialmente útil para revelar la naturaleza de un conjunto de datos es la *gráfica de cuantiles*. Igual que en el caso de la gráfica de caja y extensión (véase la sección 1.6), en el que el objetivo del analista es hacer distinciones, en la gráfica de cuantiles se pueden utilizar las ideas básicas para *comparar muestras de* *datos*. En los siguientes capítulos se presentarán más ejemplos del uso de las gráficas de cuantiles, en los que se analizará la inferencia estadística formal asociada con la comparación de muestras. En su momento, los estudios de caso mostrarán al lector tanto la inferencia formal como las gráficas diagnósticas para el mismo conjunto de datos.

Gráfica de cuantiles

El propósito de las gráficas de cuantiles consiste en describir, en forma de muestra, la función de distribución acumulada que se estudió en el capítulo 3.

Definición 8.6: Un **cuantil** de una muestra, $q(f)$, es un valor para el que una fracción específica f de los valores de los datos es menor que o igual a $q(f)$.

Evidentemente, un cuantil representa una estimación de una característica de una población o, más bien, la distribución teórica. La mediana de la muestra es $q(0.5)$. El percentil 75 (cuartil superior) es $q(0.75)$ y el cuartil inferior es $q(0.25)$.

Una **gráfica de cuantiles** simplemente *grafica los valores de los datos en el eje vertical contra una evaluación empírica de la fracción de observaciones excedidas por los valores de los datos*. Para propósitos teóricos esta fracción se calcula con

$$f_i = \frac{i - \frac{3}{8}}{n + \frac{1}{4}},$$

donde i es el orden de las observaciones cuando se ordenan de la menor a la mayor. En otras palabras, si denotamos las observaciones ordenadas como

$$y_{(1)} \le y_{(2)} \le y_{(3)} \le \cdots \le y_{(n-1)} \le y_{(n)},$$

entonces la gráfica de cuantiles describe una gráfica de $y_{(i)}$ contra f_i. En la figura 8.15 se presenta la gráfica de cuantiles para las asas de las latas de pintura analizadas con anterioridad.

A diferencia de la gráfica de caja y extensión, la gráfica de cuantiles realmente muestra todas las observaciones. Todos los cuantiles, incluidos la mediana y los cuantiles superior e inferior, se pueden aproximar de forma visual. Por ejemplo, observamos fácilmente una mediana de 35 y un cuartil superior de alrededor de 36. Las agrupaciones relativamente grandes en torno a valores específicos se indican por pendientes cercanas a cero; mientras que los datos escasos en ciertas áreas producen pendientes más abruptas. La figura 8.15 describe la dispersión de datos de los valores 28 a 30, pero una densidad relativamente alta de 36 a 38. En los capítulos 9 y 10 proseguimos con las gráficas de cuantiles mediante la ilustración de formas útiles en que es posible comparar distintas muestras.

Debería ser muy evidente para el lector que detectar si un conjunto de datos proviene o no de una distribución normal puede ser una herramienta importante para el analista de datos. Como antes indicamos en esta sección, a menudo suponemos que la totalidad o subconjuntos de las observaciones en un conjunto de datos son realizaciones de variables aleatorias normales independientes idénticamente distribuidas. Una vez más, la gráfica de diagnóstico a menudo se agrega a (con fines de presentación) una *prueba de bondad del ajuste* formal de los datos. Las pruebas de bondad del ajuste se estudiarán en el capítulo 10. Los lectores de un artículo o informe científico suelen considerar la información de diagnóstico mucho más clara, menos árida y quizá menos aburrida que un análisis formal.

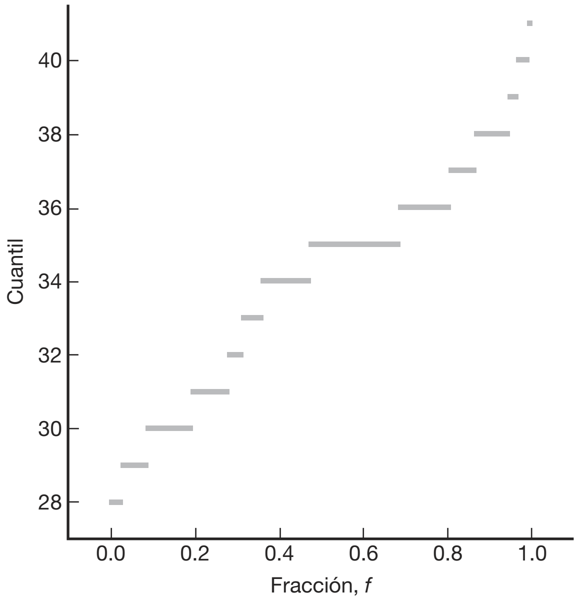

Figura 8.15: Gráfica de cuantiles para los datos de la pintura.

En los capítulos siguientes (del 9 al 13) nos enfocaremos nuevamente en los métodos de detección de desviaciones de la normalidad como un agregado de la inferencia estadística formal. Las gráficas de cuantiles son útiles para detectar los tipos de distribución. En la elaboración de modelos y en el diseño de experimentos también hay situaciones en que se utilizan las gráficas para detectar **términos** o **efectos del modelo** que están activos. En otras situaciones se utilizan para determinar si las suposiciones subyacentes que el científico o el ingeniero hicieron en la construcción del modelo son o no razonables. En los capítulos 11, 12 y 13 se incluyen muchos ejemplos con ilustraciones. La siguiente subsección brinda un análisis y un ejemplo de una gráfica de diagnóstico denominada *gráfica de cuantiles-cuantiles normales*.

Gráfica de cuantiles-cuantiles normales

La gráfica de cuantiles-cuantiles normales aprovecha lo que se conoce sobre los cuantiles de la distribución normal. La metodología incluye una gráfica de los cuantiles empíricos recién analizados, contra el cuantil correspondiente de la distribución normal. Ahora, la expresión para un cuantil de una variable aleatoria $N(\mu, \sigma)$ es muy complicada. Sin embargo, una buena aproximación es dada por

$$q_{\mu,\sigma}(f) = \mu + \sigma\{4.91[f^{0.14} - (1-f)^{0.14}]\}.$$

La expresión entre las llaves (el múltiplo de σ) es la aproximación para el cuantil correspondiente para la variable aleatoria $N(0, 1)$, es decir,

$$q_{0,1}(f) = 4.91[f^{0.14} - (1-f)^{0.14}].$$

Definición 8.7: La **gráfica de cuantiles-cuantiles normales** es una gráfica de $y_{(i)}$ (observaciones ordenadas) contra $q_{0,1}(f_i)$, donde $f_i = \frac{i - \frac{3}{8}}{n + \frac{1}{4}}$.

Una relación cercana a una línea recta sugiere que los datos provienen de una distribución normal. La intersección en el eje vertical es una estimación de la media de la población μ y la pendiente es una estimación de la desviación estándar σ. La figura 8.16 presenta una gráfica de cuantiles-cuantiles normales para los datos de las latas de pintura.

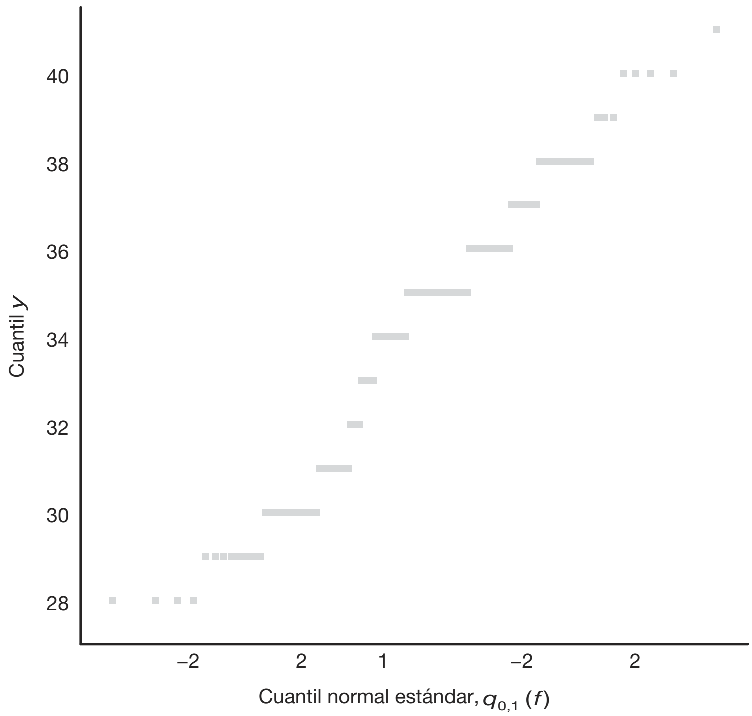

Figura 8.16: Gráfica de cuantiles-cuantiles normales para los datos de la pintura.

Graficación de la probabilidad normal

Observe cómo la desviación de la normalidad se vuelve evidente gracias a la apariencia de la gráfica. La asimetría que exhiben los datos produce cambios en la pendiente.

Las ideas para graficar la probabilidad se manifiestan en versiones diferentes de la gráfica de cuantiles-cuantiles normales que se presentó aquí. Por ejemplo, se ha puesto mucha atención a la llamada **gráfica de probabilidad normal**, en la que f se grafica contra los valores de los datos ordenados en un papel especial y la escala utilizada da como resultado una línea recta. Además, una gráfica alternativa utiliza los valores esperados de las observaciones clasificadas para la distribución normal y dibuja las observaciones clasificadas contra su valor esperado, bajo el supuesto de datos de $N(\mu, \sigma)$. Una vez más, la línea recta es el criterio gráfico que se emplea. Continuamos sugiriendo que basarse en los métodos analíticos gráficos que se describen en esta sección ayudará a comprender los métodos formales que permiten distinguir muestras diferentes de datos.

Ejemplo 8.12: Considere los datos del ejercicio 10.41 en la página 358 del capítulo 10. En el estudio "Retención de nutrientes y respuesta de comunidades de macroinvertebrados ante la presión de aguas residuales en un ecosistema fluvial", que se llevó a cabo en el departamento de zoología del Virginia Polytechnic Institute y la universidad estatal, se recabaron datos sobre mediciones de densidad (número de organismos por metro cuadrado) en dos diferentes estaciones colectoras. En el capítulo 10 se dan detalles con respecto a los métodos analíticos de comparación de muestras para determinar si ambas provienen de la misma distribución $N(\mu, \sigma)$. Los datos se presentan en la tabla 8.1.

Tabla 8.1: Datos para el ejemplo 8.12

Número de organismos por metro cuadrado			
Estación 1		**Estación 2**	
5,030	4,980	2,800	2,810
13,700	11,910	4,670	1,330
10,730	8,130	6,890	3,320
11,400	26,850	7,720	1,230
860	17,660	7,030	2,130
2,200	22,800	7,330	2,190
4,250	1,130		
15,040	1,690		

Dibuje una gráfica de cuantiles-cuantiles normales y saque conclusiones con respecto a si es razonable o no suponer que las dos muestras provienen de la misma distribución $n(x; \mu, \sigma)$.

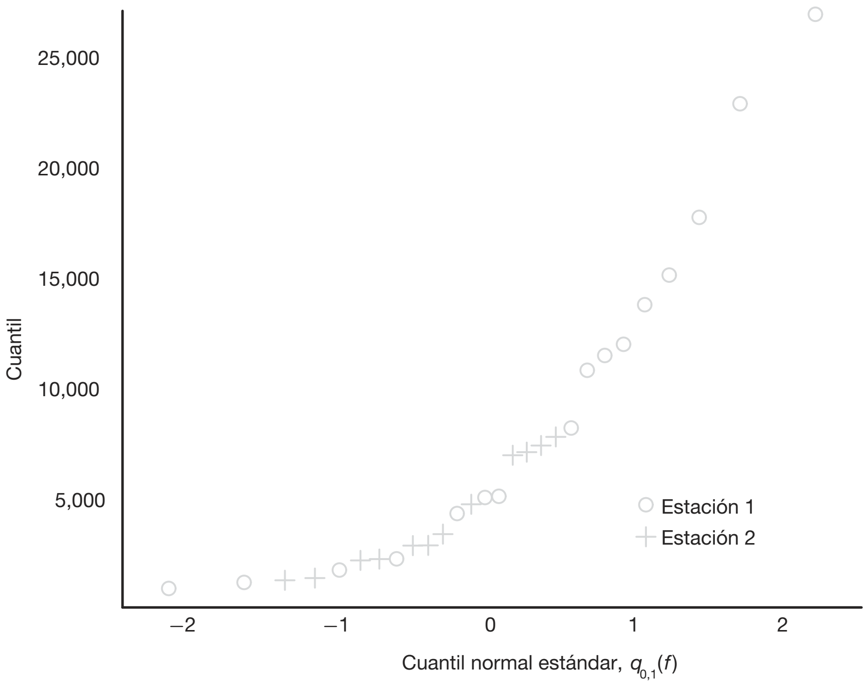

Figura 8.17: Gráfica de cuantiles-cuantiles normales para los datos de densidad del ejemplo 8.12.

***Solución*:** La figura 8.17 muestra la gráfica de cuantiles-cuantiles normales para las mediciones de densidad. La gráfica se aleja mucho de una sola línea recta. De hecho, los datos de la estación 1 reflejan pocos valores en la cola inferior de la distribución y varios en la cola superior. El "agrupamiento" de observaciones hace que parezca improbable que las dos muestras provengan de una distribución común $N(\mu, \sigma)$. ■

Aunque hemos concentrado nuestra explicación y ejemplo en las gráficas de probabilidad para distribuciones normales, podemos enfocarnos en cualquier distribución. Tan sólo necesitaríamos calcular cantidades de forma analítica para la distribución teórica en cuestión.

Ejercicios

8.37 Para una distribución chi cuadrada calcule

a) $\chi^2_{0.025}$ cuando $v = 15$;

b) $\chi^2_{0.01}$ cuando $v = 7$;

c) $\chi^2_{0.05}$ cuando $v = 24$.

8.38 Para una distribución chi cuadrada, calcule

a) $\chi^2_{0.005}$ cuando $v = 5$;

b) $\chi^2_{0.05}$ cuando $v = 19$;

c) $\chi^2_{0.01}$ cuando $v = 12$.

8.39 Para una distribución chi cuadrada calcule χ^2_α, tal que

a) $P(X^2 > \chi^2_\alpha) = 0.99$ cuando $v = 4$;

b) $P(X^2 > \chi^2_\alpha) = 0.025$ cuando $v = 19$;

c) $P(37.652 < X^2 < \chi^2_\alpha) = 0.045$ cuando $v = 25$.

8.40 Para una distribución chi cuadrada calcule χ^2_α, tal que

a) $P(X^2 > \chi^2_\alpha) = 0.01$ cuando $v = 21$;

b) $P(X^2 < \chi^2_\alpha) = 0.95$ cuando $v = 6$;

c) $P(\chi^2_\alpha < X^2 < 23.209) = 0.015$ cuando $v = 10$.

8.41 Suponga que las varianzas muestrales son mediciones continuas. Calcule la probabilidad de que una muestra aleatoria de 25 observaciones, de una población normal con varianza $\sigma^2 = 6$, tenga una varianza muestral S^2

a) mayor que 9.1;

b) entre 3.462 y 10.745.

8.42 Las calificaciones de un examen de colocación que se aplicó a estudiantes de primer año de una universidad durante los últimos cinco años tienen una distribución aproximadamente normal con una media $\mu = 74$ y una varianza $\sigma^2 = 8$. ¿Seguiría considerando que $\sigma^2 = 8$ es un valor válido de la varianza si una muestra aleatoria de 20 estudiantes, a los que se les aplica el examen de colocación este año, obtienen un valor de $s^2 = 20$?

8.43 Demuestre que la varianza de S^2 para muestras aleatorias de tamaño n de una población normal disminuye a medida que aumenta n. [*Sugerencia*: primero calcule la varianza de $(n - 1)S^2/\sigma^2$].

8.44 *a*) Calcule $t_{0.025}$ cuando $v = 14$.

b) Calcule $-t_{0.10}$ cuando $v = 10$.

c) Calcule $t_{0.995}$ cuando $v = 7$.

8.45 *a*) Calcule $P(T < 2.365)$ cuando $v = 7$.

b) Calcule $P(T > 1.318)$ cuando $v = 24$.

c) Calcule $P(-1.356 < T < 2.179)$ cuando $v = 12$.

d) Calcule $P(T > -2.567)$ cuando $v = 17$.

8.46 *a*) Calcule $P(-t_{0.005} < T < t_{0.01})$ para $v = 20$.

b) Calcule $P(T > -t_{0.025})$.

8.47 Dada una muestra aleatoria de tamaño 24 de una distribución normal, calcule k tal que

a) $P(-2.069 < T < k) = 0.965$;

b) $P(k < T < 2.807) = 0.095$;

c) $P(-k < T < k) = 0.90$.

8.48 Una empresa que fabrica juguetes electrónicos afirma que las baterías que utiliza en sus productos duran un promedio de 30 horas. Para mantener este promedio se prueban 16 baterías cada mes. Si el valor t calculado cae entre $-t_{0.025}$ y $t_{0.025}$, la empresa queda satisfecha con su afirmación. ¿Qué conclusiones debería sacar la empresa a partir de una muestra que tiene una media de $\bar{x} = 27.5$ horas y una desviación estándar de $s = 5$ horas? Suponga que la distribución de las duraciones de las baterías es aproximadamente normal.

8.49 Una población normal con varianza desconocida tiene una media de 20. ¿Es posible obtener una muestra aleatoria de tamaño 9 de esta población con una media de 24 y una desviación estándar de 4.1? Si no fuera posible, ¿a qué conclusión llegaría?

8.50 Un fabricante de cierta marca de barras de cereal con bajo contenido de grasa afirma que el contenido promedio de grasa saturada en éstas es de 0.5 gramos. En una muestra aleatoria de 8 barras de cereal de esta marca se encontró que su contenido de grasa saturada era de 0.6, 0.7, 0.7, 0.3, 0.4, 0.5, 0.4 y 0.2. ¿Estaría de acuerdo con tal afirmación? Suponga una distribución normal.

8.51 Para una distribución F calcule:

a) $f_{0.05}$ con $v_1 = 7$ y $v_2 = 15$;
b) $f_{0.05}$ con $v_1 = 15$ y $v_2 = 7$;
c) $f_{0.01}$ con $v_1 = 24$ y $v_2 = 19$;
d) $f_{0.95}$ con $v_1 = 19$ y $v_2 = 24$;
e) $f_{0.99}$ con $v_1 = 28$ y $v_2 = 12$.

8.52 Se aplican pruebas a 10 cables conductores soldados a un dispositivo semiconductor con el fin de determinar su resistencia a la tracción. Las pruebas demostraron que para romper la unión se requieren las libras de fuerza que se listan a continuación:

19.8	12.7	13.2	16.9	10.6
18.8	11.1	14.3	17.0	12.5

Otro conjunto de 8 cables conductores que forman un dispositivo se encapsuló y se probó para determinar si el encapsulado aumentaba la resistencia a la tracción. Las pruebas dieron los siguientes resultados:

24.9 22.8 23.6 22.1 20.4 21.6 21.8 22.5

Comente acerca de la evidencia disponible respecto a la igualdad de las dos varianzas de población.

8.53 Considere las siguientes mediciones de la capacidad de producción de calor del carbón producido por dos minas (en millones de calorías por tonelada):

Mina 1:	8260	8130	8350	8070	8340	
Mina 2:	7950	7890	7900	8140	7920	7840

¿Se puede concluir que las dos varianzas de población son iguales?

8.54 Dibuje una gráfica de cuantiles con los siguientes datos, que representan la vida, en horas, de cincuenta lámparas incandescentes esmeriladas de 40 watts y 110 voltios, tomados de pruebas de vida forzadas:

919	1196	785	1126	936	918
1156	920	948	1067	1092	1162
1170	929	950	905	972	1035
1045	855	1195	1195	1340	1122
938	970	1237	956	1102	1157
978	832	1009	1157	1151	1009
765	958	902	1022	1333	811
1217	1085	896	958	1311	1037
702	923				

8.55 Dibuje una gráfica de cuantiles-cuantiles normales con los siguientes datos, que representan los diámetros de 36 cabezas de remache en 1/100 de una pulgada:

6.72	6.77	6.82	6.70	6.78	6.70	6.62
6.75	6.66	6.66	6.64	6.76	6.73	6.80
6.72	6.76	6.76	6.68	6.66	6.62	6.72
6.76	6.70	6.78	6.76	6.67	6.70	6.72
6.74	6.81	6.79	6.78	6.66	6.76	6.76
6.72						

Ejercicios de repaso

8.56 Considere los datos que se presentan en el ejercicio 1.20 de la página 31. Dibuje una gráfica de caja y extensión, y comente acerca de la naturaleza de la muestra. Calcule la media muestral y la desviación estándar de la muestra.

8.57 Si $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes que tienen distribuciones exponenciales idénticas con parámetro θ, demuestre que la función de densidad de la variable aleatoria $Y = X_1 + X_2 + \cdots + X_n$ es la de una distribución gamma con parámetros $\alpha = n$ y $\beta = \theta$.

8.58 Al probar el monóxido de carbono que contiene cierta marca de cigarrillos, los datos que se obtuvieron, en miligramos por cigarrillo, se codificaron restando 12 a cada observación. Utilice los resultados del ejercicio 8.14 de la página 231 para calcular la desviación estándar del contenido de monóxido de carbono de una muestra aleatoria de 15 cigarrillos de esta marca, si las mediciones codificadas son 3.8, −0.9, 5.4, 4.5, 5.2, 5.6, −0.1, −0.3, −1.7, 5.7, 3.3, 4.4, −0.5 y 1.9.

8.59 Si S_1^2 y S_2^2 representan las varianzas de muestras aleatorias independientes de tamaños $n_1 = 8$ y $n_2 = 12$, tomadas de poblaciones normales con varianzas iguales, calcule $P(S_1^2 / S_2^2 < 4.89)$.

8.60 Una muestra aleatoria de 5 presidentes de bancos indicó sueldos anuales de \$395,000, \$521,000, \$483,000, \$479,000 y \$510,000. Calcule la varianza de este conjunto.

8.61 Si el número de huracanes que azotan cierta área del este de Estados Unidos cada año es una variable aleatoria que tiene una distribución de Poisson con $\mu = 6$, calcule la probabilidad de que esta área sea azotada por

a) exactamente 15 huracanes en 2 años;
b) a lo sumo 9 huracanes en 2 años.

8.62 Una empresa de taxis prueba una muestra aleatoria de 10 neumáticos radiales con bandas tensoras de acero de cierta marca y registra los siguientes desgastes de la banda: 48,000, 53,000, 45,000, 61,000, 59,000, 56,000, 63,000, 49,000, 53,000 y 54,000 kilómetros.

Utilice los resultados del ejercicio 8.14 de la página 231 para calcular la desviación estándar de este conjunto de datos dividiendo primero cada observación entre 1000 y después restando 55 al resultado.

8.63 Considere los datos del ejercicio 1.19 de la página 31. Dibuje una gráfica de caja y extensión. Comente y calcule la media muestral y la desviación estándar muestral.

8.64 Si S_1^2 y S_2^2 representan las varianzas de muestras aleatorias independientes de tamaños $n_1 = 25$ y $n_2 = 31$, tomadas de poblaciones normales con varianzas $\sigma_1^2 = 10$ y $\sigma_2^2 = 15$, respectivamente, calcule

$$P(S_1^2/S_2^2 > 1.26).$$

8.65 Considere el ejemplo 1.5 de la página 25. Comente acerca de cualquier valor extremo.

8.66 Considere el ejercicio de repaso 8.56. Comente acerca de cualquier valor extremo en los datos.

8.67 La resistencia a la rotura X de cierto remache que se utiliza en el motor de una máquina tiene una media de 5000 psi y una desviación estándar de 400 psi. Se toma una muestra aleatoria de 36 remaches. Considere la distribución de $\bar{X}$, la media muestral de la resistencia a la rotura.

a) ¿Cuál es la probabilidad de que la media de la muestra caiga entre 4800 psi y 5200 psi?

b) ¿Qué muestra n sería necesaria para tener

$$P(4900 < \bar{X} < 5100) = 0.99?$$

8.68 Considere la situación del ejercicio de repaso 8.62. Si la población de la cual se tomó la muestra tiene una media poblacional $\mu = 53{,}000$ kilómetros, ¿esta información de la muestra parece apoyar esa afirmación? En su respuesta calcule

$$t = \frac{\bar{x} - 53{,}000}{s/\sqrt{10}}$$

y determine, consultando la tabla A.4 (con 9 g.l.), si el valor t calculado es razonable o si parece ser un suceso raro.

8.69 Se consideran dos propulsores de combustible sólido distintos, el tipo A y el tipo B, para una actividad del programa espacial. Las velocidades de combustión en el propulsor son fundamentales. Se toman muestras aleatorias de 20 especímenes de los dos propulsores con medias muestrales de 20.5 cm/s para el propulsor A y de 24.50 cm/s para el propulsor B. Por lo general se supone que la variabilidad en la velocidad de combustión es casi igual para los dos propulsores y que es determinada por una desviación estándar de población de 5 cm/s. Suponga que la velocidad de combustión para cada propulsor es aproximadamente normal, por lo cual se debería utilizar el teorema del límite central. Nada se sabe acerca de las medias poblacionales de las dos velocidades de combustión y se espera que este experimento revele algo sobre ellas.

a) Si, de hecho, $\mu_A = \mu_B$, ¿cuál será $P(\bar{X}_B - \bar{X}_A \geq 4.0)$?

b) Utilice lo que respondió en el inciso *a*) para dar luz sobre la validez de la proposición $\mu_A = \mu_B$.

8.70 La concentración de un ingrediente activo en el producto de una reacción química es fuertemente influido por el catalizador que se usa en la reacción. Se considera que cuando se utiliza el catalizador A la concentración media de la población excede el 65%. Se sabe que la desviación estándar es $\sigma = 5\%$. Una muestra de productos tomada de 30 experimentos independientes proporciona la concentración promedio de $\bar{x}_A = 64.5\%$.

a) ¿Esta información muestral, con una concentración promedio de $\bar{x}_A = 64.5\%$, ofrece información inquietante de que quizá μ_A no sea el 65% sino menos que ese porcentaje? Respalde su respuesta con una aseveración de probabilidad.

b) Suponga que se realiza un experimento similar utilizando otro catalizador, el B. Se supone que la desviación estándar σ sigue siendo 5% y $\bar{x}_B$ resulta ser 70%. Comente si la información muestral del catalizador B sugiere con certeza que μ_B es en realidad mayor que μ_A. Respalde su respuesta calculando

$$P(\bar{X}_B - \bar{X}_A \geq 5.5 \mid \mu_B = \mu_A).$$

c) En el caso de que $\mu_A = \mu_B = 65\%$, determine la distribución aproximada de las siguientes cantidades (con la media y la varianza de cada una). Utilice el teorema del límite central.

i) $\bar{X}_B$;
ii) $\bar{X}_A - \bar{X}_B$;
iii) $\frac{\bar{X}_A - \bar{X}_B}{\sigma\sqrt{2/30}}$.

8.71 Con la información del ejercicio de repaso 8.70 calcule (suponiendo $\mu_B = 65\%$) $P(\bar{X}_B \geq 70)$.

8.72 Dada una variable aleatoria normal X con media 20 y varianza 9, y una muestra aleatoria de tamaño n tomada de la distribución, ¿qué tamaño de la muestra n se necesita para que

$$P(19.9 \leq \bar{X} \leq 20.1) = 0.95?$$

8.73 En el capítulo 9 se estudiará con detenimiento el concepto de **estimación de parámetros**. Suponga que X es una variable aleatoria con media μ y varianza $\sigma^2 = 1.0$. Además, suponga que se toma una muestra aleatoria de tamaño n y que $\bar{x}$ se utiliza como un *estimado* de μ. Cuando se toman los datos y se mide la media de la muestra, deseamos que ésta esté dentro de 0.05 unidades de la media real con una probabilidad de 0.99. Es decir, aquí queremos que haya muchas posibilidades de que la $\bar{x}$ calculada de la muestra esté "muy cerca de" la media de población (¡dondequiera que ésta se encuentre!), de manera que deseamos

$$P(|\bar{X} - \mu| > 0.05) = 0.99.$$

¿Qué tamaño de muestra se requiere?

8.74 Suponga que se utiliza una máquina para llenar envases de cartón con un líquido. La especificación que es estrictamente indispensable para el llenado de la máquina es 9 ± 1.5 onzas. El proveedor considera que cualquier envase de cartón que no cumpla con tales límites de peso en el llenado está defectuoso. Se espera que al menos 99% de los envases de cartón cumplan con la especificación. En el caso de que $\mu = 9$ y $\sigma = 1$, ¿qué proporción de envases de cartón del proceso están defectuosos? Si se hacen cambios para reducir la variabilidad, ¿cuánto se tiene que reducir σ para que haya 0.99 de probabilidades de cumplir con la especificación? Suponga una distribución normal para el peso.

8.75 Considere la situación del ejercicio de repaso 8.74. Suponga que se hace un gran esfuerzo para "estrechar" la variabilidad del sistema. Después de eso se toma una muestra aleatoria de tamaño 40 de la nueva línea de ensamble y se obtiene que la varianza de la muestra es $s^2 = 0.188$ onzas^2. ¿Tenemos evidencia numérica sólida de que σ^2 se redujo a menos de 1.0? Considere la probabilidad

$$P(S^2 \leq 0.188 \mid \sigma^2 = 1.0),$$

y dé una conclusión.

8.76 **Proyecto de grupo:** Divida al grupo en equipos de cuatro estudiantes. Cada equipo deberá ir al gimnasio de la universidad o a un gimnasio local y preguntar a cada persona que cruce el umbral cuánto mide en pulgadas. Después, cada equipo dividirá los datos de las estaturas por género y trabajará en conjunto para realizar las actividades que se indican a continuación.

a) Dibujen una gráfica de cuantiles-cuantiles normal con los datos. Si usan la gráfica como base, ¿les parecería que los datos tienen una distribución normal?

b) Utilicen la varianza muestral como un estimado de la varianza real para cada género. Supongan que la estatura media de la población de los hombres es realmente tres pulgadas más grande que la de las mujeres. ¿Cuál es la probabilidad de que la estatura promedio de los hombres sea 4 pulgadas más grande que la de las mujeres en su muestra?

c) ¿Qué factores podrían provocar que estos resultados sean engañosos?

8.9 Posibles riesgos y errores conceptuales. Relación con el material de otros capítulos

El teorema del límite central es una de las más poderosas herramientas de la estadística, y aunque este capítulo es relativamente breve, contiene gran cantidad de información fundamental acerca de las herramientas que se utilizarán en el resto del libro.

El concepto de distribución muestral es una de las ideas fundamentales más importantes de la estadística y, en este momento de su entrenamiento, el estudiante debería entenderlo con claridad antes de continuar con los siguientes capítulos, en los cuales se continuarán utilizando ampliamente las distribuciones muestrales. Suponga que se quiere utilizar el estadístico $\bar{X}$ para hacer inferencias acerca de la media de la población μ, lo cual se hace utilizando el valor observado $\bar{x}$ de una sola muestra de tamaño n. Luego, cualquier inferencia deberá hacerse tomando en cuenta no sólo el valor único, sino también la estructura teórica o la **distribución de todos los valores $\bar{x}$ que se podrían observar a partir de las muestras de tamaño n**. Como resultado de lo anterior surge el concepto de *distribución muestral*, que es la base del teorema del límite central. Las distribuciones t, χ^2 y F también se utilizan en el contexto de las distribuciones muestrales. Por ejemplo, la distribución t, que se ilustra en la figura 8.8, representa la estructura que ocurre si se forman todos los valores de $\frac{\bar{x}-\mu}{s/\sqrt{n}}$, donde $\bar{x}$ y s se toman de las muestras de tamaño n de una distribución $n(x; \mu, \sigma)$. Se pueden hacer comentarios similares en relación con χ^2 y F, y el lector no debería olvidar que la información muestral que conforma los estadísticos para todas estas distribuciones es la normal. Por lo tanto, se podría afirmar que **donde haya una t, F o χ^2 la fuente era una muestra de una distribución normal**.

Podría parecer que las tres distribuciones antes descritas se presentaron de una forma bastante aislada, sin indicar a qué se refieren. Sin embargo, aparecerán en la resolución de problemas prácticos a lo largo del texto.

Ahora bien, hay tres cuestiones que se deben tener presentes para evitar que haya confusión respecto a estas distribuciones muestrales fundamentales:

i) No se puede usar el teorema del límite central a menos que se conozca σ. Para usar el teorema del límite central cuando no se conoce σ se debe reemplazar con s, la desviación estándar de la muestra.

ii) El estadístico T **no** es un resultado del teorema del límite central y $x_1, x_2, \ldots, x_n$ deben provenir de una distribución $n(x; \mu, \sigma)$ para que $\frac{\bar{x} - \mu}{s/\sqrt{n}}$ sea una distribución t; por supuesto, s es tan sólo una estimación de σ.

iii) Aunque el concepto de **grados de libertad** es nuevo en este punto, debería ser muy intuitivo, ya que es razonable que la naturaleza de la distribución de S y también t deban depender de la cantidad de información en la muestra $x_1, x_2, \ldots, x_n$.

Capítulo 9

Problemas de estimación de una y dos muestras

9.1 Introducción

En los capítulos anteriores destacamos las propiedades del muestreo de la media y de la varianza muestrales. También destacamos las representaciones de datos en varias formas. El propósito de estas presentaciones es establecer las bases que permitan a los estadísticos sacar conclusiones acerca de los parámetros de poblaciones tomadas de datos experimentales. Por ejemplo, el teorema del límite central brinda información sobre la distribución de la media muestral $\bar{X}$. La distribución incluye la media de la población μ. Por consiguiente, cualesquiera conclusiones respecto a μ, extraídas de un promedio muestral observado, deben depender de lo que se sabe acerca de su distribución muestral. Se podría decir algo similar en lo que se refiere a S^2 y σ^2. Como es evidente, es muy probable que cualquier conclusión que saquemos acerca de la varianza de una distribución normal implique la distribución muestral de S^2.

En este capítulo comenzaremos por presentar de manera formal el propósito de la inferencia estadística. Continuaremos con el análisis del problema de la **estimación de los parámetros de la población**. Restringiremos nuestros desarrollos formales de los procedimientos de estimación específicos a problemas que impliquen una y dos muestras.

9.2 Inferencia estadística

En el capítulo 1 presentamos la filosofía general de la inferencia estadística formal. La **inferencia estadística** consta de los métodos mediante los cuales se hacen inferencias o generalizaciones acerca de una población. La tendencia actual es distinguir entre el **método clásico** de estimación de un parámetro de la población, donde las inferencias se basan estrictamente en información obtenida de una muestra aleatoria seleccionada de la población, y el **método bayesiano**, el cual utiliza el conocimiento subjetivo que ya se posee sobre la distribución de probabilidad de los parámetros desconocidos junto con la información que proporcionan los datos de la muestra. En la mayor parte de este capítulo utilizaremos los métodos clásicos para estimar los parámetros de la población desconocidos, como la media, la proporción y la varianza, mediante el cálculo de estadísticos de muestras aleatorias y la aplicación de la teoría de las distribuciones muestrales, gran parte de lo cual se estudió en el capítulo 8. La estimación bayesiana se analizará en el capítulo 18.

La inferencia estadística se puede dividir en dos áreas principales: **estimación** y **pruebas de hipótesis**. Trataremos estas dos áreas por separado: en este capítulo veremos la teoría y las aplicaciones de la estimación, y en el capítulo 10 revisaremos la prueba de hipótesis. Para distinguir claramente un área de la otra, considere los siguientes ejemplos. Un candidato a un cargo público podría estar interesado en estimar la verdadera proporción de votantes que lo favorecerán mediante la obtención de las opiniones de una muestra aleatoria de 100 de ellos. La parte de votantes en la muestra que favorecerán al candidato se podría utilizar como un estimado de la verdadera proporción en la población de votantes. El conocimiento de la distribución muestral de una proporción nos permite establecer el grado de exactitud de tal estimado. Este problema cae en el área de la estimación.

Considere ahora el caso de alguien a quien le interesa averiguar si la marca *A* de cera para piso es más resistente al desgaste que la marca *B*. Se podría plantear la hipótesis de que la marca *A* es mejor que la marca *B* y, después de la prueba adecuada, aceptar o rechazar dicha hipótesis. En este ejemplo no intentamos estimar un parámetro, sino llegar a una decisión correcta acerca de una hipótesis planteada previamente. Una vez más, dependemos de la teoría del muestreo y de utilizar datos que nos proporcionen alguna medida del grado de exactitud de nuestra decisión.

9.3 Métodos de estimación clásicos

La **estimación puntual** de algún parámetro de la población θ es un solo valor $\hat{\theta}$ de un estadístico $\hat{\Theta}$. Por ejemplo, el valor $\bar{x}$ del estadístico $\bar{X}$, que se calcula a partir de una muestra de tamaño *n,* es una estimación puntual del parámetro de la población μ. De manera similar, $\hat{p} = x/n$ es una estimación puntual de la verdadera proporción p para un experimento binomial.

No se espera que un estimador logre estimar el parámetro de la población sin error. No se espera que $\bar{X}$ estime μ con exactitud, lo que en realidad se espera es que no esté muy alejada. Para una muestra específica, la manera en que se podría obtener un estimado más cercano de μ es utilizando la mediana de la muestra $\tilde{X}$ como estimador. Considere, por ejemplo, una muestra que consta de los valores 2, 5 y 11 de una población cuya media es 4, la cual, supuestamente, se desconoce. Podríamos estimar μ para que sea $\bar{x} = 6$ usando la media muestral como nuestro estimado, o bien, $\tilde{x} = 5$ utilizando la mediana muestral. En este caso el estimador $\tilde{X}$ produce una estimación más cercana al parámetro verdadero que la que produce el estimador X^{-}. Por otro lado, si nuestra muestra aleatoria contiene los valores 2, 6 y 7, entonces $\tilde{x} = 5$ y $\bar{x} = 6$, de manera que el mejor estimador es $\bar{X}$. Cuando no conocemos el valor real de μ, tenemos que comenzar por decidir qué estimador utilizaremos, si X^{-} o $\tilde{X}$.

Estimador insesgado

¿Cuáles son las propiedades que una “buena” función de decisión debería tener para poder influir en nuestra elección de un estimador en vez de otro? Sea $\hat{\Theta}$ un estimador cuyo valor $\hat{\theta}$ es una estimación puntual de algún parámetro de la población desconocido θ. Sin duda desearíamos que la distribución muestral de $\hat{\Theta}$ tuviera una media igual al parámetro estimado. Al estimador que tuviera esta propiedad se le llamaría **estimador insesgado**.

Definición 9.1: Se dice que un estadístico $\hat{\Theta}$ es un **estimador insesgado** del parámetro θ si

$$\mu_{\hat{\Theta}} = E(\hat{\Theta}) = \theta.$$

Ejemplo 9.1: Demuestre que S^2 es un estimador insesgado del parámetro σ^2.

***Solución*:** En la sección 8.5, en la página 244, demostramos que

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2.$$

Entonces,

$$E(S^2) = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]$$
$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}E(X_i - \mu)^2 - nE(\bar{X} - \mu)^2\right] = \frac{1}{n-1}\left(\sum_{i=1}^{n}\sigma^2_{X_i} - n\sigma^2_{\bar{X}}\right).$$

Sin embargo,

$$\sigma^2_{X_i} = \sigma^2, \text{ para } i = 1, 2, \ldots, n, \text{ y } \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}.$$

Por lo tanto,

$$E(S^2) = \frac{1}{n-1}\left(n\sigma^2 - n\frac{\sigma^2}{n}\right) = \sigma^2.$$

Aunque S^2 es un estimador insesgado de σ^2, S, por otro lado, suele ser un estimador sesgado de σ, un sesgo que en el caso de muestras grandes se vuelve insignificante. Este ejemplo ilustra **por qué dividimos entre *n* – 1** en vez de entre n cuando estimamos la varianza.

Varianza de un estimador puntual

Si $\hat{\Theta}_1$ y $\hat{\Theta}_2$ son dos estimadores insesgados del mismo parámetro de la población θ, deseamos elegir el estimador cuya distribución muestral tenga la menor varianza. Por lo tanto, si $\sigma^2_{\hat{\theta}_1} < \sigma^2_{\hat{\theta}_2}$, decimos que $\hat{\Theta}_1$ es un **estimador más eficaz** de θ que $\hat{\Theta}_2$.

Definición 9.2: Si consideramos todos los posibles estimadores insesgados de algún parámetro θ, al que tiene la menor varianza lo llamamos estimador más eficaz de θ.

En la figura 9.1 se ilustran las distribuciones muestrales de tres estimadores diferentes $\hat{\Theta}_1$, $\hat{\Theta}_2$ y $\hat{\Theta}_3$, todos para θ. Es evidente que sólo $\hat{\Theta}_1$ y $\hat{\Theta}_2$ no son sesgados, ya que sus distribuciones están centradas en θ. El estimador $\hat{\Theta}_1$ tiene una varianza menor que $\hat{\Theta}_2$, por lo tanto, es más eficaz. En consecuencia, el estimador de θ que elegiríamos, entre los tres que estamos considerando, sería $\hat{\Theta}_1$.

Para poblaciones normales se puede demostrar que tanto $\bar{X}$ como $\tilde{X}$ son estimadores insesgados de la media de la población μ, pero la varianza de $\bar{X}$ es más pequeña que la varianza de $\tilde{X}$. Por consiguiente, los estimados $\bar{x}$ y $\tilde{x}$ serán, en promedio, iguales a

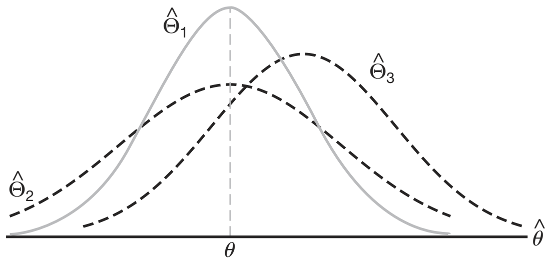

Figura 9.1: Distribuciones muestrales de diferentes estimadores de θ.

la media de la población μ, aunque podría ser que $\bar{x}$ esté más cerca de μ para una muestra dada y, por lo tanto, que $\bar{X}$ sea más eficaz que $\tilde{X}$.

Estimación por intervalo

Podría ser que ni el estimador insesgado más eficaz estime con exactitud el parámetro de la población. Es cierto que la exactitud de la estimación aumenta cuando las muestras son grandes; pero incluso así no tenemos razones para esperar que una **estimación puntual** de una muestra dada sea exactamente igual al parámetro de la población que se supone debe estimar. Hay muchas situaciones en que es preferible determinar un intervalo dentro del cual esperaríamos encontrar el valor del parámetro. Tal intervalo se conoce como **estimación por intervalo**.

Una estimación por intervalo de un parámetro de la población θ es un intervalo de la forma $\hat{\theta}_L < \theta < \hat{\theta}_U$, donde $\hat{\theta}_L$ y $\hat{\theta}_U$ dependen del valor del estadístico $\hat{\Theta}$ para una muestra específica, y también de la distribución de muestreo de $\hat{\Theta}$. Por ejemplo, una muestra aleatoria de calificaciones verbales de la prueba SAT para estudiantes universitarios de primer año produciría un intervalo de 530 a 550, dentro del cual esperamos encontrar el promedio verdadero de todas las calificaciones verbales de la prueba SAT para ese grupo. Los valores de los puntos extremos, 530 y 550, dependerán de la media muestral calculada $\bar{x}$ y de la distribución de muestreo de $\bar{X}$. A medida que aumenta el tamaño de la muestra, sabemos que $\sigma^2_{\bar{X}} = \sigma^2/n$ disminuye y, en consecuencia, cabe la posibilidad de que nuestra estimación se acerque más al parámetro μ, lo cual daría como resultado un intervalo más corto. De esta manera, el intervalo de la estimación indica, por su longitud, la precisión de la estimación puntual. Un ingeniero obtendrá información acerca de la proporción de la población de artículos defectuosos tomando una muestra y calculando la *proporción muestral defectuosa,* sin embargo, una estimación por intervalo podría ser más informativa.

Interpretación de las estimaciones por intervalo

Como muestras distintas suelen producir valores diferentes de $\hat{\Theta}$ y, por lo tanto, valores diferentes de $\hat{\theta}_L$ y $\hat{\theta}_U$, estos puntos extremos del intervalo son valores de las variables aleatorias correspondientes $\hat{\Theta}_L$ y $\hat{\Theta}_U$. De la distribución muestral de $\hat{\Theta}$ seremos capaces de determinar $\hat{\Theta}_L$ y $\hat{\Theta}_U$, de manera que $P(\hat{\Theta}_L < \theta < \hat{\Theta}_U)$ sea igual a cualquier valor positivo de una fracción que queramos especificar. Si, por ejemplo, calculamos $\hat{\Theta}_L$ y $\hat{\Theta}_U$, tales que

$$P(\hat{\Theta}_L < \theta < \hat{\Theta}_U) = 1 - \alpha,$$

para $0 < \alpha < 1$, tenemos entonces una probabilidad de $1 - \alpha$ de seleccionar una muestra aleatoria que produzca un intervalo que contenga θ. El intervalo $\hat{\theta}_L < \theta < \hat{\theta}_U$, que se calcula a partir de la muestra seleccionada, se llama entonces **intervalo de confianza** del $100(1 - \alpha)\%$, la fracción $1 - \alpha$ se denomina **coeficiente de confianza** o **grado de confianza**, y los extremos, $\hat{\theta}_L$ y $\hat{\theta}_U$, se denominan **límites de confianza** inferior y superior. Así, cuando $\alpha = 0.05$, tenemos un intervalo de confianza del 95%, y cuando $\alpha = 0.01$ obtenemos un intervalo de confianza más amplio del 99%. Cuanto más amplio sea el intervalo de confianza, más confiaremos en que contiene el parámetro desconocido. Desde luego, es mejor tener un 95% de confianza en que la vida promedio de cierto transistor de un televisor está entre los 6 y los 7 años, que tener un 99% de confianza en que esté entre los 3 y los 10 años. De manera ideal, preferimos un intervalo corto con un grado de confianza alto. Algunas veces las restricciones en el tamaño de nuestra muestra nos impiden tener intervalos cortos sin sacrificar cierto grado de confianza.

En las siguientes secciones estudiaremos los conceptos de estimación puntual y por intervalos, y en cada sección presentaremos un caso especial diferente. El lector debería notar que, aunque la estimación puntual y por intervalos representan diferentes aproximaciones para obtener información respecto a un parámetro, están relacionadas debido a que los estimadores del intervalo de confianza se basan en estimadores puntuales. En la siguiente sección, por ejemplo, veremos que $\bar{X}$ es un estimador puntual de μ muy razonable. Como resultado, el importante estimador del intervalo de confianza de μ depende del conocimiento de la distribución muestral de $\bar{X}$.

Empezaremos la siguiente sección con el caso más sencillo de un intervalo de confianza, en donde el escenario es simple pero poco realista. Nos interesa estimar una media de la población μ cuando σ todavía se desconoce. Evidentemente, si se desconoce μ es muy improbable que se conozca σ. Cualquier información histórica que produzca datos suficientes para permitir suponer que se conoce σ probablemente habría producido información similar acerca de μ. A pesar de este argumento iniciamos con este caso porque los conceptos y los mecanismos resultantes asociados con la estimación del intervalo de confianza también estarán asociados con las situaciones más realistas que presentaremos más adelante en la sección 9.4 y las siguientes.

9.4 Una sola muestra: estimación de la media

La distribución muestral de $\bar{X}$ está centrada en μ y en la mayoría de las aplicaciones la varianza es más pequeña que la de cualesquiera otros estimadores de μ. Por lo tanto, se utilizará la media muestral $\bar{x}$ como una estimación puntual para la media de la población μ. Recuerde que $\sigma_{\bar{X}}^2 = \sigma^2/n$, por lo que una muestra grande producirá un valor de $\bar{X}$ procedente de una distribución muestral con varianza pequeña. Por consiguiente, es probable que $\bar{x}$ sea una estimación muy precisa de μ cuando n es grande.

Consideremos ahora la estimación por intervalos de μ. Si seleccionamos nuestra muestra a partir de una población normal o, a falta de ésta, si n es suficientemente grande, podemos establecer un intervalo de confianza para μ considerando la distribución muestral de $\bar{X}$.

De acuerdo con el teorema del límite central, podemos esperar que la distribución muestral de $\bar{X}$ esté distribuida de forma aproximadamente normal con media $\mu_{\bar{X}} = \mu$ y desviación estándar $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. Al escribir $z_{\alpha/2}$ para el valor z por arriba del cual encontramos una área de $\alpha/2$ bajo la curva normal, en la figura 9.2 podemos ver que

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha,$$

donde

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

En consecuencia,

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha.$$

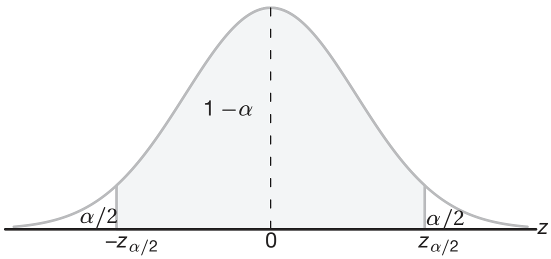

Figura 9.2: $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$.

Si multiplicamos cada término en la desigualdad por $\sigma/\sqrt{n}$ y después restamos $\bar{X}$ de cada término, y en seguida multiplicamos por − 1 (para invertir el sentido de las desigualdades), obtenemos

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Se selecciona una muestra aleatoria de tamaño n de una población cuya varianza σ^2 se conoce y se calcula la media $\bar{x}$ para obtener el intervalo de confianza $100(1 - \alpha)\%$. Es importante enfatizar que recurrimos al teorema del límite central citado anteriormente. Como resultado, es importante observar las condiciones para las aplicaciones que siguen.

Intervalo de confianza de μ cuando se conoce σ^2

Si $\bar{x}$ es la media de una muestra aleatoria de tamaño n de una población de la que se conoce su varianza σ^2, lo que da un intervalo de confianza de $100(1 - \alpha)\%$ para μ es

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

donde $z_{\alpha/2}$ es el valor z que deja una área de $\alpha/2$ a la derecha.

En el caso de muestras pequeñas que se seleccionan de poblaciones no normales, no podemos esperar que nuestro grado de confianza sea preciso. Sin embargo, para muestras de tamaño $n \geq 30$, en las que la forma de las distribuciones no esté muy sesgada, la teoría de muestreo garantiza buenos resultados.

Queda claro que los valores de las variables aleatorias $\hat{\Theta}_L$ y $\hat{\Theta}_U$, las cuales se definieron en la sección 9.3, son los límites de confianza

$$\hat{\theta}_L = \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{y} \quad \hat{\theta}_U = \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Muestras diferentes producirán valores diferentes de $\bar{x}$ y, por lo tanto, producirán diferentes estimaciones por intervalos del parámetro μ, como se muestra en la figura 9.3. Los puntos en el centro de cada intervalo indican la posición de la estimación puntual $\bar{x}$ para cada muestra aleatoria. Observe que todos los intervalos tienen el mismo ancho, pues esto depende sólo de la elección de $z_{\alpha/2}$ una vez que se determina $\bar{x}$. Cuanto más grande sea el valor de $z_{\alpha/2}$ que elijamos, más anchos haremos todos los intervalos, y podremos tener más confianza en que la muestra particular que seleccionemos producirá un intervalo que contenga el parámetro desconocido μ. En general, para una elección de $z_{\alpha/2}$, $100(1 - \alpha)\%$ de los intervalos contendrá μ.

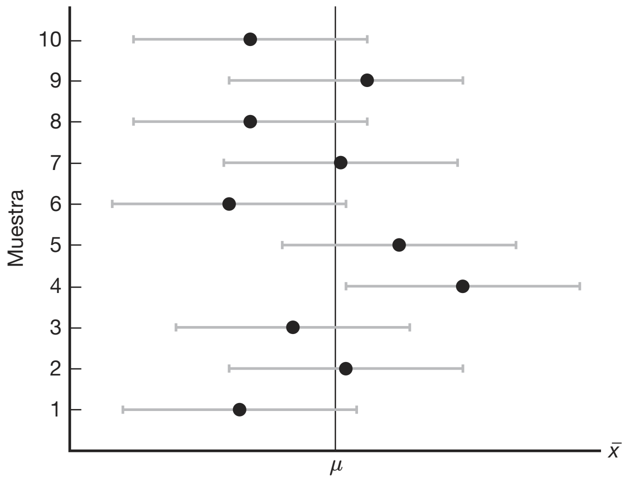

Figura 9.3: Estimaciones por intervalos de μ para muestras diferentes.

Ejemplo 9.2: Se encuentra que la concentración promedio de zinc que se obtiene en una muestra de mediciones en 36 sitios diferentes de un río es de 2.6 gramos por mililitro. Calcule los intervalos de confianza del 95% y 99% para la concentración media de zinc en el río. Suponga que la desviación estándar de la población es de 0.3 gramos por mililitro.

***Solución*:** La estimación puntual de μ es $\bar{x} = 2.6$. El valor z que deja una área de 0.025 a la derecha y, por lo tanto, una área de 0.975 a la izquierda es $z_{0.025} = 1.96$ (véase la tabla A.3). En consecuencia, el intervalo de confianza del 95% es

$$2.6 - (1.96)\left(\frac{0.3}{\sqrt{36}}\right) < \mu < 2.6 + (1.96)\left(\frac{0.3}{\sqrt{36}}\right),$$

que se reduce a $2.50 < \mu < 2.70$. Para calcular un intervalo de confianza del 99% encontramos el valor z que deja una área de 0.005 a la derecha y de 0.995 a la izquierda. Por lo tanto, usando la tabla A.3 nuevamente, $z_{0.005} = 2.575$ y el intervalo de confianza de 99% es

$$2.6 - (2.575)\left(\frac{0.3}{\sqrt{36}}\right) < \mu < 2.6 + (2.575)\left(\frac{0.3}{\sqrt{36}}\right),$$

o simplemente

$$2.47 < \mu < 2.73.$$

Ahora vemos que se requiere un intervalo más grande para estimar μ con un mayor grado de confianza. ▟

El intervalo de confianza del $100(1 - \alpha)\%$ ofrece un estimado de la precisión de nuestra estimación puntual. Si μ es realmente el valor central del intervalo, entonces $\bar{x}$ estima μ sin error. La mayoría de las veces, sin embargo, $\bar{x}$ no será exactamente igual a μ y la estimación puntual será errónea. La magnitud de este error será el valor absoluto de la diferencia entre μ y $\bar{x}$, de manera que podemos tener $100(1 - \alpha)\%$ de confianza en que esta diferencia no excederá a $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. Podemos ver esto fácilmente dibujando un diagrama de un intervalo de confianza hipotético, como el de la figura 9.4.

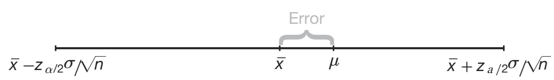

Figura 9.4: Error en la estimación de μ mediante $\bar{x}$.

Teorema 9.1: Si utilizamos $\bar{x}$ como una estimación de μ, podemos tener $100(1 - \alpha)\%$ de confianza en que el error no excederá a $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

En el ejemplo 9.2 tenemos una confianza del 95% en que la media muestral $\bar{x} = 2.6$ difiere de la media verdadera μ en una cantidad menor que $(1.96)(0.3)/\sqrt{36} = 0.1$ y 99% de confianza en que la diferencia es menor que $(2.575)(0.3)/\sqrt{36} = 0.13$.

Con frecuencia queremos saber qué tan grande necesita ser una muestra para poder estar seguros de que el error al estimar μ será menor que una cantidad específica e. Por medio del teorema 9.1 debemos elegir n de manera que $z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = e$. Al resolver esta ecuación obtenemos la siguiente fórmula para n.

Teorema 9.2: Si usamos $\bar{x}$ como una estimación de μ, podemos tener $100(1 - \alpha)\%$ de confianza en que el error no excederá a una cantidad específica e cuando el tamaño de la muestra sea

$$n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2.$$

Cuando resolvemos para la muestra con tamaño n, redondeamos todos los valores decimales al siguiente número entero. Si seguimos este principio, podemos estar seguros de que nuestro grado de confianza nunca caerá por debajo del $100(1 - \alpha)\%$.

En términos estrictos, la fórmula del teorema 9.2 sólo será aplicable si se conoce la varianza de la población de la cual se seleccionó la muestra. Si no contamos con esa información, podríamos tomar una muestra preliminar de tamaño $n \geq 30$ para proporcionar una estimación de σ. Después, usando s como aproximación para σ en el teorema 9.2, podemos determinar aproximadamente cuántas observaciones necesitamos para brindar el grado de precisión deseado.

Ejemplo 9.3: ¿Qué tan grande debe ser la muestra del ejemplo 9.2 si queremos tener 95% de confianza en que nuestra estimación de μ diferirá por menos de 0.05?

***Solución*:** La desviación estándar de la población es $\sigma = 0.3$. Entonces, por medio del teorema 9.2,

$$n = \left[\frac{(1.96)(0.3)}{0.05}\right]^2 = 138.3.$$

Por lo tanto, podemos tener 95% de confianza en que una muestra aleatoria de tamaño 139 proporcionará una estimación $\bar{x}$ que diferirá de μ en una cantidad menor que 0.05. ▟

Límites de confianza unilaterales

Los intervalos de confianza y los límites de confianza resultantes que hasta ahora hemos analizado en realidad son *bilaterales,* es decir, tienen límites superior e inferior. Sin embargo, hay muchas aplicaciones en las que sólo se requiere un límite. Por ejemplo, si a un ingeniero le interesara determinar una medida de resistencia a la tensión, la información que más le ayudaría a lograr su objetivo sería la del límite inferior, ya que éste indica el escenario del "peor caso", es decir, el de la menor resistencia. Por otro lado, si se buscara determinar una medida para la cual un valor de μ relativamente grande no fuera redituable o deseable, entonces la medida que resultaría de interés sería la del límite de confianza superior. Un ejemplo en el que la medida del límite superior sería muy informativa es el caso en el que se necesita hacer inferencias para determinar la composición media de mercurio en el agua de un río.

Los límites de confianza unilaterales se desarrollan de la misma forma que los intervalos bilaterales. Sin embargo, la fuente es un enunciado de probabilidad unilateral que utiliza el teorema del límite central:

$$P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_\alpha\right) = 1 - \alpha.$$

Entonces, es posible manipular el enunciado de probabilidad de forma muy similar a como se hizo anteriormente para obtener

$$P(\mu > \bar{X} - z_\alpha \sigma/\sqrt{n}) = 1 - \alpha.$$

Una manipulación similar de $P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > -z_\alpha\right) = 1 - \alpha$ da

$$P(\mu < \bar{X} + z_\alpha \sigma/\sqrt{n}) = 1 - \alpha.$$

Como resultado, se obtienen los siguientes límites unilaterales superior e inferior.

Límites de confianza unilaterales de μ cuando se conoce el valor de σ^2

Si $\bar{X}$ es la media de una muestra aleatoria de tamaño n a partir de una población con varianza σ^2, los límites de confianza unilaterales del $100(1-\alpha)\%$ para μ son dados por

límite unilateral superior: $\bar{x} + z_\alpha \sigma/\sqrt{n}$;

límite unilateral inferior: $\bar{x} - z_\alpha \sigma/\sqrt{n}$.

Ejemplo 9.4: En un experimento de pruebas psicológicas se seleccionan al azar 25 sujetos y se miden sus tiempos de reacción, en segundos, ante un estímulo particular. La experiencia sugiere que la varianza en los tiempos de reacción ante los diferentes tipos de estímulos es de 4 s^2 y que la distribución del tiempo de reacción es aproximadamente normal. El tiempo promedio para los sujetos fue de 6.2 segundos. Calcule un límite superior del 95% para el tiempo medio de reacción.

***Solución*:** Lo que da el límite superior del 95% es

$$\begin{aligned}\bar{x} + z_\alpha \sigma/\sqrt{n} &= 6.2 + (1.645)\sqrt{4/25} = 6.2 + 0.658\\ &= 6.858 \text{ segundos.}\end{aligned}$$

En consecuencia, tenemos un 95% de confianza en que el tiempo promedio de reacción es menor que 6.858 segundos.

El caso en que se desconoce σ

Con frecuencia debemos tratar de estimar la media de una población sin conocer la varianza. El lector debería recordar que en el capítulo 8 aprendió que, si tenemos una muestra aleatoria a partir de una *distribución normal,* entonces la variable aleatoria

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

tiene una distribución t de Student con $n - 1$ grados de libertad. Aquí S es la desviación estándar de la muestra. En esta situación, en la que se desconoce σ, se puede utilizar T para construir un intervalo de confianza para μ. El procedimiento es igual que cuando se conoce σ, sólo que en este caso σ se reemplaza con S y la distribución normal estándar se reemplaza con la distribución t. Si nos remitimos a la figura 9.5, podemos afirmar que

$$P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha,$$

donde $t_{\alpha/2}$ es el valor t con $n - 1$ grados de libertad, por arriba del cual encontramos una área de $\alpha/2$. Debido a la simetría, un área igual de $\alpha/2$ caerá a la izquierda de $-t_{\alpha/2}$. Al sustituir por T escribimos

$$P\left(-t_{\alpha/2} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2}\right) = 1 - \alpha.$$

Al multiplicar cada término en la desigualdad por $S/\sqrt{n}$ y después restar $\bar{X}$ de cada término y multiplicar por -1, obtenemos

$$P\left(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

Para nuestra muestra aleatoria particular de tamaño n se calculan la media $\bar{x}$ y la desviación estándar s, y se obtiene el siguiente intervalo de confianza $100(1 - \alpha)\%$ para μ.

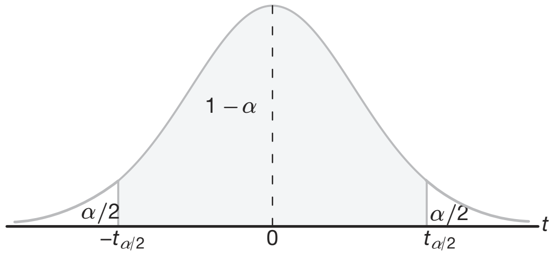

Figura 9.5: $P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha$.

Intervalo de confianza para μ cuando se desconoce σ^2

Si $\bar{x}$ y s son la media y la desviación estándar de una muestra aleatoria de una población normal de la que se desconoce la varianza σ^2, un intervalo de confianza del $100(1 - \alpha)\%$ para μ es

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}},$$

donde $t_{\alpha/2}$ es el valor t con $v = n - 1$ grados de libertad que deja una área de $\alpha/2$ a la derecha.

Hicimos una distinción entre los casos en los que se conoce σ y en los que se desconoce calculando las estimaciones del intervalo de confianza. Deberíamos resaltar que para el caso en que se conoce σ se utiliza el teorema del límite central, mientras que, para el caso en que se desconoce, se usa la distribución muestral de la variable aleatoria T. Sin embargo, el uso de la distribución t se basa en la premisa de que el muestreo es de una distribución normal. Siempre que la forma de la distribución se aproxime a la de campana, se puede utilizar la distribución t para calcular los intervalos de confianza cuando se desconoce σ^2, y se pueden esperar muy buenos resultados.

Los límites de confianza unilaterales calculados para μ con σ desconocida son como el lector esperaría, a saber:

$$\bar{x} + t_{\alpha}\frac{s}{\sqrt{n}} \quad \text{y} \quad x - t_{\alpha}\frac{s}{\sqrt{n}}.$$

Éstos son, respectivamente, los límites superior e inferior del $100(1 - \alpha)\%$. Aquí t_α es el valor t que tiene una área α a la derecha.

Ejemplo 9.5: El contenido de ácido sulfúrico de 7 contenedores similares es de 9.8, 10.2, 10.4, 9.8, 10.0, 10.2, y 9.6 litros. Calcule un intervalo de confianza del 95% para el contenido promedio de todos los contenedores suponiendo una distribución aproximadamente normal.

***Solución*:** La media muestral y la desviación estándar para los datos dados son

$$\bar{x} = 10.0 \quad \text{y} \quad s = 0.283.$$

Si usamos la tabla A.4, encontramos $t_{0.025} = 2.447$ para $v = 6$ grados de libertad. En consecuencia, el intervalo de confianza del 95% para μ es

$$10.0 - (2.447)\left(\frac{0.283}{\sqrt{7}}\right) < \mu < 10.0 + (2.447)\left(\frac{0.283}{\sqrt{7}}\right),$$

que se reduce a $9.74 < \mu < 10.26$.

Concepto de intervalo de confianza para una muestra grande

Con frecuencia los estadísticos recomiendan que incluso cuando no sea posible suponer la normalidad, se desconozca σ y $n \geq 30$, σ se puede reemplazar con s para poder utilizar el intervalo de confianza

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

A menudo se hace referencia a esto como un *intervalo de confianza para una muestra grande.* La justificación para esto reside sólo en la presunción de que, con una muestra tan grande como 30 y una distribución de la población no muy sesgada, s estará muy cerca de la σ verdadera y, de esta manera, el teorema del límite central continuará siendo válido. Se debería destacar que esto es sólo una aproximación y que la calidad de los resultados mejora a medida que aumenta el tamaño de la muestra.

Ejemplo 9.6: Se obtienen las calificaciones de matemáticas del Examen de Aptitudes Escolares (SAT, por sus siglas en inglés) de una muestra aleatoria de 500 estudiantes del último año de preparatoria del estado de Texas. Se calculan la media y la desviación estándar muestrales, que son 501 y 112, respectivamente. Calcule un intervalo de confianza del 99% de la calificación promedio de matemáticas en el SAT para los estudiantes del último año de preparatoria del estado de Texas.

***Solución*:** Como el tamaño de la muestra es grande, es razonable utilizar la aproximación normal. Si utilizamos la tabla A.3, encontramos $z_{0.005} = 2.575$. Por lo tanto, un intervalo de confianza del 99% para μ es

$$501 \pm (2.575)\left(\frac{112}{\sqrt{500}}\right) = 501 \pm 12.9,$$

que da como resultado $488.1 < \mu < 513.9$.

9.5 Error estándar de una estimación puntual

Hicimos una distinción muy clara entre los objetivos de las estimaciones puntuales y las estimaciones del intervalo de confianza. Las primeras proporcionan un solo número que se extrae de un conjunto de datos experimentales, y las segundas proporcionan un intervalo razonable para el parámetro, *dados los datos experimentales*; es decir, $100(1 - \alpha)\%$ de tales intervalos que se calcula "cubren" el parámetro.

Estos dos métodos de estimación se relacionan entre sí. El elemento en común es la distribución muestral del estimador puntual. Considere, por ejemplo, el estimador $\bar{X}$ de μ cuando se conoce σ. Indicamos antes que una medida de la calidad de un estimador insesgado es su varianza. La varianza de $\bar{X}$ es

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}.$$

De esta forma, la desviación estándar de $\bar{X}$ o *error estándar de* $\bar{X}$ es $\sigma/\sqrt{n}$. En términos simples, el error estándar de un estimador es su desviación estándar. Para el caso de $\bar{X}$ el límite de confianza que se calcula

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ se escribe como } \bar{x} \pm z_{\alpha/2} \text{ e.e.}(\bar{x}),$$

donde "e.e." es el error estándar. El punto importante es que el ancho del intervalo de confianza de μ depende de la calidad del estimador puntual a través de su error estándar. En el caso en que se desconoce σ y la muestra proviene de una distribución normal, s reemplaza a σ y se incluye el *error estándar estimado* $S/\sqrt{n}$. Por consiguiente, los límites de confianza de μ son:

Límites de confianza para μ cuando se desconoce σ^2

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} = \bar{x} \pm t_{\alpha/2} \text{ e.e.}(\bar{x})$$

De nuevo, el intervalo de confianza *no es mejor* (en términos de anchura) *que la calidad de la estimación puntual,* en este caso a través de su error estándar estimado. A menudo el software de computación se refiere a los errores estándar estimados simplemente como "errores estándar".

A medida que avanzamos a intervalos de confianza más complejos, prevalece el concepto de que el ancho de los intervalos de confianza se acorta cuando mejora la calidad de la estimación puntual correspondiente, aunque no siempre es tan sencillo como aquí se ilustra. Se puede argumentar que un intervalo de confianza es tan sólo una ampliación de la estimación puntual para tomar en cuenta la exactitud de dicha estimación.

9.6 Intervalos de predicción

La estimación puntual y la estimación por intervalos de la media que se expusieron en las secciones 9.4 y 9.5 proporcionan buena información del parámetro desconocido μ de una distribución normal, o de una distribución no normal a partir de la cual se toma una muestra grande. Algunas veces, además de la media de la población, el experimentador podría estar interesado en predecir el **valor posible de una observación futura**. Por ejemplo, en el control de calidad el experimentador podría necesitar utilizar los datos observados para predecir una nueva observación. Un proceso de manufactura de una pieza de metal se podría evaluar basándose en si la pieza cumple con las especificaciones de resistencia a la tensión. En ciertas ocasiones un cliente podría estar interesado en comprar una **sola pieza**. En este caso un intervalo de confianza de la resistencia media a la tensión no cubriría la información requerida. El cliente necesitaría una aseveración respecto a la incertidumbre de una **sola observación**. Este tipo de requerimiento se satisface muy bien construyendo un **intervalo de predicción**.

Es muy sencillo obtener un intervalo de predicción para las situaciones que hemos considerado hasta el momento. Suponga que la muestra aleatoria se tomó de una población normal con media μ desconocida y varianza σ^2 conocida. Un estimador puntual natural de una nueva observación es $\bar{X}$. En la sección 8.4 se aprendió que la varianza de $\bar{X}$ es σ^2/n. Sin embargo, para predecir una nueva observación no basta con explicar la variación debida a la estimación de la media, también tendríamos que explicar la **variación de una observación futura**. A partir de la suposición sabemos que la varianza del error aleatorio en una nueva observación es σ^2. El desarrollo de un intervalo de predicción se representa mejor empezando con una variable aleatoria normal $x_0 - \bar{x}$, donde x_0 es la nueva observación y $\bar{x}$ se toma de la muestra. Como x_0 y $\bar{x}$ son independientes, sabemos que

$$z = \frac{x_0 - \bar{x}}{\sqrt{\sigma^2 + \sigma^2/n}} = \frac{x_0 - \bar{x}}{\sigma\sqrt{1 + 1/n}}$$

es $n(z; 0, 1)$. Como resultado, si utilizamos el enunciado de probabilidad

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

con el estadístico z anterior, y si colocamos x_0 en el centro del enunciado de probabilidad, tenemos que la probabilidad de que ocurra el siguiente evento es $1 - \alpha$:

$$\bar{x} - z_{\alpha/2}\sigma\sqrt{1 + 1/n} < x_0 < \bar{x} + z_{\alpha/2}\sigma\sqrt{1 + 1/n}.$$

Como resultado, el intervalo de predicción calculado se formaliza como sigue.

Intervalo de predicción para una observación futura cuando se conoce σ^2

Para una distribución normal de mediciones con media μ desconocida y varianza σ^2 conocida, un **intervalo de predicción** del $100(1 - \alpha)\%$ de una observación futura x_0 es

$$\bar{x} - z_{\alpha/2}\sigma\sqrt{1 + 1/n} < x_0 < \bar{x} + z_{\alpha/2}\sigma\sqrt{1 + 1/n},$$

donde $z_{\alpha/2}$ es el valor z que deja una área de $\alpha/2$ a la derecha.

Ejemplo 9.7: Debido a la disminución en las tasas de interés el First Citizens Bank recibió muchas solicitudes para hipoteca. Una muestra reciente de 50 créditos hipotecarios dio como resultado un promedio en la cantidad de préstamos de \$257,300. Suponga una desviación estándar de la población de \$25,000. En el caso del siguiente cliente que llena una solicitud de crédito hipotecario calcule un intervalo de predicción del 95% para la cantidad del crédito.

***Solución*:** La predicción puntual de la cantidad del crédito del siguiente cliente es $\bar{x} = \$257{,}300$. El valor z aquí es $z_{0.025} = 1.96$. Por lo tanto, un intervalo de predicción del 95% para la cantidad de un crédito futuro es

$$257,300 - (1.96)(25{,}000)\sqrt{1 + 1/50} < x_0 < 257,300 + (1.96)(25{,}000)\sqrt{1 + 1/50},$$

que produce el intervalo (\$207,812.43, \$306,787.57). ▟

El intervalo de predicción proporciona un buen estimado de la ubicación de una observación futura, el cual es muy diferente del estimado del valor promedio de la muestra. Debe advertirse que la variación de esta predicción es la suma de la variación debida a una estimación de la media y la variación de una sola observación. Sin embargo, como antes, consideramos primero el caso en el que se conoce la varianza. En el caso en que se desconoce la varianza también es importante tratar con el intervalo de predicción de una observación futura. De hecho, en este caso se podría utilizar una distribución t de Student, como se describe en el siguiente resultado. Aquí la distribución normal simplemente se reemplaza con la distribución t.