Capítulo 8

Distribuciones de muestreo fundamentales y descripciones de datos

8.1 Muestreo aleatorio

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

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

Poblaciones y muestras

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

www.FreeLibros.me

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

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

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

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

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

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

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

Para que las inferencias que hacemos sobre la población a partir de la muestra sean válidas, debemos obtener muestras que sean representativas de ella. Con mucha

www.FreeLibros.me

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

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

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

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

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

8.2 Algunos estadísticos importantes

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

Ahora, $\hat{p}$ es una función de los valores observados en la muestra aleatoria; ya que es posible tomar muchas muestras aleatorias de la misma población, esperaríamos

www.FreeLibros.me

que $\hat{p}$ variara un poco de una a otra muestra. Es decir, $\hat{p}$ es un valor de una variable aleatoria que representamos con P. Tal variable aleatoria se llama **estadístico**.

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

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

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

a) Media muestral:

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

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

b) Mediana muestral:

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

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

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

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

0.32 0.53 0.28 0.37 0.47 0.43 0.36 0.42 0.38 0.43

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

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

www.FreeLibros.me

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

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

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

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

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

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

a) La varianza muestral:

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

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

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

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

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

Por lo tanto,

$$s^2 = \frac{1}{3}\sum_{i=1}^{4}(x_i - 16)^2 = \frac{(12-16)^2 + (15-16)^2 + (17-16)^2 + (20-16)^2}{3}$$
$$= \frac{(-4)^2 + (-1)^2 + (1)^2 + (4)^2}{3} = \frac{34}{3}.$$

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

www.FreeLibros.me

Capítulo 9

Problemas de estimación de una y dos muestras

9.1 Introducción

En los capítulos anteriores destacamos las propiedades del muestreo de la media y de la varianza muestrales. También destacamos las representaciones de datos en varias formas. El propósito de estas presentaciones es establecer las bases que permitan a los estadísticos sacar conclusiones acerca de los parámetros de poblaciones tomadas de datos experimentales. Por ejemplo, el teorema del límite central brinda información sobre la distribución de la media muestral $\bar{X}$. La distribución incluye la media de la población μ. Por consiguiente, cualesquiera conclusiones respecto a μ, extraídas de un promedio muestral observado, deben depender de lo que se sabe acerca de su distribución muestral. Se podría decir algo similar en lo que se refiere a S^2 y σ^2. Como es evidente, es muy probable que cualquier conclusión que saquemos acerca de la varianza de una distribución normal implique la distribución muestral de S^2.

En este capítulo comenzaremos por presentar de manera formal el propósito de la inferencia estadística. Continuaremos con el análisis del problema de la **estimación de los parámetros de la población**. Restringiremos nuestros desarrollos formales de los procedimientos de estimación específicos a problemas que impliquen una y dos muestras.

9.2 Inferencia estadística

En el capítulo 1 presentamos la filosofía general de la inferencia estadística formal. La **inferencia estadística** consta de los métodos mediante los cuales se hacen inferencias o generalizaciones acerca de una población. La tendencia actual es distinguir entre el **método clásico** de estimación de un parámetro de la población, donde las inferencias se basan estrictamente en información obtenida de una muestra aleatoria seleccionada de la población, y el **método bayesiano**, el cual utiliza el conocimiento subjetivo que ya se posee sobre la distribución de probabilidad de los parámetros desconocidos junto con la información que proporcionan los datos de la muestra. En la mayor parte de este capítulo utilizaremos los métodos clásicos para estimar los parámetros de la población desconocidos, como la media, la proporción y la varianza, mediante el cálculo de estadísticos de muestras aleatorias y la aplicación de la teoría de las distribuciones muestrales, gran

www.FreeLibros.me

parte de lo cual se estudió en el capítulo 8. La estimación bayesiana se analizará en el capítulo 18.

La inferencia estadística se puede dividir en dos áreas principales: **estimación** y **pruebas de hipótesis**. Trataremos estas dos áreas por separado: en este capítulo veremos la teoría y las aplicaciones de la estimación, y en el capítulo 10 revisaremos la prueba de hipótesis. Para distinguir claramente un área de la otra, considere los siguientes ejemplos. Un candidato a un cargo público podría estar interesado en estimar la verdadera proporción de votantes que lo favorecerán mediante la obtención de las opiniones de una muestra aleatoria de 100 de ellos. La parte de votantes en la muestra que favorecerán al candidato se podría utilizar como un estimado de la verdadera proporción en la población de votantes. El conocimiento de la distribución muestral de una proporción nos permite establecer el grado de exactitud de tal estimado. Este problema cae en el área de la estimación.

Considere ahora el caso de alguien a quien le interesa averiguar si la marca *A* de cera para piso es más resistente al desgaste que la marca *B*. Se podría plantear la hipótesis de que la marca *A* es mejor que la marca *B* y, después de la prueba adecuada, aceptar o rechazar dicha hipótesis. En este ejemplo no intentamos estimar un parámetro, sino llegar a una decisión correcta acerca de una hipótesis planteada previamente. Una vez más, dependemos de la teoría del muestreo y de utilizar datos que nos proporcionen alguna medida del grado de exactitud de nuestra decisión.

9.3 Métodos de estimación clásicos

La **estimación puntual** de algún parámetro de la población θ es un solo valor $\hat{\theta}$ de un estadístico $\hat{\Theta}$. Por ejemplo, el valor $\bar{x}$ del estadístico $\bar{X}$, que se calcula a partir de una muestra de tamaño n, es una estimación puntual del parámetro de la población μ. De manera similar, $\hat{p} = x/n$ es una estimación puntual de la verdadera proporción p para un experimento binomial.

No se espera que un estimador logre estimar el parámetro de la población sin error. No se espera que $\bar{X}$ estime μ con exactitud, lo que en realidad se espera es que no esté muy alejada. Para una muestra específica, la manera en que se podría obtener un estimado más cercano de μ es utilizando la mediana de la muestra $\tilde{X}$ como estimador. Considere, por ejemplo, una muestra que consta de los valores 2, 5 y 11 de una población cuya media es 4, la cual, supuestamente, se desconoce. Podríamos estimar μ para que sea $\bar{x} = 6$ usando la media muestral como nuestro estimado, o bien, $\tilde{x} = 5$ utilizando la mediana muestral. En este caso el estimador $\tilde{X}$ produce una estimación más cercana al parámetro verdadero que la que produce el estimador $\bar{X}$. Por otro lado, si nuestra muestra aleatoria contiene los valores 2, 6 y 7, entonces $\bar{x} = 5$ y $\tilde{x} = 6$, de manera que el mejor estimador es $\bar{X}$. Cuando no conocemos el valor real de μ, tenemos que comenzar por decidir qué estimador utilizaremos, si $\bar{X}$ o $\tilde{X}$.

Estimador insesgado

¿Cuáles son las propiedades que una "buena" función de decisión debería tener para poder influir en nuestra elección de un estimador en vez de otro? Sea $\hat{\Theta}$ un estimador cuyo valor $\hat{\theta}$ es una estimación puntual de algún parámetro de la población desconocido θ. Sin duda desearíamos que la distribución muestral de $\hat{\Theta}$ tuviera una media igual al parámetro estimado. Al estimador que tuviera esta propiedad se le llamaría **estimador insesgado**.

www.FreeLibros.me

Definición 9.1: Se dice que un estadístico $\hat{\Theta}$ es un **estimador insesgado** del parámetro θ si

$$\mu_{\hat{\Theta}} = E(\hat{\Theta}) = \theta.$$

Ejemplo 9.1: Demuestre que S^2 es un estimador insesgado del parámetro σ^2.

***Solución*:** En la sección 8.5, en la página 244, demostramos que

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2.$$

Entonces,

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

Sin embargo,

$$\sigma_{X_i}^2 = \sigma^2, \text{ para } i = 1, 2, \ldots, n, \text{ y } \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}.$$

Por lo tanto,

$$E(S^2) = \frac{1}{n-1}\left(n\sigma^2 - n\frac{\sigma^2}{n}\right) = \sigma^2.$$

Aunque S^2 es un estimador insesgado de σ^2, S, por otro lado, suele ser un estimador sesgado de σ, un sesgo que en el caso de muestras grandes se vuelve insignificante. Este ejemplo ilustra **por qué dividimos entre $n - 1$** en vez de entre n cuando estimamos la varianza.

Varianza de un estimador puntual

Si $\hat{\Theta}_1$ y $\hat{\Theta}_2$ son dos estimadores insesgados del mismo parámetro de la población θ, deseamos elegir el estimador cuya distribución muestral tenga la menor varianza. Por lo tanto, si $\sigma_{\hat{\theta}_1}^2 < \sigma_{\hat{\theta}_2}^2$, decimos que $\hat{\Theta}_1$ es un **estimador más eficaz** de θ que $\hat{\Theta}_2$.

Definición 9.2: Si consideramos todos los posibles estimadores insesgados de algún parámetro θ, al que tiene la menor varianza lo llamamos estimador más eficaz de θ.

En la figura 9.1 se ilustran las distribuciones muestrales de tres estimadores diferentes $\hat{\Theta}_1$, $\hat{\Theta}_2$ y $\hat{\Theta}_3$, todos para θ. Es evidente que sólo $\hat{\Theta}_1$ y $\hat{\Theta}_2$ no son sesgados, ya que sus distribuciones están centradas en θ. El estimador $\hat{\Theta}_1$ tiene una varianza menor que $\hat{\Theta}_2$, por lo tanto, es más eficaz. En consecuencia, el estimador de θ que elegiríamos, entre los tres que estamos considerando, sería $\hat{\Theta}_1$.

Para poblaciones normales se puede demostrar que tanto $\bar{X}$ como $\tilde{X}$ son estimadores insesgados de la media de la población μ, pero la varianza de $\bar{X}$ es más pequeña que la varianza de $\tilde{X}$. Por consiguiente, los estimados $\bar{x}$ y $\tilde{x}$ serán, en promedio, iguales a

www.FreeLibros.me

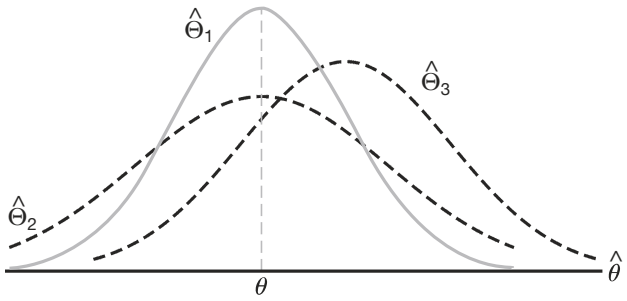

Figura 9.1: Distribuciones muestrales de diferentes estimadores de θ.

la media de la población μ, aunque podría ser que $\bar{x}$ esté más cerca de μ para una muestra dada y, por lo tanto, que $\bar{X}$ sea más eficaz que $\tilde{X}$.

Estimación por intervalo

Podría ser que ni el estimador insesgado más eficaz estime con exactitud el parámetro de la población. Es cierto que la exactitud de la estimación aumenta cuando las muestras son grandes; pero incluso así no tenemos razones para esperar que una **estimación puntual** de una muestra dada sea exactamente igual al parámetro de la población que se supone debe estimar. Hay muchas situaciones en que es preferible determinar un intervalo dentro del cual esperaríamos encontrar el valor del parámetro. Tal intervalo se conoce como **estimación por intervalo**.

Una estimación por intervalo de un parámetro de la población θ es un intervalo de la forma $\hat{\theta}_L < \theta < \hat{\theta}_U$, donde $\hat{\theta}_L$ y $\hat{\theta}_U$ dependen del valor del estadístico $\hat{\Theta}$ para una muestra específica, y también de la distribución de muestreo de $\hat{\Theta}$. Por ejemplo, una muestra aleatoria de calificaciones verbales de la prueba SAT para estudiantes universitarios de primer año produciría un intervalo de 530 a 550, dentro del cual esperamos encontrar el promedio verdadero de todas las calificaciones verbales de la prueba SAT para ese grupo. Los valores de los puntos extremos, 530 y 550, dependerán de la media muestral calculada $\bar{x}$ y de la distribución de muestreo de $\bar{X}$. A medida que aumenta el tamaño de la muestra, sabemos que $\sigma^2_{\bar{X}} = \sigma^2/n$ disminuye y, en consecuencia, cabe la posibilidad de que nuestra estimación se acerque más al parámetro μ, lo cual daría como resultado un intervalo más corto. De esta manera, el intervalo de la estimación indica, por su longitud, la precisión de la estimación puntual. Un ingeniero obtendrá información acerca de la proporción de la población de artículos defectuosos tomando una muestra y calculando la *proporción muestral defectuosa,* sin embargo, una estimación por intervalo podría ser más informativa.

Interpretación de las estimaciones por intervalo

Como muestras distintas suelen producir valores diferentes de $\hat{\Theta}$ y, por lo tanto, valores diferentes de $\hat{\theta}_L$ y $\hat{\theta}_U$, estos puntos extremos del intervalo son valores de las variables aleatorias correspondientes $\hat{\Theta}_L$ y $\hat{\Theta}_U$. De la distribución muestral de $\hat{\Theta}$ seremos capaces de determinar $\hat{\Theta}_L$ y $\hat{\Theta}_U$, de manera que $P(\hat{\Theta}_L < \theta < \hat{\Theta}_U)$ sea igual a cualquier

www.FreeLibros.me

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

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

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

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

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

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

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

Calcule

a) la media de la diferencia de puntos;

b) la mediana de las diferencias de puntos.

8.3 Distribuciones muestrales

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

Inferencias sobre la población a partir de información de la muestra

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

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

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

www.FreeLibros.me

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

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

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

8.4 Distribución muestral de medias y el teorema del límite central

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

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

tiene una distribución normal con media

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

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

www.FreeLibros.me

El teorema del límite central

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

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

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

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

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

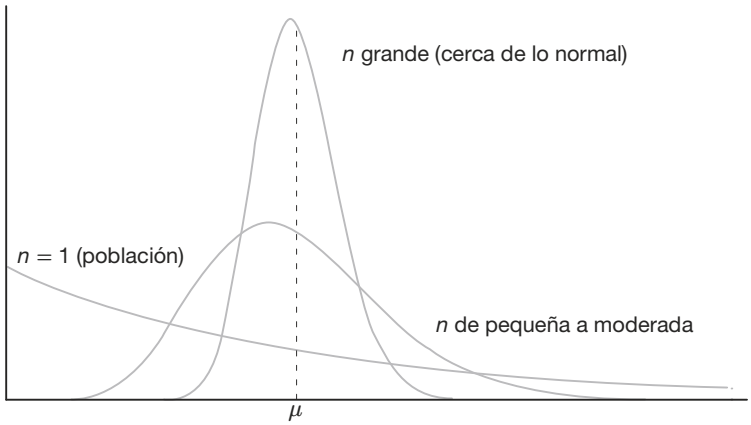

Figura 8.1: Ejemplo del teorema del límite central (distribución de $\bar{X}$ para $n = 1$, n moderada y n grande).

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

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

www.FreeLibros.me

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

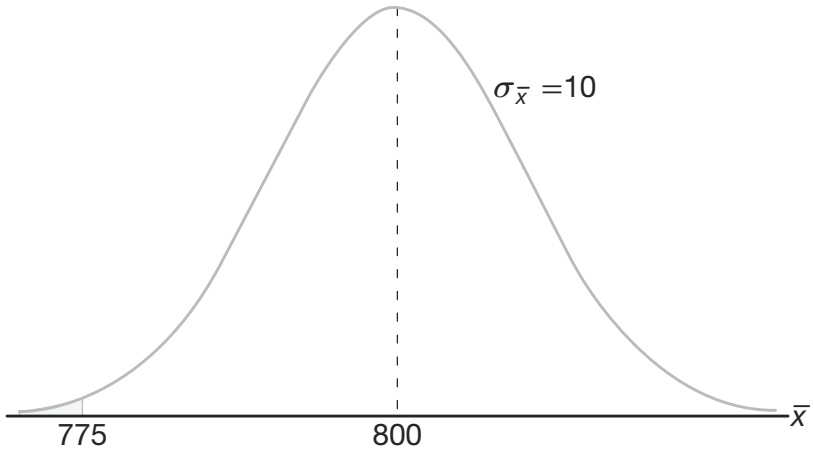

Figura 8.2: Área para el ejemplo 8.4.

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

y, por lo tanto,

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

Inferencias sobre la media de la población

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

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

Estudio de caso 8.1: Partes para automóviles. Un importante proceso de fabricación produce partes de componentes cilíndricos para la industria automotriz. Es importante que el proceso produzca partes que tengan un diámetro medio de 5.0 milímetros. El ingeniero implicado asume

www.FreeLibros.me

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

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

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

a) Utilice el teorema del límite central para determinar

$$P(\bar{X}_B - \bar{X}_A \geq 0.2)$$

bajo la condición de que $\mu_A = \mu_B$.

b) ¿Los experimentos mencionados parecen, de cualquier forma, apoyar consistentemente la suposición de que las medias de población de las dos máquinas son diferentes? Explique utilizando la respuesta que encontró en el inciso *a*.

8.33 El benceno es una sustancia química altamente tóxica para los seres humanos. Sin embargo, se utiliza en la fabricación de medicamentos, de tintes y de recubrimientos, así como en la peletería. Las regulaciones del gobierno establecen que el contenido de benceno en el agua que resulte de cualquier proceso de producción en el que participe esta sustancia no debe exceder 7950 partes por millón (ppm). Para un proceso particular de interés, un fabricante recolectó una muestra de agua 25 veces de manera aleatoria y el promedio muestral $\bar{x}$ fue de 7960 ppm. A partir de los datos históricos, se sabe que la desviación estándar σ es 100 ppm.

a) ¿Cuál es la probabilidad de que el promedio muestral en este experimento exceda el límite establecido por el gobierno, si la media de la población es igual al límite? Utilice el teorema del límite central.

b) ¿La $\bar{x} = 7960$ observada en este experimento es firme evidencia de que la media de la población en este proceso excede el límite impuesto por el gobierno? Responda calculando

$$P(\bar{X} \geq 7960 \mid \mu = 7950).$$

Suponga que la distribución de la concentración de benceno es normal.

8.34 En la fabricación de cierto producto de acero se están utilizando dos aleaciones, la *A* y la *B*. Se necesita diseñar un experimento para comparar las dos aleaciones en términos de su capacidad de carga máxima en toneladas, es decir, la cantidad máxima de carga que pueden soportar sin romperse. Se sabe que las dos desviaciones estándar de la capacidad de carga son iguales a 5 toneladas cada una. Se realiza un experimento en el que se prueban 30 especímenes de cada aleación (*A* y *B*) y se obtienen los siguientes resultados:

$$\bar{x}_A = 49.5, \quad \bar{x}_B = 45.5; \qquad \bar{x}_A - \bar{x}_B = 4.$$

Los fabricantes de la aleación *A* están convencidos de que esta evidencia demuestra de forma concluyente que $\mu_A > \mu_B$ y, por lo tanto, que su aleación es mejor. Los fabricantes de la aleación *B* afirman que el experimento fácilmente podría haber resultado $\bar{x}_A - \bar{x}_B = 4$, *incluso si* las dos medias de población fueran iguales. En otras palabras, "¡los resultados no son concluyentes!".

a) Encuentre un argumento que ponga en evidencia el error de los fabricantes de la aleación *B*. Para ello calcule

$$P(\bar{X}_A - \bar{X}_B > 4 \mid \mu_A = \mu_B).$$

b) ¿Considera que estos datos apoyan fuertemente a la aleación *A*?

8.35 Considere la situación del ejemplo 8.4 de la página 234. ¿Los resultados que se obtuvieron allí lo llevan a cuestionar la premisa de que $\mu = 800$ horas? Proporcione un resultado probabilístico que indique qué tan *raro* es el evento $\bar{X} \leq 775$ cuando $\mu = 800$. Por otro lado, ¿qué tan raro sería si μ fuera, verdaderamente, digamos, $\neq 760$ horas?

8.36 Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución que sólo puede adoptar valores positivos. Utilice el teorema del límite central para argumentar que si n es tan grande como se requiere, entonces $Y = X_1 X_2 \ldots X_n$ tiene aproximadamente una distribución logarítmica normal.

8.5 Distribución muestral de S^2

En la sección anterior aprendimos acerca de la distribución muestral de $\bar{X}$. El teorema del límite central nos permitió utilizar el hecho de que

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

www.FreeLibros.me

tiende a $N(0, 1)$ a medida que crece el tamaño de la muestra. *Las distribuciones muestrales de estadísticos importantes* nos permiten conocer información sobre los parámetros. Por lo general, los parámetros son las contrapartes del estadístico en cuestión. Por ejemplo, si un ingeniero se interesa en la resistencia media de la población de cierto tipo de resistencia, sacará provecho de la distribución muestral de $\bar{X}$ una vez que reúna la información de la muestra. Por otro lado, si está estudiando la variabilidad en la resistencia, evidentemente utilizará la distribución muestral de S^2 para conocer la contraparte paramétrica, la varianza de la población σ^2.

Si se extrae una muestra aleatoria de tamaño n de una población normal con media μ y varianza σ^2, y se calcula la varianza muestral, se obtiene un valor del estadístico S^2. Procederemos a considerar la distribución del estadístico $(n-1)S^2/\sigma^2$.

Mediante la suma y la resta de la media muestral $\bar{X}$ es fácil ver que

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\mu)^2 &= \sum_{i=1}^{n}[(X_i-\bar{X})+(\bar{X}-\mu)]^2\\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2+\sum_{i=1}^{n}(\bar{X}-\mu)^2+2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\bar{X})\\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2+n(\bar{X}-\mu)^2.\end{aligned}$$

Al dividir cada término de la igualdad entre σ^2 y sustituir $(n-1)S^2$ por $\sum_{i=1}^{n}(X_i-\bar{X})^2$, obtenemos

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2=\frac{(n-1)S^2}{\sigma^2}+\frac{(\bar{X}-\mu)^2}{\sigma^2/n}.$$

Ahora, de acuerdo con el corolario 7.1 de la página 222, sabemos que

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

es una variable aleatoria chi cuadrada con n grados de libertad. Tenemos una variable aleatoria chi cuadrada con n grados de libertad dividida en dos componentes. Observe que en la sección 6.7 demostramos que una distribución chi cuadrada es un caso especial de la distribución gamma. El segundo término del lado derecho es Z^2, que es una variable aleatoria chi cuadrada con 1 grado de libertad, y resulta que $(n-1)S^2/\sigma^2$ es una variable aleatoria chi cuadrada con $n-1$ grados de libertad. Formalizamos esto en el siguiente teorema.

Teorema 8.4: Si S^2 es la varianza de una muestra aleatoria de tamaño n que se toma de una población normal que tiene la varianza σ^2, entonces el estadístico

$$\chi^2=\frac{(n-1)S^2}{\sigma^2}=\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{\sigma^2}$$

tiene una distribución chi cuadrada con $v=n-1$ grados de libertad.

www.FreeLibros.me

Los valores de la variable aleatoria χ^2 se calculan de cada muestra mediante la fórmula

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}.$$

La probabilidad de que una muestra aleatoria produzca un valor χ^2 mayor que algún valor específico es igual al área bajo la curva a la derecha de este valor. El valor χ^2 por arriba del cual se encuentra un área de α por lo general se representa con χ^2_α. Esto se ilustra mediante la región sombreada de la figura 8.7.

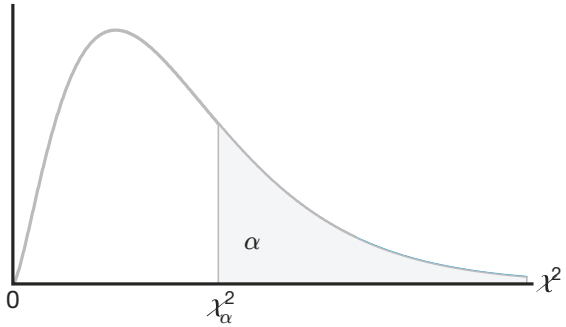

Figura 8.7: La distribución chi cuadrada.

La tabla A.5 da los valores de χ^2_α para diversos valores de α y v. Las áreas, α, son los encabezados de las columnas; los grados de libertad, v, se dan en la columna izquierda, y las entradas de la tabla son los valores χ^2. En consecuencia, el valor χ^2 con 7 grados de libertad, que deja un área de 0.05 a la derecha, es $\chi^2_{0.05} = 14.067$. Debido a la falta de simetría, para encontrar $\chi^2_{0.95} = 2.167$ para $v = 7$ también debemos usar las tablas.

Exactamente 95% de una distribución chi cuadrada cae entre $\chi^2_{0.975}$ y $\chi^2_{0.025}$. Un valor χ^2 que cae a la derecha de $\chi^2_{0.025}$ no tiene probabilidades de ocurrir, a menos que el valor de σ^2 que supusimos sea demasiado pequeño. Lo mismo sucede con un valor χ^2 que cae a la izquierda de $\chi^2_{0.975}$, el cual tampoco es probable que ocurra, a menos que el valor de σ^2 que supusimos sea demasiado grande. En otras palabras, es posible tener un valor χ^2 a la izquierda de $\chi^2_{0.975}$ o a la derecha de $\chi^2_{0.025}$ cuando el valor de σ^2 es correcto; pero si esto sucediera, lo más probable es que el valor de σ^2 que se supuso sea un error.

Ejemplo 8.7: Un fabricante de baterías para automóvil garantiza que su producto durará, en promedio, 3 años con una desviación estándar de 1 año. Si cinco de estas baterías tienen duraciones de 1.9, 2.4, 3.0, 3.5 y 4.2 años, ¿el fabricante continuará convencido de que sus baterías tienen una desviación estándar de 1 año? Suponga que las duraciones de las baterías siguen una distribución normal.

***Solución*:** Primero se calcula la varianza de la muestra usando el teorema 8.1,

$$s^2 = \frac{(5)(48.26) - (15)^2}{(5)(4)} = 0.815.$$

Entonces,

$$\chi^2 = \frac{(4)(0.815)}{1} = 3.26$$

www.FreeLibros.me

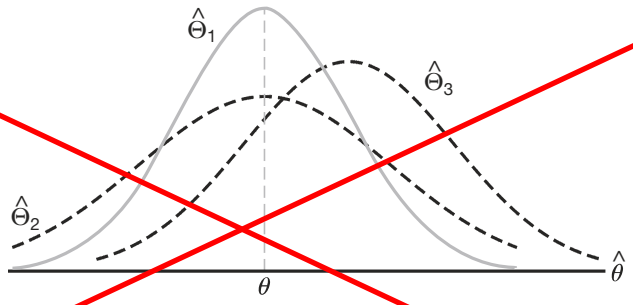

Figura 9.1: Distribuciones muestrales de diferentes estimadores de θ.

la media de la población μ, aunque podría ser que $\bar{x}$ esté más cerca de μ para una muestra dada y, por lo tanto, que $\bar{X}$ sea más eficaz que $\tilde{X}$.

Estimación por intervalo

Podría ser que ni el estimador insesgado más eficaz estime con exactitud el parámetro de la población. Es cierto que la exactitud de la estimación aumenta cuando las muestras son grandes; pero incluso así no tenemos razones para esperar que una **estimación puntual** de una muestra dada sea exactamente igual al parámetro de la población que se supone debe estimar. Hay muchas situaciones en que es preferible determinar un intervalo dentro del cual esperaríamos encontrar el valor del parámetro. Tal intervalo se conoce como **estimación por intervalo**.

Una estimación por intervalo de un parámetro de la población θ es un intervalo de la forma $\hat{\theta}_L < \theta < \hat{\theta}_U$, donde $\hat{\theta}_L$ y $\hat{\theta}_U$ dependen del valor del estadístico $\hat{\Theta}$ para una muestra específica, y también de la distribución de muestreo de $\hat{\Theta}$. Por ejemplo, una muestra aleatoria de calificaciones verbales de la prueba SAT para estudiantes universitarios de primer año produciría un intervalo de 530 a 550, dentro del cual esperamos encontrar el promedio verdadero de todas las calificaciones verbales de la prueba SAT para ese grupo. Los valores de los puntos extremos, 530 y 550, dependerán de la media muestral calculada $\bar{x}$ y de la distribución de muestreo de $\bar{X}$. A medida que aumenta el tamaño de la muestra, sabemos que $\sigma^2_{\bar{X}} = \sigma^2/n$ disminuye y, en consecuencia, cabe la posibilidad de que nuestra estimación se acerque más al parámetro μ, lo cual daría como resultado un intervalo más corto. De esta manera, el intervalo de la estimación indica, por su longitud, la precisión de la estimación puntual. Un ingeniero obtendrá información acerca de la proporción de la población de artículos defectuosos tomando una muestra y calculando la *proporción muestral defectuosa,* sin embargo, una estimación por intervalo podría ser más informativa.

Interpretación de las estimaciones por intervalo

Como muestras distintas suelen producir valores diferentes de $\hat{\Theta}$ y, por lo tanto, valores diferentes de $\hat{\theta}_L$ y $\hat{\theta}_U$, estos puntos extremos del intervalo son valores de las variables aleatorias correspondientes $\hat{\Theta}_L$ y $\hat{\Theta}_U$. De la distribución muestral de $\hat{\Theta}$ seremos capaces de determinar $\hat{\Theta}_L$ y $\hat{\Theta}_U$, de manera que $P(\hat{\Theta}_L < \theta < \hat{\Theta}_U)$ sea igual a cualquier

www.FreeLibros.me

valor positivo de una fracción que queramos especificar. Si, por ejemplo, calculamos $\hat{\Theta}_L$ y $\hat{\Theta}_U$, tales que

$$P(\hat{\Theta}_L < \theta < \hat{\Theta}_U) = 1 - \alpha,$$

para $0 < \alpha < 1$, tenemos entonces una probabilidad de $1 - \alpha$ de seleccionar una muestra aleatoria que produzca un intervalo que contenga θ. El intervalo $\hat{\theta}_L < \theta < \hat{\theta}_U$, que se calcula a partir de la muestra seleccionada, se llama entonces **intervalo de confianza** del $100(1 - \alpha)\%$, la fracción $1 - \alpha$ se denomina **coeficiente de confianza** o **grado de confianza**, y los extremos, $\hat{\theta}_L$ y $\hat{\theta}_U$, se denominan **límites de confianza** inferior y superior. Así, cuando $\alpha = 0.05$, tenemos un intervalo de confianza del 95%, y cuando $\alpha = 0.01$ obtenemos un intervalo de confianza más amplio del 99%. Cuanto más amplio sea el intervalo de confianza, más confiaremos en que contiene el parámetro desconocido. Desde luego, es mejor tener un 95% de confianza en que la vida promedio de cierto transistor de un televisor está entre los 6 y los 7 años, que tener un 99% de confianza en que esté entre los 3 y los 10 años. De manera ideal, preferimos un intervalo corto con un grado de confianza alto. Algunas veces las restricciones en el tamaño de nuestra muestra nos impiden tener intervalos cortos sin sacrificar cierto grado de confianza.

En las siguientes secciones estudiaremos los conceptos de estimación puntual y por intervalos, y en cada sección presentaremos un caso especial diferente. El lector debería notar que, aunque la estimación puntual y por intervalos representan diferentes aproximaciones para obtener información respecto a un parámetro, están relacionadas debido a que los estimadores del intervalo de confianza se basan en estimadores puntuales. En la siguiente sección, por ejemplo, veremos que $\overline{X}$ es un estimador puntual de μ muy razonable. Como resultado, el importante estimador del intervalo de confianza de μ depende del conocimiento de la distribución muestral de $\overline{X}$.

Empezaremos la siguiente sección con el caso más sencillo de un intervalo de confianza, en donde el escenario es simple pero poco realista. Nos interesa estimar una media de la población μ cuando σ todavía se desconoce. Evidentemente, si se desconoce μ es muy improbable que se conozca σ. Cualquier información histórica que produzca datos suficientes para permitir suponer que se conoce σ probablemente habría producido información similar acerca de μ. A pesar de este argumento iniciamos con este caso porque los conceptos y los mecanismos resultantes asociados con la estimación del intervalo de confianza también estarán asociados con las situaciones más realistas que presentaremos más adelante en la sección 9.4 y las siguientes.

9.4 Una sola muestra: estimación de la media

La distribución muestral de $\overline{X}$ está centrada en μ y en la mayoría de las aplicaciones la varianza es más pequeña que la de cualesquiera otros estimadores de μ. Por lo tanto, se utilizará la media muestral $\bar{x}$ como una estimación puntual para la media de la población μ. Recuerde que $\sigma^2_{\bar{X}} = \sigma^2/n$, por lo que una muestra grande producirá un valor de $\overline{X}$ procedente de una distribución muestral con varianza pequeña. Por consiguiente, es probable que $\bar{x}$ sea una estimación muy precisa de μ cuando n es grande.

Consideremos ahora la estimación por intervalos de μ. Si seleccionamos nuestra muestra a partir de una población normal o, a falta de ésta, si n es suficientemente grande, podemos establecer un intervalo de confianza para μ considerando la distribución muestral de $\overline{X}$.

www.FreeLibros.me

De acuerdo con el teorema del límite central, podemos esperar que la distribución muestral de $\bar{X}$ esté distribuida de forma aproximadamente normal con media $\mu_{\bar{X}} = \mu$ y desviación estándar $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. Al escribir $z_{\alpha/2}$ para el valor z por arriba del cual encontramos una área de $\alpha/2$ bajo la curva normal, en la figura 9.2 podemos ver que

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha,$$

donde

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

En consecuencia,

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha.$$

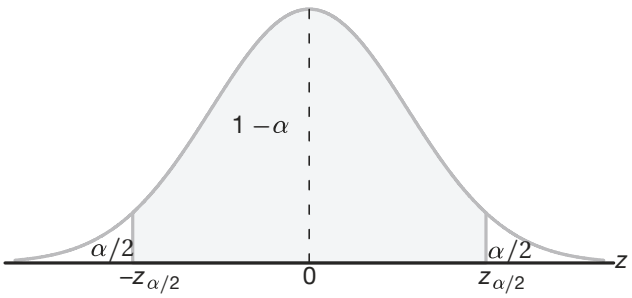

Figura 9.2: $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$.

Si multiplicamos cada término en la desigualdad por $\sigma/\sqrt{n}$ y después restamos $\bar{X}$ de cada término, y en seguida multiplicamos por − 1 (para invertir el sentido de las desigualdades), obtenemos

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Se selecciona una muestra aleatoria de tamaño n de una población cuya varianza σ^2 se conoce y se calcula la media $\bar{x}$ para obtener el intervalo de confianza $100(1 - \alpha)\%$. Es importante enfatizar que recurrimos al teorema del límite central citado anteriormente. Como resultado, es importante observar las condiciones para las aplicaciones que siguen.

Intervalo de confianza de μ cuando se conoce σ^2

Si $\bar{x}$ es la media de una muestra aleatoria de tamaño n de una población de la que se conoce su varianza σ^2, lo que da un intervalo de confianza de $100(1 - \alpha)\%$ para μ es

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

donde $z_{\alpha/2}$ es el valor z que deja una área de $\alpha/2$ a la derecha.

En el caso de muestras pequeñas que se seleccionan de poblaciones no normales, no podemos esperar que nuestro grado de confianza sea preciso. Sin embargo, para muestras

www.FreeLibros.me

de tamaño $n \geq 30$, en las que la forma de las distribuciones no esté muy sesgada, la teoría de muestreo garantiza buenos resultados.

Queda claro que los valores de las variables aleatorias $\hat{\Theta}_L$ y $\hat{\Theta}_U$, las cuales se definieron en la sección 9.3, son los límites de confianza

$$\hat{\theta}_L = \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{y} \quad \hat{\theta}_U = \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Muestras diferentes producirán valores diferentes de $\bar{x}$ y, por lo tanto, producirán diferentes estimaciones por intervalos del parámetro μ, como se muestra en la figura 9.3. Los puntos en el centro de cada intervalo indican la posición de la estimación puntual $\bar{x}$ para cada muestra aleatoria. Observe que todos los intervalos tienen el mismo ancho, pues esto depende sólo de la elección de $z_{\alpha/2}$ una vez que se determina $\bar{x}$. Cuanto más grande sea el valor de $z_{\alpha/2}$ que elijamos, más anchos haremos todos los intervalos, y podremos tener más confianza en que la muestra particular que seleccionemos producirá un intervalo que contenga el parámetro desconocido μ. En general, para una elección de $z_{\alpha/2}$, $100(1-\alpha)\%$ de los intervalos contendrá μ.

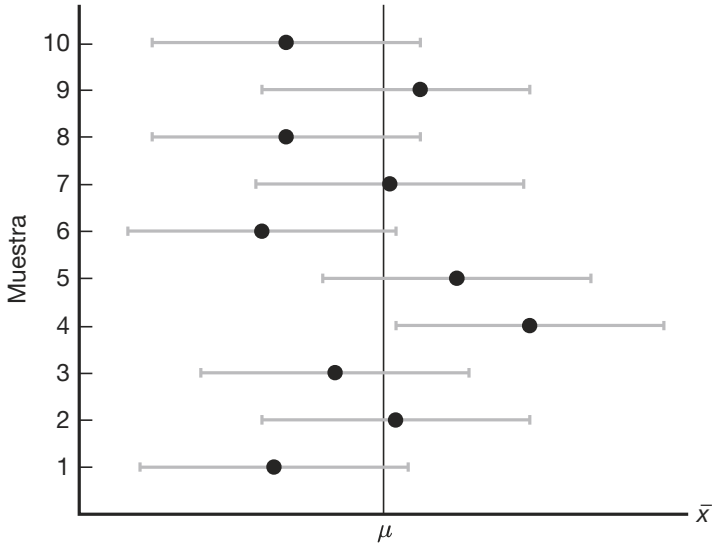

Figura 9.3: Estimaciones por intervalos de μ para muestras diferentes.

Ejemplo 9.2: Se encuentra que la concentración promedio de zinc que se obtiene en una muestra de mediciones en 36 sitios diferentes de un río es de 2.6 gramos por mililitro. Calcule los intervalos de confianza del 95% y 99% para la concentración media de zinc en el río. Suponga que la desviación estándar de la población es de 0.3 gramos por mililitro.

***Solución*:** La estimación puntual de μ es $\bar{x} = 2.6$. El valor z que deja una área de 0.025 a la derecha y, por lo tanto, una área de 0.975 a la izquierda es $z_{0.025} = 1.96$ (véase la tabla A.3). En consecuencia, el intervalo de confianza del 95% es

$$2.6 - (1.96)\left(\frac{0.3}{\sqrt{36}}\right) < \mu < 2.6 + (1.96)\left(\frac{0.3}{\sqrt{36}}\right),$$

www.FreeLibros.me

que se reduce a $2.50 < \mu < 2.70$. Para calcular un intervalo de confianza del 99% encontramos el valor z que deja una área de 0.005 a la derecha y de 0.995 a la izquierda. Por lo tanto, usando la tabla A.3 nuevamente, $z_{0.005} = 2.575$ y el intervalo de confianza de 99% es

$$2.6 - (2.575)\left(\frac{0.3}{\sqrt{36}}\right) < \mu < 2.6 + (2.575)\left(\frac{0.3}{\sqrt{36}}\right),$$

o simplemente

$$2.47 < \mu < 2.73.$$

Ahora vemos que se requiere un intervalo más grande para estimar μ con un mayor grado de confianza.

El intervalo de confianza del $100(1-\alpha)\%$ ofrece un estimado de la precisión de nuestra estimación puntual. Si μ es realmente el valor central del intervalo, entonces $\bar{x}$ estima μ sin error. La mayoría de las veces, sin embargo, $\bar{x}$ no será exactamente igual a μ y la estimación puntual será errónea. La magnitud de este error será el valor absoluto de la diferencia entre μ y $\bar{x}$, de manera que podemos tener $100(1-\alpha)\%$ de confianza en que esta diferencia no excederá a $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. Podemos ver esto fácilmente dibujando un diagrama de un intervalo de confianza hipotético, como el de la figura 9.4.

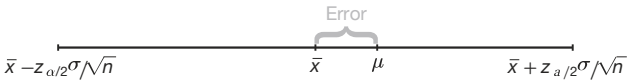

Figura 9.4: Error en la estimación de μ mediante $\bar{x}$.

Teorema 9.1: Si utilizamos $\bar{x}$ como una estimación de μ, podemos tener $100(1-\alpha)\%$ de confianza en que el error no excederá a $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

En el ejemplo 9.2 tenemos una confianza del 95% en que la media muestral $\bar{x} = 2.6$ difiere de la media verdadera μ en una cantidad menor que $(1.96)(0.3)/\sqrt{36} = 0.1$ y 99% de confianza en que la diferencia es menor que $(2.575)(0.3)/\sqrt{36} = 0.13$.

Con frecuencia queremos saber qué tan grande necesita ser una muestra para poder estar seguros de que el error al estimar μ será menor que una cantidad específica e. Por medio del teorema 9.1 debemos elegir n de manera que $z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = e$. Al resolver esta ecuación obtenemos la siguiente fórmula para n.

Teorema 9.2: Si usamos $\bar{x}$ como una estimación de μ, podemos tener $100(1-\alpha)\%$ de confianza en que el error no excederá a una cantidad específica e cuando el tamaño de la muestra sea

$$n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2.$$

Cuando resolvemos para la muestra con tamaño n, redondeamos todos los valores decimales al siguiente número entero. Si seguimos este principio, podemos estar seguros de que nuestro grado de confianza nunca caerá por debajo del $100(1-\alpha)\%$.

www.FreeLibros.me

9.63 Se llevará a cabo un estudio para estimar el porcentaje de ciudadanos de una ciudad que están a favor de tener agua fluorada. ¿Qué tan grande debería ser la muestra si se desea tener al menos 95% de confianza en que el estimado esté dentro del 1% del porcentaje verdadero?

9.64 Se realizará un estudio para estimar la proporción de residentes de cierta ciudad y sus suburbios que está a favor de que se construya una planta de energía nuclear cerca de la ciudad. ¿Qué tan grande debería ser la muestra, si se desea tener al menos un 95% de confianza en que el estimado esté dentro del 0.04 de la verdadera proporción de residentes que están a favor de que se construya la planta de energía nuclear?

9.65 A cierto genetista le interesa determinar la proporción de hombres y mujeres de la población que padecen cierto trastorno sanguíneo menor. En una muestra aleatoria de 1000 hombres encuentra que 250 lo padecen; mientras que de 1000 mujeres examinadas, 275 parecen padecerlo. Calcule un intervalo de confianza del 95% para la diferencia entre la proporción de hombres y mujeres que padecen el trastorno sanguíneo.

9.66 Se encuestan 10 escuelas de ingeniería de Estados Unidos. La muestra contiene a 250 ingenieros eléctricos, de los cuales 80 son mujeres; y 175 ingenieros químicos, de los cuales 40 son mujeres. Calcule un intervalo de confianza del 90% para la diferencia entre la proporción de mujeres en estos dos campos de la ingeniería. ¿Hay una diferencia significativa entre las dos proporciones?

9.67 Se llevó a cabo una prueba clínica para determinar si cierto tipo de vacuna tiene un efecto sobre la incidencia de cierta enfermedad. Una muestra de 1000 ratas, 500 de las cuales recibieron la vacuna, se mantuvo en un ambiente controlado durante un periodo de un año. En el grupo que no fue vacunado, 120 ratas presentaron la enfermedad, mientras que en el grupo inoculado 98 ratas la contrajeron. Si p_1 es la probabilidad de incidencia de la enfermedad en las ratas sin vacuna y p_2 es la probabilidad de incidencia en las ratas inoculadas, calcule un intervalo de confianza del 90% para $p_1 - p_2$.

9.68 En el estudio *Germination and Emergence of Broccoli,* realizado por el Departamento de horticultura del Virginia Tech, un investigador encontró que a 5°C, de 20 semillas de brócoli germinaron 10; en tanto que a 15°C, de 20 semillas germinaron 15. Calcule un intervalo de confianza del 95% para la diferencia en la proporción de semillas que germinaron a las dos temperaturas y decida si esta diferencia es significativa.

9.69 Una encuesta de 1000 estudiantes reveló que 274 eligen al equipo profesional de beisbol *A* como su equipo favorito. En 1991 se realizó una encuesta similar con 760 estudiantes y 240 de ellos también eligieron a ese equipo como su favorito. Calcule un intervalo de confianza del 95% para la diferencia entre la proporción de estudiantes que favorecen al equipo *A* en las dos encuestas. ¿Hay una diferencia significativa?

9.70 De acuerdo con el *USA Today* (17 de marzo de 1997), las mujeres constituían el 33.7% del personal de redacción en las estaciones locales de televisión en 1990 y el 36.2% en 1994. Suponga que en 1990 y en 1994 se contrataron 20 nuevos empleados para el personal de redacción.

a) Estime el número de trabajadores que habrían sido mujeres en 1990 y en 1994, respectivamente.

b) Calcule un intervalo de confianza del 95% para saber si hay evidencia de que la proporción de mujeres contratadas para el equipo de redacción fue mayor en 1994 que en 1990.

9.12 Una sola muestra: estimación de la varianza

Si extraemos una muestra de tamaño n de una población normal con varianza σ^2 y calculamos la varianza muestral s^2, obtenemos un valor del estadístico S^2. Esta varianza muestral calculada se utiliza como una estimación puntual de σ^2. En consecuencia, al estadístico S^2 se le denomina estimador de σ^2

Se puede establecer una estimación por intervalos de σ^2 utilizando el estadístico

$$X^2 = \frac{(n-1)S^2}{\sigma^2}.$$

De acuerdo con el teorema 8.4, cuando las muestras se toman de una población normal el estadístico X^2 tiene una distribución chi cuadrada con $n - 1$ grados de libertad. Podemos escribir (véase la figura 9.7)

$$P(\chi^2_{1-\alpha/2} < X^2 < \chi^2_{\alpha/2}) = 1 - \alpha,$$

www.FreeLibros.me

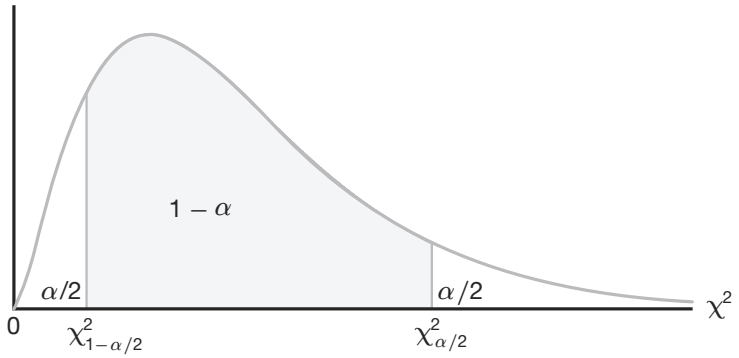

Figura 9.7: $P\left(\chi^2_{1-\alpha/2} < X^2 < \chi^2_{\alpha/2}\right) = 1-\alpha.$

donde $\chi^2_{1-\alpha/2}$ y $\chi^2_{\alpha/2}$ son valores de la distribución chi cuadrada con $n-1$ grados de libertad, que dejan áreas de $1-\alpha/2$ y $\alpha/2$, respectivamente, a la derecha. Al sustituir para X^2 escribimos

$$P\left[\chi^2_{1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}\right] = 1-\alpha.$$

Si dividimos cada término de la desigualdad entre $(n-1)S^2$, y después invertimos cada término (lo que cambia el sentido de las desigualdades), obtenemos

$$P\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right] = 1-\alpha.$$

Para una muestra aleatoria de tamaño *n,* tomada de una población normal, se calcula la varianza muestral s^2 y se obtiene el siguiente intervalo de confianza del $100(1-\alpha)\%$ para σ^2.

Intervalo de confianza para σ^2

Si s^2 es la varianza de una muestra aleatoria de tamaño n de una población normal, un intervalo de confianza del $100(1-\alpha)\%$ para σ^2 es

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}},$$

donde $\chi^2_{\alpha/2}$ y $\chi^2_{1-\alpha/2}$ son valores χ^2 con $v = n-1$ grados de libertad, que dejan áreas de $\alpha/2$ y $1-\alpha/2$, respectivamente, a la derecha.

Un intervalo de confianza aproximado a $100(1-\alpha)\%$ para σ se obtiene tomando la raíz cuadrada de cada extremo del intervalo para σ^2.

Ejemplo 9.18: Los siguientes son los pesos, en decagramos, de 10 paquetes de semillas de pasto distribuidas por cierta empresa: 46.4, 46.1, 45.8, 47.0, 46.1. 45.9, 45.8, 46.9, 45.2 y 46.0. Calcule un intervalo de confianza del 95% para la varianza de todos los pesos de este tipo de paquetes de semillas de pasto distribuidos por la empresa. Suponga una población normal.

***Solución*:** Primero calculamos

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

$$= \frac{(10)(21{,}273.12) - (461.2)^2}{(10)(9)} = 0.286.$$

www.FreeLibros.me

Para obtener un intervalo de confianza del 95% elegimos $\alpha = 0.05$. Después, usando la tabla A.5 con $v = 9$ grados de libertad, encontramos $\chi^2_{0.025} = 19.023$ y $\chi^2_{0.975} = 2.700$. Por lo tanto, el intervalo de confianza del 95% para σ^2 es

$$\frac{(9)(0.286)}{19.023} < \sigma^2 < \frac{(9)(0.286)}{2.700},$$

o simplemente $0.135 < \sigma^2 < 0.953$. ■

9.13 Dos muestras: estimación de la proporción de dos varianzas

Una estimación puntual de la proporción de dos varianzas de la población σ_1^2/σ_2^2 es dada por la proporción s_1^2/s_2^2 de las varianzas muestrales. En consecuencia, el estadístico S_1^2/S_2^2 se conoce como un estimador de σ_1^2/σ_2^2.

Si σ_1^2 y σ_2^2 son las varianzas de poblaciones normales, podemos establecer una estimación por intervalos de σ_1^2/σ_2^2 usando el estadístico

$$F = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}.$$

De acuerdo con el teorema 8.8, la variable aleatoria F tiene una distribución F con $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$ grados de libertad. Por lo tanto, podemos escribir (véase la figura 9.8)

$$P[f_{1-\alpha/2}(v_1, v_2) < F < f_{\alpha/2}(v_1, v_2)] = 1 - \alpha,$$

donde $f_{1-\alpha/2}(v_1, v_2)$ y $f_{\alpha/2}(v_1, v_2)$ son los valores de la distribución F con v_1 y v_2 grados de libertad, que dejan áreas de $1 - \alpha/2$ y $\alpha/2$, respectivamente, a la derecha.

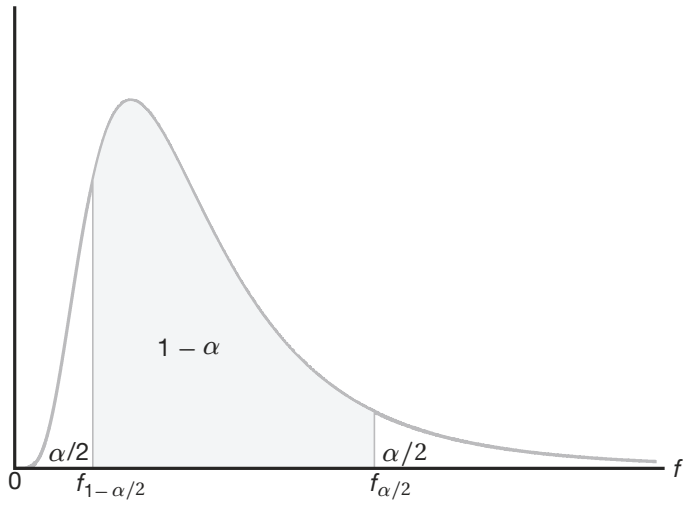

Figura 9.8: $P[f_{1-\alpha/2}(v_1, v_2) < F < f_{\alpha/2}(v_1, v_2)] = 1 - \alpha$.

www.FreeLibros.me

Al sustituir para *F,* escribimos

$$P\left[f_{1-\alpha/2}(v_1, v_2) < \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} < f_{\alpha/2}(v_1, v_2)\right] = 1 - \alpha.$$

Si multiplicamos cada término de la desigualdad por S_2^2/S_1^2, y después invertimos cada término, obtenemos

$$P\left[\frac{S_1^2}{S_2^2}\frac{1}{f_{\alpha/2}(v_1, v_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\frac{1}{f_{1-\alpha/2}(v_1, v_2)}\right] = 1 - \alpha.$$

Los resultados del teorema 8.7 nos permiten reemplazar la cantidad $f_{1-\alpha/2}(v_1, v_2)$ por $1/f_{\alpha/2}(v_2, v_1)$. Por lo tanto,

$$P\left[\frac{S_1^2}{S_2^2}\frac{1}{f_{\alpha/2}(v_1, v_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} f_{\alpha/2}(v_2, v_1)\right] = 1 - \alpha.$$

Para cualesquiera dos muestras aleatorias independientes de tamaño n_1 y n_2 que se seleccionan de dos poblaciones normales, se calcula la proporción de las varianzas muestrales s_1^2/s_2^2 y se obtiene el siguiente intervalo de confianza del $100(1-\alpha)\%$ para σ_1^2/σ_2^2.

Intervalo de confianza para σ_1^2/σ_2^2

Si s_1^2 y s_2^2 son las varianzas de muestras independientes de tamaño n_1 y n_2, respectivamente, tomadas de poblaciones normales, entonces un intervalo de confianza del $100(1-\alpha)\%$ para σ_1^2/σ_2^2 es

$$\frac{s_1^2}{s_2^2}\frac{1}{f_{\alpha/2}(v_1, v_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} f_{\alpha/2}(v_2, v_1),$$

donde $f_{\alpha/2}(v_1, v_2)$ es un valor f con $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$ grados de libertad que deja una área de $\alpha/2$ a la derecha, y $f_{\alpha/2}(v_2, v_1)$ es un valor f similar con $v_2 = n_2 - 1$ y $v_1 = n_1 - 1$ grados de libertad.

Como vimos en la sección 9.12, tomando la raíz cuadrada de cada extremo del intervalo para σ_1^2/σ_2^2, se obtiene un intervalo de confianza del $100(1-\alpha)\%$ para σ_1/σ_2.

Ejemplo 9.19: En el ejemplo 9.12 de la página 290 se construyó un intervalo de confianza para la diferencia en el contenido medio de ortofósforo de dos estaciones ubicadas sobre el río James, medido en miligramos por litro, suponiendo que las varianzas normales de la población son diferentes. Justifique esta suposición construyendo intervalos de confianza del 98% para σ_1^2/σ_2^2 y para σ_1/σ_2, donde σ_1^2 y σ_2^2 son las varianzas de la población del contenido de ortofósforo en la estación 1 y en la estación 2, respectivamente.

***Solución*:** Del ejemplo 9.12 tenemos $n_1 = 15$, $n_2 = 12$, $s_1 = 3.07$ y $s_2 = 0.80$. Para un intervalo de confianza del 98%, $\alpha = 0.02$. Al interpolar en la tabla A.6 encontramos $f_{0.01}(14,11) \approx 4.30$ y $f_{0.01}(11,14) \approx 3.87$. Por lo tanto, el intervalo de confianza del 98% para σ_1^2/σ_2^2 es

$$\left(\frac{3.07^2}{0.80^2}\right)\left(\frac{1}{4.30}\right) < \frac{\sigma_1^2}{\sigma_2^2} < \left(\frac{3.07^2}{0.80^2}\right)(3.87),$$

www.FreeLibros.me

que se simplifica a $3.425 < \frac{\sigma_1^2}{\sigma_2^2} < 56.991$. Al calcular las raíces cuadradas de los límites de confianza encontramos que un intervalo de confianza del 98% para σ_1/σ_2 es

$$1.851 < \frac{\sigma_1}{\sigma_2} < 7.549.$$

Como este intervalo no permite la posibilidad de que σ_1/σ_2 sea igual a 1, es correcto suponer que $\sigma_1 \neq \sigma_2$ o $\sigma_1^2 \neq \sigma_2^2$ en el ejemplo 9.12. ■

Ejercicios

9.71 Un fabricante de baterías para automóvil afirma que sus baterías durarán, en promedio, 3 años con una varianza de 1 año. Suponga que 5 de estas baterías tienen duraciones de 1.9, 2.4, 3.0, 3.5 y 4.2 años y con base en esto construya un intervalo de confianza del 95% para σ^2, después decida si la afirmación del fabricante de que $\sigma^2 = 1$ es válida. Suponga que la población de duraciones de las baterías se distribuye de forma aproximadamente normal.

9.72 Una muestra aleatoria de 20 estudiantes obtuvo una media de $\bar{x} = 72$ y una varianza de $s^2 = 16$ en un examen universitario de colocación en matemáticas. Suponga que las calificaciones se distribuyen normalmente y con base en esto construya un intervalo de confianza del 98% para σ^2.

9.73 Construya un intervalo de confianza del 95% para σ^2 en el ejercicio 9.9 de la página 283.

9.74 Construya un intervalo de confianza del 99% para σ^2 en el ejercicio 9.11 de la página 283.

9.75 Construya un intervalo de confianza del 99% para σ en el ejercicio 9.12 de la página 283.

9.76 Construya un intervalo de confianza del 90% para σ en el ejercicio 9.13 de la página 283.

9.77 Construya un intervalo de confianza del 98% para σ_1/σ_2 en el ejercicio 9.42 de la página 295, donde σ_1 y σ_2 son, respectivamente, las desviaciones estándar para las distancias recorridas por litro de combustible de los camiones compactos Volkswagen y Toyota.

9.78 Construya un intervalo de confianza del 90% para σ_1^2/σ_2^2 en el ejercicio 9.43 de la página 295. ¿Se justifica que supongamos que $\sigma_1^2 \neq \sigma_2^2$ cuando construimos nuestro intervalo de confianza para $\mu_1 - \mu_2$?

9.79 Construya un intervalo de confianza del 90% para σ_1^2/σ_2^2 en el ejercicio 9.46 de la página 295. ¿Deberíamos suponer que $\sigma_1^2 = \sigma_2^2$ cuando construimos nuestro intervalo de confianza para $\mu_I - \mu_{II}$?

9.80 Construya un intervalo de confianza del 95% para σ_A^2/σ_B^2 en el ejercicio 9.49 de la página 295. ¿Tendría que utilizar la suposición de la igualdad de la varianza?

9.14 Estimación de la máxima verosimilitud (opcional)

A menudo los estimadores de parámetros han tenido que recurrir a la intuición. El estimador $\bar{X}$ ciertamente parece razonable como estimador de una media de la población μ. La virtud de S^2 como estimador de σ^2 se destaca en el estudio de estimadores insesgados de la sección 9.3. El estimador para un parámetro binomial p es simplemente una proporción de la muestra que, desde luego, es un *promedio* y recurre al sentido común. Sin embargo, hay muchas situaciones en las que no es del todo evidente cuál debería ser el estimador adecuado. Como resultado, el estudiante de estadística tiene mucho que aprender respecto a las diferentes filosofías que producen distintos métodos de estimación. En esta sección estudiaremos el **método de máxima verosimilitud**.

La estimación por máxima verosimilitud representa uno de los métodos de estimación más importantes en toda la estadística inferencial. No explicaremos el método de manera detallada; más bien, intentaremos transmitir la filosofía de la máxima verosimilitud e ilustrarla con ejemplos que la relacionan con otros problemas de estimación que se examinan en este capítulo.

www.FreeLibros.me

$\bar{x}_B = 30.47$, $s_A = 1.57$ y $s_B = 1.74$. Suponga que los datos de cada batería se distribuyen normalmente y que $\sigma_A = \sigma_B$.

a) Calcule un intervalo de confianza del 95% para $\mu_A - \mu_B$.

b) Del inciso *a*) saque algunas conclusiones que le ayuden a la empresa a decidir si debería utilizar la batería *A* o la *B*.

9.49 Se considera usar dos marcas diferentes de pintura vinílica. Se seleccionaron 15 especímenes de cada tipo de pintura, para los cuales los tiempos de secado en horas fueron los siguientes:

Pintura *A*					**Pintura *B***				
3.5	2.7	3.9	4.2	3.6	4.7	3.9	4.5	5.5	4.0
2.7	3.3	5.2	4.2	2.9	5.3	4.3	6.0	5.2	3.7
4.4	5.2	4.0	4.1	3.4	5.5	6.2	5.1	5.4	4.8

Suponga que el tiempo de secado se distribuye normalmente, con $\sigma_A = \sigma_B$. Calcule un intervalo de confianza del 95% de $\mu_B - \mu_A$, donde μ_A y μ_B son los tiempos medios de secado.

9.50 A dos grupos de ratas diabéticas se les suministran dos niveles de dosis de insulina (alto y bajo) para verificar la capacidad de fijación de esta hormona. Se obtuvieron los siguientes datos.

Dosis baja:	$n_1 = 8$	$\bar{x}_1 = 1.98$	$s_1 = 0.51$
Dosis alta:	$n_2 = 13$	$\bar{x}_2 = 1.30$	$s_2 = 0.35$

Suponga que las varianzas son iguales. Determine un intervalo de confianza del 95% para la diferencia en la capacidad promedio verdadera de fijación de la insulina entre las dos muestras.

9.10 Una sola muestra: estimación de una proporción

El estadístico $\hat{P} = X/n$, en donde X representa el número de éxitos en n ensayos, provee un estimador puntual de la proporción p en un experimento binomial. Por lo tanto, la proporción de la muestra $\hat{p} = x/n$ se utilizará como el estimador puntual del parámetro p.

Si no se espera que la proporción p desconocida esté demasiado cerca de 0 o de 1, se puede establecer un intervalo de confianza para p considerando la distribución muestral de $\hat{P}$. Si en cada ensayo binomial asignamos el valor 0 a un fracaso y el valor 1 a un éxito, el número de éxitos, x, se puede interpretar como la suma de n valores que consta sólo de ceros y unos, y $\hat{p}$ es sólo la media muestral de esos n valores. En consecuencia, por el teorema del límite central, para n suficientemente grande $\hat{P}$ está distribuida de forma casi normal con media

$$\mu_{\hat{P}} = E(\hat{P}) = E\left(\frac{X}{n}\right) = \frac{np}{n} = p$$

y varianza

$$\sigma^2_{\hat{P}} = \sigma^2_{X/n} = \frac{\sigma^2_X}{n^2} = \frac{npq}{n^2} = \frac{pq}{n}.$$

Por lo tanto, podemos afirmar que

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha, \text{ con } Z = \frac{\hat{P} - p}{\sqrt{pq/n}},$$

y $z_{\alpha/2}$ es el valor por arriba del cual encontramos una área de $\alpha/2$ debajo de la curva normal estándar. Al sustituir para Z escribimos

$$P\left(-z_{\alpha/2} < \frac{\hat{P} - p}{\sqrt{pq/n}} < z_{\alpha/2}\right) = 1 - \alpha.$$

Cuando n es grande se introduce un error muy pequeño sustituyendo el estimado puntual $\hat{p} = x/n$ para la p debajo del signo de radical. Entonces podemos escribir

$$P\left(\hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{P} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right) \approx 1 - \alpha.$$

www.FreeLibros.me

Por otro lado, al resolver para p en la desigualdad cuadrática anterior,

$$-z_{\alpha/2} < \frac{\hat{P} - p}{\sqrt{pq/n}} < z_{\alpha/2},$$

obtenemos otra forma del intervalo de confianza para p con los siguientes límites:

$$\frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n}}{1 + \frac{z^2_{\alpha/2}}{n}} \pm \frac{z_{\alpha/2}}{1 + \frac{z^2_{\alpha/2}}{n}} \sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z^2_{\alpha/2}}{4n^2}}.$$

Para una muestra aleatoria de tamaño n se calcula la proporción muestral $\hat{p} = x/n$ y se pueden obtener los siguientes intervalos de confianza aproximados del $100(1-\alpha)\%$ para p.

Intervalos de confianza para p de una muestra grande

Si $\hat{p}$ es la proporción de éxitos en una muestra aleatoria de tamaño n, y $\hat{q} = 1 - \hat{p}$, un intervalo de confianza aproximado del $100(1-\alpha)\%$ para el parámetro binomial p se obtiene por medio de (método 1)

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

o mediante (método 2)

$$\frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n}}{1 + \frac{z^2_{\alpha/2}}{n}} - \frac{z_{\alpha/2}}{1 + \frac{z^2_{\alpha/2}}{n}} \sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z^2_{\alpha/2}}{4n^2}} < p < \frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n}}{1 + \frac{z^2_{\alpha/2}}{n}} + \frac{z_{\alpha/2}}{1 + \frac{z^2_{\alpha/2}}{n}} \sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z^2_{\alpha/2}}{4n^2}},$$

donde $z_{\alpha/2}$ es el valor z que deja una área de $\alpha/2$ a la derecha.

Cuando n es pequeña y se cree que la proporción desconocida p se acerca a 0 o a 1, el procedimiento del intervalo de confianza que se establece aquí no es confiable y, por lo tanto, no se debería emplear. Para estar seguros se requiere que tanto $n\hat{p}$ como $n\hat{q}$ sean mayores que o iguales a 5. Los métodos para calcular un intervalo de confianza para el parámetro binomial p también se pueden aplicar cuando se está utilizando la distribución binomial con el fin de aproximar la distribución hipergeométrica; es decir, cuando n es pequeña respecto a N, como se ilustra en el ejemplo 9.14.

Observe que, aunque el método 2 produce resultados más precisos, su cálculo es más complicado, y la ventaja en precisión que brinda disminuye cuando el tamaño de la muestra es lo suficientemente grande. Debido a esto en la práctica es más común utilizar el método 1.

Ejemplo 9.14: En una muestra aleatoria de $n = 500$ familias que tienen televisores en la ciudad de Hamilton, Canadá, se encuentra que $x = 340$ están suscritas a HBO. Calcule un intervalo de confianza del 95% para la proporción real de familias que tienen televisores en esta ciudad y están suscritas a HBO.

***Solución*:** La estimación puntual de p es $\hat{p} = 340/500 = 0.68$. Si usamos la tabla A.3, encontramos que $z_{0.025} = 1.96$. Por lo tanto, si utilizamos el método 1, el intervalo de confianza del 95% para p es

$$0.68 - 1.96\sqrt{\frac{(0.68)(0.32)}{500}} < p < 0.68 + 1.96\sqrt{\frac{(0.68)(0.32)}{500}},$$

que se simplifica a $0.6391 < p < 0.7209$.

www.FreeLibros.me

Si utilizamos el segundo método, obtenemos

$$\frac{0.68 + \frac{1.96^2}{(2)(500)}}{1 + \frac{1.96^2}{500}} \pm \frac{1.96}{1 + \frac{1.96^2}{500}} \sqrt{\frac{(0.68)(0.32)}{500} + \frac{1.96^2}{(4)(500^2)}} = 0.6786 \pm 0.0408,$$

que se simplifica a $0.6378 < p < 0.7194$. Aparentemente, cuando n es grande (500 en este caso) ambos métodos producen resultados muy similares.

Si p es el valor central de un intervalo de confianza del $100(1 - \alpha)\%$, entonces $\hat{p}$ estima p sin error. Sin embargo, la mayoría de las veces $\hat{p}$ no será exactamente igual a p y el estimado puntual será erróneo. El tamaño de este error será la diferencia positiva que separa a p de $\hat{p}$, y podemos tener una confianza del $100(1 - \alpha)\%$ de que tal diferencia no excederá a $z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}$. Si dibujamos un diagrama de un intervalo de confianza típico, como el de la figura 9.6, podemos ver esto fácilmente. En este caso utilizamos el método 1 para estimar el error.

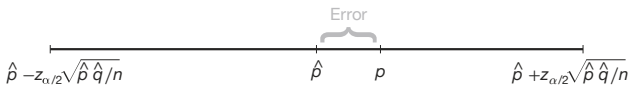

Figura 9.6: Error en la estimación de p por medio de $\hat{p}$.

Teorema 9.3: Si $\hat{p}$ se utiliza como un estimado de p, podemos tener un $100(1 - \alpha)\%$ de confianza en que el error no excederá a $z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}$.

En el ejemplo 9.14 tenemos un 95% de confianza en que la proporción de la muestra $\hat{p} = 0.68$ difiere de la verdadera proporción p en una cantidad que no excede a 0.04.

Selección del tamaño de la muestra

Determinemos ahora qué tan grande debe ser una muestra para poder estar seguros de que el error al estimar p será menor que una cantidad específica e. Por medio del teorema 9.3, debemos elegir una n tal que $z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n} = e$.

Teorema 9.4: Si $\hat{p}$ se utiliza como un estimado de p, podemos tener un $100(1 - \alpha)\%$ de confianza en que el error será menor que una cantidad específica e cuando el tamaño de la muestra sea aproximadamente

$$n = \frac{z_{\alpha/2}^2 \hat{p}\hat{q}}{e^2}.$$

El teorema 9.4 es algo engañoso, pues debemos utilizar $\hat{p}$ para determinar el tamaño n de la muestra, pero $\hat{p}$ se calcula a partir de la muestra. Si se puede hacer una estimación burda de p sin tomar una muestra, se podría usar este valor para determinar n. A falta de tal estimado, podríamos tomar una muestra preliminar de tamaño $n \geq 30$ para proporcionar un estimado de p. Si utilizamos el teorema 9.4 podríamos determinar aproximadamente cuántas observaciones se necesitan para proporcionar el grado de precisión deseado. Observe que los valores fraccionarios de n se redondean al siguiente número entero mayor.

www.FreeLibros.me

Ejemplo 9.15: ¿Qué tan grande debe ser una muestra en el ejemplo 9.14 si queremos tener un 95% de confianza en que la estimación de p esté dentro de 0.02 del valor verdadero?

***Solución*:** Tratemos a las 500 familias como una muestra preliminar que proporciona una estimación $\hat{p} = 0.68$. Entonces, mediante el teorema 9.4,

$$n = \frac{(1.96)^2(0.68)(0.32)}{(0.02)^2} = 2089.8 \approx 2090.$$

Por lo tanto, si basamos nuestra estimación de p en una muestra aleatoria de tamaño 2090, podemos tener un 95% de confianza en que nuestra proporción muestral no diferirá de la proporción verdadera en más de 0.02. ▟

Ocasionalmente será poco práctico obtener una estimación de p que se utilice para determinar el tamaño muestral para un grado específico de confianza. Si esto sucede, se establece un límite superior para n al notar que $\hat{p}\hat{q} = \hat{p}(1 - \hat{p})$, que debe ser a lo sumo $1/4$, ya que $\hat{p}$ debe caer entre 0 y 1. Este hecho se verifica completando el cuadrado. Por consiguiente,

$$\hat{p}(1 - \hat{p}) = -(\hat{p}^2 - \hat{p}) = \frac{1}{4} - \left(\hat{p}^2 - \hat{p} + \frac{1}{4}\right) = \frac{1}{4} - \left(\hat{p} - \frac{1}{2}\right)^2,$$

que siempre es menor que $1/4$ excepto cuando $\hat{p} = 1/2$ y entonces $\hat{p}\hat{q} = 1/4$. Por lo tanto, si sustituimos $\hat{p} = 1/2$ en la fórmula para n del teorema 9.4, cuando, de hecho, p difiere de $1/2$, entonces n se agrandará más de lo necesario para el grado de confianza específico y, como resultado, se incrementará nuestro grado de confianza.

Teorema 9.5: Si utilizamos $\hat{p}$ como un estimado de p, podemos tener, **al menos**, un $100(1 - \alpha)\%$ de confianza en que el error no excederá a una cantidad específica e cuando el tamaño de la muestra sea

$$n = \frac{z_{\alpha/2}^2}{4e^2}.$$

Ejemplo 9.16: ¿Qué tan grande debe ser una muestra en el ejemplo 9.14 si queremos tener al menos un 95% de confianza en que nuestra estimación de p está dentro de 0.02 del valor verdadero?

***Solución*:** A diferencia del ejemplo 9.15, supondremos ahora que no se tomó una muestra preliminar para obtener una estimación de p. En consecuencia, podemos tener al menos un 95% de confianza en que nuestra proporción de la muestra no diferirá de la proporción verdadera en más de 0.02, si elegimos una muestra de tamaño

$$n = \frac{(1.96)^2}{(4)(0.02)^2} = 2401.$$

Si comparamos los resultados de los ejemplos 9.15 y 9.16, vemos que la información concerniente a p, proporcionada por una muestra preliminar, o quizás obtenida a partir de la experiencia, nos permite elegir una muestra más pequeña a la vez que mantenemos el grado de precisión requerido. ▟

www.FreeLibros.me

Capítulo 10

Pruebas de hipótesis de una y dos muestras

10.1 Hipótesis estadísticas: conceptos generales

Como se expuso en el capítulo 9, a menudo el problema al que se enfrentan el científico o el ingeniero no es tanto la estimación de un parámetro de la población, sino la formación de un procedimiento de decisión que se base en los datos y que pueda producir una conclusión acerca de algún sistema científico. Por ejemplo, un investigador médico puede decidir con base en evidencia experimental si beber café incrementa el riesgo de cáncer en los seres humanos; un ingeniero quizá tenga que decidir con base en datos muestrales si hay una diferencia entre la precisión de un tipo de medidor y la de otro; o tal vez un sociólogo desee reunir los datos apropiados que le permitan decidir si el tipo de sangre y el color de ojos de un individuo son variables independientes. En cada uno de estos casos el científico o el ingeniero *postulan* o *conjeturan* algo acerca de un sistema. Además, cada uno debe utilizar datos experimentales y tomar decisiones basadas en ellos. En cada caso la conjetura se puede expresar en forma de hipótesis estadística. Los procedimientos que conducen a la aceptación o al rechazo de hipótesis estadísticas como éstas comprenden una área importante de la inferencia estadística. Empecemos por definir con precisión lo que entendemos por **hipótesis estadística**.

Definición 10.1: Una **hipótesis estadística** es una aseveración o conjetura respecto a una o más poblaciones.

La verdad o falsedad de una hipótesis estadística nunca se sabe con absoluta certeza, a menos que se examine toda la población, lo cual, por supuesto, sería poco práctico en la mayoría de las situaciones. En vez de eso se toma una muestra aleatoria de la población de interés y se utilizan los datos contenidos en ella para proporcionar evidencia que respalde o no la hipótesis. La evidencia de la muestra que es inconsistente con la hipótesis planteada conduce al rechazo de la misma.

www.FreeLibros.me

El papel que desempeña la probabilidad en la prueba de hipótesis

Debería quedar claro al lector que un procedimiento de toma de decisiones debe implicar la conciencia de la *probabilidad de llegar a una conclusión errónea*. Por ejemplo, suponga que la hipótesis que postuló el ingeniero es que la fracción p de artículos defectuosos en cierto proceso es 0.10. El experimento consiste en observar una muestra aleatoria del producto en cuestión. Suponga que se prueban 100 artículos y que se encuentran 12 defectuosos. Es razonable concluir que esta evidencia no rechaza la condición de que el parámetro binomial $p = 0.10$, por lo que puede provocar que no se rechace la hipótesis. Sin embargo, también puede provocar que no se refute $p = 0.12$, o quizá incluso $p = 0.15$. Como resultado, el lector se debe acostumbrar a la idea de que **el rechazo de una hipótesis implica que fue refutada por la evidencia de la muestra**. En otras palabras, **el rechazo significa que existe una pequeña probabilidad de obtener la información muestral observada cuando, de hecho, la hipótesis es verdadera**. Por ejemplo, en la hipótesis de la proporción de artículos defectuosos, una muestra de 100 artículos que revela que hay 20 defectuosos es ciertamente evidencia para el rechazo. ¿Por qué? Si en realidad $p = 0.10$, la probabilidad de obtener 20 o más artículos defectuosos es aproximadamente de 0.002. Con el pequeño riesgo resultante de llegar a una conclusión errónea parecería seguro **rechazar la hipótesis** de que $p = 0.10$. En otras palabras, el rechazo de una hipótesis tiende a casi "descartar" la hipótesis. Por otro lado, es muy importante enfatizar que la aceptación o, más bien, la falta de rechazo no descarta otras posibilidades. Como resultado, *el analista de datos establece una conclusión firme cuando se rechaza una hipótesis*.

En el planteamiento formal de una hipótesis a menudo influye la estructura de la probabilidad de una conclusión errónea. Si el científico está interesado en *apoyar firmemente* un argumento, espera llegar a éste en la forma del rechazo de una hipótesis. Si el investigador médico desea mostrar evidencia sólida a favor del argumento de que beber café aumenta el riesgo de contraer cáncer, la hipótesis a probar debería tener la forma "el riesgo de desarrollar cáncer no aumenta como consecuencia de beber café". Como resultado, el argumento se obtiene mediante un rechazo. De manera similar, para apoyar la afirmación de que un tipo de medidores es más preciso que otro, el ingeniero prueba la hipótesis de que no hay diferencia en la precisión de los dos tipos de medidores.

Lo anterior implica que cuando el analista de datos formaliza la evidencia experimental con base en la prueba de hipótesis, es muy importante el **planteamiento formal de la hipótesis**.

La hipótesis nula y la hipótesis alternativa

La estructura de la prueba de hipótesis se establece usando el término **hipótesis nula**, el cual se refiere a cualquier hipótesis que se desea probar y se denota con H_0. El rechazo de H_0 conduce a la aceptación de una **hipótesis alternativa**, que se denota con H_1. La comprensión de las diferentes funciones que desempeñan la hipótesis nula (H_0) y la hipótesis alternativa (H_1) es fundamental para entender los principios de la prueba de hipótesis. La hipótesis alternativa H_1 por lo general representa la *pregunta que se responderá o la teoría que se probará*, por lo que su especificación es muy importante. La hipótesis nula H_0 *anula o se opone a* H_1 y a menudo es el complemento lógico de H_1. A medida que el lector aprenda más sobre la prueba de hipótesis notará que el analista llega a una de las siguientes dos conclusiones:

www.FreeLibros.me

rechazar H_0 a favor de H_1 debido a evidencia suficiente en los datos o
no rechazar H_0 debido a evidencia insuficiente en los datos.

Observe que las *conclusiones no implican una "aceptación de H_0" formal y literal.* La aseveración de H_0 a menudo representa el "status quo" contrario a una nueva idea, conjetura, etcétera, enunciada en H_1; en tanto que no rechazar H_0 representa la conclusión adecuada. En nuestro ejemplo binomial la cuestión práctica podría ser el interés en que la probabilidad histórica de artículos defectuosos de 0.10 ya no sea verdadera. De hecho, la conjetura podría ser que p excede a 0.10. Entonces podríamos afirmar que

$$H_0: p = 0.10,$$
$$H_1: p > 0.10.$$

Ahora, 12 artículos defectuosos de cada 100 no refutan $p = 0.10$, por lo que la conclusión es "no rechazar H_0". Sin embargo, si los datos revelan 20 artículos defectuosos de cada 100, la conclusión sería "rechazar H_0" a favor de H_1: $p > 0.10$.

Aunque las aplicaciones de la prueba de hipótesis son muy abundantes en trabajos científicos y de ingeniería, quizás el mejor ejemplo para un principiante sea el dilema que enfrenta el jurado en un juicio. Las hipótesis nula y alternativa son

H_0: el acusado es inocente,
H_1: el acusado es culpable.

La acusación proviene de una sospecha de culpabilidad. La hipótesis H_0 (el status quo) se establece en oposición a H_1 y se mantiene a menos que se respalde H_1 con evidencia "más allá de una duda razonable". Sin embargo, en este caso "no rechazar H_0" no implica inocencia, sino sólo que la evidencia fue insuficiente para lograr una condena. Por lo tanto, el jurado no necesariamente *acepta* H_0 sino que *no rechaza* H_0.

10.2 Prueba de una hipótesis estadística

Para ilustrar los conceptos que se utilizan al probar una hipótesis estadística acerca de una población considere el siguiente ejemplo. Se sabe que, después de un periodo de dos años, cierto tipo de vacuna contra un virus que produce resfriado ya sólo es 25% eficaz. Suponga que se eligen 20 personas al azar y se les aplica una vacuna nueva, un poco más costosa, para determinar si protege contra el mismo virus durante un periodo más largo. (En un estudio real de este tipo el número de participantes que reciben la nueva vacuna podría ascender a varios miles. Aquí la muestra es de 20 sólo porque lo único que se busca es demostrar los pasos básicos para realizar una prueba estadística). Si más de 8 individuos de los que reciben la nueva vacuna superan el lapso de 2 años sin contraer el virus, la nueva vacuna se considerará superior a la que se usa en la actualidad. El requisito de que el número exceda a 8 es algo arbitrario, aunque parece razonable, ya que representa una mejoría modesta sobre las 5 personas que se esperaría recibieran protección si fueran inoculadas con la vacuna que actualmente está en uso. En esencia probamos la hipótesis nula de que la nueva vacuna es igual de eficaz después de un periodo de 2 años que la que se utiliza en la actualidad. La hipótesis alternativa es que la nueva vacuna es

www.FreeLibros.me

mejor, y esto equivale a poner a prueba la hipótesis de que el parámetro binomial para la probabilidad de un éxito en un ensayo dado es $p = ¼$, contra la alternativa de que $p > ¼$. Esto por lo general se escribe como se indica a continuación:

$$H_0: p = 0.25,$$
$$H_1: p > 0.25.$$

El estadístico de prueba

El **estadístico de prueba** en el cual se basa nuestra decisión es X, el número de individuos en nuestro grupo de prueba que reciben protección de la nueva vacuna durante un periodo de al menos 2 años. Los valores posibles de X, de 0 a 20, se dividen en dos grupos: los números menores o iguales que 8 y aquellos mayores que 8. Todos los posibles valores mayores que 8 constituyen la **región crítica**. El último número que observamos al pasar a la región crítica se llama **valor crítico**. En nuestro ejemplo el valor crítico es el número 8. Por lo tanto, si $x > 8$, rechazamos H_0 a favor de la hipótesis alternativa H_1. Si $x \leq 8$, no rechazamos H_0. Este criterio de decisión se ilustra en la figura 10.1.

Figura 10.1: Criterio de decisión para probar $p = 0.25$ contra $p > 0.25$.

La probabilidad de un error tipo 1

El procedimiento de toma de decisiones recién descrito podría conducir a cualquiera de dos conclusiones erróneas. Por ejemplo, es probable que la nueva vacuna no sea mejor que la que se usa en la actualidad (H_0 verdadera) y, sin embargo, en este grupo específico de individuos seleccionados aleatoriamente más de 8 pasan el periodo de 2 años sin contraer el virus. Si rechazáramos H_0 a favor de H_1 cuando, de hecho, H_0 es verdadera, cometeríamos un error que se conoce como **error tipo I**.

Definición 10.2: El rechazo de la hipótesis nula cuando es verdadera se denomina **error tipo I**.

Si 8 o menos miembros del grupo superan exitosamente el periodo de 2 años y no concluimos que la nueva vacuna es mejor cuando en realidad sí lo es (H_1 verdadera), cometemos un segundo tipo de error, el de no rechazar la hipótesis H_0 cuando en realidad es falsa. A este error se le conoce como **error tipo II**.

Definición 10.3: No rechazar la hipótesis nula cuando es falsa se denomina **error tipo II**.

Al probar cualquier hipótesis estadística, hay cuatro situaciones posibles que determinan si nuestra decisión es correcta o errónea. Estas cuatro situaciones se resumen en

www.FreeLibros.me

la tabla 10.1.

Tabla 10.1: Situaciones posibles al probar una hipótesis estadística.

	H_0 es verdadera	H_0 es falsa
No rechazar H_0	Decisión correcta	Error tipo II
Rechazar H_0	Error tipo I	Decisión correcta

La probabilidad de cometer un error tipo I, también llamada **nivel de significancia**, se denota con la letra griega α. En nuestro ejemplo un error tipo I ocurriría si más de 8 individuos inoculados con la nueva vacuna superan el periodo de 2 años sin contraer el virus y los investigadores concluyen que la nueva vacuna es mejor, cuando en realidad es igual a la vacuna que se utiliza en la actualidad. Por lo tanto, si X es el número de individuos que permanecen sin contraer el virus por al menos dos años,

$$\alpha = P(\text{error tipo I}) = P\left(X > 8 \text{ cuando } p = \frac{1}{4}\right) = \sum_{x=9}^{20} b\left(x; 20, \frac{1}{4}\right)$$

$$= 1 - \sum_{x=0}^{8} b\left(x; 20, \frac{1}{4}\right) = 1 - 0.9591 = 0.0409.$$

Decimos que la hipótesis nula, $p = 1/4$, se prueba al nivel de significancia $\alpha = 0.0409$. En ocasiones el nivel de significancia se conoce como **tamaño de la prueba**. Una región crítica de tamaño 0.0409 es muy pequeña y, por lo tanto, es poco probable que se cometa un error de tipo I. En consecuencia, sería poco probable que más de 8 individuos permanecieran inmunes a un virus durante 2 años utilizando una vacuna nueva que en esencia es equivalente a la que actualmente está en el mercado.

La probabilidad de un error tipo II

La probabilidad de cometer un error tipo II, que se denota con β, es imposible de calcular a menos que tengamos una hipótesis alternativa específica. Si probamos la hipótesis nula $p = 1/4$ contra la hipótesis alternativa $p = 1/2$, entonces podremos calcular la probabilidad de no rechazar H_0 cuando es falsa. Simplemente calculamos la probabilidad de obtener 8 o menos en el grupo que supera el periodo de 2 años cuando $p = 1/2$. En este caso,

$$\beta = P\ (\text{error tipo II}) = P\ \left(X \leq 8 \text{ cuando } p = \frac{1}{2}\right)$$

$$= \sum_{x=0}^{8} b\left(x; 20, \frac{1}{2}\right) = 0.2517.$$

Se trata de una probabilidad elevada que indica un procedimiento de prueba en el cual es muy probable que se rechace la nueva vacuna cuando, de hecho, es mejor a la que está actualmente en uso. De manera ideal, es preferible utilizar un procedimiento de prueba con el cual haya pocas probabilidades de cometer el error tipo I y el error tipo II.

Es posible que el director del programa de prueba esté dispuesto a cometer un error tipo II si la vacuna más costosa no es significativamente mejor. De hecho, la única

www.FreeLibros.me

ocasión en la que desea evitar un error tipo II es cuando el verdadero valor de p es de al menos 0.7. Si $p = 0.7$, este procedimiento de prueba da

$$\beta = P(\text{error tipo II}) = P(X \leq 8 \text{ cuando } p = 0.7)$$
$$= \sum_{x=0}^{8} b(x; 20, 0.7) = 0.0051.$$

Con una probabilidad tan pequeña de cometer un error tipo II es muy improbable que se rechace la nueva vacuna cuando tiene una efectividad de 70% después de un periodo de 2 años. A medida que la hipótesis alternativa se aproxima a la unidad, el valor de β tiende a disminuir hasta cero.

El papel que desempeñan α, β y el tamaño de la muestra

Supongamos que el director del programa de prueba no está dispuesto a cometer un error tipo II cuando la hipótesis alternativa $p = 1/2$ es verdadera, aun cuando se encuentre que la probabilidad de tal error es $\beta = 0.2517$. Siempre es posible reducir β aumentando el tamaño de la región crítica. Por ejemplo, considere lo que les sucede a los valores de α y β cuando cambiamos nuestro valor crítico a 7, de manera que todos los valores mayores que 7 caigan en la región crítica y aquellos menores o iguales que 7 caigan en la región de no rechazo. Así, al probar $p = 1/4$ contra la hipótesis alternativa $p = 1/2$, encontramos que

$$\alpha = \sum_{x=8}^{20} b\left(x; 20, \frac{1}{4}\right) = 1 - \sum_{x=0}^{7} b\left(x; 20, \frac{1}{4}\right) = 1 - 0.8982 = 0.1018$$

$$\beta = \sum_{x=0}^{7} b\left(x; 20, \frac{1}{2}\right) = 0.1316.$$

Al adoptar un nuevo procedimiento de toma de decisiones, reducimos la probabilidad de cometer un error tipo II a costa de aumentar la probabilidad de cometer un error tipo I. Para un tamaño muestral fijo, una disminución en la probabilidad de un error por lo general tendrá como resultado un incremento en la probabilidad del otro error. Por fortuna, **la probabilidad de cometer ambos tipos de errores se puede reducir aumentando el tamaño de la muestra**. Considere el mismo problema usando una muestra aleatoria de 100 individuos. Si más de 36 miembros del grupo superan el periodo de 2 años, rechazamos la hipótesis nula de $p = 1/4$ y aceptamos la hipótesis alternativa de $p > 1/4$. El valor crítico ahora es 36. Todos los valores posibles mayores de 36 constituyen la región crítica y todos los valores posibles menores o iguales que 36 caen en la región de aceptación.

Para determinar la probabilidad de cometer un error tipo I debemos utilizar la aproximación a la curva normal con

$$\mu = np = (100)\left(\frac{1}{4}\right) = 2 \quad \text{y} \quad \sigma = \sqrt{npq} = \sqrt{(100)(1/4)(3/4)} = 4.33.$$

Con respecto a la figura 10.2, necesitamos el área bajo la curva normal a la derecha de $x = 36.5$. El valor z correspondiente es

$$z = \frac{36.5 - 25}{4.33} = 2.66.$$

www.FreeLibros.me

Capítulo 11

Regresión lineal simple y correlación

11.1 Introducción a la regresión lineal

En la práctica a menudo se requiere resolver problemas que implican conjuntos de variables de las cuales se sabe que tienen alguna relación inherente entre sí. Por ejemplo, en una situación industrial quizá se sepa que el contenido de alquitrán en el flujo de salida de un proceso químico está relacionado con la temperatura en la entrada. Podría ser de interés desarrollar un método de pronóstico, es decir, un procedimiento que permita estimar el contenido de alquitrán para varios niveles de temperatura de entrada a partir de información experimental. Desde luego, es muy probable que para muchos ejemplos concretos en los que la temperatura de entrada sea la misma, por ejemplo 130°C, el contenido de alquitrán de salida no sea el mismo. Esto es muy similar a lo que ocurre cuando se estudian varios automóviles con un motor del mismo volumen; no todos tienen el mismo rendimiento de combustible. No todas las casas ubicadas en la misma zona del país, con la misma superficie de construcción, se venden al mismo precio. El contenido de alquitrán, el rendimiento del combustible (en millas por galón) y el precio de las casas (en miles de dólares) son **variables dependientes** naturales o respuestas en los tres escenarios. La temperatura en la entrada, el volumen del motor (pies cúbicos) y los metros cuadrados de superficie de construcción son, respectivamente, **variables independientes** naturales o **regresores**. Una forma razonable de relación entre la **respuesta** Y y el regresor x es la relación lineal,

$$Y = \beta_0 + \beta_1 x,$$

en la que, por supuesto, β_0 es la **intersección** y β_1 es la **pendiente**. Esta relación se ilustra en la figura 11.1.

Si la relación es exacta y no contiene ningún componente aleatorio o probabilístico, entonces se trata de una relación **determinista** entre dos variables científicas. Sin embargo, en los ejemplos que se mencionaron, así como en muchos otros fenómenos científicos y de ingeniería, la relación no es determinista, es decir, una x dada no siempre produce el mismo valor de Y. Como resultado, los problemas importantes en este caso son de naturaleza probabilística, toda vez que la relación anterior no puede considerarse exacta. El concepto de **análisis de regresión** se refiere a encontrar la mejor relación entre Y y x

www.FreeLibros.me

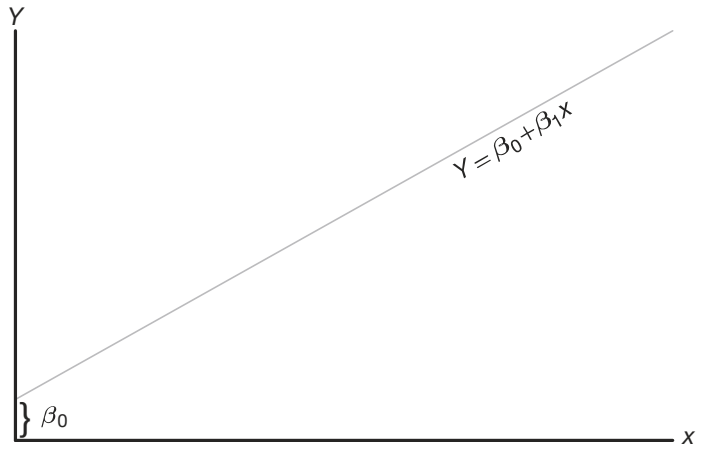

Figura 11.1: Una relación lineal; β_0: intersección; β_1: pendiente.

cuantificando la fuerza de esa relación, y empleando métodos que permitan predecir los valores de la respuesta dados los valores del regresor x.

En muchas aplicaciones habrá más de un regresor, es decir, más de una variable independiente **que ayude a explicar a *Y***. Por ejemplo, si se tratara de explicar las razones para el precio de una casa, se esperaría que una de ellas fuera su antigüedad, en cuyo caso la estructura múltiple de la regresión se podría escribir como

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

donde Y es el precio, x_1 son los metros cuadrados y x_2 es la antigüedad de la casa en años. En el capítulo siguiente se estudiarán problemas con regresores múltiples. El análisis resultante se denomina **regresión múltiple**; en tanto que el análisis del caso con un solo regresor recibe el nombre de **regresión simple**. En un segundo ejemplo de la regresión múltiple, un ingeniero químico podría estar interesado en la cantidad de hidrógeno que se ha perdido en las muestras de un metal específico que se tiene almacenado. En este caso habría dos entradas, x_1, el tiempo de almacenamiento en horas, y x_2, la temperatura de almacenamiento en grados centígrados. De modo que la respuesta sería Y, la pérdida de hidrógeno en partes por millón.

En este capítulo estudiaremos el tema de la **regresión lineal simple**, que trata el caso de una sola variable regresora, en el que la relación entre x y y es lineal. Para el caso en el que hay más de una variable regresora el lector debe consultar el capítulo 12. Denotemos una muestra aleatoria de tamaño n mediante el conjunto $\{(x_i, y_i); i = 1, 2,..., n\}$. Si se tomaran muestras adicionales utilizando exactamente los mismos valores de x, se esperaría que los valores de y variaran. Así, el valor y_i en el par ordenado (x_i, y_i) es el valor de cierta variable aleatoria Y_i.

11.2 El modelo de regresión lineal simple (RLS)

Hemos limitado el uso del término *análisis de regresión* a los casos en los que las relaciones entre las variables no son deterministas, es decir, no son exactas. En otras palabras, debe existir un **componente aleatorio** en la ecuación que relaciona las variables. Este componente aleatorio toma en cuenta consideraciones que no son medibles o, de

www.FreeLibros.me

hecho, que los científicos o los ingenieros no comprenden. En realidad, en la mayoría de aplicaciones de la regresión, la ecuación lineal, digamos, $Y = \beta_0 + \beta_1 x$ es una aproximación que representa de manera simplificada algo desconocido y mucho más complicado. Por ejemplo, en el caso que implica la respuesta $Y =$ contenido de alquitrán y $x =$ temperatura de entrada es probable que $Y = \beta_0 + \beta_1 x$ sea una aproximación razonable que podría funcionar dentro de un rango limitado de x. La mayoría de las veces los modelos que son simplificaciones de estructuras más complicadas y desconocidas son de naturaleza lineal, es decir, lineales en los **parámetros** β_0 y β_1 o, en el caso del modelo que implica el precio, el tamaño y la antigüedad de la casa, lineal en los **parámetros** β_0, β_1 y β_2. Estas estructuras lineales son sencillas y de naturaleza empírica, por lo que se denominan **modelos empíricos**.

Un análisis de la relación entre x y Y requiere el planteamiento de un **modelo estadístico**. Con frecuencia un estadístico utiliza un modelo como representación de un **ideal** que, en esencia, define cómo percibimos que el sistema en cuestión generó los datos. El modelo debe incluir al conjunto $\{(x_i, y_i); i = 1, 2,..., n\}$ de datos que implica n pares de valores (x, y). No debemos olvidar que el valor de y_i depende de x_i por medio de una estructura lineal que también incluye el componente aleatorio. La base para el uso de un modelo estadístico se relaciona con la manera en que la variable aleatoria Y cambia con x y el componente aleatorio. El modelo también incluye lo que se asume acerca de las propiedades estadísticas del componente aleatorio. A continuación se presenta el modelo estadístico para la regresión lineal simple. La respuesta Y se relaciona con la variable independiente x a través de la ecuación

Modelo de regresión lineal simple

$$Y = \beta_0 + \beta_1 x + \epsilon.$$

en la cual β_0 y β_1 son los parámetros desconocidos de la intersección y la pendiente, respectivamente, y ϵ es una variable aleatoria que se supone está distribuida con $E(\epsilon) = 0$ y $\text{Var}(\epsilon) = \sigma^2$. Es frecuente que a la cantidad σ^2 se le denomine varianza del error o varianza residual.

En el modelo anterior hay varias cuestiones evidentes. La cantidad Y es una variable aleatoria, ya que ϵ es aleatoria. El valor x de la variable regresora no es aleatorio y, de hecho, se mide con un error despreciable. La cantidad ϵ, que a menudo recibe el nombre de **error aleatorio** o **alteración aleatoria**, tiene varianza constante. Es común que a esta parte se le denomine **suposición de varianza homogénea**. La presencia de este error aleatorio ϵ evita que el modelo se convierta tan sólo en una ecuación determinista. Ahora, el hecho de que $E(\epsilon) = 0$ implica que para una x específica, los valores de y se distribuyen alrededor de la **recta verdadera** o **recta de regresión** de la población $y = \beta_0 + \beta_1 x$. Si se elige bien el modelo, es decir, si no hay otros regresores de importancia y la aproximación lineal es buena dentro de los rangos de los datos, entonces son razonables los errores positivos y negativos que rodean a la regresión verdadera. Debe recordarse que en la práctica β_0 y β_1 se desconocen y que deben estimarse a partir de los datos. Además, el modelo que se acaba de describir es de naturaleza conceptual. Como resultado, en la práctica nunca se observan los valores ϵ reales, por lo que nunca se puede trazar la verdadera recta de regresión, aunque suponemos que ahí está. Sólo es posible dibujar una recta estimada. En la figura 11.2 se ilustra la naturaleza de los datos (x, y) hipotéticos dispersos alrededor de la verdadera recta de regresión para un caso en que sólo se dispone de $n = 5$ observaciones. Debemos destacar que lo que observamos en la figura 11.2 no es la recta que utilizan el científico o ingeniero. En vez de esa recta, ¡lo

www.FreeLibros.me

que describe la ilustración es el significado de las suposiciones! Ahora describiremos la regresión que el usuario tiene a su disposición.

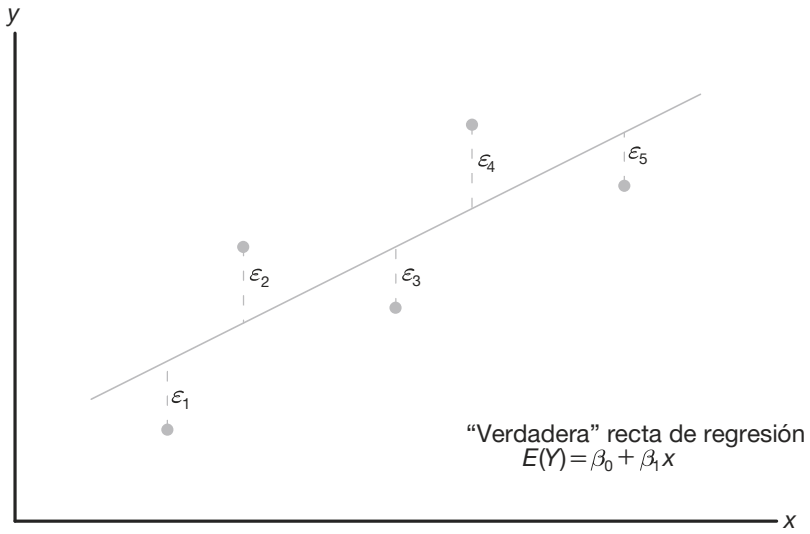

Figura 11.2: Datos (x, y) hipotéticos dispersos alrededor de la verdadera recta de regresión para $n = 5$.

La recta de regresión ajustada

Un aspecto importante del análisis de regresión es, en términos sencillos, estimar los parámetros β_0 y β_1, es decir, estimar los llamados **coeficientes de regresión**. En la sección siguiente se estudiará el método para estimarlos. Suponga que denotamos los estimados b_0 para β_0 y b_1 para β_1. Entonces, la recta de **regresión ajustada**, o estimada, es dada por

$$\hat{y} = b_0 + b_1 x,$$

donde $\hat{y}$ es el valor pronosticado o ajustado. Es evidente que la recta ajustada es un estimado de la verdadera recta de regresión. Se espera que la recta ajustada esté más cerca de la verdadera línea de regresión cuando se dispone de una gran cantidad de datos. En el ejemplo siguiente se ilustra la recta ajustada para un estudio sobre contaminación en la vida real.

Uno de los problemas más desafiantes que enfrenta el campo del control de la contaminación del agua lo representa la industria de la peletería, ya que sus desechos son químicamente complejos; se caracterizan por valores elevados de la demanda de oxígeno químico, sólidos volátiles y otras medidas de contaminación. Considere los datos experimentales de la tabla 11.1, que se obtuvieron de 33 muestras de desechos tratados químicamente en un estudio realizado en Virginia Tech. Se registraron los valores de x, la reducción porcentual de los sólidos totales, y de y, el porcentaje de disminución de la demanda de oxígeno químico.

Los datos de la tabla 11.1 aparecen graficados en un **diagrama de dispersión** en la figura 11.3. Al inspeccionar dicho diagrama se observa que los puntos se acercan mucho a una línea recta, lo cual indica que la suposición de linealidad entre las dos variables parece ser razonable.

www.FreeLibros.me

Tabla 11.1: Medidas de la reducción de los sólidos y de la demanda de oxígeno químico

Reducción de sólidos, *x* (%)	**Reducción de la demanda de oxígeno, *y* (%)**	**Reducción de sólidos, *x* (%)**	**Reducción de la demanda de oxígeno, *y* (%)**
3	5	36	34
7	11	37	36
11	21	38	38
15	16	39	37
18	16	39	36
27	28	39	45
29	27	40	39
30	25	41	41
30	35	42	40
31	30	42	44
31	40	43	37
32	32	44	44
33	34	45	46
33	32	46	46
34	34	47	49
36	37	50	51
36	38		

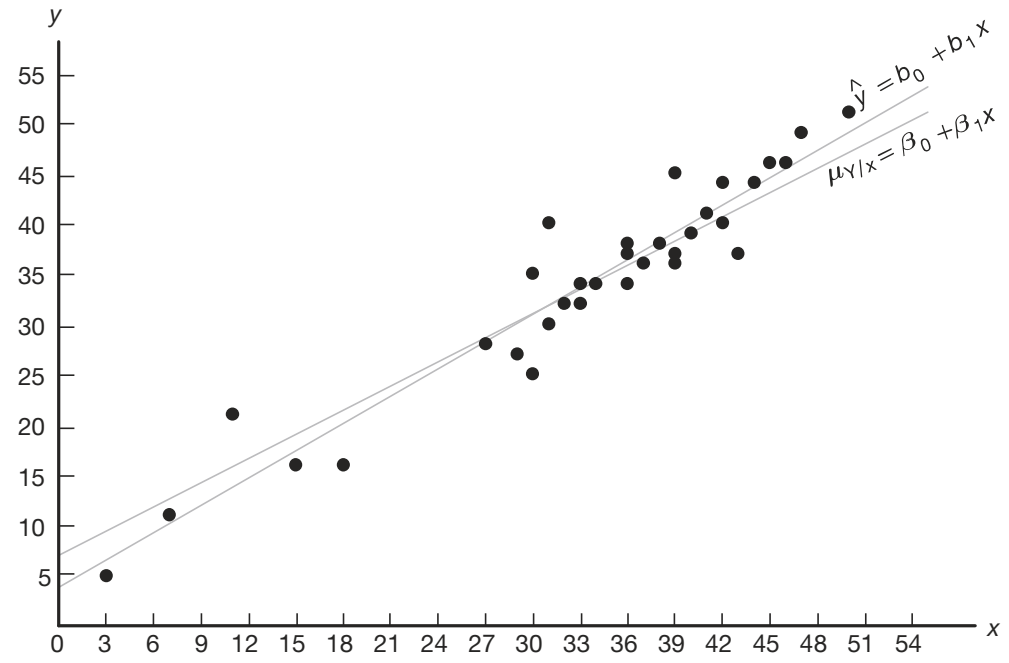

Figura 11.3: Diagrama de dispersión con rectas de regresión.

En el diagrama de dispersión de la figura 11.3 se ilustra la recta de regresión ajustada y *una recta hipotética de regresión verdadera*. Más adelante, en la sección 11.3, en la cual estudiaremos el método de estimación, revisaremos este ejemplo.

www.FreeLibros.me