10

Prueba de hipótesis

En nuestro estudio de los intervalos de confianza, determinamos la distribución de los niveles de colesterol en la sangre de la población masculina de Estados Unidos hipertensos y fumadores. Esta distribución es aproximadamente normal con una media desconocida μ. Sin embargo, si sabemos que el nivel medio de colesterol en la sangre de la población general de varones de 2 a 74 años de edad es de 211 mg/100 ml [1]. Por tanto, podríamos preguntarnos si el nivel medio de colesterol en la sangre de la subpoblación de hombres hipertensos fumadores es también de 211 mg/100 ml. Si elegimos una muestra aleatoria de 25 varones de este grupo cuyo nivel medio de colesterol en la sangre sea de $\bar{x} = 220$ mg/100 ml, ¿es esta media de muestreo compatible con una media hipotética de 211 mg/100 ml? Sabemos que cabe esperar cierto grado de variabilidad aleatoria. ¿Qué hay si la media de muestreo es de 230 mg/100 ml o 250 mg/100 ml? ¿Cuán alejada de 211 debe estar $\bar{x}$ antes de que podamos concluir que μ realmente es igual a algún otro valor?

10.1 Conceptos generales

De nuevo centramos nuestra atención en la obtención de alguna conclusión sobre un parámetro de población —en este caso la media de una variable aleatoria continua— con la información contenida en una muestra de observaciones. Como vimos en el capítulo anterior, una aproximación consistiría en construir un intervalo de confianza para μ; otra, en construir una *prueba de hipótesis estadística.*

Para llevar a cabo dicha prueba, comenzamos por afirmar que la media de la población es igual a algún valor dado μ_0. Esta declaración sobre el valor de un parámetro de población se denomina *hipótesis nula,* o H_0. Si quisiéramos probar si el nivel medio de colesterol en la sangre de la subpoblación de fumadores hipertensos es igual a la media de la población general de hombres de 20 a 74 años, por ejemplo, la hipótesis nula sería

$$H_0: \mu = \mu_0 = 211 \text{ mg/100 ml.}$$

La *hipótesis alterna*, representada por H_A, es una segunda afirmación que contradice a H_0. En este caso, tenemos

$$H_A: \mu \neq 211 \text{ mg/100 ml.}$$

Juntas, las hipótesis nula y la alterna cubren todos los posibles valores de la media de población μ; en consecuencia, una de las dos declaraciones debe ser verdadera.

Inmediatamente después de formular la hipótesis, tomamos una muestra al azar de tamaño n de la población de interés. En el caso de los fumadores hipertensos, elegimos una muestra de tamaño 12. Comparamos la media de esta muestra, $\bar{x}$, con la media propuesta μ_0; específicamente, deseamos saber si la diferencia entre la media de muestreo y la media hipotética es demasiado grande para atribuirla a la pura casualidad.

Si existe evidencia de que la muestra **no** puede provenir de una población con media μ_0, rechazamos la hipótesis nula. Esto ocurre cuando, en el supuesto que H_0 sea verdadera, la probabilidad de obtener una media de muestreo tan extrema o más que el valor observado $\bar{x}$ —"más extrema" significa más alejada del valor μ_0— es suficientemente pequeña. En este caso, los datos no son compatibles con la hipótesis nula; apoyan más a la hipótesis alterna. Por tanto, concluimos que la media de población no puede ser μ_0. Adhiriéndonos a la fraseología tradicional, se dice que dicho resultado de la prueba es *estadísticamente significativo*. Observe que el significado estadística* no implica significado clínico o científico; el resultado de la prueba podría en realidad tener pocas consecuencias prácticas.

Si no existe suficiente evidencia para dudar de la validez de la hipótesis nula, no podemos rechazar esta afirmación. En lugar de eso, aceptamos que la media de población *puede* ser igual a μ_0. Sin embargo, no decimos que aceptamos H_0; la prueba no demuestra la hipótesis nula. Todavía es posible que la media de población tenga un valor distinto de μ_0, pero que la muestra aleatoria elegida no confirme este hecho. Esto puede suceder, por ejemplo, cuando la muestra elegida es demasiado pequeña. Este punto se analiza con detalle más adelante en este capítulo.

Hemos establecido anteriormente que si la probabilidad de obtener una media de muestreo tan extrema o más extrema que la $\bar{x}$ observada es suficientemente pequeña, rechazamos la hipótesis nula. Ahora bien, ¿qué entendemos por probabilidad "suficientemente pequeña"? En la mayoría de las aplicaciones se elige 0.05 2. [2] Así, rechazamos H_0 cuando la probabilidad de que la muestra pudiera provenir de una población con media μ_0 sea menor o igual a 5%. Esto implica que rechazamos incorrectamente 5% de las veces; dadas varias pruebas repetidas de significancia, 5 veces de 100 rechazaremos erróneamente la hipótesis nula cuando sea verdadera. Con el fin de ser más cautos, a veces se elige una probabilidad de 0.01. En este caso, equivocadamente rechazamos H_0 cuando es verdadera sólo 1% de las ocasiones. Si deseamos ser menos cautos, podríamos utilizar una probabilidad de 0.10. La probabilidad que elegimos —ya sea 0.05, 0.01 o algún otro valor— se conoce con el nombre de *nivel de significancia* de la prueba de hipótesis. El nivel de significancia se denota con la letra griega α y debe especificarse **antes** de que la prueba se lleve efectivamente a cabo.

La prueba de hipótesis es comparable en gran medida a un proceso penal llevado a cabo por un jurado en Estados Unidos. El individuo sometido a juicio es inocente o culpable, pero la ley supone su inocencia. Después de que se han presentado las pruebas relacionadas con el caso, el jurado encuentra al acusado culpable o no culpable. Si el acusado es inocente y la decisión del jurado es que éste no es culpable, se ha alcanzado el veredicto

*En este texto nos referiremos a ese significado como *significancia*.

justo. También resulta correcto el veredicto si el acusado es culpable y condenado por el crimen.

Veredicto del jurado	Acusado	
	Inocente	**Culpable**
No culpable	Correcto	Incorrecto
Culpable	Incorrecto	Correcto

Análogamente, la verdadera media de población es μ_0 o no es μ_0. Comenzamos por suponer que la hipótesis nula

$$H_0: \mu = \mu_0$$

es correcta, y consideremos la "evidencia" presentada en la forma de una muestra de tamaño n. Con base en nuestros hallazgos, la hipótesis nula se rechaza o no se rechaza. De nuevo, existen dos situaciones en las cuales la conclusión derivada es correcta: cuando la media de población es μ_0 y la hipótesis nula no se rechaza, y cuando la media poblacional no es μ_0 y H_0 se rechaza.

Resultado de la prueba	Población	
	$\mu = \mu_0$	$\mu \neq \mu_0$
No rechazada	Correcto	Incorrecto
Rechazada	Incorrecto	Correcto

Así como en el caso del sistema jurídico, el proceso de prueba de hipótesis no es perfecto. Existen dos tipos de errores que pueden cometerse. En particular, podríamos rechazar la hipótesis nula cuando μ es igual μ_0, o errar en rechazarla cuando μ no es igual a μ_0. Estos dos tipos de errores —que tienen mucho en común con los resultados falso positivo y falso negativo que ocurren en las pruebas de diagnóstico— se analizan con más detalles en la sección 10.4.

La probabilidad de obtener una media tan extrema o más extrema que la muestra observada $\bar{x}$ en el supuesto de que la hipótesis nula

$$H_0: \mu = \mu_0$$

sea verdadera, se denomina *valor p* de la prueba o sencillamente *p*. El valor *p* se compara con el nivel de significancia predeterminado α para decidir si la hipótesis nula debería rechazarse. Si *p* es menor o igual a α, rechazamos H_0. Si *p* es mayor que α, no rechazamos H_0. Además de la conclusión de la prueba, el propio valor de *p* a menudo figura en la bibliografía.

10.2 Pruebas de hipótesis bilaterales

Para llevar a cabo una prueba de hipótesis, de nuevo aprovechamos nuestro conocimiento de la distribución del promedio muestral. Supongamos que la variable aleatoria continua X posee la media μ_0 y desviación estándar conocida σ. Así, de acuerdo con el teorema del límite central,

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

tiene una distribución normal estándar aproximada si el valor de n es suficientemente grande. Para una muestra dada con media $\bar{x}$, podemos calcular el resultado correspondiente de Z, denominado *estadístico de prueba*. Entonces utilizamos ya sea un programa de computadora o una tabla de la curva normal estándar —como la tabla A.3 en el apéndice A— para determinar la probabilidad de obtener un valor de Z que sea tan extremo o más extremo que el observado. Con "más extremo" queremos decir más alejado de μ_0 en la dirección de la hipótesis alterna. Puesto que se apoya en la distribución normal estándar, una prueba de este tipo se denomina *prueba z*.

Cuando no se conoce la desviación estándar de población, sustituimos el valor de muestreo s por σ. Si la población estudiada está normalmente distribuida, la variable aleatoria

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

tiene una distribución t con $n - 1$ grados de libertad. En este caso, podemos calcular el resultado de t correspondiente a una $\bar{x}$ dada y consultar nuestro programa de cómputo o la tabla A.4 para determinar la probabilidad de obtener una media de muestreo más extrema que la observada. Este procedimiento se conoce con el nombre de *prueba t*.

Para ilustrar el procedimiento de la prueba de hipótesis, considere de nuevo la distribución de los niveles de colesterol en la sangre en adultos varones en Estados Unidos que son hipertensos y fuman. La desviación estándar de esta distribución se supone que es $\sigma =$ 46 mg/100 ml; la hipótesis nula que se ha de probar es

$$H_0: \mu = 211 \text{ mg/100 ml},$$

donde $\mu_0 = 211$ mg/100 ml es el nivel medio de colesterol en la sangre de todos los hombres de 20 a 74 años de edad. Puesto que la media de la subpoblación de fumadores hipertensos podría ser mayor o menor que μ_0, nos interesan las desviaciones que ocurran en cualquier dirección. Como resultado, llevamos a cabo lo que se conoce como prueba bilateral en el nivel de significancia $\alpha = 0.05$. La hipótesis alterna para la prueba bilateral es

$$H_A: \mu \neq 211 \text{ mg/100 ml}.$$

La muestra aleatoria antes mencionada de 12 fumadores hipertensos tiene un nivel medio de colesterol en la sangre de $\bar{x} = 217$ mg/100 ml [3]. ¿Es probable que esta muestra pro-

venga de una población con una media de 211 mg/100 ml? Para responder esta pregunta, calculamos el estadístico de prueba

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$
$$= \frac{217 - 211}{46/\sqrt{12}}$$
$$= 0.45.$$

Si la hipótesis nula es verdadera, este estadístico es el resultado de una variable aleatoria normal estándar. De acuerdo con la tabla A.3, el área a la derecha de $z = 0.45$ —que es la probabilidad de observar $Z = 0.45$ o cualquier valor más grande, dado que H_0 es verdadera— es 0.326. El área a la izquierda de $z = -0.45$ es también 0.326. Así, el área en las dos secciones de la distribución normal estándar suma 0.652. Éste es el valor p de la prueba. Puesto que $p > 0.05$, no rechazamos la hipótesis nula. Con base en esta muestra, la evidencia es insuficiente para concluir el hecho de que el nivel medio de colesterol en la sangre de la población de fumadores hipertensos sea diferente de 211 mg/100 ml.

Aunque no parezca muy obvio, en realidad existe una equivalencia matemática entre los intervalos de confianza y las pruebas de hipótesis. Puesto que llevamos a cabo una prueba bilateral, cualquier valor de z entre –1.96 y 1.96 daría como resultado un valor p mayor que 0.05. (El resultado 0.45 es sólo uno de dichos valores.) En todos estos casos, la hipótesis nula no se rechazaría. Por otra parte, H_0 se rechazaría para cualquier valor de z ya sea menor que –1.96 o mayor que 1.96. Debido a que estos indican cuándo debemos rechazar, los números –1.96 y 1.96 se denominan *valores críticos* del estadístico de prueba.

Otra forma de ver este hecho consiste en notar que no será posible rechazar la hipótesis nula cuando μ_0 sea cualquier valor ubicado en el intervalo de confianza de 95% para μ. Recuerde que en el capítulo 9 vimos que un intervalo de confianza de 95% del nivel medio de colesterol en la sangre de fumadores hipertensos era

(191, 243).

Cualquier valor de μ_0 localizado en este intervalo daría como resultado un estadístico de prueba localizado entre –1.96 y 1.96. Por tanto, si la hipótesis nula hubiera sido

H_0: $\mu = 240$ mg/100 ml.

H_0 no se habría rechazado. De manera similar, la hipótesis nula no sería rechazada para $\mu_0 = 195$ mg/100 ml. En contraste, cualquier valor de μ_0 localizado fuera del intervalo de confianza de 95% para μ —como $\mu_0 = 260$ mg/100 ml—, daría como resultado un rechazo de la hipótesis nula en el nivel $\alpha = 0.05$. Estos valores producen estadísticos de prueba ya sea menores que –1.96 o mayores que 1.96.

Aunque los intervalos de confianza y las pruebas de hipótesis nos conducen a las mismas conclusiones, la información que cada uno proporciona es de alguna manera diferente. El intervalo de confianza suministra un rango de valores razonables para el parámetro μ y nos

dice algo sobre la incertidumbre en nuestra estimación puntual $\bar{x}$. La prueba de hipótesis nos ayuda a decidir si el valor propuesto de la media es posiblemente correcto y proporciona un valor p específico.

De regreso a la prueba misma, se eligió el valor $\mu_0 = 211$ mg/100 ml para la hipótesis nula porque es el nivel medio de colesterol en la sangre de la población de todos los hombres de 20 a 74 años de edad. En consecuencia, H_0 indica que el nivel medio de colesterol en la sangre de hombres fumadores hipertensos es idéntico al nivel medio de colesterol de la población masculina general. La hipótesis se estableció con el interés de obtener evidencia para rechazarla en favor de la hipótesis alterna; un rechazo habría implicado que el nivel medio de colesterol en la sangre de varones fumadores hipertensos **no** es igual a la media de la población en conjunto.

En un segundo ejemplo, consideremos la muestra aleatoria de diez niños elegidos de la población infantil que reciben antiácidos con aluminio. La distribución real de niveles de aluminio en la sangre de esta población es aproximadamente normal con una media desconocida μ y una desviación estándar σ. Sin embargo, sabemos que el nivel medio de aluminio en la sangre de la muestra de tamaño 10 es $\bar{x} = 37.20$ μg/l y que su desviación estándar es $s = 7.13$ μg/l [4]. Además, el nivel medio de aluminio en la sangre en la población de niños que no reciben antiácidos es de 4.13 μg/l. ¿Es posible que los datos en nuestra muestra provinieran de una población con media $\mu_0 = 4.13$ μg/l? Para averiguarlo, se realiza una prueba de hipótesis; la hipótesis nula es

$$H_0: \mu = 4.13 \ \mu\text{g/l},$$

y la hipótesis alterna es

$$H_A: \mu \neq 4.13 \ \mu\text{g/l}.$$

Estamos interesados en desviaciones de la media que pudieran ocurrir en cualquier dirección; quisiéramos saber si μ es realmente mayor que 4.13 o menor. Por tanto, ejecutamos una prueba bilateral en el nivel de significancia $\alpha = 0.05$.

Como no conocemos la desviación estándar de la población σ, utilizamos una prueba t en lugar de una prueba z. El estadístico de prueba es

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}},$$

o

$$t = \frac{37.20 - 4.13}{7.13/\sqrt{10}}$$
$$= 14.67.$$

Si la hipótesis nula es verdadera, este resultado tiene una distribución t con $10 - 1 = 9$ gl. Al consultar la tabla A.4, observamos que el área total a la derecha de $t_9 = 14.67$ y a la izquier-

da de $t_9 = -14.67$ es menor que $2(0.0005) = 0.001$. Por tanto, $p < 0.05$, y rechazamos la hipótesis nula

$$H_0: \mu = 4.13 \ \mu g/l.$$

Esta muestra de niños ofrece evidencia de que el nivel medio de aluminio en la sangre de los niños que reciben antiácidos no es igual al nivel medio de aluminio de niños que no los reciben. De hecho, puesto que la media de muestreo $\bar{x}$ es mayor que μ_0, el verdadero nivel medio de aluminio es más alto que 4.13 μg/l.

10.3 Pruebas de hipótesis unilaterales

Antes de realizar una prueba de hipótesis, debemos decidir si nos interesan desviaciones de μ_0 que pudieran ocurrir en ambas direcciones —lo cual significa más altas o más bajas que μ_0— o en una sola dirección. Esta elección determina si consideramos el área en dos extremos de la curva apropiada cuando calculamos un valor p o el área en un solo extremo. La decisión debe tomarse **antes** de elegir una muestra aleatoria; en ésta no debería influir el resultado de la muestra. Si el conocimiento previo indica que μ no puede ser menor que μ_0, los únicos valores de $\bar{x}$ que proporcionarán evidencia contra la hipótesis nula

$$H_0: \mu = \mu_0$$

son aquellos que son mucho mayores que μ_0. En una situación así, la hipótesis nula se expresa más apropiadamente de la siguiente manera:

$$H_0: \mu \leq \mu_0$$

y la hipótesis alterna como

$$H_A: \mu > \mu_0.$$

Por ejemplo, la mayoría de la gente concordaría en que no es razonable creer que exponerse a una sustancia tóxica —como el monóxido de carbono o dióxido de azufre en la atmósfera— podría resultar benéfico para los seres humanos. Por tanto, sólo anticipamos efectos dañinos y realizamos una prueba unilateral. Una prueba bilateral siempre es la elección más cauta; en general, el valor p de una prueba bilateral es dos veces más amplio que el valor p de una prueba unilateral.

Considere la distribución de los niveles de hemoglobina de la población de niños menores de 6 años de edad que se han expuesto a niveles altos de plomo. Esta distribución tiene una media desconocida μ; se supone una desviación estándar $\sigma = 0.85$ g/100 ml [5]. Quizá desearíamos saber si el nivel medio de hemoglobina de esta población es igual a la media de la población general de niños menores de 6 años de edad, $\mu = 12.29$ g/100 ml. Creemos que si los niveles de hemoglobina de niños expuestos difieren de aquellos de niños que no se han expuesto, en promedio deben ser más bajos. Por tanto, sólo nos interesan las desviaciones de la media menores de μ_0. La hipótesis nula para la prueba es

$$H_0: \mu \geq 12.29 \text{ g/100 ml}$$

y la alterna unilateral es

$$H_A: \mu < 12.29 \text{ g/100 ml.}$$

H_0 se rechazaría para valores de $\bar{x}$ menores de 12.29, pero no para los más altos. Llevamos a cabo una prueba unilateral en el nivel de significancia $\alpha = 0.05$. Debido a que σ se conoce, utilizamos la distribución normal en lugar de la distribución t.

Una muestra aleatoria de 74 niños expuestos a niveles altos de plomo tiene un nivel medio de hemoglobina de $\bar{x} = 10.6$ g/100 ml. [6] Por tanto, el estadístico de prueba adecuado es

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

$$= \frac{10.6 - 12.29}{0.85/\sqrt{74}}$$

$$= -17.10.$$

De acuerdo con la tabla A.3, el área a la izquierda de $z = -17.10$ es menor de 0.001. Puesto que este valor p es menor que $\alpha = 0.05$, rechazamos la hipótesis nula

$$H_0: \mu \geq 12.29 \text{ g/100 ml}$$

en favor de la hipótesis alterna. Como ésta es una prueba unilateral, cualquier valor de z menor o igual al valor crítico -1.645 nos llevaría a rechazar la hipótesis nula. (Observe también que 12.29 se ubica arriba de 10.8, el límite de confianza superior unilateral de 95% para μ, calculado en el capítulo 9.)

En este ejemplo, se eligió H_0 para probar la afirmación de que el nivel medio de hemoglobina de la población de niños expuestos al plomo es la misma que la referente a la población general, 12.29 g/100 ml. Al rechazar H_0, concluimos que éste no es el caso; el nivel medio de hemoglobina de niños que se han expuesto al plomo es, de hecho, más bajo que la media para niños que no se han expuesto.

La elección entre una prueba unilateral y una bilateral puede resultar extremadamente polémica. Es frecuente que una prueba unilateral tenga significancia cuando una prueba bilateral no la consiga. En consecuencia, la decisión a menudo se efectúa sobre bases no científicas. En respuesta a esta práctica, algunos editores científicos se muestran reacios a publicar estudios que emplean pruebas unilaterales. Esto puede ser una reacción extrema a algo que el lector inteligente es capaz de discernir. En cualquier caso, no contribuiremos a ampliar la discusión sobre esta controversia.

10.4 Tipos de errores

Como se observó en la sección 10.1, se pueden cometer dos tipos de errores cuando llevamos a cabo una prueba de hipótesis. El primero se denomina *error de tipo I*; también se le conoce como *error de rechazo* o *error* α. Un error tipo I se comete si rechazamos la hipótesis nula

$$H_0: \mu = \mu_0$$

cuando H_0 es verdadera. La probabilidad de cometer un error tipo I se determina por el nivel de significancia de la prueba. Recuerde que

$$\alpha = \text{P(rechaza } H_0 \mid H_0 \text{ es verdadera).}$$

Si lleváramos a cabo repetidas pruebas de hipótesis independientes estableciendo el nivel de significancia en 0.05, rechazaríamos erróneamente una hipótesis nula verdadera 5% de las ocasiones.

Considere el caso de un fármaco que ha probado ser efectivo para reducir la presión arterial alta. Después de recibir un tratamiento con este medicamento durante un periodo determinado, una población de individuos que padece hipertensión tiene una presión arterial diastólica μ_d, valor clínicamente menor que la presión arterial diastólica media de individuos hipertensos que no se sometieron al tratamiento. Ahora suponga que otra compañía elabora una versión genérica del mismo medicamento. Quisiéramos saber si el medicamento genérico es tan efectivo para reducir la presión arterial alta como la versión de marca de fábrica. Para determinarlo, examinamos la distribución de presiones arteriales diastólicas de una muestra de individuos que han recibido tratamiento con el fármaco genérico. Si μ es la media de esta población, utilizamos la muestra para probar la hipótesis nula

$$H_0: \mu = \mu_d.$$

¿Qué hay si el fabricante del medicamento genérico en realidad somete el producto de marca de fábrica a una prueba en lugar de su propia versión? Se dice que Vitarine Pharmaceuticals, una compañía farmacéutica con sede en Nueva York, ha hecho esto presisamente [7]. La compañía ha efectuado sustituciones semejantes en cuatro diferentes ocasiones. En un caso como éste, sabemos que la hipótesis nula debe ser verdadera: estamos probando el medicamento que establece la norma. Por tanto, si la prueba de hipótesis nos lleva a rechazar H_0 y a sancionar que el medicamento "genérico" es más o menos eficaz que la versión de marca de fábrica, se ha cometido un error de tipo I.

El segundo tipo de error que puede cometerse en una prueba de hipótesis es un *error de tipo II,* también conocido como *error de aceptación o error ß.* Un error de tipo II se comete si erramos al rechazar la hipótesis nula

$$H_0: \mu = \mu_0$$

cuando H_0 es falsa. La probabilidad de cometer un error de tipo II se representa por medio de la letra griega β, donde

$$\beta = \text{P(no se rechaza } H_0 \mid H_0 \text{ es falsa).}$$

Si $\beta = 0.10$, por ejemplo, la probabilidad de que no rechacemos la hipótesis nula cuando $\mu \neq \mu_0$ es 0.10, o 10%. Los dos tipos de errores que pueden cometerse se resumen en seguida.

Resultado de la prueba	**Población**	
	$\mu = \mu_0$	$\mu \neq \mu_0$
No rechazada	Correcto	Error de tipo II
Rechazada	Error de tipo I	Correcto

Recuerde la distribución de niveles de colesterol en la sangre de todos los varones de 20 a 74 años de edad en Estados Unidos. La media de esta población es $\mu = 211$ mg/100 ml, y la desviación estándar es $\sigma = 46$ mg/100 ml. Suponga que no conocemos la verdadera media de esta población. No obstante, conocemos el hecho de que el nivel medio de colesterol en la sangre de la subpoblación de hombres de 20 a 24 años de edad es de 180 mg/100 ml. Puesto que los hombres mayores tienden a tener niveles de colesterol más altos que los más jóvenes en promedio, esperaríamos que el nivel medio de colesterol de la población de individuos de 20 a 74 años de edad fuera más alto que 180 mg/100 ml. (Y de hecho lo es, aunque pretendemos no saberlo.) Por tanto, si lleváramos a cabo una prueba unilateral de la hipótesis nula

$$H_0: \mu \leq 180 \text{ mg/100 ml}$$

contra la hipótesis alterna

$$H_A: \mu > 180 \text{ mg/100 ml},$$

esperaríamos que H_0 fuera rechazada. Sin embargo, es posible que no fuera así. La probabilidad de llegar a esta conclusión incorrecta —un error de tipo II— es β.

¿Cuál es el valor de β asociado a la prueba de la hipótesis nula

$$H_0: \mu \leq 180 \text{ mg/100 ml},$$

suponiendo que elegimos una muestra de tamaño 25? Para determinarlo, primero encontramos el nivel medio de colesterol en la sangre que nuestra muestra debe tener para que H_0 sea rechazada. Puesto que estamos llevando a cabo una prueba unilateral en el nivel de significancia $\alpha = 0.05$, H_0 sería rechazada para $z \geq 1.645$; éste es el valor crítico de la prueba. Al escribir el estadístico de prueba

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}},$$

tenemos

$$1.645 = \frac{\bar{x} - 180}{46/\sqrt{25}},$$

y, al resolver para $\bar{x}$,

$$\bar{x} = 180 + \frac{1.645(46)}{\sqrt{25}}$$
$$= 195.1.$$

Como se muestra en la figura 10.1, el área a la derecha de $\bar{x} = 195.1$ corresponde al 5% superior de la distribución del promedio muestral de muestras de tamaño 25 si $\mu = 180$. Por tanto, la hipótesis nula

$$H_0: \mu \leq 180 \text{ mg/100 ml}$$

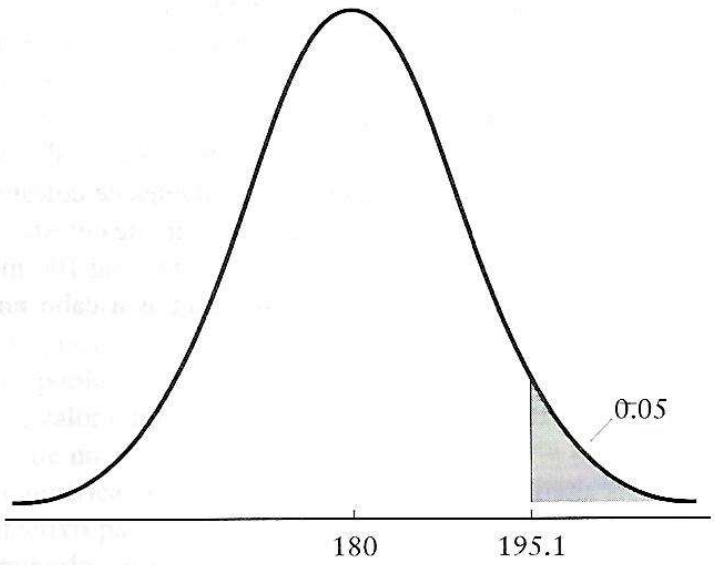

FIGURA 10.1
Distribución de medias de muestras de tamaño 25 para los niveles de colesterol de varones de 20 a 74 años de edad, $\mu = 180$ mg/100 ml.

sería rechazada si nuestra muestra tiene una media $\bar{x}$ mayor o igual a 195.1 mg/100 ml. Una muestra con una media menor no aportaría suficiente evidencia para rechazar H_0 en favor de H_A en el nivel de significancia de 0.05.

Recuerde que la probabilidad de cometer un error de tipo II, o β, es la probabilidad de no rechazar la hipótesis nula dado que H_0 es falsa. Por tanto, se trata de la posibilidad de obtener una media de muestreo menor que 195.1 mg/100 ml en caso de que la verdadera media de población no sea 180 sino $\mu_1 = 211$ mg/100 ml. Para determinar el valor de β, de nuevo consideramos la distribución del promedio muestral de muestras de tamaño 25; esta vez, sin embargo, tomamos $\mu = 211$. Esta distribución se muestra al lado derecho de la figura 10.2. Puesto que una media de muestreo menor que $\bar{x} = 195.1$ mg/100 ml implica que no rechazamos H_0, quisiéramos conocer la proporción de esta nueva distribución centrada en 211 mg/100 ml que se localiza debajo de 195.1. Observe que

$$z = \frac{195.1 - 211.0}{46/\sqrt{25}}$$

$$= -1.73.$$

El área bajo la curva normal estándar que se localiza a la izquierda de $z = -1.73$ es 0.042. Por tanto, β —la probabilidad de fallar en el rechazo de

$$H_0 : \mu \leq 180 \text{ mg/100 ml}$$

cuando la verdadera media poblacional es $\mu_1 = 211$ mg/100 ml — es igual a 0.042.

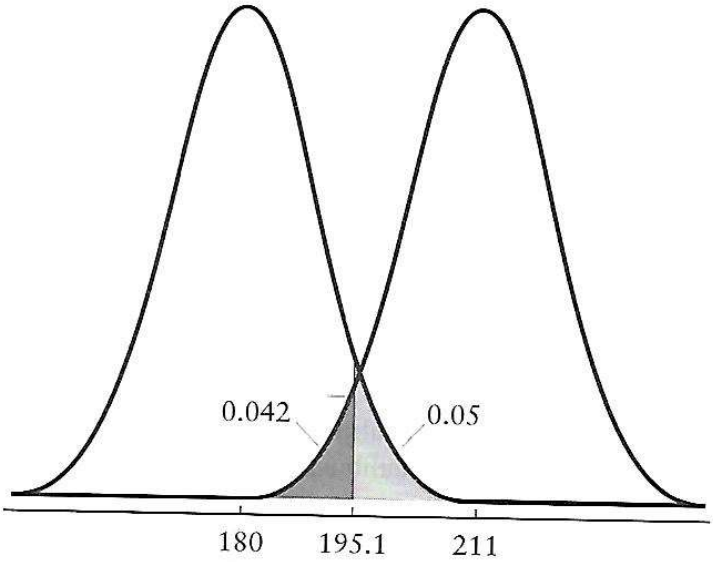

FIGURA 10.2
Distribución de medias de muestras de tamaño 25 para los niveles de colesterol en la sangre de varones de 20 a 74 años de edad, para $\mu = 180$ mg/100 ml y para $\mu = 211$ mg/100 ml.

Mientras que α, la probabilidad de cometer un error de tipo I, se determina observando el caso en el cual H_0 es verdadera y μ es igual a μ_0, β se determina al considerar la situación en la que H_0 es falsa y μ no es igual a μ_0. No obstante, si μ no es igual a μ_0, existe una infinidad de valores posibles que μ puede tomar. El error de tipo II se calcula para uno **solo de estos** valores, μ_1; en el ejemplo anterior, μ_1 se eligió con el valor de 211 mg/100 ml. (Elegimos 211 porque en este ejemplo poco común conocíamos que este valor era la media de población verdadera.) Si hubiésemos elegido una media de población diferente, habríamos calculado un valor diferente para β. Mientras más cerca se encuentre μ_1 de μ_0, más difícil será rechazar la hipótesis nula.

10.5 Eficiencia

Si β es la probabilidad de cometer un error de tipo II, $1 - \beta$ se denomina *eficiencia* de la prueba de hipótesis. La eficiencia es la probabilidad de rechazar la hipótesis nula cuando H_0 es falsa. En otras palabras, es la probabilidad de evitar el error de tipo II:

$$\text{eficiencia} = P(\text{rechazo de } H_0 \mid H_0 \text{ es falsa}).$$

La eficiencia puede también considerarse como la posibilidad de que un estudio particular detecte una desviación de la hipótesis nula en el supuesto de que exista. Igual que β, la eficiencia debe calcularse para una media de población alterna particular μ_1.

En el ejemplo anterior del nivel de colesterol en la sangre, la eficiencia de la prueba de hipótesis unilateral es

$$\begin{aligned} 1-\beta &= 1-0.042 \\ &= 0.958. \end{aligned}$$

En consecuencia, por cada prueba llevada a cabo en el nivel de significancia de 0.05 y con una muestra de tamaño 25, hay 95.8% de posibilidades de rechazar la hipótesis nula

$$H_0: \mu \leq 180 \text{ mg/100 ml}$$

en el supuesto de que H_0 sea falsa y que la verdadera media de población sea $\mu_1 = 211$ mg/100 ml. Observe que esto también podría haberse expresado de la siguiente manera:

$$\begin{aligned} \text{eficiencia} &= P(\text{rechazo } \mu \leq 180 \mid \mu = 211) \\ &= P(\bar{X} \geq 195.1 \mid \mu = 211) \\ &= P(Z \geq -1.73) \\ &= 1 - P(Z < -1.73) \\ &= 1 - 0.042 \\ &= 0.958. \end{aligned}$$

La cantidad $1-\beta$ habría tomado un valor diferente de haber igualado μ_1 a 200 mg/100 ml, y aun otro valor si hubiéramos supuesto que μ_1 es 220 mg/100 ml. Si trazamos una gráfica de los valores de $1-\beta$ en función de todos las medias de población alterna, terminaríamos con lo que se conoce como *curva de eficiencia.* La curva de eficiencia de la prueba de la hipótesis nula

$$H_0 : \mu \leq 180 \text{ mg/100 ml}$$

aparece en la figura 10.3. Observe que cuando $\mu_1 = 180$,

$$\begin{aligned} \text{eficiencia} &= P(\text{rechazo } \mu \leq 180 \mid \mu = 180) \\ &= P(\text{rechazo } \mu \leq 180 \mid H_0 \text{ es verdadera}) \\ &= \alpha \\ &= 0.05. \end{aligned}$$

La eficiencia de la prueba se aproxima a 1 conforme la media se aleja cada vez más del valor nulo de 180 mg/100 ml.

Los investigadores generalmente intentan diseñar pruebas de hipótesis de tal forma que consigan una alta eficiencia. No es suficiente saber que tenemos una pequeña probabilidad de rechazar H_0 cuando es verdadera; también quisiéramos contar con una alta probabilidad de rechazar la hipótesis nula cuando sea falsa. En la mayoría de las aplicaciones prácticas, una eficiencia menor a 80% se considera insuficiente. Una forma de incrementar la eficiencia de una prueba consiste en elevar el nivel de significancia α. Si incrementamos α, separamos una pequeña porción de la cola de la distribución de muestreo centrada en μ_1. De

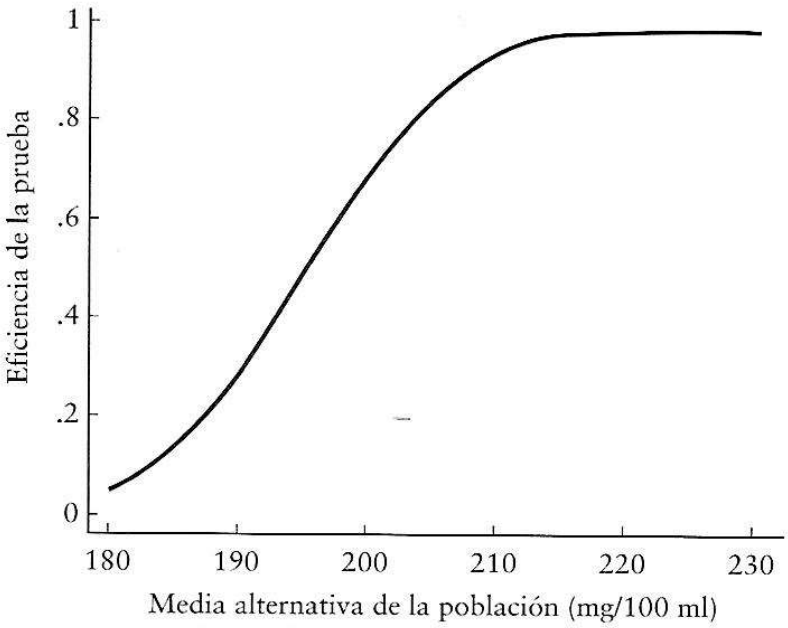

FIGURA 10.3
Curva de eficiencia de $\mu_0 = 180$, $\alpha = 0.05$ y $n = 25$.

forma correspondiente, β se reduce y la eficiencia, $1 - \beta$, aumenta. Si α hubiese sido igual a 0.10 para la prueba de la hipótesis nula

$$H_0: \mu \leq 180 \text{ mg/100 ml},$$

por ejemplo, β habría sido 0.018 y la eficiencia 0.982. Esta situación se ilustra en la figura 10.4. Compárese con la figura 10.2, donde α era igual a 0.05. Sin embargo, mantenga en mente el hecho de que si se eleva α, incrementamos la probabilidad de cometer un error de tipo I.

Este balance entre α y β es semejante al que se observa entre la sensibilidad y la especificidad de una prueba de diagnóstico. Recuerde que el incremento de la sensibilidad de una prueba automáticamente reduce su especificidad; al contrario, si se incrementa la especificidad, disminuye la sensibilidad. Lo mismo se puede decir de α y β. El equilibrio entre los dos tipos de error es delicado y su importancia relativa varía dependiendo de la época y la situación. En 1692, durante los procesos de brujería de Salem, Increase Mather publicó un sermón, firmado por él y otras 14 personas, donde decía [8]:

> Sería preferible que diez sospechosas de brujería escaparan a que una persona inocente fuera condenada.

En el siglo XVIII, Benjamín Franklin dijo,

> Es preferible que 100 culpables escapen a que un inocente sufra.

Sin embargo, más recientemente un editorial sobre el abuso de menores declaró que tiene "igual importancia" identificar y castigar a los abusadores de menores que exonerar a los que han sido acusados falsamente [9].

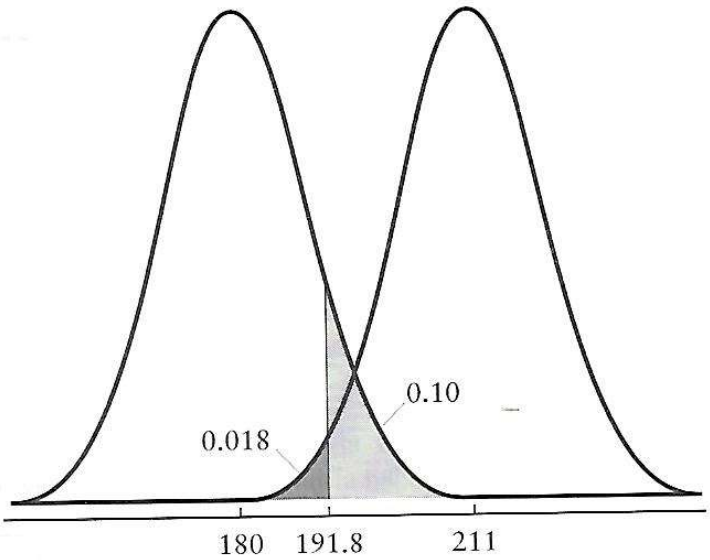

FIGURA 10.4
Distribuciones de medias de muestras de tamaño 25 para los niveles de colesterol en la sangre de hombres de 20 a 74 años de edad, para μ = 180 mg/100 ml y para μ = 211 mg/100 ml.

Cuanta más información tengamos —es decir cuanto más amplia sea nuestra muestra—, menos probabilidad tenemos de cometer un error de cualquier tipo. No obstante, sin importar nuestra decisión, siempre existe la posibilidad de haber cometido un error.

El único camino para disminuir α y β simultáneamente consiste en reducir el grado de superposición en las dos distribuciones normales: una centrada en μ_0 y la otra en μ_1. Una forma en que esto se puede llevar a cabo consiste en considerar sólo desviaciones grandes de μ_0. Mientras más se aparten los valores de μ_0 y μ_1, mayor es la eficiencia de la prueba. Una alternativa consiste en incrementar el tamaño n de la muestra. Si se incrementa n, el error estándar $\sigma/\sqrt{n}$ disminuye; esto provoca que las dos distribuciones de muestreo se tornen más estrechas, lo que, a su vez, reduce el grado de superposición. El error estándar también disminuye si reducimos la desviación estándar de la población estudiada σ, pero esto normalmente no es posible. Otra opción que aún no mencionamos consiste en determinar un estadístico de prueba "más eficiente". Este tema se analiza con más detalles en el capítulo 13.

10.6 Estimación del tamaño de muestra

En la sección anterior bosquejamos un método para calcular la eficiencia de una prueba llevada a cabo en el nivel α con una muestra de tamaño n. No obstante, en las primeras etapas de planeación de un estudio, los investigadores normalmente desean invertir la situación y determinar el tamaño de muestra necesario para proporcionar una eficiencia específica. Por ejemplo, supongamos que deseamos probar la hipótesis nula

$$H_0: \mu \leq 180 \text{ mg/100 ml}$$

en el nivel de significancia $\alpha = 0.01$. Una vez más, μ es el nivel medio de colesterol en la sangre de la población de varones de 20 a 74 años de edad en Estados Unidos; la desviación estándar es $\sigma = 46$ mg/100 ml. Si la verdadera media poblacional tiene un valor de 211 mg/100 ml, deseamos arriesgarnos sólo 5% de posibilidades de fallar en el rechazo de la hipótesis nula; en consecuencia, igualamos β a 0.05 y la eficiencia de la prueba a 0.95. En estas circunstancias, ¿qué tamaño de muestra se requiere?

Puesto que $\alpha = 0.01$ en lugar de 0.05, comenzamos por observar que H_0 sería rechazada para $z \geq 2.32$. Al Sustituir $(\bar{x} - 180)/(46/\sqrt{n})$ en lugar de la desviación normal z, establecemos que

$$z = 2.32$$
$$= \frac{\bar{x} - 180}{46/\sqrt{n}}.$$

Al resolver para $\bar{x}$,

$$\bar{x} = 180 + 2.32\left(\frac{46}{\sqrt{n}}\right).$$

Por tanto, rechazaríamos la hipótesis nula si la media de muestreo $\bar{x}$ tomara cualquier valor mayor o igual a $180 + 2.32(46/\sqrt{n})$.

Ahora considere la eficiencia de la prueba que se busca. Si el verdadero nivel medio de colesterol en la sangre fuera en realidad $\mu_1 = 211$ mg/100 ml —de tal forma que la desviación normal z pudiera expresarse como $(\bar{x} - 211)/(46/\sqrt{n})$—, querríamos rechazar la hipótesis nula con una probabilidad de $1 - \beta = 1 - 0.05 = 0.95$. El valor de z que corresponde a $\beta = 0.05$ es $z = -1.645$; por tanto,

$$z = -1.645$$
$$= \frac{\bar{x} - 211}{46/\sqrt{n}}$$

y

$$\bar{x} = 211 - 1.645\left(\frac{46}{\sqrt{n}}\right).$$

Al igual las dos expresiones para la media de muestreo $\bar{x}$,

$$180 + 2.32\left(\frac{46}{\sqrt{n}}\right) = 211 - 1.645\left(\frac{46}{\sqrt{n}}\right).$$

Al multiplicar ambos miembros de la igualdad por $\sqrt{n}$ y agrupar términos,

$$\sqrt{n}(211 - 180) = [2.32 - (-1.645)](46)$$

y

$$n = \left[\frac{(2.32 + 1.645)(46)}{(211 - 180)}\right]^2$$
$$= 34.6.$$

Por convención, siempre redondeamos hacia arriba las cifras en el cálculo del tamaño de muestra. Por tanto, se requeriría una muestra de 35 hombres.

Con la notación introducida en el capítulo 9, es posible escribir una fórmula más general para calcular el tamaño de muestra. Recuerde que z_α representa el valor que divide un área de α en el extremo superior de la distribución, mientras que $-z_\alpha$ es el valor que separa un área de α en el extremo inferior de la distribución. Si llevamos a cabo una prueba unilateral de la hipótesis nula

$$H_0: \mu \leq \mu_0$$

contra la alternativa

$$H_0: \mu > \mu_0$$

en el nivel de significancia α, H_0 sería rechazada por cualquier estadístico de prueba que tomara un valor $z \geq z_\alpha$. De igual forma, si consideramos la eficiencia deseada de la prueba $1 - \beta$, el valor genérico de z que corresponde a una probabilidad β es $z = -z_\beta$. Las dos diferentes expresiones para $\bar{x}$ son

$$\bar{x} = \mu_0 + z_\alpha\left(\frac{\sigma}{\sqrt{n}}\right)$$

y

$$\bar{x} = \mu_1 - z_\beta\left(\frac{\sigma}{\sqrt{n}}\right),$$

y al igualarlos obtenemos

$$n = \left[\frac{[z_\alpha - (-z_\beta)](\sigma)}{(\mu_1 - \mu_0)}\right]^2 = \left[\frac{(z_\alpha + z_\beta)(\sigma)}{(\mu_1 - \mu_0)}\right]^2.$$

Este es el tamaño de muestra necesario para conseguir una eficiencia de $1 - \beta$ cuando llevamos a cabo una prueba unilateral en el nivel α.

Varios factores influyen en el tamaño de n. Si reducimos el error α de tipo I, entonces z_α —el punto de separación para rechazar H_0— se incrementaría en valor; esto daría como resultado un tamaño de muestra mayor. De forma similar, si reducimos el error de tipo β o si incrementamos la eficiencia, entonces $-z_\beta$ se reduce, es decir, se torna más negativo. De nuevo, esto daría como resultado un valor de n mayor. Si consideramos una media de población alterna más próxima al valor hipotético, la diferencia $\mu_1 - \mu_0$ disminuiría y el tamaño de la muestra aumentaría. Tiene sentido el hecho de que necesitemos un tamaño de muestra mayor para detectar una diferencia menor. Finalmente, mientras mayor sea la variabilidad de la población estudiada σ, mayor será el tamaño de la muestra que se requiere.

En el ejemplo del nivel de colesterol en la sangre, sabíamos que la media de población hipotética μ_0 tenía que ser menor que la media alterna μ_1. En consecuencia, llevamos a cabo

una prueba unilateral. Si no se sabe si μ_0 es mayor o menor que μ_1, una prueba bilateral resulta adecuada. En este caso, debemos modificar el valor crítico de z que provocaría el rechazo de la hipótesis nula. Por ejemplo, cuando $\alpha = 0.01$,

$$H_0: \mu = 180 \text{ mg/100 ml}$$

sería rechazada para $z \geq 2.58$, no para $z \geq 2.32$. Al sustituir este valor en la ecuación anterior,

$$n = \left[\frac{(2.58 + 1.645)(46)}{(211 - 180)}\right]^2$$
$$= 39.3,$$

y donde se requeriría una muestra de tamaño 40. Más generalmente, H_0 será rechazada al nivel α para $z \geq z_{\alpha/2}$ (y también para $z \leq -z_{\alpha/2}$, y la fórmula del tamaño de muestra se convierte en

$$n = \left[\frac{(z_{\alpha/2} + z_{\beta})(\sigma)}{(\mu_1 - \mu_0)}\right]^2.$$

Observe que el tamaño de muestra para una prueba bilateral siempre es mayor que el tamaño de muestra para la prueba unilateral correspondiente.

10.7 Aplicaciones adicionales

Considere nuevamente la distribución de alturas para la población de individuos de 12 a 40 años de edad que padecen del síndrome de alcoholismo fetal. Esta distribución es aproximadamente normal con una media μ desconocida; su desviación estándar es $\sigma = 6$ centímetros [10]. Quizá deseemos conocer si la estatura media de esta población es igual a la estatura media de individuos en el mismo grupo por edades que no padecen el síndrome de alcoholismo fetal.

El primer paso al llevar a cabo una prueba de hipótesis consiste en hacer una declaración formal sobre el valor de μ_0. Puesto que la estatura media de los individuos de 12 a 40 años de edad que no padecen el síndrome de alcoholismo fetal es de 160.0 centímetros, la hipótesis nula es

$$H_0: \mu = 160.0 \text{ cm.}$$

Nos interesan las desviaciones de μ_0 que pudieran ocurrir en cualquier dirección; así, llevamos a cabo una prueba bilateral en el nivel de significancia $\alpha = 0.05$. La hipótesis alterna es

$$H_A: \mu \neq 160.0 \text{ cm.}$$

Para una muestra aleatoria de tamaño 31 elegida de la población de individuos de 12 a 40 años que padecen el síndrome de alcoholismo fetal, la estatura media es $\bar{x} = 147.4$ cm. Si

la verdadera estatura media de este grupo es $\mu = 160.0$ cm, ¿cuál es la probabilidad de elegir una muestra con una media tan baja como 147.4? Para responder esta pregunta, calculamos el estadístico de prueba

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}},$$

$$= \frac{147.4 - 160.0}{6/\sqrt{31}}$$

$$= -11.69.$$

Utilizamos una prueba z en lugar de una prueba t ya que el valor de σ se conoce. Puesto que z es el resultado de una variable aleatoria normal estándar, consultamos la tabla A.3 y encontramos que el área a la izquierda de $z = -11.69$ y a la derecha de $z = 11.69$ es mucho menos de 0.001. Por tanto, ya que $p < 0.05$, rechazamos la hipótesis nula en el nivel de significancia 0.05. La muestra aleatoria proporciona evidencia de que la estatura media de la población de individuos que padecen el síndrome de alcoholismo fetal es diferente de la estatura media de la población de individuos que no lo padecen; las personas que padecen este mal tienden a ser más bajos en promedio.

En lugar de realizar nosotros mismos los cálculos, podríamos haber usado una computadora para la prueba de la hipótesis. La Tabla 10.1 muestra la salida pertinente de Minitab, la salida repite las hipótesis nula y alternativa y la desviación estándar de población supuesta σ; enseguida proporciona varias medidas de resumen, el estadístico de prueba z y el valor p de la prueba. Sin embargo, Minitab no nos ofrece ninguna conclusión: eso lo tenemos que hacer nosotros mismos.

Otra forma de abordar el problema habría sido construir un intervalo de confianza para μ, la verdadera estatura media de la población de individuos de 12 a 40 años que padecen el síndrome de alcoholismo fetal. En el capítulo 9 vimos que un intervalo de confianza bilateral de 95% para μ era

(145.3, 149.5).

Puesto que este intervalo no contiene el valor 160.0, sabemos que la hipótesis nula

H_0: $\mu = 160.0$ cm

sería rechazada en favor de H_A en el nivel de significancia 0.05.

Cuando la desviación estándar de una población no se conoce, empleamos la desviación estándar s en lugar de σ para llevar a cabo una prueba de hipótesis. Considere la distribución de la concentración de benceno —una sustancia que se supone dañina para los seres

TABLA 10.1
Salida de Minitab para la prueba z

```
TEST OF MU = 160.0 VS MU N.E. 160.0
THE ASSUMED SIGMA = 6.0

            N     MEAN    STDEV   SE MEAN        Z      P VALUE
HEIGHT     31    147.4    6.000     1.078   -11.69        0.000
```

humanos— en una marca determinada de puros. Esta distribución es aproximadamente normal con una media μ desconocida y una desviación estándar σ. Se nos dice que la concentración media de benceno en una marca de cigarillos que se emplea como norma es de 81 μg/g de tabaco [11], y quisiéramos saber si la concentración media de benceno en los puros es igual a la de los cigarrillos. Para determinar esto, probamos la hipótesis nula

$$H_0: \mu = 81 \ \mu\text{g/g}.$$

Estamos interesados en desviaciones de la media que pudieran ocurrir en cualquier dirección, así que llevamos a cabo una prueba bilateral en el nivel de significancia $\alpha = 0.05$. La hipótesis alterna es

$$H_A: \mu \neq 81 \ \mu\text{g/g}.$$

Una muestra aleatoria de siete puros tiene una concentración media de benceno de $\bar{x} = 151$ μg/g y una desviación estándar de $s = 9$ μg/g. ¿Es posible que estas observaciones pudieran provenir de una población con media $\mu = 81$ μg/g? Para determinar esto, calculamos el estadístico de prueba

$$\begin{aligned} t &= \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \\ &= \frac{151 - 81}{9/\sqrt{7}} \\ &= 20.6. \end{aligned}$$

El estadístico t es el resultado de una variable aleatoria que tiene una distribución t con $7 - 1 = 6$ grados de libertad. Al consultar la tabla A.4, observamos que el área total bajo la curva ubicada a la izquierda de -20.6 y a la derecha de 20.6 es de menos de 0.001. Puesto que $p < 0.05$, rechazamos la hipótesis nula. La muestra aleatoria de tamaño 7 sugiere que los puros contienen una concentración de benceno más alta que la que tienen los cigarrillos.

La tabla 10.2 muestra la salida relevante de Stata. Observe que la porción inferior de la salida proporciona el estadístico de prueba y valores p para tres diferentes hipótesis alternas. La información en el centro es para la prueba bilateral, mientras que la información en cualquiera de los lados es para las dos posibles pruebas unilaterales. (Recuerde que debimos haber determinado con anticipación qué prueba en particular nos interesaba.) Además, Stata presenta el intervalo de confianza de 95% para la verdadera media de población μ.

Ahora recuerde la distribución de niveles de hemoglobina en la población de niños menores de 6 años que se expusieron a niveles altos de plomo. La media de esta población es $\mu = 10.60$ g/100 ml, y su desviación estándar es $\sigma = 0.85$ g/100 ml. Supongamos que no conocemos la verdadera media de población μ; no obstante, sabemos que el nivel medio de hemoglobina de la población general de niños menores de 6 años es de 12.29 g/100 ml. Si lleváramos a cabo una prueba de la hipótesis nula

$$H_0: \mu = 12.29 \text{ g/100 ml},$$

esperaríamos que esta hipótesis falsa fuera rechazada. Suponiendo que elegimos una muestra aleatoria muy pequeña de tamaño 5 de la población de niños expuestos al plomo, ¿cuál es

TABLA 10.2
Salida de Stata para la prueba *t*

```
Prueba t de una muestra                                  Number of obs = 7

------------------------------------------------------------------------------
Variable |  Mean   Std. Err.      t    P>|t|    [95% Conf. Interval]
---------+--------------------------------------------------------------------
benceno  |   151    3.40168   44.3898   0.0000    142.6764   159.3236
------------------------------------------------------------------------------
Degrees of freedom: 6

                    Ho: mean(benzene) = 81

Ha: mean < 81               Ha: mean ~= 81               Ha: mean > 81
    t = 20.5781                 t = 20.5781                  t = 20.5781
P < t = 1.0000            P > |t| = 0.0000             P > t = 0.0000
```

la probabilidad de que cometamos un error de tipo II —fallemos en rechazar H_0 cuando es falsa —supuesto que la verdadera media poblacional es $\mu_1 = 10.60$ g/100 ml?

Para responder esta pregunta, comenzamos por determinar el nivel medio de hemoglobina que la muestra debe tener para que H_0 sea rechazada. Creemos que los niveles de hemoglobina de niños expuestos al plomo en promedio deben ser más bajos que los de niños no expuestos. Si llevamos a cabo una prueba unilateral en el nivel de significancia $\alpha = 0.05$, la hipótesis nula sería rechazada para $z \leq -1.645$. Puesto que

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}},$$

tenemos

$$z = -1.645$$
$$= \frac{\bar{x} - 12.29}{0.85/\sqrt{5}}$$

y

$$\bar{x} = 12.29 - \frac{1.645(0.85)}{\sqrt{5}}$$
$$= 11.66.$$

Por tanto, la hipótesis nula

$$H_0: \mu \geq 12.29 \text{ g/100 ml}$$

sería rechazada en favor de la hipótesis alterna

$$H_A: \mu < 12.29 \text{ g/100 ml}$$

si la muestra de tamaño 5 tiene una media $\bar{x}$ que es menor o igual a 11.66 g/100 ml. Esta área corresponde al 5% inferior de la distribución del promedio muestral de muestras de tamaño 5 cuando $\mu = 12.29$ g/100 ml. Este hecho se ilustra en la figura 10.5.

La cantidad β es la probabilidad de cometer un error de tipo II, es decir, de fallar en el rechazo de H_0 en el supuesto de que sea falsa y de que la verdadera media poblacional sea $\mu_1 = 10.60$ g/100 ml. Para determinar β, consideramos la distribución del promedio muestral de muestras de tamaño 5 cuando $\mu = 10.60$ g/100 ml. Puesto que una media de muestreo mayor que 11.66 g/100 ml implica que no rechazamos H_0, debemos determinar qué proporción de la distribución centrada en 10.60 g/100 ml se ubica a la derecha de 11.66. Observe que

$$z = \frac{11.66 - 10.60}{0.85/\sqrt{5}}$$

$$= 2.79.$$

Al consultar la tabla A.3, encontramos que el área bajo la curva normal estándar que se localiza a la derecha de $z = 2.79$ es 0.003. Por tanto, β es igual a 0.003.

La eficiencia de la prueba —la probabilidad de rechazar la hipótesis nula dado que H_0 es falsa y que la verdadera media de población es $\mu_1 = 10.60$ g/100 ml— es

$$1 - \beta = 1 - 0.003$$
$$= 0.997.$$

Aun con un tamaño de muestra de sólo 5, casi estamos seguros de rechazar H_0. Esto se debe, en parte, a que la desviación estándar de la población estudiada es muy pequeña en comparación con la diferencia de las medias $\mu_1 - \mu_0$.

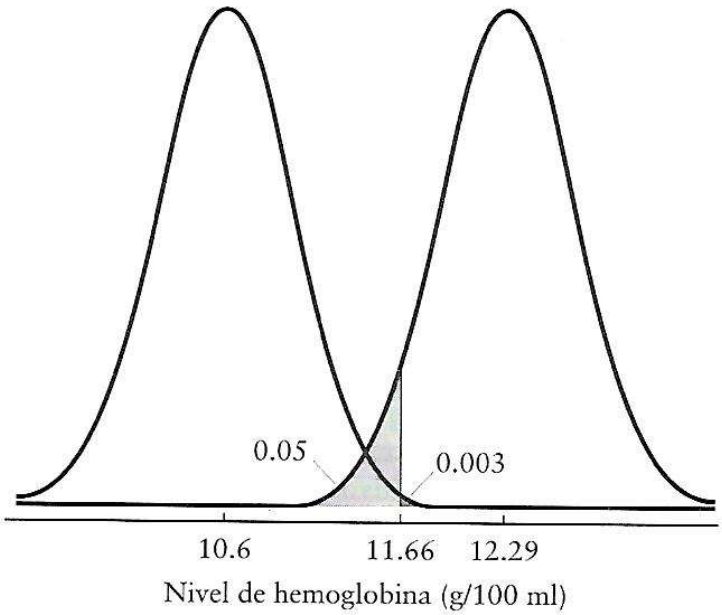

FIGURA 10.5
Distribución de medias de muestras de tamaño 5 de los niveles de hemoglobina de niños menores de 6 años, para $\mu = 10.60$ g/100 ml y para $\mu = 12.29$ g/100 ml.

Supongamos ahora que planeamos un nuevo estudio para intentar determinar el nivel medio de hemoglobina para la población de niños de menos de 6 años de edad que se han expuesto a niveles altos de plomo. Se nos dice que la población general de niños en este grupo de edades tiene un nivel medio de hemoglobina $\mu = 12.29$ g/100 ml y una desviación estándar $\sigma = 0.85$ g/100 ml. Si el verdadero nivel medio de hemoglobina para niños expuestos es de 0.5 g/100 ml más bajo que el de niños no expuestos, deseamos que la prueba tenga una eficiencia de 80% para detectar esta diferencia. Planeamos llevar a cabo una prueba unilateral con un nivel de significancia de 0.05. ¿Qué tamaño de muestra se requerirá en este estudio?

Comenzamos por reunir todas las piezas necesarias para efectuar un cálculo del tamaño de la muestra. Puesto que se llevará a cabo una prueba unilateral en el nivel $\alpha = 0.05$, $z_\alpha = 1.645$. Deseamos una eficiencia de 0.80; por tanto, $\beta = 0.20$ y $z_\beta = 0.84$. La media hipotética de la población es $\mu_0 = 12.29$ g/100 ml, y la media alterna es 0.5 unidades menor que esto, o $\mu_1 = 11.79$ g/100 ml. No conocemos la desviación estándar de los niveles de hemoglobina de niños expuestos; sin embargo, estamos dispuestos a suponer que se trata de la misma que la de los niños no expuestos. Por tanto, $\sigma = 0.85$ g/100 ml. Al unir las piezas,

$$\begin{aligned} n &= \left[\frac{(z_\alpha + z_\beta)(\sigma)}{(\mu_1 - \mu_0)}\right]^2 \\ &= \left[\frac{(1.645 + 0.84)(0.85)}{(11.79 - 12.29)}\right]^2 \\ &= 17.8. \end{aligned}$$

Así, se requeriría una muestra de tamaño 18.

10.8 Ejercicios de repaso

1. ¿Cuál es el propósito de una prueba de hipótesis?
2. ¿Demuestra una prueba de hipótesis la hipótesis nula? Explique.
3. ¿Qué es un valor *p*? ¿Qué significa en palabras valor *p*?
4. Explique brevemente la relación entre intervalos de confianza y prueba de hipótesis.
5. ¿En qué circunstancias podría usted emplear una prueba unilateral de hipótesis en lugar de una prueba bilateral?
6. Describa los dos tipos de errores que pueden cometerse al llevar a cabo una prueba de hipótesis.
7. Explique la analogía entre el tipo de error I y el tipo de error II en una prueba de hipótesis y los resultados positivo falso y negativo falso que ocurren en las pruebas de diagnóstico.
8. Enliste cuatro factores que afectan la eficiencia de una prueba.
9. La distribución de presiones arteriales diastólicas para la población de mujeres diabéticas entre las edades de 30 y 34 años tiene una media desconocida μ_d y una desviación estándar $\sigma_d = 9.1$ mm Hg. Puede resultar útil a los médicos conocer si la media de esta

población es igual a la presión arterial diastólica media de la población general de mujeres en este grupo por edades, 74.4 mm Hg [12].

a) ¿Cuál es la hipótesis nula de la prueba adecuada?

b) ¿Cuál es la hipótesis alterna?

c) Se elige una muestra de diez mujeres diabéticas, cuya presión arterial diastólica es $\bar{x}_d = 84$ mm Hg. Con esta información, lleve a cabo una prueba bilateral en el nivel de significancia $\alpha = 0.05$. ¿Cuál es el valor p de la prueba?

d) ¿Qué concluye usted de los resultados de la prueba?

e) ¿Habría sido su conclusión diferente de haber elegido $\alpha = 0.01$ en lugar de $\alpha = 0.05$?

10. La infección por *E. canis* es una enfermedad canina transmitida por la garrapata, que algunas veces contraen los seres humanos. Entre los humanos infectados, la distribución de recuentos de glóbulos blancos tiene una media μ y una desviación estándar σ desconocidas. En la población general, el recuento medio de glóbulos blancos es 7250/mm^3 [13]. Se cree que las personas infectadas con *E. canis* deben tener en promedio recuentos de glóbulos blancos más bajos.

a) ¿Cuál es la hipótesis nula y la hipótesis alterna para una prueba unilateral?

b) Para una muestra de 15 personas infectadas, el recuento medio de glóbulos blancos es de $\bar{x} = 4767$/mm^3 y la desviación estándar es $s = 3204$/mm^3 [14]. Lleve a cabo la prueba en el nivel de significancia = 0.05.

c) ¿Qué concluye usted?

11. El índice de masa corporal se calcula dividiendo el peso del individuo entre el cuadrado de su estatura. Se trata de una medida de la cantidad de sobrepeso de una persona. Para la población de hombres de mediana edad que más tarde contrajeron diabetes mellitus, la distribución de índices de masa corporal representativa es aproximadamente normal con media μ y desviación estándar σ desconocidos. Una muestra de 58 hombres elegidos de este grupo tiene una media $\bar{x} = 25.0$ Kg/m^2 y una desviación estándar $s = 2.7$ kg/m^2 [15].

a) Construya un intervalo de confianza de 95% para la media de población μ.

b) En el nivel de significancia de 0.05, pruebe si el índice de masa corporal medio representativo de la población de hombres de mediana edad que contrajeron diabetes es igual a 24.0 kg/m^2, la media para la población de hombres que no la contrajeron. ¿Cuál es el valor p de la prueba?

c) ¿Qué concluye usted?

d) Con base en un intervalo de confianza de 95%, ¿esperaría usted rechazar o no la hipótesis nula? ¿Por qué?

12. La población de trabajadores industriales varones en Londres que jamás han padecido un problema coronario serio tiene una presión arterial sistólica de 136 mm Hg y una presión arterial diastólica media de 84 mm Hg [16]. Quizá le interesaría determinar si estos valores son los mismos que los de la población de trabajadores de la industria que **han** padecido un problema coronario.

a) Una muestra de 86 trabajadores que han sufrido un problema coronario serio tiene una presión arterial sistólica media de $\bar{x}_s = 143$ mm Hg y una desviación estándar de $s_s = 24.4$ mm Hg. Pruebe la hipótesis nula de que la presión arterial sistólica media para la población de trabajadores industriales que han padecido dicho problema es idéntica a la media de los trabajadores que no lo ha padecido, con una prueba bilateral en el nivel $\alpha = 0.10$.

b) La misma muestra de hombres tiene una presión arterial diastólica media de $\bar{x}_d = 87$ mm Hg una y desviación estándar de $s_d = 16.0$ mm Hg. Pruebe la hipótesis nula de que la presión arterial diastólica media para la población de trabajadores que han padecido un problema coronario serio es idéntica a la media para los trabajadores que no lo han padecido.

c) ¿Cómo se comparan los dos grupos de trabajadores?

13. A través de los años, la Food and Drugs Administration de Estados Unidos (FDA) ha trabajado arduamente para evitar cometer errores de tipo II. Un error de tipo II sucede cuando la FDA aprueba un medicamento que no es seguro y efectivo al mismo tiempo. No obstante, pese a los esfuerzos de la dependencia, los malos medicamentos se filtran a veces hasta el público. Por ejemplo, Omniflox, un antibiótico, tuvo que retirarse menos de seis meses después de ser aprobado debido a informes de reacciones adversas severas, que incluían la muerte. Asimismo, el Fenoterol, un medicamento para inhalar que pretendía aliviar los ataques de asma, incrementaba el riesgo de muerte en lugar de reducirlo [17]. ¿Hay alguna forma en que la FDA pueda eliminar completamente errores de tipo II? Explique.

14. Los datos del estudio Framingham nos permiten comparar las distribuciones de niveles iniciales de colesterol en la sangre de dos poblaciones de hombres: los que contraen enfermedades coronarias y los que no. El nivel medio de colesterol en la sangre de la población masculina que no desarrolla enfermedades coronarias es $\mu = 219$ mg/100 ml y la desviación estándar es $\sigma = 41$ mg/100 ml [18]. Sin embargo, suponga que usted no conoce la verdadera media de población; en su lugar, usted supone que μ es igual a 244 mg/100 ml. Éste es el nivel medio inicial de colesterol en la sangre de hombres que finalmente contraen la enfermedad. Puesto que se supone que el nivel medio de colesterol en la sangre de varones que no desarrollan enfermedades coronarias no puede ser más alto que el nivel medio de quienes las contraen, resulta apropiado llevar a cabo una prueba unilateral con nivel de significancia $\alpha = 0.05$.

a) ¿Cuál es la probabilidad de cometer un error de tipo I?

b) Si se elige una muestra de tamaño 25 de una población de hombres que no desarrollan alguna enfermedad coronaria, ¿cuál es la probabilidad de cometer un error de tipo II?

c) ¿Cuál es la eficiencia de la prueba?

d) ¿Podría usted incrementar la eficiencia?

e) Usted desea probar la hipótesis nula

H_0: $= \mu \geq 244$ mg/100 ml

en función de la hipótesis alterna

H_A: $= \mu < 244$ mg/100 ml

en el nivel de significancia $\alpha = 0.05$. Si la verdadera media de población es tan baja como 219 mg/100 ml, usted desea arriesgar sólo 5% de posibilidades de fallar en el rechazo de H_0. ¿Qué tamaño de muestra se requiere?

f) ¿Cómo cambiaría el tamaño de la muestra si usted deseara arriesgar 10% de posibilidades de fallar en el rechazo de la hipótesis nula falsa?

15. En Noruega, la distribución de pesos al nacer de niños que cumplen su periodo de gestación de 40 semanas es aproximadamente normal con una media de $\mu = 3500$ gramos y

una desviación estándar de $\sigma = 430$ gramos [19]. Un investigador planea llevar a cabo un estudio para determinar si los pesos de niños al nacer que llegan al término del embarazo cuyas madres fumaban durante ese periodo tienen la misma media. Si el verdadero peso medio al nacer de niños cuyas madres fumaban es tan bajo como 3200 gramos (o tan alto como 3800 gramos), el investigador desea arriesgar sólo 10% de posibilidades de fallar en detectar esta diferencia. Se aplicará una prueba bilateral con un nivel de significancia 0.05. ¿Qué tamaño de muestra se requiere en este estudio?

16. Las escalas de Bayley del desarrollo infantil arrojan resultados en dos índices —el índice de desarrollo psicomotor (Psichomotor Development Index, PDI) y el índice de desarrollo mental (Mental Development Index, MDI) que pueden emplearse para evaluar el grado del niño respecto del funcionamiento de cada una de estas áreas a la edad aproximada de un año. Entre los niños saludables, ambos índices tienen un valor medio de 100. Como parte de un estudio para evaluar el desarrollo y estado neurológico de niños sometidos a cirugía reconstructiva de corazón durante los primeros tres meses de vida, las escalas de Bayley se aplicaron a una muestra de niños de un año de edad con enfermedades de corazón congénitas. Los datos se encuentran en la serie de datos `heart` [20] (apéndice B, tabla B.12); los resultados de PDI se almacenan en la variable denominada `pdi`, mientras que los resultados MDI, en la variable `mdi`.

a) En el nivel de significancia 0.05, pruebe la hipótesis nula de que el resultado medio PDI de niños nacidos con alguna enfermedad congénita de corazón que se sometieron a cirugía reconstructiva de corazón durante los primeros tres meses de vida es igual a 100, el resultado medio para niños saludables. Utilice una prueba bilateral. ¿Cuál es el valor *p*? ¿Qué concluye usted?

b) Lleve a cabo la prueba de hipótesis análoga para el resultado medio de MDI. ¿Qué concluye usted?

c) Construya intervalos de confianza de 95% del verdadero resultado medio de PDI y el verdadero resultado medio de MDI para esta población de niños con alguna enfermedad de corazón congénita. ¿Contiene cualquiera de estos intervalos el valor 100? ¿Habría esperado usted que así fuera?

Bibliografía

[1] National Center for Health Statistics, Fulwood, R., Kalsbeek, W., Rifkind, B., Russell-Briefel, R., Muesing, R., La Rosa, J. y K. Lippel, "Total Serum Cholesterol Levels of Adults 20-74 Years of Age: United States, 1976-1980", *Vital and Health Statistics,* serie 11, núm. 236, mayo de 1986.

[2] Gauvreau, K. y M. Pagano, "Why 5%?", *Nutrition,* vol. 10, 1994, pp. 93-94.

[3] Kaplan, N. M., "Strategies to Reduce Risk Factors in Hypertensive Patients Who Smoke", *American Heart Journal,* vol. 115, enero de 1988, pp. 288-294.

[4] Tsou, V. M., Young, R. M., Hart, M. H. y J. A. Vanderhoof, "Elevated Plasma Aluminum Levels in Normal Infants Receiving Antacids Containing Aluminum", *Pediatrics,* vol. 87, febrero de 1991, pp. 148-151.

[5] National Center for Health Statistics, Fulwood, R., Johnson, C. L., Bryner, J. D., Gunter, E. W. y C. R. McGrath, "Hematological and Nutritional Biochemistry Reference Data for Persons 6 Months-74 Years of Age: United States, 1976-1980", *Vital and Health Statistics,* serie 11, núm. 232, diciembre de 1982.

[6] Clark, M., Royal, J. y R. Seeler, "Interaction of Iron Deficiency and Lead and the Hematologic Findings in Children with Severe Lead Poisoning", *Pediatrics,* vol. 81, febrero de 1988, pp. 247-253.

[7] "Firm Admits Usign Rival's Drug in Tests", *The Boston Globe,* 1 de julio de 1989, p. 41.

[8] Davidson, J. W. y M. H. Lytle, *After the Fact: The Art of Historical Detection,* vol. 1, Nueva York, McGraw-Hill, 1992, p. 26.

[9] "Child Abuse—and Trial Abuse", *The New York Times,* 20 de enero de 1990, p. 24.

[10] Streissguth, A. P., Aase, J. M., Clarren, S. K., Randels, S. P., LaDue, R. A. y D. F. Smith, "Fetal Alcohol Syndrome in Adolescents and Adults", *Journal of the American Medical Association,* vol. 265, 17 de abril de 1991, pp. 1961-1967.

[11] Appel, B. R., Guirguis, G., Kim, I., Garbin, O., Fracchia, M., Flessel, C. P., Kizer, K. W., Book, S. A. y T. E. Warriner, "Benzene, Benzo(a)Pyrene, and Lead in Smoke from Tobacco Products Other Than Cigarettes", *American Journal of Public Health,* vol. 80, mayo de 1990, pp. 560-564.

[12] Klein, B. E. K., Klein, R. y S. E. Moss, "Blood Pressure in a Population of Diabetic Persons Diagnosed After 30 Years of Age", *American Journal of Public Health,* vol. 74, abril de 1984, pp. 336-339.

[13] National Center for Health Statistics, Fulwood, R., Johnson, C. L., Bryner, J. D., Gunter, E. W. y C. R. McGrath, "Hematological and Nutritional Biochemistry Reference Data for Persons 6 Months-74 Years of Age: Unites States, 1976-1980", *Vital and Health Statistics,* serie 11, núm. 232, diciembre de 1982.

[14] Rohrbrach, B. W., Harkess, J. R., Ewing, S. A., Kudlac, J., McKee, G. L. y G. R. Istre, "Epidemiologic and Clinical Characteristics of Persons with Serologic Evidence of *E. canis* Infection", *American Journal of Public Health,* vol. 80, abril de 1990, pp. 442-445.

[15] Feskens, E. J. M. y D. Kromhout, "Cardiovascular Risk Factors and the 25 Year Incidence of Diabetes Mellitus in Middle-Aged Men", *American Journal of Epidemiology,* vol. 130 diciembre de 1989, pp. 1101-1108.

[16] Meade, T. W., Cooper, J. A. y W. S. Peart, "Plasma Renin Activity and Ischemic Heart Disease", *The New England Journal of Medicine,* vol. 329, 26 de agosto de 1993, pp. 616-619.

[17] Burkholz, H., *The FDA Follies,* Nueva York, Basic Books, 1994, pp. 107-113.

[18] MacMahon, S. W. y G. J. MacDonald, "A Population at Risk: Prevalence of High Cholesterol Levels in Hypertensive Patients in the Framingham Study", *American Journal of Medicine Supplement,* vol. 80, 14 de febrero de 1986, pp. 40-47.

[19] Wilcox, A. J. y R. Skjaerven, "Birth Weight and Perinatal Mortality: The Effect of Gestational Age", *American Journal of Public Health,* vol. 82, marzo de 1992, pp. 378-382.

[20] Bellinger, D. C., Jonas, R. A., Rappaport, L. A., Wypij, D., Wernovsky, G., Kuban, K. C. K., Barnes, P. D., Holmes, G. L., Hichey, P. R., Strand, R. D., Walsh, A. Z., Helmers, S. L., Constantinou, J. E., Carrazana, E. J., Mayer, J. E., Hanley, F. L., Castaneda, A. R., Ware, J. H. y J. W. Newburger, "Developmental and Neurologic Status of Children After Heart Surgery with Hipothermic Circulatory Arrest of Low-Flow Cardiopulmonary Bypass", *The New England Journal of Medicine,* vol. 332, 2 de marzo de 1995, pp. 549-555.

11

Comparación de dos medias

En el capítulo anterior empleamos una prueba de hipótesis estadística para comparar la media desconocida de una sola población con algún valor fijo conocido μ_0. Sin embargo, en aplicaciones prácticas es mucho más común comparar las medias de dos diferentes poblaciones cuyas medias son desconocidas. A menudo, los dos grupos en cuestión han recibido tratamientos distintos o han sido expuestos a diferentes entornos.

Por siglos, ha estado presente la idea de comparar poblaciones para derivar una conclusión sobre sus semejanzas y diferencias. Por ejemplo, en el siglo XV1 se creía que las heridas provocadas por armas de fuego eran susceptibles de infectarse y, por lo tanto, requerían cauterización. Uno de los primeros usos extendidos de la pólvora ocurrió durante la invasión francesa a Italia en 1537. Esta expedición resultó ser la primera para un cirujano del ejército francés llamado Ambroise Paré. En su informe sobre un ataque a Turín [1] escribe:

> Entonces, los mencionados soldados en Chateau, viendo el coraje de nuestros hombres, hicieron cuanto pudieron por defenderse y mataron e hirieron un gran número de soldados nuestros con picas, arcabuces y piedras; de ahí que los cirujanos tenían mucho trabajo propio de su profesión. En aquel entonces yo era un soldado novato, y jamás había tenido que prestar los primeros auxilios en heridas provocadas por un disparo. Es verdad que había leído en el capítulo ocho del libro de Jean de Virgo, *Sobre las heridas en general,* que las heridas hechas por armas de fuego estaban envenenadas a causa de la pólvora, y que para curarlas recomienda cauterizarlas con aceite de saúco hirviendo, mezclado con un poco de triaca, y con el fin de no errar antes de aplicar dicho aceite, y sabiendo que este procedimiento podría provocar mucho dolor al paciente, quise conocer primero qué hacían los otros médicos para aplicar los primeros auxilios con dicho aceite tan caliente como fuera posible a la herida con mecha y sedal, de quienes saqué valor para hacerlo como ellos. Finalmente se agotó mi aceite y me vi obligado a aplicar en su lugar un digestivo elaborado a base de yemas de huevo, aceite de rosas y trementina. Esa noche no pude dormir con tranquilidad. El temor de que por la falta de cauterización encontraría muerto o envenenado al herido en el que no había aplicado el aceite me obligó a levantarme muy temprano para visitarlos, y, para mi sorpresa, encontré a aquellos a quienes había aplicado el medicamento digestivo con menos dolor y sus heridas sin inflamación o hinchazón,

> y que habían descansado tranquilamente toda la noche; en los demás a quienes les había aplicado el aceite hirviendo, encontré fiebre, mucho dolor e hinchazón alrededor de las heridas. Entonces me decidí a no quemar jamás a estos hombres cruelmente heridos con arma de fuego.

Los resultados de esta comparación —uno de los primeros casos clínicos documentados— fueron totalmente convincentes. Lo mismo podría decirse de los estudios sobre el uso de la penicilina en el tratamiento de enfermedades bacterianas. Por desgracia, dichos casos son la excepción más que la regla; el progreso por lo general se mide con más lentitud.

Este capítulo muestra un procedimiento para decidir si las diferencias que observamos entre medias de muestreo son suficientemente grandes para atribuírselas sólo al azar. Una prueba de hipótesis que implique dos muestras se asemeja en varios aspectos a una prueba llevada a cabo en una sola muestra. Comenzaremos por especificar una hipótesis nula: en la mayoría de los casos nos interesa probar si las dos medias poblacionales son iguales. Enseguida calculamos la probabilidad de obtener una par de medias de muestreo que difieran tanto o más de las medias observadas en el caso de que la hipótesis nula sea verdadera. Si esta probabilidad es suficientemente pequeña, rechazaremos la hipótesis nula y concluiremos que las dos medias de población son diferentes. Como antes, debemos especificar un nivel de significancia α y definir si nos interesa una prueba unilateral o una bilateral. La forma específica del análisis depende de la naturaleza de los dos conjuntos de observaciones implicados. En particular, debemos determinar si los datos provienen de muestras pareadas o independientes.

11.1 Muestras pareadas

La característica distintiva de las muestras pareadas consiste en que para cada observación en el primer grupo hay una observación correspondiente en el segundo grupo. En la técnica conocida como *autoapareamiento,* se toman medidas de de un solo individuo en dos diferentes momentos. Un ejemplo común de autoapareamiento es el experimento "antes y después", en el que cada individuo se somete a examen antes de aplicar cierto tratamiento, y se le somete a examen de nuevo, una vez concluido el tratamiento. Un segundo tipo de apareamiento ocurre cuando un investigador relaciona a los individuos de un grupo con los de un segundo grupo de tal forma que los miembros de una pareja se parezcan tanto como sea posible en lo que se refiere a una característica importante, como la edad y género.

Con frecuencia se emplea el apareamiento con la intención de controlar fuentes de variación ajenas que podrían, por otra parte, influir en los resultados de la comparación. Si se efectúan las mediciones en el mismo sujeto en lugar de realizarlas en dos diferentes individuos, se elimina cierto grado de variabilidad biológica. Queremos eliminar la necesidad de preocuparnos acerca del hecho de que un individuo sea mayor que el otro, o que uno sea hombre y el otro mujer. Por tanto, la intención del apareamiento consiste en llevar a cabo una comparación más precisa.

Considere los datos tomados de un estudio en el que cada uno de los 63 hombres adultos con arteriosclerosis se someten a una serie de pruebas de ejercicios en diferentes ocasiones. Un día, un paciente se somete primero a una prueba de ejercicio en una trotadora; se registra el tiempo que pasa desde el inicio de la prueba hasta que el paciente experimenta angina de pecho —dolor o espasmos en el pecho—. Enseguida se le ubica en un ambiente de aire común durante una hora aproximadamente. Al concluir este periodo, lleva a cabo una segunda prueba de ejercicio. De nuevo se registra el tiempo desde el inicio del ejerci-

cio hasta la aparición de la angina. La observación de interés es la disminución porcentual del tiempo hasta la aparición de la angina entre la primera y segunda pruebas. Si, por ejemplo, durante la primera prueba un hombre sufre un ataque de angina después de 983 segundos, y durante la segunda prueba sufre un ataque después de 957 segundos, su disminución porcentual hasta la aparición de la angina es

$$\frac{983 - 957}{983} = 0.026$$
$$= 2.6\%.$$

La media de población desconocida de esta distribución de disminución porcentual es μ_1. Para los 63 pacientes de la muestra, la disminución media porcentual observada es $\bar{x}_1 = 0.96\%$ [2].

Otro día, el mismo paciente se somete a una serie de pruebas similares. Esta vez, sin embargo, se le expone a un ambiente compuesto de una mezcla de aire y monóxido de carbono durante el intervalo entre las pruebas. Se pretende que la cantidad de monóxido de carbono que se añade al aire incremente el nivel de carboxihemoglobina del paciente —una medida biológica de exposición— a 4%. Este nivel es más bajo que el que normalmente soportan los fumadores, pero es muy cercano el que experimentaría un individuo en medio de un tráfico vehicular pesado en un área poco ventilada. De nuevo, la observación de interés es la disminución porcentual del tiempo hasta la aparición de la angina entre la primera y la segunda pruebas. La media desconocida de esta distribución es μ_2. La media de la muestra para el grupo de 63 individuos es $\bar{x}_2 = 7.59\%$.

Para los primeros diez pacientes del estudio, las disminuciones porcentuales del tiempo hasta la aparición de la angina para cada una de las dos ocasiones se muestran en la figura 11.1. Observe que la medida se incrementa en el caso de ocho de los hombres; en el caso de los otros dos, disminuye. Suponga que queremos determinar si existe alguna evidencia acerca de la diferencia en la disminución porcentual del tiempo hasta la aparición de la angina, entre la prueba en la que se expone a los individuos a un ambiente con monóxido de car-

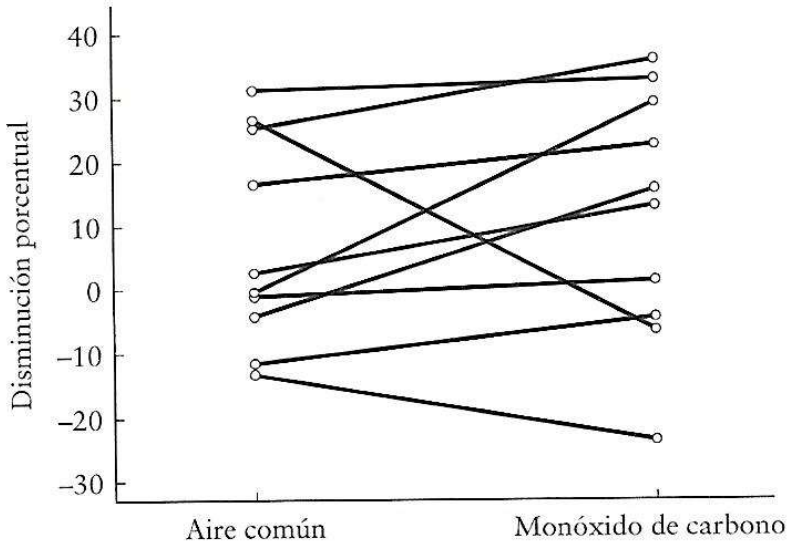

FIGURA 11.1
Disminución porcentual en el tiempo hasta la aparición de la angina en dos diferentes ocasiones para cada uno de los diez hombres con arteriosclerosis.

bono y la prueba en la que no se les expone a él. Como sabemos que la exposición excesiva al monóxido de carbono resulta nociva para la salud de una persona, sólo tomaremos en cuenta las desviaciones que ocurren en una sola dirección. Por tanto, llevaremos a cabo una prueba unilateral con un nivel de significancia de $\alpha = 0.05$. La hipótesis nula es

$$H_0 : \mu_1 \geq \mu_2,$$

o

$$H_0 : \mu_1 - \mu_2 \geq 0,$$

y la alterna es

$$H_A : \mu_1 - \mu_2 < 0.$$

En este estudio, cada paciente se somete a la misma serie de pruebas, tanto con exposición como sin exposición al monóxido de carbono. Este autoapareamiento elimina todas las posibles distorsiones o sesgos que pudieran surgir cuando se comparan pacientes que difieren en edad, peso y gravedad de arteriosclerosis. Debido a que los datos consisten en muestras pareadas, el método adecuado de análisis es la *prueba pareada t.*

En lugar de considerar los dos conjuntos de observaciones como muestras distintas, nos concentraremos en la **diferencia** en las mediciones entre cada pareja. Supongamos que nuestros dos grupos de observaciones son los siguientes:

Muestra 1	**Muestra 2**
x_{11}	x_{12}
x_{21}	x_{22}
x_{31}	x_{32}
$\vdots$	$\vdots$
x_{n1}	x_{n2}

En estas muestras, x_{11} y x_{12} constituyen una pareja, x_{21} y x_{22} constituyen una pareja, y así sucesivamente. Empleamos estos datos para crear un nuevo conjunto de observaciones que representen las diferencias entre cada pareja:

$$\begin{aligned} d_1 &= x_{11} - x_{12} \\ d_2 &= x_{21} - x_{22} \\ d_3 &= x_{31} - x_{32} \\ &\vdots \\ d_n &= x_{n1} - x_{n2} \end{aligned}$$

En lugar de analizar las observaciones individuales, utilizamos la diferencia entre los miembros de cada pareja como la variable de interés. Puesto que la diferencia es una sola medición, nuestro análisis se reduce al problema de una muestra y aplicamos el procedimiento de la prueba de hipótesis del capítulo 10.

Para hacerlo, primero notemos que la media del conjunto de diferencias es

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n};$$

esta media de la muestra proporciona una estimación puntual para la verdadera diferencia en las medias poblacionales $\mu_1 - \mu_2$. La desviación estándar de la diferencia es

$$s_d = \sqrt{\frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1}}.$$

Si denotamos la verdadera diferencia en las medias poblacionales

$$\delta = \mu_1 - \mu_2$$

y deseamos probar si estas dos medias son iguales, podemos escribir la hipótesis nula de la siguiente manera:

$$H_0 : \delta = 0$$

y la hipótesis alterna:

$$H_A : \delta \neq 0.$$

Si suponemos que la población de diferencias está distribuida normalmente, H_0 puede probarse al calcular el estadístico

$$\mathrm{t} = \frac{\bar{d} - \delta}{s_d/\sqrt{n}};$$

Observe que $s_d/\sqrt{n}$ es el error estándar de $\bar{d}$. Si la hipótesis nula es verdadera, esta cantidad tiene una distribución t con $n - 1$ grados de libertad. Comparemos el resultado de t para los valores de la tabla A.4 en el apéndice A para determinar p, la probabilidad de observar una diferencia media tan grande o más que $\bar{d}$, en el supuesto de que $\delta = 0$. (O, como siempre, podemos utilizar un programa de cómputo para efectuar los cálculos.) Si $p \leq \alpha$, rechazamos H_0. Si $p > \alpha$, no rechazamos la hipótesis nula.

Volvamos al estudio de los hombres con arteriosclerosis y concentrémonos en la diferencia de mediciones en el caso de un individuo determinado. Para cada uno de los 63 hombres del estudio, calculemos la disminución porcentual en el tiempo hasta la aparición de la angina cuando se expone al monóxido de carbono, menos la disminución porcentual cuando se expone al aire no adulterado. La media de estas diferencias es

$$\bar{d} = \frac{\sum_{i=1}^{63} d_i}{63}$$
$$= -6.63,$$

y la desviación estándar es

$$s_d = \sqrt{\frac{\sum_{i=1}^{63}(d_i - \bar{d})^2}{63 - 1}}$$
$$= 20.29.$$

Como se observa en la figura 11.2, las diferencias son más o menos simétricas y puede considerarse que están muy cercanas a la distribución normal. Por tanto, si escribimos de nuevo la hipótesis nula de la prueba de la forma

$$H_0 : \delta \geq 0,$$

podemos evaluar H_0 con el estadístico

$$t = \frac{\bar{d} - \delta}{s_d / \sqrt{n}};$$

o

$$t = \frac{-6.63 - 0}{20.29/\sqrt{63}}$$
$$= -2.59.$$

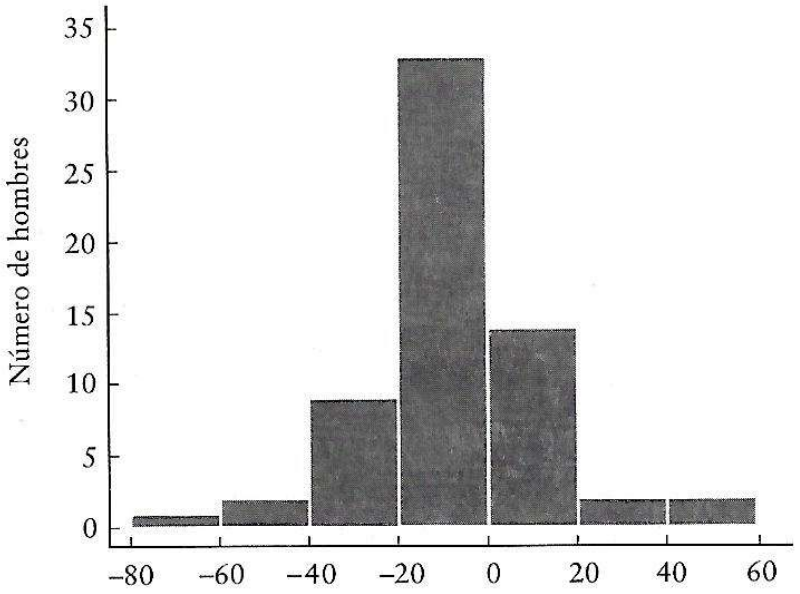

FIGURA 11.2
Diferencias de la disminución porcentual del tiempo hasta la angina en una muestra de 63 hombres con arteriosclerosis.

Al consultar la tabla A.4, observamos que para una distribución t con $63 - 1 = 62$ grados de libertad, el área bajo la curva a la izquierda de $t_{62} = -2.59$ se encuentra entre 0.005 y 0.01. Por tanto, $0.005 < p < 0.01$. Al rechazar la hipótesis nula en el nivel 0.05, concluimos que hay una diferencia significativa entre la disminución porcentual media del tiempo hasta la aparición de la angina, cuando los pacientes quedan expuestos al monóxido de carbono y la disminución cuando quedan expuestos al aire normal. La exposición al monóxido de carbono incrementa la disminución porcentual del tiempo hasta la angina; en otras palabras: los pacientes expuestos tienden a desarrollar angina más rápidamente.

Como hemos dicho, la media de la muestra, $\bar{d}$, proporciona una estimación puntual de la verdadera diferencia en las medias de población $\delta = \mu_1 - \mu_2$. Sin embargo, podría interesarnos calcular el límite de confianza superior para δ. De nuevo nos basamos en la técnica de una muestra. Para el caso de una distribución t con 62 grados de libertad, 95% de las observaciones se ubican arriba de –1.671. Por tanto,

$$\begin{aligned} P(t \geq -1.671) &= P\left(\frac{\bar{d} - \delta}{s_d/\sqrt{n}} \geq -1.671\right) \\ &= 0.95. \end{aligned}$$

Un intervalo de confianza unilateral de 95% para δ es

$$\begin{aligned} \delta &\leq \bar{d} + 1.671 \frac{s_d}{\sqrt{n}} \\ &= -6.63 + 1.671 \frac{20.29}{\sqrt{63}} \\ &= -2.36. \end{aligned}$$

Tenemos 95% de confianza en que la verdadera diferencia en las medias de población es menor o igual a –2.36%. En otras palabras, tenemos 95% de seguridad de que la **disminución** en el tiempo hasta la aparición de la angina después de la exposición al monóxido de carbono es de por lo menos 2.36%.

11.2 Muestras independientes

Ahora supongamos que se nos proporcionan las mediciones del nivel de hierro en la sangre para dos muestras de niños: un grupo de niños se encuentra saludable y el otro grupo padece fibrosis quística, una enfermedad congénita de las glándulas. Las dos poblaciones bajo estudio son independientes y están normalmente distribuidas. Si la población de niños enfermos tiene un nivel medio de hierro en la sangre μ_1 y la población de niños saludables tiene una media μ_2, podría de nuevo interesarnos probar la hipótesis nula de que las dos medias de población son idénticas. Esta hipótesis puede expresarse ya sea como

$$H_0 : \mu_1 - \mu_2 = 0$$

o

$$H_0 : \mu_1 = \mu_2.$$

La hipótesis alterna es

$$H_A : \mu_1 \neq \mu_2.$$

A partir de la población normal con media μ_1 y desviación estándar σ_1, tomamos una muestra aleatoria de tamaño n_1. La media de esta muestra se denota $\bar{x}_1$ y su desviación estándar s_1. Asimismo, seleccionamos una muestra aleatoria de tamaño n_2 de la población normal con media μ_2 y desviación estándar σ_2. La media de esta muestra está representada por $\bar{x}_2$ y su desviación estándar por s_2. Observe que el número de observaciones en las dos muestras —n_1 y n_2— no necesariamente debe ser el mismo.

		Grupo 1	Grupo 2
Población	Media	μ_1	μ_2
	Desviación estándar	σ_1	σ_2
Muestra	Media	$\bar{x}_1$	$\bar{x}_2$
	Desviación estándar	s_1	s_2
	Tamaño de muestra	n_1	n_2

Surgen dos diferentes situaciones en la comparación de muestras independientes. En la primera, se sabe que las varianzas de las poblaciones bajo estudio son iguales entre sí o al menos se suponen iguales. Esto nos lleva a la *prueba t de dos muestras,* la cual puede encontrarse en cualquier texto. En la segunda, las varianzas no son las mismas. En este caso, la prueba t estándar ya no puede aplicarse. Antes de que se lleve a cabo una prueba de medias, muchos piensan que debería efectuarse una prueba preliminar de varianzas con el fin de distinguir entre estas dos opciones. Otros objetan la prueba sobre bases filosóficas: resulta sumamente sensible al supuesto de normalidad y tiene poca eficiencia en muchos casos donde debería evitarse la prueba t [3]. Además, se ha demostrado que una modificación de la prueba de dos muestras llevada a cabo sin esta verificación inicial resulta más eficaz en escenarios en los que no se conoce si las varianzas de la población fundamental son iguales [4]. Debido a que por lo general es innecesaria o ineficaz, no recomendamos la aplicación de una prueba preliminar de varianzas en este texto.

11.2.1 Varianzas iguales

Primero consideraremos el caso en el que o se sabe o se supone que las dos varianzas de población son idénticas. Recuerde que para el caso de una sola población normal con media μ y desviación estándar σ, el teorema del límite central establece que la media muestral $\bar{X}$ tiene una distribución muy cercana a la normal —suponiendo que n es suficientemente grande— con una media μ y un error estándar $\sqrt{\sigma^2/n} = \sigma/\sqrt{n}$. Por tanto,

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

es el resultado de una variable aleatoria normal estándar. Cuando tratamos con muestras tomadas de dos poblaciones normales independientes, una extensión del teorema del límite central dice que la diferencia entre las medias de muestreo $\bar{X}_1 - \bar{X}_2$ es aproximadamente normal con una media $\mu_1 - \mu_2$ y un error estándar $\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$. Puesto que se supone que

las varianzas de población son iguales, sustituimos el valor común σ^2 por σ_1^2 y σ_2^2. En consecuencia, sabemos que

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2/n_1 + \sigma^2/n_2}}$$
$$= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2[(1/n_1) + (1/n_2)]}}$$

es el resultado de una variable aleatoria normal estándar. Si se conoce el valor de la varianza de población σ^2, puede emplearse este estadístico para probar la hipótesis nula

$$H_0 : \mu_1 = \mu_2.$$

Como hemos hecho notar anteriormente, es más frecuente que el valor verdadero de σ^2 se desconozca. En este caso, utilizamos el estadístico de prueba

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2[(1/n_1) + (1/n_2)]}}$$

en su lugar. La cantidad s_p^2 es una estimación combinada de la varianza común σ^2. Con la hipótesis nula de que las medias de población son idénticas, $\mu_1 - \mu_2$ es igual a 0, y el estadístico de prueba t tiene una distribución t con $(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$ grados de libertad. Podemos comparar el valor de este estadístico con los valores críticos en la tabla A.4 para determinar p, la probabilidad de observar una discrepancia tan amplia como $\bar{x}_1 - \bar{x}_2$, supuesto que μ_1 es igual a μ_2. Si $p \leq \alpha$, rechazamos la hipótesis nula. Si $p > \alpha$ no rechazamos H_0.

La estimación combinada de la varianza, s_p^2, combina información de ambas muestras para producir una estimación más confiable de σ^2. Ésta se puede calcular de dos formas diferentes. Si conocemos los valores de todas las observaciones de las muestras, aplicamos la fórmula

$$s_p^2 = \frac{\sum_{i=1}^{n_1}(x_{i1} - \bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_{j2} - \bar{x}_2)^2}{n_1 + n_2 - 2}.$$

Si sólo se nos proporciona s_1 y s_2, debemos utilizar

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Esta segunda fórmula demuestra que s_p^2 es en realidad un promedio ponderado de las dos varianzas de muestreo s_1^2 y s_2^2, donde cada varianza se encuentra ponderada por los grados de libertad asociados con ella. Si n_1 es igual a n_2, s_p^2 es simplemente un promedio aritmético; de otra manera se da más peso a la varianza de la muestra más grande. Si recordamos que

$$s_1^2 = \frac{\sum_{i=1}^{n_1}(x_{i1} - \bar{x}_1)^2}{n_1 - 1}$$

y

$$s_2^2 = \frac{\sum_{j=1}^{n_2} (x_{j2} - \bar{x}_2)^2}{n_2 - 1}.$$

podemos ver que las dos fórmulas para calcular s_p^2 son matemáticamente equivalentes.

Como ilustración de la prueba t de dos muestras, considere las distribuciones de niveles de hierro en la sangre en la población de niños saludables y la población de niños con fibrosis quística. Ambas distribuciones son aproximadamente normales. Denotemos con μ_1 el nivel de hierro en la sangre en los niños saludables y con μ_2 el nivel medio de hierro en la sangre en niños con la enfermedad. Las desviaciones estándares de las dos poblaciones —σ_1 y σ_2— se desconocen, pero con base en estudios previos se supone que son iguales. Quisiéramos determinar si los niños con fibrosis quística tienen un nivel normal de hierro en sus flujos sanguíneos en promedio. Por tanto, probamos la hipótesis nula que dice que las dos medias poblacionales son idénticas,

$$H_0 : \mu_1 = \mu_2.$$

Se elige una muestra aleatoria de cada población. La muestra de $n_1 = 9$ niños saludables tiene un nivel medio de hierro en la sangre de $\bar{x}_1 = 18.9$ μmol/1 y una desviación estándar de $s_1 = 5.9$ μmol/1, la muestra de $n_2 = 13$ niños con fibrosis quística tiene un nivel medio de $\bar{x}_2 = 11.9$ μmol/1 y una desviación estándar de $s_2 = 6.3$ μmol/1 [5]. ¿Es probable que la diferencia observada en las medias de muestreo —18.9 en comparación con 11.9 μmol/1— sean el resultado de una variación al azar, o deberíamos concluir que la discrepancia se debe a una verdadera diferencia en las medias poblacionales?

En algunos casos, un investigador comenzará un análisis construyendo un intervalo de confianza distinto para la media de cada población individual. Como ilustración, los intervalos de confianza de 95% de los niveles medios de hierro en la sangre de niños con fibrosis quística y sin ella aparecen en la figura 11.3. En general, el hecho de que los dos intervalos

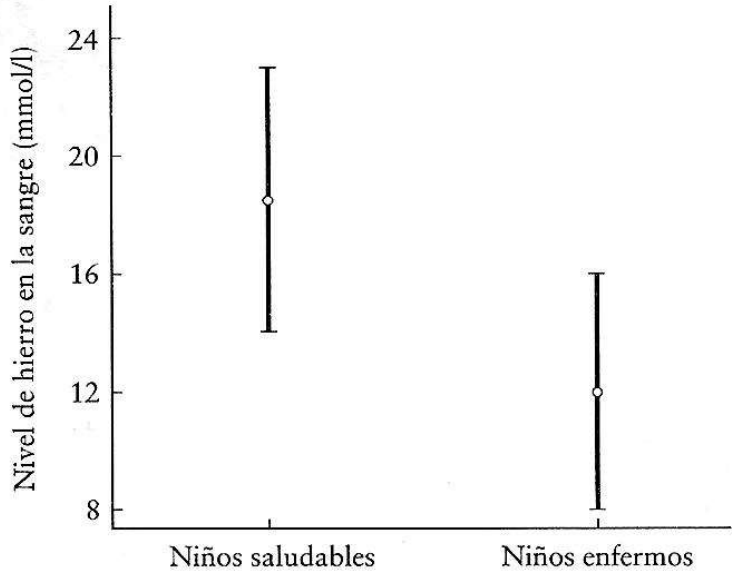

FIGURA 11.3
Intervalos de confianza de 95% de los niveles medios de hierro en la sangre de niños saludables y niños con fibrosis quística.

no se superpongan sugiere que las medias de población son, por tanto, diferentes. Sin embargo, debemos recordar que esta técnica no constituye una prueba de hipótesis formal. En nuestro ejemplo, hay un pequeño grado de superposición entre los intervalos; como consecuencia, no es posible obtener ninguna conclusión significativa.

Observe que las dos muestras de niños se eligieron al azar de distintas poblaciones normales. Además, se supone que las varianzas de población son iguales. Así, la prueba t de dos muestras es la técnica adecuada que debe aplicarse. La hipótesis nula establece que no existe diferencia en los niveles medios de hierro de la población fundamental de los dos grupos de niños. Puesto que deseamos detectar una diferencia que pudiera darse en cualquier sentido —nos interesaría saber si los niños con fibrosis quística tienen una media más alta o más baja que los niños que no padecen la enfermedad—, llevamos a cabo una prueba bilateral y fijamos el nivel de significancia en $\alpha = 0.05$. La hipótesis alterna es

$$H_A : \mu_1 \neq \mu_2.$$

Con el fin de efectuar la prueba, comenzamos por calcular la estimación combinada de la varianza

$$\begin{aligned} s_p^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \\ &= \frac{(9-1)(5.9)^2 + (13-1)(6.3)^2}{9 + 13 - 2} \\ &= \frac{(8)(34.81) + (12)(39.69)}{20} \\ &= 37.74. \end{aligned}$$

Enseguida calculamos el estadístico de prueba

$$\begin{aligned} t &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2[(1/n_1) + (1/n_2)]}} \\ &= \frac{(18.9 - 11.9) - 0}{\sqrt{(37.74)[(1/9) + (1/13)]}} \\ &= 2.63. \end{aligned}$$

Al consultar la tabla A.4, observamos que, para una distribución t con $n_1 + n_2 - 2 = 9 + 13 - 2 = 20$ grados de libertad, el área total bajo la curva ubicada a la derecha de $t_{20} = 2.63$ se encuentra entre 0.005 y 0.01. Por tanto, la suma de las áreas localizadas a la derecha de $t_{20} = 2.63$ y a la izquierda de $t_{20} = -2.63$ deben encontrarse entre 0.01 y 0.02. Puesto que p es menor que 0.05, rechazamos la hipótesis nula

$$H_0 : \mu_1 = \mu_2$$

en el nivel de significancia 0.05. La diferencia entre el nivel medio de hierro en la sangre en niños saludables y el nivel medio en niños con fibrosis quística es estadísticamente significativa; con base en estas muestras, parece ser cierto que los niños con fibrosis quística padecen una insuficiencia de hierro.

La cantidad $\bar{x}_1 - \bar{x}_2$ proporciona una estimación puntual para la verdadera diferencia en las medias de población $\mu_1 - \mu_2$; sin embargo, una vez más podríamos construir un intervalo de confianza. Observe que para el caso de una distribución t con 20 grados de libertad, 95% de las observaciones se encuentran entre –2.086 y 2.086. Como resultado,

$$P\left(-2.086 \leq \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2[(1/n_1) + (1/n_2)]}} \leq 2.086\right) = 0.95.$$

Al reordenar términos, encontramos que los límites del intervalo de confianza de 95% $\mu_1 - \mu_2$ son

$$(\bar{x}_1 - \bar{x}_2) \pm (2.086)\sqrt{s_p^2\left[\frac{1}{n_1} + \frac{1}{n_2}\right]},$$

o

$$(18.9 - 11.9) \pm (2.086)\sqrt{(37.74)\left[\frac{1}{9} + \frac{1}{13}\right]}.$$

Por lo tanto, tenemos 95% de confianza de que el intervalo

(1.4, 12.6)

cubre $\mu_1 - \mu_2$, la verdadera diferencia de los niveles medios de hierro en la sangre de las dos poblaciones de niños. A diferencia de los intervalos separados que se muestran en la figura 11.3, este intervalo de confianza de la diferencia en las medias es matemáticamente equivalente a la prueba de dos muestras de la hipótesis que se llevó a cabo con un nivel del 0.05. Observe que el intervalo no contiene el valor 0.

11.2.2 Varianzas desiguales

Ahora dirigimos nuestra atención al caso en que no se supone que las varianzas de las dos poblaciones sean iguales. En este caso, debe emplearse una modificación de la prueba t de dos muestras. En lugar de emplear s_p^2 como estimación de la varianza común σ^2, sustituimos s_1^2 por σ_1^2 y s_2^2 por σ_2^2. Por tanto, el estadístico de prueba adecuado es

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}.$$

Este caso es diferente de aquel en que teníamos varianzas iguales, por la dificultad que supone obtener la distribución exacta de t. Por tanto, es necesario utilizar una aproximación [6]. Comenzamos por calcular la cantidad

$$v = \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{[(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)]};$$

enseguida se redondea el valor de v al entero menor más próximo. Con la hipótesis nula, podemos aproximar la distribución de t por una distribución t con v grados de libertad. Como siempre, comparamos el valor de la estadística con los valores críticos de la tabla A.4 —o empleamos un programa de computadora— para decidir si deberíamos rechazar H_0.

Supongamos que nos interesa conocer los efectos de un tratamiento con medicamentos antihipertensores en personas de más de 60 años que padecen hipertensión sistólica aislada. Por definición, los individuos con este padecimiento tienen una presión arterial sistólica mayor de 160 mm Hg, mientras que su presión arterial diastólica se encuentra por debajo de los 90 mm Hg. Antes de iniciar el estudio, las personas que habían sido elegidas al azar para tomar el medicamento activo y a las que se eligió para tomar un placebo presentaban medidas de presión arterial sistólica comparables. El estudio tuvo una duración de un año. A la presión arterial sistólica media en pacientes a los que se les administró el medicamento se denota con μ_1 y la media de los que tomaron el placebo con μ_2. Se desconocen las desviaciones estándares de las dos poblaciones, y no tenemos indicio alguno que nos haga suponer que son iguales. Puesto que queramos determinar si las presiones arteriales sistólicas medias de los pacientes de estos dos grupos distintos siguen siendo las mismas, probamos la hipótesis nula

$$H_0 : \mu_1 = \mu_2.$$

Comenzamos por eligir una muestra aleatoria de cada uno de los dos grupos. La muestra de $n_1 = 2308$ individuos que reciben el tratamiento con el medicamento activo tiene una presión arterial sistólica media de $\bar{x}_1 = 142.5$ mm Hg y una desviación estándar de $s_1 = 15.7$ mm Hg; la muestra de $n_2 = 2293$ personas que toman el placebo tiene una media de $\bar{x}_2 = 156.5$ mm Hg y una desviación estándar de $s_2 = 17.3$ mm Hg [7]. Nos interesa detectar diferencias que pudieran ocurrir en cualquier sentido, y, por tanto, llevamos a cabo una prueba bilateral con un nivel de significancia de $\alpha = 0.05$. La hipótesis alterna es

$$H_0 : \mu_1 \neq \mu_2.$$

Puesto que los dos grupos de pacientes se eligieron de poblaciones normales independientes y no se supone que las varianzas sean iguales, debería aplicarse la prueba de las dos muestras modificada. (Observe que la prueba modificada no supone que las varianzas sean desiguales; sólo supone que no sean las mismas.) En este caso, el estadístico de prueba es

$$\begin{aligned} t &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}} \\ &= \frac{(142.5 - 156.5) - 0}{\sqrt{[(15.7)^2/2308] + [(17.3)^2/2293]}} \\ &= -28.74. \end{aligned}$$

Enseguida calculamos los grados de libertad aproximados; puesto que $s_1^2 = (15.7)^2 = 246.49$ y $s_2^2 = (17.3)^2 = 299.29$,

$$\begin{aligned} v &= \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{[(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)]} \\ &= \frac{[(246.49/2308) + (299.29/2293)]^2}{[(246.49/2308)^2/(2308 - 1) + (299.29/2293)^2/(2293 - 1)]} \\ &= 4550.5. \end{aligned}$$

Al redondear al entero próximo inferior, $v = 4550$. Debido a que una distribución t con 4550 grados de libertad es, para fines prácticos, idéntica a la distribución normal estándar, podemos consultar la tabla A.3 o la tabla A.4. En ambos casos, encontramos que p es menor de 0.001. Como resultado, rechazamos la hipótesis nula

$$H_0 : \mu_1 = \mu_2$$

en el nivel de significancia 0.05. Después de un año, los individuos que reciben el tratamiento con el medicamento activo tienen una presión arterial sistólica media más baja que los que reciben el placebo.

Una vez más, podríamos construir un intervalo de confianza para la verdadera diferencia de las medias de población $\mu_1 - \mu_2$. Para una distribución t con 4550 grados de libertad —o la distribución normal estándar—, 95% de las observaciones se ubican entre –1.96 y 1.96. En consecuencia,

$$P\left(-1.96 \leq \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}} \leq 1.96\right) = 0.95.$$

Si reordenamos términos, los límites de confianza de 95% para $\mu_1 - \mu_2$ son

$$(\bar{x}_1 - \bar{x}_2) \pm (1.96)\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

o

$$(142.5 - 156.5) \pm (1.96)\sqrt{\frac{(15.7)^2}{2308} + \frac{(17.3)^2}{2293}}.$$

Por tanto, estamos 95% seguros de que el intervalo

$$(-15.0, -13.0)$$

cubre $\mu_1 - \mu_2$, la verdadera diferencia de las presiones arteriales sistólicas medias de las dos poblaciones. Observe que este intervalo no contiene el valor 0, y por tanto es congruente con los resultados de la prueba de dos muestras modificada.

11.3 Aplicaciones adicionales

Ahora volvemos al estudio de los efectos de la exposición al monóxido de carbono en pacientes con arteriosclerosis. Antes comparamos la disminución porcentual media en el tiempo hasta la aparición de la angina, cuando un grupo de 63 adultos se expueso a un nivel de monóxido de carbono, con lo cual se pretendía elevar sus niveles de carboxihemoglobina a 4% respecto de la disminución porcentual media cuando no se habían expuesto al aire no

contaminado. Ahora deseamos hacer comparaciones semejantes, pero esta vez se expone a los mismos pacientes a un ambiente con monóxido de carbono con el fin de incrementar sus niveles de carboxihemoglobina sólo a 2%. Por lo que en esta etapa del estudio se expone a cada paciente a concentraciones menores de monóxido de carbono. Las medias de población de las disminuciones porcentuales asociadas con la exposición al aire y la exposición al monóxido de carbono se encuentran representadas por μ_1 y μ_2, respectivamente.

De nuevo, nos interesa saber si las dos medias de población μ_1 y μ_2 son idénticas. Debido a que sabemos que la exposición al monóxido de carbono resulta perjudicial para la salud de un individuo, nos interesan las desviaciones que pudieran presentarse en un solo sentido. Por tanto, llevamos a cabo una prueba unilateral con un nivel de significancia de $\alpha = 0.05$. La hipótesis nula es

$$H_0 : \mu_1 \geq \mu_2,$$

o

$$H_0 : \delta \geq 0,$$

donde $\delta = \mu_1 - \mu_2$, y la hipótesis alterna es

$$H_A : \delta < 0.$$

En lugar de trabajar con los dos conjuntos individuales de observaciones, nos concentraremos de la diferencia de la disminución porcentual del tiempo hasta la angina calculada para cada individuo. De esta manera estamos en posibilidades de llevar a cabo una prueba de una muestra. La media de estas diferencias —una estimación puntual de la verdadera diferencia en las medias de población, δ— es

$$\begin{aligned} \bar{d} &= \frac{\sum_{i=1}^{62} d_i}{62} \\ &= -4.95, \end{aligned}$$

y su desviación estándar es

$$\begin{aligned} s_d &= \sqrt{\frac{\sum_{i=1}^{62} (d_i - \bar{d})^2}{62 - 1}} \\ &= 19.05 \end{aligned}$$

(A un hombre no le fue posible participar el día en que sería expuesto al nivel bajo de monóxido de carbono; por tanto, el tamaño de la muestra es de 62 en lugar de 63.) El estadístico de prueba para la prueba t de apareamiento es

$$t = \frac{\bar{d} - \delta}{s_d / \sqrt{n}},$$

o

$$t = \frac{-4.95 - 0}{19.05/\sqrt{62}}$$

$$= -2.05.$$

Para una distribución t con 61 grados de libertad, $0.01 < p < 0.025$. Por tanto, rechazamos la hipótesis nula en el nivel 0.05. Las muestras pareadas sugieren que la disminución porcentual media del tiempo hasta la aparición de la angina, cuando el paciente se expone a un nivel bajo de monóxido de carbono, es mayor que la disminución porcentual media cuando el paciente no se expone. Una vez más, los pacientes expuestos tienden a desarrollar angina de pecho con mayor rapidez.

En lugar de efectuar los cálculos de la prueba de apareamiento t nosotros mismos utilizamos una computadora. En la mayoría de los programas estadísticos hay dos formas de hacerlo: podemos llevar a cabo la prueba con los conjuntos originales de observaciones y permitir que la computadora genere las diferencias, o calcularlas nosotros mismos y llevar a cabo una prueba de una muestra. En la tabla 11.1 se muestra la salida adecuada de Stata, con las mediciones originales. Además de las estadísticas de resumen de las disminuciones porcentuales asociadas con la exposición al monóxido de carbono y la exposición al aire puro, también figuran los resúmenes de las diferencias. La salida también muestra la hipótesis nula y las tres opciones posibles, junto con un estadístico de prueba y un valor p para cada una. En este caso, nos interesa la hipótesis alternativa localizada a la izquierda. Así, $p = 0.0226$. Observe que la computadora nos proporciona un valor p más preciso que la tabla A.4.

Ahora consideremos un estudio diferente, diseñado para analizar los efectos del consumo de lactosa en la absorción de energía en carbohidratos en niños prematuros. En particular, nos interesa determinar si una reducción del consumo de lactosa —un azúcar que se encuentra en la leche— incrementa o disminuye la absorción de energía. En este estudio, un grupo de recién nacidos se alimentó con leche materna; otro grupo recibió una fórmula que contenía sólo la mitad de la lactosa presente en la leche materna. Las distribuciones de ab-

TABLA 11.1
Salida de Stata para la prueba t de apareamiento.

Prueba t de apareamiento — Number of obs = 62

Variable	Mean	Std. Err.	t	P>|t|	[95% Conf.	Interval]
monóxido de C.	.9254365	2.232755	.414482	0.6800	-3.539232	5.390105
aire	5.873768	1.801627	3.26026	0.0018	2.271192	9.476344
diff	-4.948331	2.418982	-2.04563	0.0451	-9.785384	-.111278

Degrees of freedom: 61

Ho: mean diff = 0

Ha: diff < 0	Ha: diff ~= 0	Ha: diff > 0
t = -2.046	t = -2.046	t = -2.046
P < t = 0.0226	P > |t| = 0.0451	P > t = 0.9774

sorción de energía en carbohidratos de las dos poblaciones son aproximadamente normales. Suponemos que estas distribuciones tienen varianzas iguales, y nos gustaría saber si también tienen medias idénticas. Puesto que nos interesan las desviaciones que pudieran aparecer en cualquier sentido, efectuamos la prueba de la hipótesis nula

$$H_0 : \mu_1 = \mu_2$$

en función de la hipótesis alterna bilateral

$$H_A : \mu_1 \neq \mu_2.$$

Una muestra aleatoria de $n_1 = 8$ niños que se alimentaron con leche materna tiene una absorción de energía media de $\bar{x}_1 = 87.38\%$ y una desviación estándar de $s_1 = 4.56\%$; una muestra de $n_2 = 10$ recién nacidos a los que se les administró la fórmula tiene una media de $\bar{x}_2 = 90.14\%$ y una desviación estándar de $s_2 = 4.58\%$ [8]. Puesto que las muestras son independientes y las varianzas de las poblaciones estudiadas se suponen iguales —suposición que parece razonablemente fundada en los valores de s_1 y s_2—, aplicamos la prueba t para dos muestras.

Comencemos por calcular la estimación combinada de la varianza,

$$\begin{aligned} s_p^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \\ &= \frac{(8-1)(4.56)^2 + (10-1)(4.58)^2}{8 + 10 - 2} \\ &= 20.90. \end{aligned}$$

El valor s_p^2 combina información de ambas muestras de niños para dar como resultado una estimación más confiable de la varianza común σ^2. El estadístico de prueba es

$$\begin{aligned} t &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2 [(1/n_1) + (1/n_2)]}} \\ &= \frac{(87.38 - 90.14) - 0}{\sqrt{(20.90)\,[(1/8) + (1/10)]}} \\ &= -1.27. \end{aligned}$$

Para una distribución t con $8 + 10 - 2 = 16$ grados de libertad, el área total bajo la curva a la izquierda de -1.27 y a la derecha de 1.27 es mayor que $2(0.10) = 0.20$. Por tanto, no rechazamos la hipótesis nula. Con base en estas muestras, la ingestión de lactosa en recién nacidos no parece tener efecto alguno en la absorción de energía en carbohidratos.

Una vez más, utilizamos una computadora para llevar a cabo la prueba de la hipótesis. La salida de Stata aparece en la tabla 11.2. Como nos interesa una prueba bilateral, nos concentramos en la información en el centro de la parte inferior de la salida. Mientras que la tabla A.4 nos permitió afirmar que $p > 0.20$, la computadora nos dice que $p = 0.2213$.

TABLA 11.2
Salida de Stata que muestra la prueba *t* de dos muestras, suponiendo que las varianzas son iguales.

Prueba *t* de dos muestras con varianzas iguales — M: Number of obs = 8; F: Number of obs = 10

Variable	Mean	Std. Err.	t	P > |t|	[95% Conf.	Interval]
fórmula	87.38	1.612203	54.1991	0.0000	83.56774	91.19226
láctea	90.14	1.448323	62.2375	0.0000	86.86367	93.41633
diff	−2.76	2.168339	−1.27286	0.2213	−7.356674	1.836674

Degrees of freedom: 16

Ho: mean(x) − mean(y) = diff = 0

Ha: diff < 0	Ha: diff ~= 0	Ha: diff > 0
t = −1.2729	t = −1.2729	t = −1.2729
P < t = 0.1106	P > |t| = 0.2213	P > t = 0.8894

En un estudio llevado a cabo para analizar los factores de riesgo de una enfermedad del corazón entre pacientes que padecen diabetes, una de las características que se examinó fue el índice de masa corporal. El índice de masa corporal es una medida del grado de sobrepeso de un individuo. Deseamos determinar si el índice medio de masa corporal de hombres diabéticos es igual a la media en mujeres diabéticas. En cada grupo, la distribución de índices es aproximadamente normal. No tenemos razones para creer que las varianzas deben ser iguales, por lo que esta suposición queda fuera de lugar. En consecuencia, efectuamos la prueba de la hipótesis nula

$$H_0 : \mu_1 = \mu_2$$

en función de la hipótesis alterna bilateral

$$H_A : \mu_1 \neq \mu_2$$

con la versión modificada de la prueba de dos muestras.

Se elige una muestra aleatoria de cada población. Los $n_1 = 207$ varones diabéticos tienen un índice medio de masa corporal de $\bar{x}_1 = 26.4$ Kg/m^2 y una desviación estándar de $s_1 = 3.3$ Kg/m^2. Las $n_2 = 127$ mujeres diabéticas tienen un índice medio de masa corporal de $\bar{x}_2 = 25.4$ Kg/m^2 y una desviación estándar de $s_2 = 5.2$ Kg /m^2 [9]. La estadística de prueba es

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)]}}$$

$$= \frac{(26.4 - 25.4) - 0}{\sqrt{[(3.3)^2/207] + [(5.2)^2/127]}}$$

$$= -1.94.$$

Debido a que $s_1^2 = (3.3)^2 = 10.89$ y $s_2^2 = (5.2)^2 = 27.04$, encontramos que los grados de libertad aproximados son

$$v = \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{[(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)]}$$

$$= \frac{[(10.89/207) + (27.04/127)]^2}{[(10.89/207]^2/(207 - 1) + (27.04/127)^2/(127 - 1)]}$$

$$= 188.9.$$

Al redondear al entero próximo inferior, $v = 188$. Para una distribución t con 188 grados de libertad, $0.05 < p < 0.10$. Los resultados de esta prueba se ubican en el límite; aunque rechazaríamos la hipótesis nula

$$H_0 : \mu_1 = \mu_2$$

con un nivel de significancia de 0.10, no la rechazaríamos con un nivel de 0.05. Parece que los varones que padecen diabetes tienden a tener índices de masa corporal ligeramente más altos —y, por tanto, a estar más pasados de peso— que las mujeres que padecen esta enfermedad.

De nuevo, pudimos emplear una computadora para efectuar estos cálculos. La salida adecuada de SAS aparece en la tabla 11.3. Además de las estadísticas de resumen para cada grupo independiente, la salida muestra la estadística de prueba, los grados de libertad y el valor p de la prueba que supone que las varianzas son iguales así como la de la prueba que supone que no lo son. (También incluye una prueba preliminar de varianzas, aunque no se pida.) El valor p de la prueba modificada es aproximadamente igual a 0.05 si redondeamos a dos decimales hacia abajo. De nuevo concluimos que los varones diabéticos tienden a estar más pasados de peso que las mujeres diabéticas, pero ahora sabemos que p se encuentra mucho más próximo a 0.05 de lo que se encuentra de 0.10. Observe que el valor p de la prueba

TABLA 11.3
Salida SAS que muestra la prueba t de dos muestras, tanto para varianzas iguales como para varianzas distintas.

TTEST PROCEDURE

Variable: BMI

GROUP	N	Mean	Std Dev	Std Error	Minimum	Maximum
M	207	26.4	3.3	0.229366	19.7	32.8
F	127	25.4	5.2	0.461425	17.5	35.2

Variances	T	DF	Prob> |T|
Unequal	1.9407	188.9	0.0538
Equal	2.1505	332.0	0.0322

For H0: Variances are equal, F = 0.403 DF = (126,206)
Prob > F = 0.000

que supone varianzas iguales es en realidad un poco menor de 0.05. Sin embargo, puesto que no teníamos razón para creer que las varianzas deberían ser iguales —y, de hecho, las desviaciones estándares de muestreo s_1 y s_2 sugieren que probablemente no sean las mismas— resulta más seguro emplear la prueba modificada. Aunque esta prueba es menos precisa que la prueba tradicional t para dos muestras si las varianzas en verdad son iguales, es más confiable si no lo son.

11.4 Ejercicios de repaso

1. ¿Cuál es la principal diferencia entre muestras pareadas e independientes?
2. Explique el propósito de los datos pareados. En ciertos casos, ¿cuál sería la ventaja de utilizar muestras pareadas en lugar de muestras independientes?
3. ¿Cuándo se debería utilizar una prueba t de dos muestras? ¿Cuándo debería aplicarse la versión modificada de la prueba?
4. ¿Cuál es la razón fundamental para utilizar una estimación combinada de la varianza en la prueba t de dos muestras?
5. Se llevó a cabo un estudio cruzado para investigar si el cereal de avena integral contribuye a reducir los niveles de colesterol en la sangre en el caso de varones hipercolesterolémicos. Se sometió a 14 de estos individuos a una dieta que incluía avena integral u hojuelas de maíz; después de dos semanas, se registraron sus niveles colesterol lipoproteínico de baja densidad (LDL). Después se impuso a cada sujeto la otra dieta. Después de un segundo periodo de dos semanas, se registró de nuevo el nivel de colesterol LDL de cada individuo. Los datos del estudio aparecen en la siguiente tabla [10].

	LDL (mmol/l)	
Individuo	**Hojuelas de maíz**	**Avena integral**
1	4.61	3.84
2	6.42	5.57
3	5.40	5.85
4	4.54	4.80
5	3.98	3.68
6	3.82	2.96
7	5.01	4.41
8	4.34	3.72
9	3.80	3.49
10	4.56	3.84
11	5.35	5.26
12	3.89	3.73
13	2.25	1.84
14	4.24	4.14

a) ¿Son las dos muestras de datos pareadas o independientes?

b) ¿Cuáles son las hipótesis nula y alterna de una prueba bilateral?

c) Lleve a cabo una prueba con un nivel de significancia de 0.05. ¿Cuál es el valor *p*?

d) ¿Qué concluye usted?

6. Suponga que le interesa determinar si la exposición al DDT —una molécula orgánica con cloro—, que se ha utilizado como insecticida por muchos años, está asociada con el cáncer de mama. Como parte de un estudio sobre el tema, se tomó sangre de una muestra de mujeres a las que se les diagnosticó cáncer de mama por un periodo de seis años y de una muestra de un grupo testigo de pacientes saludables relacionadas con los pacientes de cáncer en lo que se refiere a edad, estado de menopausia y fecha de donación de sangre. Se midió el nivel en la sangre de DDE —un importante derivado del DDT en el cuerpo humano— en cada mujer, y se calculó la diferencia de niveles de cada paciente y su control asociado. Una muestra de 171 de dichas diferencias tiene una media de $\bar{d} = 2.7$ ng/ml y una desviación estándar $s_d = 15.9$ ng/ml [11].

a) Aplique la prueba de la hipótesis nula de que los niveles medios de sangre de DDE son idénticos en mujeres con cáncer de mama y en mujeres saludables del grupo de control. ¿Qué concluye usted?

b) ¿Esperaría usted que un intervalo de confianza de 95% para la verdadera diferencia en los niveles medios de DDE en las poblaciones incluyera el valor 0? Explique.

7. Los siguientes datos provienen de un estudio que analiza la eficacia de cotinina en la saliva como indicador de la exposición al humo del tabaco. En una parte del estudio, se pidió a siete individuos —ninguno fumador habitual, y quienes se habían abstenido de fumar por lo menos una semana antes del estudio— que fumaran un solo cigarrillo. Se tomaron las muestras de saliva de todos los individuos 2, 12, 24 y 48 horas después de fumar. A continuación se muestran los niveles de cotinina a las 12 y 24 horas [12].

	Niveles de cotinina (nmol/l)	
Individuo	**Después de 12 horas**	**Después de 24 horas**
1	73	24
2	58	27
3	67	49
4	93	59
5	33	0
6	18	11
7	147	43

μ_{12} representa el nivel de cotinina medio de la población 12 horas después de fumar el cigarrillo y μ_{24} el nivel de cotinina media 24 horas después de fumar. Se cree que μ_{24} debe ser más bajo que μ_{12}.

a) Construya un intervalo de confianza unilateral de 95% para la verdadera diferencia en las medias de población $\mu_{12} - \mu_{24}$.

b) Aplique la prueba de la hipótesis nula de que las medias de población son idénticas en el nivel de significancia $\alpha = 0.05$. ¿Qué concluye usted?

8. Se llevó a cabo un estudio para determinar si el humo del cigarrillo de una futura madre tiene algún efecto en el contenido mineral óseo de su hijo por lo demás, saludable. Una muestra de 77 recién nacidos cuyas madres fumaron durante el embarazo tiene un contenido mineral óseo medio de $\bar{x}_1 = 0.098$ g/cm y una desviación estándar de $s_1 = 0.026$ g/cm; una muestra de 161 niños cuyas madres no fumaban tiene una media de $\bar{x}_2 = 0.095$ g/cm y una desviación estándar de $s_2 = 0.025$ g/cm [13]. Suponga que las varianzas de las poblaciones estudiadas son iguales.

a) ¿Son las dos muestras de datos pareadas o independientes?

b) Formule las hipótesis nula y alterna de la prueba bilateral.

c) Lleve a cabo una prueba con el nivel de significancia 0.05. ¿Qué concluye usted?

9. En una investigación sobre la hipertensión inducida por el embarazo, un grupo de mujeres con este padecimiento se sometieron a un tratamiento con dosis bajas de aspirina, y un segundo grupo recibió un placebo. Una muestra de 23 mujeres que tomaron aspirina tiene una presión arterial media de 111 mm Hg y una desviación estándar de 8 mm Hg; una muestra de 24 mujeres a quienes se les administró el placebo tiene una presión arterial media de 109 mm Hg y una desviación estándar de 8 mm Hg [14].

a) Con un nivel de significancia de 0.01, aplique la prueba de la hipótesis nula de que las dos poblaciones de mujeres tienen la misma presión arterial media.

b) Construya un intervalo de confianza de 99% para la verdadera diferencia en medias de población. ¿Contiene este intervalo el valor 0?

10. Como parte de la Prueba de Salud para Mujeres de Estados Unidos, se exhortó a un grupo de mujeres a adoptar una dieta baja en grasa mientras que un segundo grupo no recibió consejos dietéticos. Un año más tarde, las mujeres del grupo que participó mantuvieron exitosamente sus dietas. En aquel entonces, se llevó a cabo un estudio para determinar si sus esposos también tenían un nivel reducido de ingestión de grasa [15].

a) En el grupo de intervención, una muestra de 156 esposos tienen un consumo medio de grasa diario de $\bar{x}_1 = 54.8$ gramos y una desviación estándar de $s_1 = 28.1$ gramos. En el grupo control, una muestra de 148 esposos tiene una ingestión media de $\bar{x}_2 = 69.5$ gramos y una desviación estándar de $s_2 = 34.7$ gramos. Calcule intervalos de confianza separados de 95% para los verdaderos consumos de grasa medios de varones en cada grupo. Utilice estos intervalos para construir una gráfica como la de la figura 11.3. ¿Sugiere la gráfica que las dos medias de población probablemente sean iguales?

b) Pruebe formalmente la hipótesis nula de que los dos grupos de hombres tienen el mismo consumo dietético medio de grasa, con una prueba bilateral. ¿Qué concluye usted?

c) Construya un intervalo de confianza de 95% para la verdadera diferencia en las medias de población.

d) Un investigador podría también estar interesado en saber si los hombres difieren respecto de la ingestión de otros tipos de alimentos, como proteínas o carbohidratos. En el grupo de intervención, los esposos tienen una ingestión media diaria de carbohidratos de $\bar{x}_1 = 172.5$ gramos y una desviación estándar de $s_1 = 68.8$ gramos. En el grupo control, los varones tienen una ingestión media de carbohidratos de $\bar{x}_2 = 185.5$ gramos u una desviación estándar de $s_2 = 69.0$ gramos. Pruebe la hipótesis nula de que las dos poblaciones tienen la misma ingestión media de carbohidratos. ¿Qué concluye usted?

11. La siguiente tabla compara los niveles de carboxihemoglobina de un grupo de no fumadores y un grupo de fumadores de cigarrillos. Se muestran las medias muestrales y des-

viaciones estándares [16]. Se supone que el nivel medio de carboxihemoglobina de los fumadores debe ser más alto que el nivel medio de los no fumadores. No hay razón para suponer que las varianzas de población sean idénticas.

Grupo	n	Carboxihemoglobina (%)
No fumadores	121	$\bar{x} = 1.3, s = 1.3$
Fumadores	75	$\bar{x} = 4.1, s = 2.0$

a) ¿Cuáles son las hipótesis nula y alterna de la prueba unilateral?

b) Lleve a cabo la prueba con un nivel de significancia de 0.05. ¿Qué concluye usted?

12. Suponga que desea comparar las características de la meningitis tuberculosa en pacientes infectados con VIH y en los que no están infectados. En particular, desearía determinar si las dos poblaciones tienen la misma edad media. Una muestra de 37 pacientes infectados tiene una edad media de $\bar{x}_1 = 27.9$ años y una desviación estándar de $s_1 = 5.6$ años. Una muestra de 19 pacientes que no se encuentran infectados tiene una edad media de $\bar{x}_2 = 38.8$ años y una desviación estándar de $s_2 = 21.7$ años [17].

a) Pruebe la hipótesis nula de que las dos poblaciones de pacientes tienen la misma edad media con un nivel de significancia de 0.05.

b) ¿Espera usted que un intervalo de confianza de 95% para la verdadera diferencia en las medias de población contendría el valor 0? ¿Por qué?

13. Considere las cantidades de camas de hospitales comunitarios por cada 100 individuos disponibles en cada estado de Estados Unidos y en el distrito de Columbia. Los datos de 1980 y 1986 se encuentran en una serie de datos denominada *bed* [18] (apéndice B, tabla B.13) o en el sitio de la red mencionado en las páginas preliminares de este texto. Los valores de 1980 se encuentran en la variable *bed80*; los valores de 1986, en *bed86*. Una segunda serie de datos, llamada *bed2*, contiene la misma información en un diferente formato. Los números de camas por cada 100 individuos para ambos años se encuentran en la variable denominada *bed*, y un indicador de año en la variable *year*.

a) Genere estadísticos descriptivos para las cantidades de camas de hospital en cada año.

b) Puesto que hay dos observaciones para cada estado —una para 1980 y otra para 1986—, los datos son en realidad pareados. Un error común al analizar este tipo de datos consiste en ignorar la paridad y suponer que las muestras son independientes. Compare la cantidad media de camas de hospitales comunitarios por cada 1000 individuos en 1980 con la cantidad media de camas en 1986 utilizando la prueba t para dos muestras. ¿Qué concluye usted?

c) Ahora compare la cantidad media de camas en 1980 con la cantidad media en 1986 utilizando la prueba t pareada.

d) Comente las diferencias entre las dos pruebas. ¿Llega usted a la misma conclusión en cada caso?

e) Genere un intervalo de confianza de 95% para la verdadera diferencia en la cantidad media de camas de hospital en 1980 y 1986.

14. El conjunto de datos `lowbwt` contiene información de una muestra de 100 niños de bajo peso al nacer en dos hospitales generales en Boston, Massachusetts [19] (apéndice B, ta-

bla B.7). Las mediciones de presión arterial sistólica se encuentran en la variable denominada `sbp`, y los indicadores de género —donde 1 representa un hombre y 0 una mujer— con el nombre `sex`.

a) Construya un histograma de mediciones de presión arterial sistólica para esta muestra. Con base en esta gráfica, ¿cree usted que la presión arterial es muy cercana a la distribución normal?

b) Pruebe la hipótesis nula de que, entre los niños de bajo peso al nacer, la presión arterial sistólica media de los niños es igual a la media de las niñas. Aplique una prueba bilateral con un nivel de significancia de 0.05. ¿Qué concluye usted?

15. Las Escalas de Bayley de Desarrollo Infantil proporcionan resultados de dos índices —el Índice de Desarrollo Psicomotor (PDI) y el Índice de Desarrollo Mental (MDI)— que pueden emplearse para evaluar el nivel de funcionamiento de un niño a la edad de un año aproximadamente. Como parte de un estudio que analiza el estado de desarrollo y neurológico de niños que se sometieron a cirugía correctiva de corazón durante los primeros tres meses de vida, las escalas de Bayley se aplicaron a una muestra de niños de un año de edad con insuficiencia cardiaca congénita. Los niños se habían distribuido al azar en uno de los dos diferentes grupos de tratamiento, conocidos como "paro circulatorio" y "desviación de bajo flujo". Estos grupos difieren en la forma específica en que se realizó la cirugía correctiva. A diferencia del paro circulatorio, la desviación de bajo flujo mantiene continua la circulación hasta el cerebro; aunque parece que la prefieren algunos médicos, también tiene su propio riesgo de daño cerebral. Los datos de este estudio se encuentran almacenados en la serie de datos `heart` [20] (apéndice B, tabla B.12). Los resultados de PDI se guardan en la variable denominada `pdi`, los resultados de MDI, en `mdi`, y los indicadores de grupo de tratamiento en `trtment`. Para esta variable, 0 representa paro circulatorio y 1 desviación de bajo flujo.

a) Con un nivel de significancia de 0.05, pruebe la hipótesis nula de que el resultado medio de PDI a la edad de un año para el grupo de tratamiento de paro circulatorio es igual al resultado medio de PDI para el grupo de bajo flujo. ¿Cuál es el valor p para esta prueba?

b) Pruebe la hipótesis nula que indica que los resultados medios de MDI son idénticos en los dos grupos de tratamiento. ¿Cuál es el valor p?

c) ¿Qué sugieren estas pruebas acerca de la relación entre un grupo de tratamiento quirúrgico para un niño durante los primeros tres meses de vida y su estado de desarrollo subsecuente a la edad de un año?

Bibliografía

[1] Packard, F. R., *The Life and Times of Ambroise Paré,* 1510-1590, Nueva York, Paul B. Hoeber, 1921.

[2] Allred, E. N., Bleecker, E. R., Chaitman, B. R., Dahms, T. E., Gottlieb, S. O., Hackney, J. D., Hayes, D., Pagano, M., Selvester, R. H., Walden, S. M. y J. Warren, "Acute Effects of Carbon Monoxide Exposure on Individuales with Coronary Artery Disease", *Health Effects Institute Research Report Number 25,* noviembre de 1989.

[3] Markowski, C. A. y E. P. Markowski, "Conditions for the Effectiveness of a Preliminary Test of Variance", *The American Statistician,* vol. 44, noviembre de 1990, pp. 322-326.

[4] Moser, B. K. y G. R. Stevens, "Homogeneity of Variance in the Two-Sample Means Test", *The American Statistician,* vol. 46, febrero de 1992, pp. 19-21.

[5] Zempsky, W. T., Rosenstein, B. J., Carroll, J. A. y F. A. Oski, "Effect of Pancreatic Enzyme Supplements on Iron Absorption", *American Journal of Diseases of Children,* vol. 143, agosto de 1989, pp. 966-972.

[6] Satterthwaite, F. W., "An Approximate Distribution of Estimates of Variance Componentes", *Biometrics Bulletin,* vol. 2, diciembre de 1946, pp. 110-114.

[7] SHEP Cooperative Research Group, "Prevention of Stroke by Antihypertensive Drug Treatment in Older Persons with Isolated Systolic Hypertension: Final Results of the Systolic Hypertension in the Elderly Program (SHEP)", *Journal of the American Medical Association,* vol. 265, 26 de junio de 1991, pp. 3255-3264.

[8] Kien, C. L., Liechty, E. A. y M. D. Mullett, "Effects of Lactose Intake on Nutritional Status in Premature Infants", *Journal of Pediatrics,* vol. 116, marzo de 1990, pp. 446-449.

[9] Barrett-Connor, E. L., Cohn, B. A., Wingard, D. L. y Edelstein, S. L., "Why Is Diabetes Mellitus a Stronger Risk Factor for Fatal Ischemic Heart Disease in Women Than in Men?", *Journal of the American Medical Association,* vol. 265, 6 de febrero de 1991, pp. 627-631.

[10] Anderson, J. W., Spencer, D. B., Hamilton, C. C., Smith, S. F., Tietyen, J., Bryant, C. A. y P. Oeltgen, "Oat-Bran Cereal Lowers Serum Total and LDL Cholesterol in Hypercholesterolemic Men", *American Journal of Clinical Nutrition,* vol. 52, septiembre de 1990, pp. 495-499.

[11] Wolff, M. S., Toniolo, P. G., Lee, E. W., Rivera, M. y N. Dubin, "Blood Levels of Organichlorine Residues and Risk of Breast Cancer", *Journal of the National Cancer Institute,* vol. 85, 21 de abril de 1993, pp. 648-652.

[12] DiGiusto, E., e I. Eckard, "Some Properties of Saliva Cotinine Measurements in Indicating Exposure to Tobacco Smoking", *American Journal of Public Health,* vol. 76, octubre de 1986, pp. 1245-1246.

[13] Venkataraman, P. S. y J. C. Duke, "Bone Mineral Content of Healthy, Full-term Neonates: Effect of Race, Gender, and Maternal Cigarette Smoking", *American Journal of Diseases of Children,* vol. 145, noviembre de 1991, pp. 1310-1312.

[14] Schiff, E., Barkai, G., Ben-Baruch, G. y S. Mashiach, S., "Low-Dose Aspirin Does Not Influence the Clinical Course of Women with Mild Pregnancy-Induced Hypertension", *Obstetrics and Gynecology,* vol. 76, noviembre de 1990, pp. 742-744.

[15] Shattuck, A. L., White, E. y A. R. Kristal, "How Women's Adopted Low-Fat Diets Affect Their Husbands", *American Journal of Public Health,* vol. 82, septiembre de 1992, pp. 1244-1250.

[16] Jarvis, M. J., Tunstall-Pedoe, H. Feyerabend, C., Vesey, C. y Y. Saloojee, "Comparison of Tests Used to Distinguish Smokers from Nonsmokers", *American Journal of Public Health,* vol. 77, noviembre de 1987, pp. 1435-1438.

[17] Berenguer, J., Moreno, S., Laguna, F., Vicente, T., Adrados, M., Ortega, A., González-LaHoz, J. y Bouza, E., "Tuberculosis Meningitis in Patients Infected with the Human Immunodeficiency Virus", *The New England Journal of Medicine,* vol. 326, 5 de marzo de 1992, pp. 668-672.

[18] National Center for Health Statistics, *Health United States 1998,* Public Health Service, Hyattsville, Maryland; marzo de 1989.

[19] Leviton, A., Fenton, T., Kuban, K. C. K. y M. Pagano, "Labor and Delivery Characteristics and the Risk of Germinal Matrix Hemorrhage in Low Birth Weight Infants", *Journal of Child Neurology,* vol. 6, octubre de 1991, pp. 35-40.

[20] Bellinger, D. C., Jonas, R. A., Rappaport, L. A., Wypij, D., Wernovsky, G., Kuban, K. C. K., Barnes, P. D., Holmes, G. L., Hickey, P. R., Strand, R. D., Walsh, A. Z., Helmers, S. L., Constantinou, J. E., Carrazana, E. J., Mayer, J. E., Hanley, F. L., Castaneda, A. R., Ware, J. H. y J. W. Newburger, "Developmental and Neurologic Status of Children After Heart Surgery with Hipothermic Circulatory Arrest or Low-Flow Cardiopulmonary Bypass", *The New England Journal of Medicine,* vol. 332, 2 de marzo de 1995, pp. 549-555.