3

Estimación de parámetros

De manera general, un proceso de inferencia estadística hace referencia a la extracción de una muestra aleatoria de una población, la cual tiene una distribución de probabilidad que contiene uno o varios parámetros desconocidos, sobre los cuales es posible realizar dos tipos de inferencia: *i)* estimación puntual o por intervalo y *ii)* contrastes o pruebas de hipótesis sobre cada uno de los parámetros. En el presente capítulo se desarrollan procesos de inferencia estadística focalizados en la estimación de parámetros; se inicia con algunos conceptos básicos asociados a la estimación puntual y, luego, se determina un intervalo de confianza para estimar algunos de los parámetros usuales en estadística inferencial básica, entre ellos, el intervalo para estimar la media y proporción poblacionales, la diferencia de medias y de proporciones poblacionales, la varianza y el cociente de varianzas. El capítulo cuatro se dedica la prueba de hipótesis.

3.1. Estimación puntual

Varios de los aspectos teóricos considerados en esta sección son asumidos o adaptados de los conceptos expuestos al respecto por diversos autores, entre ellos: Bickel & Doksum (1977), Canavos (1988), Devore (2008), Freund y Miller (2000), Gutiérrez, *et al.* (2008), Lindgren (1993), Meeker & Escobar (1998), Mayorga (2003), Macchi *et al.* (2014) y Walpole, Myers, Myers & Ye (2007). A continuación se indican algunos conceptos y propiedades alusivos a los estimadores, y se describe el método de estimación de máxima verosimilitud.

3.1.1 Propiedades de los estimadores

Recuérdese que un estimador T es una función de las variables que conforman una muestra aleatoria, pero que no incluye ningún parámetro θ; de hecho, T es otra variable aleatoria y, por lo tanto, es posible determinar su valor esperado $E(T)$ y su varianza $Var(T)$. La varianza de T se expresa de la siguiente manera:

$$Var(T) = E\left(T - E(T)\right)^2$$

Se define el error cuadrático medio del estimador T de la siguiente manera:

$$e.c.m(T) = E\left(T - \theta\right)^2$$

Ahora, la anterior expresión se puede descomponer así:

$$e.c.m(T) = E\left(T - \theta\right)^2 = E\left(T - E(T) + (E(T) - \theta)\right)^2$$

$$e.c.m(T) = E\left(T - E(T)\right)^2 + 2E\left((T - E(T))(E(T) - \theta)\right) + \left(E(T) - \theta\right)^2$$

$$e.c.m(T) = E\left(T - E(T)\right)^2 + 2\left((E(T) - E(T))(E(T) - \theta)\right) + \left(E(T) - \theta\right)^2$$

Como el doble producto se anula, entonces, la anterior expresión toma la forma:

$$e.c.m(T) = E\left(T - E(T)\right)^2 + \left(E(T) - \theta\right)^2$$

Al valor

$$B(T) = E(T) - \theta$$

se le denomina el sesgo del estimador T.

Enseguida se indican algunas características susceptibles de presentarse en un determinado estimador.

3.1.1.1 Insesgamiento

Se dice que un estimador T para el parámetro θ es insesgado si su sesgo es cero, es decir, T es insegado si

$$B(T) = E(T) - \theta = 0$$

En otras palabras, T es insesgado si el valor esperado del estimador T es igual al parámetro θ; luego para verificar si T es insesgado se ha de verificar la siguiente igualdad:

$$E(T) = \theta$$

Ahora, si T es un estimador insesgado, entonces el error cuadrático medio de T es igual a la varianza de T, es decir:

$$e.c.m(T) = E\left(T - E(T)\right)^2 + \left(E(T) - \theta\right)^2 = E\left(T - E(T)\right)^2 = Var(T)$$

Ejemplo 3.1. Analizar si el promedio muestral $\bar{X}$ es insesgado.

Efectivamente, como se cumple que:

$$E(\bar{X}) = \mu$$

entonces, se concluye que $\bar{X}$ es un estimador insesgado para el parámetro media poblacional μ.

Ejemplo 3.2. Determinar si la proporción muestral $\hat{p}$ es un estimador insesgado.

En efecto,

$$E(\hat{p}) = p$$

Luego, entonces, se concluye que $\hat{p}$ es un estimador insesgado para el parámetro proporción poblacional p.

Ejemplo 3.3. Establecer si la varianza corregida $\hat{S}^2$ es un estimador insesgado.

Debido a que se satisface que

$$E(\hat{S}^2) = \sigma^2$$

se concluye que $\hat{S}^2$ es un estimador insesgado para el parámetro varianza poblacional σ^2.

Ejemplo 3.4. Analizar si la varianza muestral S^2 es un estimador insesgado.

Puesto que, por definición, la varianza muestral y cuasivarianza se relacionan mediante la siguiente igualdad:

$$\hat{S}^2 = \frac{n}{n-1} S^2 \qquad \text{o} \qquad S^2 = \frac{n-1}{n} \hat{S}^2$$

entonces, resulta que

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

Lo anterior indica que S^2 es un estimador sesgado o no insesgado para el parámetro varianza poblacional σ^2.

3.1.1.2 Insesgamiento asintótico

Un estimador T del parámetro θ es asintóticamente insesgado si

$$\lim_{n\to\infty} E(T) = \theta$$

Ejemplo 3.5. Analizar si la varianza muestral S^2 es un estimador asintóticamente insesgado.

La varianza muestral y la cuasivarianza se relacionan mediante la siguiente igualdad:

$$S^2 = \frac{n-1}{n}\hat{S}^2$$

Entonces,

$$\lim_{n\to\infty} E(S^2) = \lim_{n\to\infty} E\left(\frac{n-1}{n}\hat{S}^2\right) = \lim_{n\to\infty} \frac{n-1}{n} E\left(\hat{S}^2\right) = \lim_{n\to\infty} \frac{n-1}{n}\sigma^2 = \sigma^2$$

Por lo tanto,

$$\lim_{n\to\infty} E(S^2) = \sigma^2$$

Lo anterior indica que S^2 es un estimador asintóticamente insesgado para el parámetro varianza poblacional σ^2.

3.1.1.3 Eficiencia relativa

Si T_1 y T_2 son estimadores para el parámetro θ y se cumple la siguiente desigualdad:

$$Var(T_1) < Var(T_2)$$

entonces, el estimador T_1 es relativamente más eficiente que el estimador T_2. En consecuencia, el estimador con menor varianza es el más eficiente; además, de la definición se tiene que

$$\frac{Var(T_1)}{Var(T_2)} < 1$$

3.1.1.4 Consistencia

Antes de indicar la propiedad de consistencia, se presenta la denominada desigualdad de *Tchebychev*. Sea X una variable aleatoria con varianza finita,

entonces, para todo $\varepsilon > 0$ se satisface la siguiente desigualdad (Blanco, 2004):

$$P\left(\left|X - E(X)\right| \geq \varepsilon\right) \leq \frac{Var(X)}{\varepsilon^2}$$

La propiedad de consistencia se enuncia de la siguiente forma: si la sucesión de estimadores $T_1, T_2, \ldots, T_n$ del parámetro θ, entonces, T_n es consistente si

$$\lim_{n \to \infty} P\left(\left|T_n - \theta\right| \geq \varepsilon\right) = 0$$

Ejemplo 3.6. Analizar si el promedio muestral es un estimador consistente para el parámetro μ.

Se sabe que para cualquier muestra de tamaño *n*,

$$E(\bar{X}_n) = \mu$$

Aplicando la desigualdad de *Tchebychev* a la variable promedio muestral, resulta:

$$P\left(\left|\bar{X}_n - \mu\right| \geq \varepsilon\right) \leq \frac{Var(\bar{X}_n)}{\varepsilon^2}$$

debido a que

$$Var(\bar{X}_n) = \frac{\sigma^2}{n}$$

al tomar el límite en ambos miembros de la última desigualdad y reemplazar el valor de la varianza del promedio resulta:

$$\lim_{n \to \infty} P\left(\left|\bar{X}_n - \mu\right| \geq \varepsilon\right) \leq \lim_{n \to \infty} \frac{\sigma^2}{n\varepsilon^2}$$

El límite de la parte derecha de la desigualdad es cero, y las probabilidades del lado izquierdo resultan mayores o iguales que cero, en consecuencia:

$$0 \leq \lim_{n \to \infty} P\left(\left|\bar{X}_n - \mu\right| \geq \varepsilon\right) \leq 0$$

Por lo tanto,

$$\lim_{n \to \infty} P\left(\left|\bar{X}_n - \mu\right| \geq \varepsilon\right) = 0$$

Lo anterior indica que $\bar{X}_n$ es un estimador consistente para el parámetro media poblacional μ.

3.1.1.5 Robustez

Intuitivamente, un estimador T es robusto si es insensible a datos extremos, es decir, no se deja afectar por la presencia de datos extremos.

Ejemplo 3.7. Ilustrar si el promedio muestral es un estimador robusto para el parámetro μ.

Sea *X:* el peso de unos vacunos en kilogramos

Los datos recolectados para una muestra de tamaño cinco fueron: 400, 399, 401, 395, 405. El promedio muestral es 400 y la mediana también es de 400 kg.

Ahora, se ha tomado una segunda muestra, resultando un dato extremo; esta muestra presenta los siguientes datos: 400, 399, 401, 405, 300.

En este caso, la muestra apenas ha cambiado un valor correspondiente al peso; sin embargo, el peso promedio en esta muestra es de 381, "ha disminuido de manera importante, se ha dejado afectar por el dato extremo 300"; en cambio, la mediana sigue siendo 400 kg; lo anterior indica que la mediana muestral es un estimador robusto, y el promedio muestral es un estimador no robusto.

3.1.2 Estimación de parámetros por el método de máxima verosimilitud

En el proceso de estimación puntual de parámetros es posible utilizar diversos métodos: el de *máxima verosimilitud*, el de los *momentos*, el del *pivote*, por *analogía*, la *estimación bayesiana*, entre otros (Aubone & Wöhler, 2000; Cao & Van Keilegom, 2006). Por ser el más usual, a continuación se trata el método de estimación máximo verosímil y se proporcionan algunos ejemplos alusivos.

Si $X_1, X_2, ..., X_n$ es una muestra aleatoria para estudiar la variable aleatoria X con función de densidad de probabilidad:

$$f_X(x_1, x_2, ..., x_n, \theta)$$

donde θ es un parámetro o un vector en el espacio de parámetros $\Theta \subseteq R^n$, entonces la función de verosimilitud se denota y se define de la siguiente manera:

$$L(\theta) = L(x_1, x_2, ..., x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

Se dice que un estimador $T = t(X_1, X_2, ..., X_n)$ es un estimador máximo verosímil del parámetro θ si el valor particular de $t = t(x_1, x_2, ..., x_n)$ es tal que el supremo de L siguiente:

$$Sup\{L(\theta) / \theta \in \Theta\}$$

se alcanza cuando $t = \theta$. En este caso t recibe el nombre de estimador máximo

verosímil de θ (Mayorga, 2003). Es de anotar que si $t=\theta$ maximiza a $L(\theta)$, entonces $t=\theta$ también maximiza a $Ln(L(\theta))$.

Ejemplo 3.8. Enseguida se obtiene el estimador para el parámetro *p* del modelo de probabilidad de *Bernoulli* (Burbano *et al.*, 2014), por el método de máxima verosimilitud.

Para una variable aleatoria *X*, cuya función de probabilidad está dada por:

$$f(x,p)=p^{x}(1-p)^{1-x}$$

donde p es el parámetro de la variable aleatoria X del modelo de *Bernoulli*, esta toma los valores 0, 1; si se considera una muestra aleatoria $X_1, X_2, ..., X_n$ de una población con $f(x,p)=p^{x}(1-p)^{1-x}$, la función de verosimilitud es:

$$L(p)=\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

Donde los x_i son observaciones correspondientes a las variables aleatorias $X_1, X_2, ..., X_n$.

Se trata de encontrar un valor del parámetro de tal manera que se maximice la función de verosimilitud:

$$L(p)=p^{x_1}(1-p)^{1-x_1}.p^{x_2}(1-p)^{1-x_2}....p^{x_n}(1-p)^{1-x_n}.$$

$$L(p)=p^{x_1+x_2+...+x_n}(1-p)^{1+1+...+1-(x_1+x_2+...+x_n)}$$

$$L(p)=p^{\sum_{i=1}^{n}x_i}(1-p)^{n-\sum_{i=1}^{n}x_i}$$

$$Ln\left(L(p)\right)=Ln\left(p^{\sum_{i=1}^{n}x_i}(1-p)^{n-\sum_{i=1}^{n}x_i}\right)$$

$$Ln\left(L(p)\right)=\sum_{i=1}^{n}x_i Ln(p)+\left(n-\sum_{i=1}^{n}x_i\right)Ln(1-p)$$

Ahora, se hace uso del cálculo diferencial para derivar parcialmente con respecto al parámetro p:

$$\frac{\partial}{\partial p}\left(Ln\left(L(p)\right)\right)=\frac{\partial}{\partial p}\left(\sum_{i=1}^{n}x_i Ln(p)\right)+\frac{\partial}{\partial p}\left(\left(n-\sum_{i=1}^{n}x_i\right)Ln(1-p)\right)$$

$$\frac{\partial}{\partial p}\left(Ln\left(L(p)\right)\right)=\sum_{i=1}^{n}x_i\left(\frac{1}{p}\right)+\left(n-\sum_{i=1}^{n}x_i\right)\left(\frac{-1}{1-p}\right)$$

Igualando a cero se tiene:

$$\frac{\partial}{\partial p}\left(Ln\left(L(p)\right)\right)=\sum_{i=1}^{n}x_i\left(\frac{1}{p}\right)+\left(n-\sum_{i=1}^{n}x_i\right)\left(\frac{-1}{1-p}\right)=0$$

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

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

$$\sum_{i=1}^{n}x_i-p\sum_{i=1}^{n}x_i=np-p\sum_{i=1}^{n}x_i$$

$$np=\sum_{i=1}^{n}x_i$$

$$p=\frac{\sum_{i=1}^{n}x_i}{n}$$

$$\hat{p}=\frac{\sum_{i=1}^{n}x_i}{n}$$

Esta última expresión corresponde al estimador de máxima verosimilitud del parámetro p de la distribución de Bernoulli.

Ejemplo 3.9. Para la variable aleatoria X con distribución de *Poisson*, determinar el estimador máximo verosímil. La función de probabilidad correspondiente está dada por:

$$f(x,\theta)=\begin{cases}\dfrac{\lambda^{x}e^{-\lambda}}{x!} & \text{si } x=0,1,2,3,....,...\\ 0 & \text{en otro caso.}\end{cases}$$

donde $\lambda=\theta$ es el parámetro de la variable aleatoria X con modelo de *Poisson;* si se toma una muestra aleatoria $X_1,X_2,...,X_n$, la función de verosimilitud es:

$$L(\lambda)=\prod_{i=1}^{n}\frac{\lambda^{x_i}e^{-\lambda}}{x_i!}$$

donde los x_i son observaciones correspondientes a las variables aleatorias $X_1,X_2,...,X_n$.

Se trata de encontrar un valor del parámetro de tal manera que se maximice la función de verosimilitud:

$$L(\lambda)=\left(\frac{\lambda^{x_1}e^{-\lambda}}{x_1!}\right).\left(\frac{\lambda^{x_2}e^{-\lambda}}{x_2!}\right)....\left(\frac{\lambda^{x_n}e^{-\lambda}}{x_n!}\right)$$

$$L(\lambda)=\frac{\lambda^{x_1+x_2+...+x_n}e^{-n\lambda}}{x_1!x_2!...x_n!}$$

Ahora, aplicando el logaritmo natural, resulta:

$$Ln\left(L(\lambda)\right)=Ln(\lambda)\sum_{i=1}^{n}x_i+Ln\left(e^{-n\lambda}\right)-Ln(x_1!x_2!...x_n!)$$

A continuación, se realiza el cálculo de la derivada parcial con respecto al parámetro λ:

$$\frac{\partial Ln\left(L(\lambda)\right)}{\partial\lambda}=\frac{1}{\lambda}\sum_{i=1}^{n}x_i-n$$

Igualando a cero, resulta:

$$\frac{1}{\lambda}\sum_{i=1}^{n}x_i-n=0$$

Por lo tanto,

$$\lambda=\frac{1}{n}\sum_{i=1}^{n}x_i$$

$$\hat{\lambda}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

Esta última expresión corresponde al estimador de máxima verosimilitud del parámetro λ de una variable aleatoria con distribución de *Poisson*.

Ejemplo 3.10. Para la variable aleatoria *X,* con posible función de densidad de probabilidad, definida por:

$$f(x,\theta)=\begin{cases}\theta x^{\theta-1} & \text{si } x\in(0,1)\\ 0 & \text{en otro caso.}\end{cases}$$

donde $\theta>0$ es el parámetro de la población.

Como se trata de un modelo no usual, inicialmente se ha de analizar si *f* corresponde a una función de densidad para la variable aleatoria *X*.

En efecto, si $\theta > 0$ y la variable aleatoria X toma valores x en el intervalo $(0,1)$, entonces,

$$f(x,\theta) = \theta x^{\theta-1} \geq 0$$

Enseguida se ha de probar que

$$\int_{-\infty}^{\infty} f(x,\theta)dx = 1$$

Esta integral se descompone de la siguiente manera:

$$\int_{-\infty}^{\infty} f(x,\theta)dx = \int_{-\infty}^{0} f(x,\theta)dx + \int_{0}^{1} f(x,\theta)dx + \int_{1}^{\infty} f(x,\theta)dx$$

$$\int_{-\infty}^{\infty} f(x,\theta)dx = \int_{-\infty}^{0} 0dx + \int_{0}^{1} \theta x^{\theta-1}dx + \int_{1}^{\infty} 0dx = x^{\theta}\Big|_{0}^{1} = 1-0 = 1$$

En consecuencia, la función f sí es una densidad de probabilidad para la variable aleatoria X. Para la muestra aleatoria $X_1, X_2, ..., X_n$, la función de verosimilitud es:

$$L(\theta) = \prod_{i=1}^{n} \theta x_i^{\theta-1} = (\theta x_1^{\theta-1}).(\theta x_2^{\theta-1})...(\theta x_n^{\theta-1})$$

Donde los x_i son observaciones correspondientes a las variables aleatorias $X_1, X_2, ..., X_n$.

Se trata de encontrar un valor del parámetro de tal manera que se maximice la función de verosimilitud,

$$L(\theta) = \theta^n (x_1.x_2...x_n)^{\theta-1}$$

Ahora, aplicando el logaritmo natural, resulta:

$$Ln\left(L(\theta)\right) = nLn\theta + (\theta - 1)Ln(x_1.x_2...x_n)$$

Luego, se realiza el cálculo de la derivada parcial con respecto al parámetro θ

$$\frac{\partial Ln\left(L(\theta)\right)}{\partial\theta} = \frac{n}{\theta} + Ln(x_1.x_2...x_n)$$

Igualando a cero, se tiene:

$$\frac{n}{\theta} + Ln(x_1.x_2...x_n) = 0$$

Por consiguiente,

$$\theta = \frac{-n}{Ln(x_1.x_2...x_n)}$$

$$\hat{\theta} = \frac{-n}{Ln(x_1.x_2...x_n)}$$

Esta última expresión corresponde al estimador de máxima verosimilitud del parámetro θ.

3.2 Estimación por intervalo

Se trata de determinar un intervalo de la forma *(a,b)* que contenga al parámetro θ de interés con una alta probabilidad $(1-\alpha)$; a esta se le denomina nivel de confianza o nivel de confiabilidad. En esta sección se obtienen los intervalos de confianza para estimar la media y la proporción poblacional, la diferencia de medias y la diferencia de proporciones, el intervalo para la varianza poblacional y para el cociente de varianzas poblacionales.

De manera sintética, se trata de determinar los valores *a* y *b* tales que

$$P(a \leq \theta \leq b) = 1-\alpha$$

Una representación gráfica de esta situación se observa en la Figura 3.1.

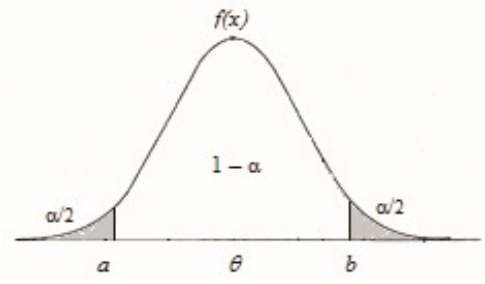

Figura 3.1 Intervalo de confianza

Fuente: los autores con la ayuda del *software* libre R.

3.2.1 Intervalos de confianza para estimar la media poblacional

Ahora, se especifica un intervalo de la forma *(a,b)* que contenga el parámetro μ con un nivel de confianza de $(1-\alpha)$. Se busca determinar los valores *a* y *b* tales que

$$P(a \leq \mu \leq b) = 1-\alpha$$

Al valor *a* se le denomina límite inferior del intervalo, y al valor *b* se le llama límite superior. Una representación gráfica de esta situación se visualiza en la Figura 3.2.

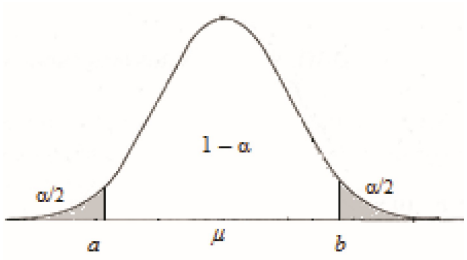

Figura 3.2 Intervalo de confianza para la media poblacional

Fuente: los autores con la ayuda del *software* libre R.

Caso 1. De la expresión 2.1 se tiene que

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Luego, en concordancia con la distribución normal estándar, se trata de determinar

$$P(Z_{\frac{\alpha}{2}} \leq Z \leq Z_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

$$P(Z_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq Z_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

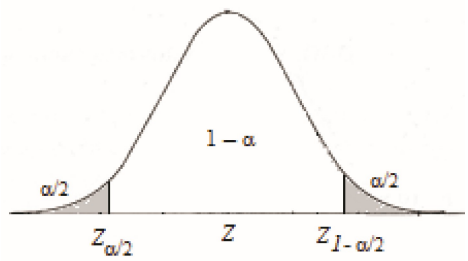

Figura 3.3 Intervalo de confianza sobre la curva normal

Fuente: los autores con la ayuda del *software* libre R.

En concordancia con la Figura 3.3, la función de densidad de una variable aleatoria *Z* con distribución normal estándar permite escribir:

$$P(Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

De aquí se deduce que

$$P(-\bar{X} + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

Multiplicando por -1 en cada uno de los miembros de la desigualdad del evento al que se le calcula la probabilidad, resulta:

$$P(\bar{X} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \geq \mu \geq \bar{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

La anterior expresión se escribe de la siguiente forma:

$$P(\bar{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

Debido a que la curva normal es simétrica, resulta que:

$$Z_{\frac{\alpha}{2}} = -Z_{1-\frac{\alpha}{2}}$$

Usando la igualdad anterior en la expresión que le precede, se tiene que:

$$P(\bar{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

Por lo tanto, el intervalo de confianza para estimar la media poblacional cuando se conoce la desviación estándar poblacional está dado por la expresión 3.1, para cuando se muestrea de una poblacional infinita.

$$\mu \in (\bar{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) \qquad (3.1)$$

Al valor $\frac{\sigma}{\sqrt{n}}$ se le denomina error estándar, para estimar la media poblacional usando el promedio muestral (Gutiérrez *et al.*, 2008).

Al valor $Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$ se le denomina error de estimación, porque hasta ese valor

puede diferir el estimador puntual denominado media muestral del parámetro media poblacional (Gutiérrez *et al.*, 2008).

Caso 2. De la expresión 2.3 se tiene que

$$t = \frac{\bar{X} - \mu}{\frac{\hat{S}}{\sqrt{n}}}$$

tiene distribución t-student con $n - 1$ grados de libertad, luego se trata de determinar

$$P(t_{\frac{\alpha}{2}} \leq t \leq t_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

$$P(t_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\frac{\hat{S}}{\sqrt{n}}} \leq t_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

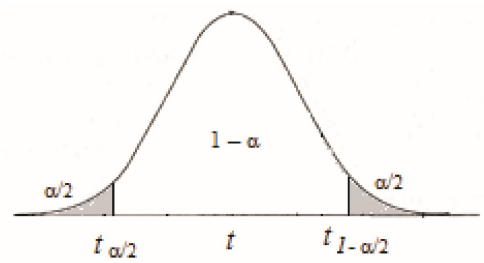

Figura 3.4 Intervalo de confianza sobre la curva t-student

Fuente: los autores con la ayuda del *software* libre R.

De acuerdo con la Figura 3.4, la función de densidad de una variable aleatoria t con distribución *t-student* permite escribir:

$$P(t_{\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}} \leq \bar{X} - \mu \leq t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}}) = 1 - \alpha$$

De aquí se deduce que

$$P(-\bar{X} + t_{\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}} \leq -\mu \leq -\bar{X} + t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}}) = 1 - \alpha$$

Multiplicando por -1 en cada uno de los miembros de la desigualdad del evento involucrado en el cálculo de la probabilidad, resulta

$$P(\bar{X} - t_{\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}} \geq \mu \geq \bar{X} - t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}}) = 1 - \alpha$$

La anterior expresión se escribe de la siguiente manera:

$$P(\bar{X} - t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}} \leq \mu \leq \bar{X} - t_{\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}}) = 1 - \alpha$$

Puesto que la curva t-student es simétrica, se tiene que:

$$t_{\frac{\alpha}{2}} = -t_{1-\frac{\alpha}{2}}$$

Usando la igualdad anterior en la expresión que le precede, se deduce que:

$$P(\bar{X} - t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}}) = 1 - \alpha$$

Por lo tanto, el intervalo de confianza para estimar la media poblacional cuando se desconoce la desviación estándar poblacional y el tamaño de la muestra es inferior a 30; cuando se muestrea de una población infinita está dado por la expresión 3.2:

$$\mu \in (\bar{X} - t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}}, \bar{X} + t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}}) \tag{3.2}$$

Caso 3. Al proceder de manera similar a como se obtuvo el intervalo de confianza presentado por medio de la expresión 3.1, pero utilizando la expresión 2.2, que estableció que

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}\left(\sqrt{\frac{N-n}{N-1}}\right)} \sim N(0,1)$$

se deduce que el intervalo de confianza para estimar la media poblacional cuando se conoce la desviación estándar poblacional corresponde a la expresión 3.3 para cuando se muestrea de una poblacional finita de tamaño *N*:

$$\mu \in (\bar{X} - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}, \bar{X} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}) \tag{3.3}$$

Caso 4. Al realizar un proceso semejante al desarrollado para obtener el intervalo de confianza indicado a través de la expresión 3.2, pero usando la expresión 2.4, se estableció que:

$$t = \frac{\bar{X} - \mu}{\frac{\hat{S}}{\sqrt{n}}\left(\sqrt{\frac{N-n}{N-1}}\right)}$$

tenía una distribución *t-student* con $n-1$ grados de libertad; por consiguiente, se deduce que el intervalo de confianza para estimar la media poblacional cuando se desconoce la desviación estándar poblacional corresponde a la expresión 3.4 para cuando se muestrea de una poblacional finita de tamaño N.

$$\mu \in (\bar{X} - t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}, \bar{X} + t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}) \quad (3.4)$$

Ejemplo 3.11. En la empresa de lácteos T&R, la cantidad de leche depositada por la máquina A en cada una de las bolsas se distribuye normalmente con desviación estándar de 5 mililitros; se toma una muestra aleatoria de 16 bolsas de tal máquina y se encuentra un contenido promedio de 900 mililitros. Construir un intervalo de confianza del 98 % para estimar la verdadera media de llenado en la producción que se haga a través de la máquina A.

En este caso, se tiene:

$$n = 16$$
$$\bar{X} = 900$$
$$\sigma = 5$$

Además,

$$1-\alpha = 0.98 \rightarrow \alpha = 0.02$$
$$\frac{\alpha}{2} = 0.01$$

En la Figura 3.5 se observa el valor de Z obtenido al leer una tabla normal estándar para un 99 % de probabilidad.

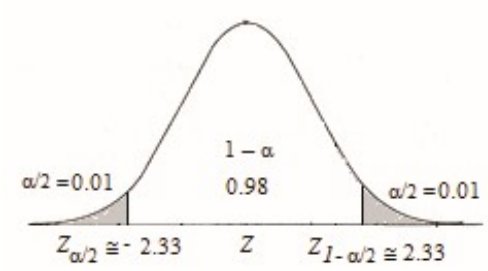

Figura 3.5 Valor de Z en una curva normal estándar

Fuente: los autores con la ayuda del *software* libre R.

$$Z_{1-\frac{\alpha}{2}} = Z_{0.99} \cong 2.33$$

Luego el intervalo de confianza es:

$$\mu \in (\bar{X} - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$$

Reemplazando los valores, resulta:

$$\mu \in \left(900 - 2.33(\frac{5}{\sqrt{16}}), 900 + 2.33(\frac{5}{\sqrt{16}}) \right)$$

Realizando las operaciones de tipo aritmético se obtiene:

$$\mu \in (900 - 2.33(1.25), 900 + 2.33(1.25))$$

Luego,

$$\mu \in (900 - 2.9125, 900 + 2.9125)$$

Finalmente,

$$\mu \in (897.08, 902.91)$$

En conclusión, con un nivel de confianza del 98 % se puede afirmar que el promedio poblacional de llenado de leche de las bolsas producidas por la máquina A está ente 897.08 mililitros y 902.91 mililitros, aproximadamente. Lo anterior indica que de cada 100 muestras que se tomen, en 98 se encuentra el parámetro μ.

Ejemplo 3.12. En una muestra aleatoria de 10 latas de un producto se obtuvo un peso neto promedio de 184 g, con una desviación estándar corregida de 3 g; si se asume que los datos provienen de una distribución normal, determinar un intervalo de confianza del 95 % para estimar el verdadero peso promedio de las latas del producto en la población.

$$\hat{S} = 3 \qquad 1 - \alpha = 95\% = 0.95$$
$$n = 10 \qquad \alpha = 0.05$$
$$\bar{X} = 184 \qquad \frac{\alpha}{2} = 0.025$$

En la Figura 3.6 se observa el valor de *t* obtenido al leer una tabla de la distribución t-student con n – 1 = 10 – 1 = 9 grados de libertad para un 0.975 = 97.5 % de probabilidad.

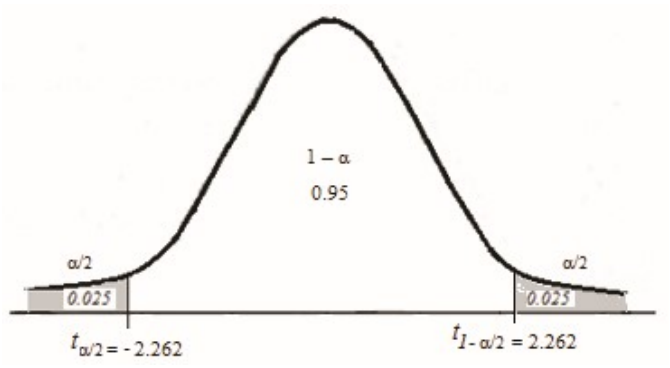

Figura 3.6 Valor de t en una curva t-student con $n = 9$ grados
Fuente: los autores con la ayuda del *software* libre R.

$$t_{1-\frac{\alpha}{2}} = t_{0.975,9} = 2.262$$

Así, el intervalo de confianza es:

$$\mu \in (\overline{X} - t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}}, \overline{X} + t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}})$$

Al reemplazar los valores se tiene:

$$\mu \in \left(184 - 2.262(\frac{3}{\sqrt{10}}), 184 + 2.262(\frac{3}{\sqrt{10}}) \right)$$

Haciendo las operaciones de tipo aritmético se obtiene:

$$\mu \in (184 - 2.262(0.9487), 184 + 2.262(0.9487))$$

Entonces,

$$\mu \in (184 - 2.1459, 184 + 2.1459)$$

Por consiguiente,

$$\mu \in (181.85, 186.15)$$

En conclusión, con un nivel de confianza del 95 % se infiere que el peso promedio de las latas del producto en la poblacional está entre 181.85 g y 186.15 g, aproximadamente. Lo anterior indica que en 95 de cada 100 muestras tomadas se encuentra el parámetro μ.

Ejemplo 3.13. Este ejemplo es adaptado de Gutiérrez *et al.* (2008). En un proceso de inyección de plástico, una característica de calidad del producto

(disco) es su grosor, que ha de ser de 1.21 mm, con una tolerancia de ±0.09 mm. En este contexto, el grosor del disco debe estar en un rango de 1.12 y 1.30 mm, para aceptar que el proceso de inyección resultó satisfactorio. Para evaluar esta característica de calidad, durante una semana se ha realizado un muestreo sistemático en una de las líneas de producción, seleccionando cuatro muestras de tamaño 20 y una de tamaño 21, bajo normalidad, para finalmente obtener una muestra de n = 101; de allí se obtienen la media muestral y la desviación estándar corregida, que tuvieron los siguientes valores: $\bar{X} = 1.18$ mm y $\hat{S} = 0.0259$ mm. Hallar el error estándar para estimar la media, construir un intervalo de confianza del 95 % para estimar la media poblacional y determinar el error de estimación.

En primera instancia, el error estándar para la media es:

$$\frac{\hat{S}}{\sqrt{n}} = \frac{0.0259}{\sqrt{101}} = \frac{0.0259}{10.0498} = 0.002577$$

Ahora, para construir el intervalo de confianza se tiene en cuenta que:

$$1 - \alpha = 95\% = 0.95$$
$$\alpha = 0.05$$
$$\frac{\alpha}{2} = 0.025$$

En la Figura 3.7 se observa el valor de *t* obtenido al leer una tabla de la distribución *t-student* con *n – 1* = 101 – 1 =100 grados de libertad para un 0.975 = 97.5 % de probabilidad.

$$t_{1-\frac{\alpha}{2}} = t_{0.975,100} = 1.984$$

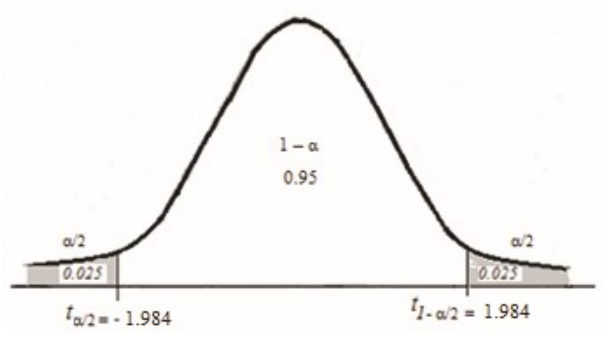

Figura 3.7 Valor de t en una curva t-student con *n* = 100 grados

Fuente: los autores con la ayuda del *software* libre R.

Así, el intervalo de confianza es:

$$\mu \in (\overline{X} - t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}}, \overline{X} + t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}})$$

Al reemplazar los valores se tiene:

$$\mu \in \left(1.18 - 1.984(\frac{0.0259}{\sqrt{101}}), 1.18 + 1.984(\frac{0.0259}{\sqrt{101}}) \right)$$

Haciendo las operaciones de tipo aritmético se obtiene:

$$\mu \in (1.18 - 1.984(0.002577), 1.18 + 1.984(0.002577))$$

Entonces,

$$\mu \in (1.18 - 0.00511, 1.18 + 0.00511)$$

Por lo tanto,

$$\mu \in (1.17489, 1.18511)$$

En conclusión, con un nivel de confianza del 95 % se infiere que el grosor promedio de los discos en la poblacional está entre 1.17489 mm y 1.18511 mm, aproximadamente; estos resultados indican que tal grosor se encuentra dentro de los valores especificados de calidad. Lo anterior significa que en 95 de cada 100 muestras tomadas se encuentra el parámetro μ, correspondiente al grosor promedio en la población de discos.

Finalmente, el error de estimación es:

$$t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}} = 1.984(\frac{0.0259}{\sqrt{101}}) = 0.00511$$

El valor 0.00511 indica que hasta ese valor puede diferir el estimador puntual $\overline{X}$ del parámetro μ.

3.2.2 Intervalos de confianza para estimar la proporción poblacional

Nuevamente se trata de construir un intervalo de la forma *(a,b)* que contenga el parámetro p con un nivel de confianza de $(1 - \alpha)$. Se busca determinar los valores *a* y *b* tales que:

$$P(a \leq p \leq b) = 1 - \alpha$$

Una representación gráfica de esta situación se observa en la Figura 3.8:

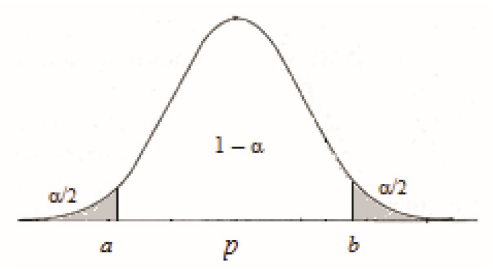

Figura 3.8 Intervalo de confianza para la proporción poblacional

Fuente: los autores con la ayuda del *software* libre R.

Caso 1. De la expresión 2.5 se tiene que

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1)$$

Para un tamaño de muestra *n* lo suficientemente grande, se asume que

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}\hat{q}}{n}}} \sim N(0,1)$$

Así entonces, de acuerdo con la distribución normal estándar (ver Figura 3.3), se tiene:

$$P(Z_{\frac{\alpha}{2}} \leq Z \leq Z_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

$$P(Z_{\frac{\alpha}{2}} \leq \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}\hat{q}}{n}}} \leq Z_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

Esta expresión se escribe de la siguiente forma:

$$P(Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}} \leq \hat{p} - p \leq Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}) = 1 - \alpha$$

De aquí se establece que

$$P(-\hat{p} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}} \leq -p \leq -\hat{p} + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}) = 1 - \alpha$$

Multiplicando por -1 en cada uno de los miembros de la desigualdad del evento al que se le calcula la probabilidad, resulta:

$$P(\hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}} \geq p \geq \hat{p} - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}) = 1 - \alpha$$

Esta expresión se escribe de la siguiente manera:

$$P(\hat{p} - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}} \leq p \leq \hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}) = 1 - \alpha$$

Puesto que la curva normal es simétrica, resulta que

$$Z_{\frac{\alpha}{2}} = -Z_{1-\frac{\alpha}{2}}$$

Usando la igualdad anterior en la expresión que le precede, se tiene que

$$P(\hat{p} - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}} \leq p \leq \hat{p} + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}) = 1 - \alpha$$

Por lo tanto, el intervalo de confianza para estimar la proporción poblacional cuando se muestrea de una poblacional infinita se especifica en la expresión 3.5.

$$p \in \left(\hat{p} - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}} \ , \ \hat{p} + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}} \right) \tag{3.5}$$

Al valor $\sqrt{\frac{\hat{p}\hat{q}}{n}}$ se le denomina error estándar para estimar la proporción poblacional usando la proporción muestral, y al valor $Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}$ se le llama error de estimación.

Caso 2. Ahora, si se muestrea de una población finita de tamaño N y el tamaño de la muestra es "grande", entonces se deduce que

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}}} \sim N(0,1) \tag{3.6}$$

Para un tamaño de muestra n lo suficientemente grande, se asume que

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}\hat{q}}{n}}\sqrt{\frac{N-n}{N-1}}} \sim N(0,1)$$

Realizando un procedimiento similar al efectuado para el *caso 1*, se establece que el intervalo de confianza para estimar la proporción poblacional cuando

se muestrea de una población finita de tamaño N y el tamaño de la muestra es "grande" es aquel que se indica en la expresión 3.7.

$$p \in \left(\hat{p} - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} \sqrt{\frac{N-n}{N-1}} \ , \ \hat{p} + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} \sqrt{\frac{N-n}{N-1}} \right) \tag{3.7}$$

Ejemplo 3.14. Un jugador de baloncesto produce 140 aciertos en 400 lanzamientos al aro en entrenamientos seleccionados de manera aleatoria. Construir un intervalo de confianza del 99 % para estimar la verdadera proporción de la efectividad del jugador.

En primera instancia, se define así *X:* número de aciertos en la muestra de 400 lanzamientos, luego

$$n = 400$$

$$\hat{p} = \frac{x}{n} = \frac{140}{400} = 0.35$$

$$\hat{q} = 1 - 0.35 = 0.65$$

Adicionalmente,

$$1 - \alpha = 0.99 \rightarrow \alpha = 0.01$$

$$\frac{\alpha}{2} = 0.005$$

En la Figura 3.9 se observa el valor de Z obtenido al leer una tabla normal estándar para un 0.995 = 99.5 % de probabilidad.

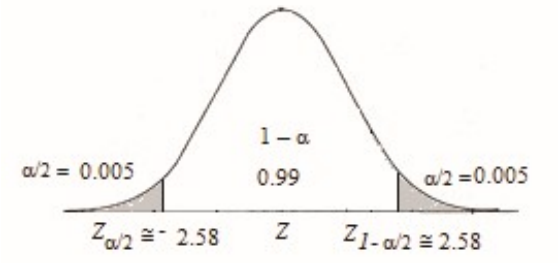

Figura 3.9 Valor de Z en una curva normal estándar
Fuente: los autores con la ayuda del *software* libre R.

$$Z_{1-\frac{\alpha}{2}} = Z_{0.995} \cong 2.58$$

Luego el intervalo de confianza es:

$$p \in \left(\hat{p} - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} \ , \ \hat{p} + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} \right)$$

Reemplazando los valores, resulta:

$$p \in \left(0.35 - 2.58 \sqrt{\frac{0.35 * 0.65}{400}}, 0.35 + 2.58 \sqrt{\frac{0.35 * 0.65}{400}} \right)$$

Realizando las operaciones de tipo aritmético se obtiene:

$$p \in (0.35 - 2.58(0.0238) \ , \ 0.35 + 2.58(0.0238))$$

Luego

$$p \in (0.35 - 0.0614 \ , \ 0.35 + 0.0614)$$

En consecuencia,

$$p \in (0.2886 \ , \ 0.4114)$$

En conclusión, con un nivel de confianza del 99 % es posible afirmar que la verdadera proporción de la efectividad del jugador (en la población de entrenamientos) se encuentra entre el 28.86 % y 41.14 %. Lo anterior indica que en 99 de cada 100 muestras tomadas se encuentra el parámetro p.

3.2.3 Intervalos de confianza para estimar la diferencia de proporciones poblacionales

Ahora se pretende construir un intervalo de la forma *(a,b)* que contenga al parámetro $p_1 - p_2$ con un nivel de confianza de $(1 - \alpha)$. Se busca determinar los valores *a* y *b* tales que:

$$P(a \leq p_1 - p_2 \leq b) = 1 - \alpha$$

Una representación gráfica de esta situación se presenta en la Figura 3.10.

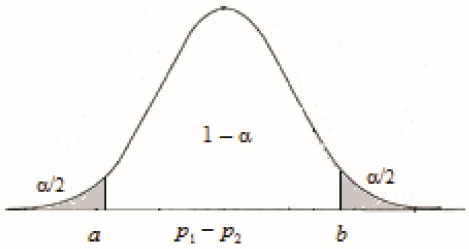

Figura 3.10 Intervalo de confianza para la diferencia de proporciones poblacionales

Fuente: los autores con la ayuda del *software* libre R.

De la expresión 2.6 se tiene que:

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \sim N(0,1)$$

Para tamaños de las dos muestras n_1 y n_2 grandes, se asume que:

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}} \sim N(0,1)$$

En concordancia con la distribución normal estándar (ver Figura 3.3), resulta:

$$P(Z_{\frac{\alpha}{2}} \leq Z \leq Z_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

$$P(Z_{\frac{\alpha}{2}} \leq \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}} \leq Z_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

La anterior expresión se escribe de la siguiente manera:

$$P\left(Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \leq \hat{p}_1 - \hat{p}_2 - (p_1 - p_2) \leq Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \right) = 1 - \alpha$$

De aquí se establece que:

$$P\left(-(\hat{p}_1 - \hat{p}_2) + Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \leq -(p_1 - p_2) \leq -(\hat{p}_1 - \hat{p}_2) + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \right) = 1 - \alpha$$

Multiplicando por -1 en cada uno de los miembros de la desigualdad del evento involucrado en el cálculo de la probabilidad, resulta:

$$P\left((\hat{p}_1 - \hat{p}_2) - Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \geq (p_1 - p_2) \geq (\hat{p}_1 - \hat{p}_2) - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \right) = 1 - \alpha$$

Esta expresión se escribe de la siguiente forma:

$$P\left((\hat{p}_1 - \hat{p}_2) - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \leq p_1 - p_2 \leq (\hat{p}_1 - \hat{p}_2) - Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \right) = 1 - \alpha$$

Puesto que la curva normal es simétrica, resulta que:

$$Z_{\frac{\alpha}{2}} = -Z_{1-\frac{\alpha}{2}}$$

Usando la igualdad anterior en la expresión que le precede, se tiene que:

$$P\left((\hat{p}_1-\hat{p}_2)-Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}} \le p_1-p_2 \le (\hat{p}_1-\hat{p}_2)+Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}}\right)=1-\alpha$$

Por lo tanto, el intervalo de confianza para estimar la diferencia de proporciones poblacionales se especifica en la expresión 3.8.

$$p_1-p_2 \in \left((\hat{p}_1-\hat{p}_2)-Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}}\ ,\ (\hat{p}_1-\hat{p}_2)+Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}}\right) \qquad (3.8)$$

Ejemplo 3.15. Un candidato a la presidencia de la república de Colombia tiene, en una muestra aleatoria de 500 ciudadanos seleccionados en el departamento de Boyacá, el 60 % de la intención de voto, y en una muestra aleatoria de 450 ciudadanos tomada en el departamento de Cundinamarca logra un 56 % de la intención de voto; construir un intervalo de confianza del 98 % para estimar la verdadera diferencia de proporciones.
En este caso se tiene:

Boyacá	Cundinamarca
$n_1 = 500$	$n_2 = 450$
$\hat{p}_1 = 0.6$	$\hat{p}_2 = 0.56$
$\hat{q}_1 = 1-0.6 = 0.4$	$\hat{q}_2 = 1-0.56 = 0.44$

Además,

$$1-\alpha = 0.98 \rightarrow \alpha = 0.02$$

$$\frac{\alpha}{2} = 0.01$$

En la Figura 3.11 se observa el valor de Z obtenido al leer una tabla normal estándar para un 0.99 = 99 % de probabilidad.

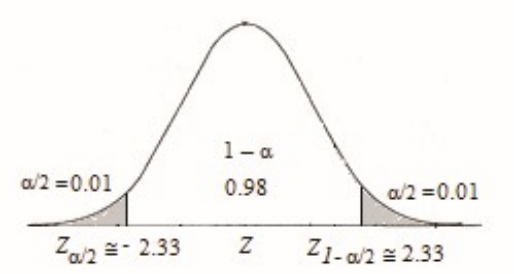

Figura 3.11 Valor de Z en una curva normal estándar
Fuente: los autores con la ayuda del *software* libre R.

$$Z_{1-\frac{\alpha}{2}} = Z_{0.99} \cong 2.33$$

Luego, en concordancia con la expresión 3.8, el intervalo de confianza es:

$$p_1 - p_2 \in \left((\hat{p}_1 - \hat{p}_2) - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} \ , \ (\hat{p}_1 - \hat{p}_2) + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} \right)$$

Reemplazando los valores, resulta:

$$\left((0.6-0.56) - 2.33\sqrt{\frac{(0.6)(0.4)}{500} + \frac{(0.56)(0.44)}{450}} \ , \ (0.6-0.56) + 2.33\sqrt{\frac{(0.6)(0.4)}{500} + \frac{(0.56)(0.44)}{450}} \right)$$

Realizando las operaciones de tipo aritmético se obtiene:

$$p_1 - p_2 \in \left((0.6-0.56) - 2.33\sqrt{0.000480 + 0.000547} \ , \ (0.6-0.56) + 2.33\sqrt{0.000480 + 0.000547} \right)$$

Luego:

$$p_1 - p_2 \in \left((0.6-0.56) - 2.33(0.0320) \ , \ (0.6-0.56) + 2.33(0.032) \right)$$

En consecuencia,

$$p_1 - p_2 \in \left(0.04 - 0.0746 \ , \ 0.04 + 0.0746 \right)$$

Finalmente,

$$p_1 - p_2 \in \left(-0.0346 \ , \ 0.1146 \right)$$

En conclusión, con un nivel de confianza del 98% se establece que la verdadera diferencia de proporciones poblacionales se encuentra entre -3.46% y 11.46%. Lo anterior indica que en 98 de cada 100 muestras tomadas se encuentra el parámetro $p_1 - p_2$.

Adicionalmente, en este ejemplo se infieren las tres situaciones siguientes:

i) $p_1 - p_2 = 0 \Rightarrow p_1 = p_2$;

ii) $p_1 - p_2 > 0 \Rightarrow p_1 > p_2$;

iii) $p_1 - p_2 < 0 \Rightarrow p_1 < p_2$

3.2.4 Intervalos de confianza para estimar la diferencia de medias poblacionales

En este apartado se determina un intervalo de la forma *(a,b),* de modo que este contenga el parámetro $\mu_1 - \mu_2$ con un nivel de confianza de $(1 - \alpha)$. Se busca especificar los valores *a* y *b* tales que:

$$P(a \leq \mu_1 - \mu_2 \leq b) = 1 - \alpha$$

Una representación gráfica de esta situación se observa en la Figura 3.12.

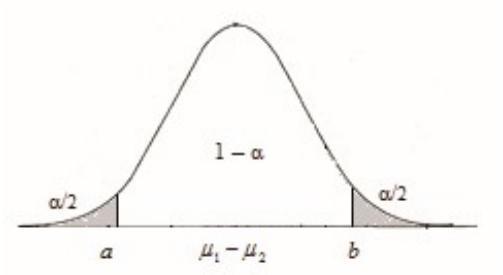

Figura 3.12 Intervalo de confianza para la diferencia de medias poblacionales
Fuente: los autores con la ayuda del *software* libre R.

Caso 1. De la expresión 2.7 se sigue que:

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Mediante la distribución normal estándar, se trata de determinar:

$$P(Z_{\frac{\alpha}{2}} \leq Z \leq Z_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

$$P(Z_{\frac{\alpha}{2}} \leq \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq Z_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

La anterior expresión se escribe así:

$$P\left(Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \le \bar{X}_1-\bar{X}_2-(\mu_1-\mu_2) \le Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \right)=1-\alpha$$

De aquí se deduce que:

$$P\left(-(\bar{X}_1-\bar{X}_2)+Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \le -(\mu_1-\mu_2) \le -(\bar{X}_1-\bar{X}_2)+Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \right)=1-\alpha$$

Multiplicando por -1 en cada uno de los miembros de la desigualdad del evento para el cálculo de la probabilidad, resulta:

$$P\left((\bar{X}_1-\bar{X}_2)-Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \ge (\mu_1-\mu_2) \ge (\bar{X}_1-\bar{X}_2)-Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \right)=1-\alpha$$

Esta expresión se escribe de la siguiente manera:

$$P\left((\bar{X}_1-\bar{X}_2)-Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \le \mu_1-\mu_2 \le (\bar{X}_1-\bar{X}_2)-Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \right)=1-\alpha$$

Debido a que la curva normal es simétrica, se tiene que:

$$Z_{\frac{\alpha}{2}} = -Z_{1-\frac{\alpha}{2}}$$

Usando la igualdad anterior en la expresión que le precede, se deduce que:

$$P\left((\bar{X}_1-\bar{X}_2)-Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \le \mu_1-\mu_2 \le (\bar{X}_1-\bar{X}_2)+Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \right)=1-\alpha$$

Por lo tanto, el intervalo de confianza para estimar la diferencia de medias poblacionales cuando se conocen las respectivas desviaciones estándar poblacionales está dado por la expresión 3.9.

$$\mu_1-\mu_2 \in \left((\bar{X}_1-\bar{X}_2)-Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \ , \ (\bar{X}_1-\bar{X}_2)+Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \right)=1-\alpha \qquad (3.9)$$

Al valor $\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$ se le denomina error estándar para estimar la diferencia de medias poblacionales. Al valor $Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$ se le denomina error de estimación.

Caso 2. De la expresión 2.8 se sigue que

$$t = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

tiene distribución t-student con $n_1 + n_2 - 2$ grados de libertad, donde de S_p se obtiene mediante la siguiente expresión, a partir de datos provenientes de muestras independientes de poblaciones normales:

$$S_p = \sqrt{\frac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{n_1 + n_2 - 2}}$$

Luego, se trata de determinar:

$$P(t_{\frac{\alpha}{2}} \leq t \leq t_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

$$P(t_{\frac{\alpha}{2}} \leq \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

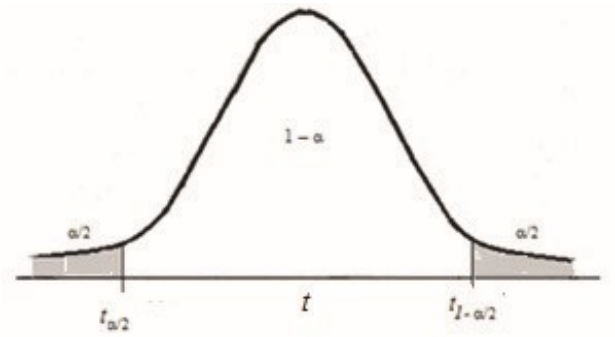

Figura 3.13 Intervalo de confianza sobre la curva t-student

Fuente: los autores con la ayuda del *software* libre R.

De acuerdo con la Figura 3.13, la función de densidad de una variable aleatoria t con distribución *t-student* permite escribir:

$$P\left(t_{\frac{\alpha}{2}} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2) \leq t_{1-\frac{\alpha}{2}} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) = 1 - \alpha$$

De aquí se deduce que:

$$P\left(-(\bar{X}_1-\bar{X}_2)+t_{\frac{\alpha}{2}}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\le-(\mu_1-\mu_2)\le-(\bar{X}_1-\bar{X}_2)+t_{1-\frac{\alpha}{2}}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)=1-\alpha$$

Multiplicando por -1 en cada uno de los miembros de la desigualdad del evento involucrado en el cálculo de la probabilidad, resulta:

$$P\left((\bar{X}_1-\bar{X}_2)-t_{\frac{\alpha}{2}}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\ge(\mu_1-\mu_2)\ge(\bar{X}_1-\bar{X}_2)-t_{1-\frac{\alpha}{2}}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)=1-\alpha$$

La anterior expresión se escribe de la siguiente manera:

$$P\left((\bar{X}_1-\bar{X}_2)-t_{1-\frac{\alpha}{2}}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\le\mu_1-\mu_2\le(\bar{X}_1-\bar{X}_2)-t_{\frac{\alpha}{2}}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)=1-\alpha$$

Puesto que la curva t-student es simétrica, se tiene que:

$$t_{\frac{\alpha}{2}}=-t_{1-\frac{\alpha}{2}}$$

Usando la igualdad anterior en la expresión que le precede, se establece que:

$$P\left((\bar{X}_1-\bar{X}_2)-t_{1-\frac{\alpha}{2}}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\le\mu_1-\mu_2\le(\bar{X}_1-\bar{X}_2)+t_{1-\frac{\alpha}{2}}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)=1-\alpha$$

Por lo tanto, el intervalo de confianza para estimar la diferencia de medias poblacionales, cuando se desconocen sus correspondientes desviaciones estándar poblacionales y el tamaño de las muestras, es inferior a 30; bajo el supuesto de varianzas poblacionales iguales (homocedasticidad) está dado por la expresión 3.10.

$$\mu_1-\mu_2\in\left((\bar{X}_1-\bar{X}_2)-t_{1-\frac{\alpha}{2}}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\ ,\ (\bar{X}_1-\bar{X}_2)+t_{1-\frac{\alpha}{2}}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right) \quad (3.10)$$

Caso 3. Al proceder de manera similar, pero utilizando la expresión 2.9, se obtiene que:

$$t=\frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{\sqrt{\frac{\hat{S}_1^2}{n_1}+\frac{\hat{S}_2^2}{n_2}}}$$

con *g* grados de libertad. Se deduce que el intervalo de confianza para estimar la diferencia de medias poblacionales cuando se tienen muestras inferiores a

30 y se desconocen las desviaciones estándar poblacionales, pero considerando varianzas poblacionales desiguales (heterocedasticidad), corresponde a la expresión 3.11.

$$\mu_1-\mu_2 \in \left((\bar{X}_1-\bar{X}_2)-t_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{S}_1^2}{n_1}+\frac{\hat{S}_2^2}{n_2}}\ ,\ (\bar{X}_1-\bar{X}_2)+t_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{S}_1^2}{n_1}+\frac{\hat{S}_2^2}{n_2}} \right) \quad (3.11)$$

Caso 4. Al realizar un proceso semejante, pero usando la expresión 2.10, se estableció que:

$$Z=\frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{\sqrt{\frac{\hat{S}_1^2}{n_1}+\frac{\hat{S}_2^2}{n_2}}} \sim N(0,1)$$

Luego se deduce que el intervalo de confianza para estimar la diferencia de medias poblacionales cuando se desconocen las desviaciones estándar poblacionales, pero las muestras tienen tamaños superiores a 30, está dado por la expresión 3.12.

$$\mu_1-\mu_2 \in \left((\bar{X}_1-\bar{X}_2)-Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{S}_1^2}{n_1}+\frac{\hat{S}_2^2}{n_2}}\ ,\ (\bar{X}_1-\bar{X}_2)+Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{S}_1^2}{n_1}+\frac{\hat{S}_2^2}{n_2}} \right) \quad (3.12)$$

Ejemplo 3.16. En una muestra aleatoria conformada por 22 bolsas de jugo de la marca A se obtiene un contenido medio de 10 g de fibra con una desviación estándar corregida de 1.2 g; en otra, conformada por 25 bosas de esta clase de jugo, pero de la marca B, se obtuvo un contenido promedio de fibra de 9.8 g, con una desviación estándar de 0.95 g; se supone que las muestras provienen de poblaciones normales. Construir un intervalo de confianza del 95 % para estimar la diferencia de contenido medio de fibra entre las marcas A y B del mencionado jugo.

En este caso se tiene:

Marca A	Marca B
$n_1=22$	$n_2=25$
$\bar{X}_1=10$	$\bar{X}_2=9.8$
$\hat{S}_1=1.2$	$\hat{S}_2=0.9695$

El valor de la desviación estándar corregida para la marca B se obtiene de la siguiente manera:

$$\hat{S}_2^2 = \frac{n_2}{n_2 - 1} S_2^2 = \frac{25}{24}(0.95)^2 = 0.9401 \rightarrow \hat{S}_2 = 0.9695$$

Además,

$$1 - \alpha = 0.95 \rightarrow \alpha = 0.05$$

$$\frac{\alpha}{2} = 0.025$$

En la Figura 3.14 se observa el valor aproximado de *t* obtenido al leer una tabla *t-student* con $n_1+n_2-2=22+25-2=45$ grados de libertad, para un 0.975=97.5 % de probabilidad. En este caso se obtendrá un intervalo de confianza considerando que las varianzas poblacionales son iguales y se recurrirá a la expresión 3.10.

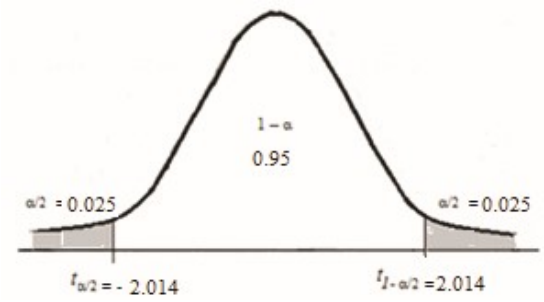

Figura 3.14 Intervalo de confianza sobre la curva t-student

Fuente: los autores con la ayuda del *software* libre R.

$$t_{1-\frac{\alpha}{2}} = t_{0.975,45} = 2.014$$

Ahora se calcula S_p así:

$$S_p = \sqrt{\frac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{21(1.2)^2 + 24(0.9695)^2}{22 + 25 - 2}} = \sqrt{\frac{52.8024}{45}} = \sqrt{1.1733} = 1.083$$

Luego, usando:

$$\mu_1 - \mu_2 \in \left((\bar{X}_1 - \bar{X}_2) - t_{1-\frac{\alpha}{2}} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \ , \ (\bar{X}_1 - \bar{X}_2) + t_{1-\frac{\alpha}{2}} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

Y reemplazando los valores, resulta:

$$\mu_1 - \mu_2 \in \left((10 - 9.8) - 2.014(1.083)\sqrt{\frac{1}{22} + \frac{1}{25}} \ , \ (10 - 9.8) + 2.014(1.083)\sqrt{\frac{1}{22} + \frac{1}{25}} \right)$$

Efectuando las operaciones de tipo aritmético se obtiene:

$$\mu_1 - \mu_2 \in \left((10-9.8) - 2.014(1.083)(0.2923)\ ,\ (10-9.8) + 2.014(1.083)(0.2923)\right)$$

Luego,

$$\mu_1 - \mu_2 \in \left(0.2 - 0.6375\ ,\ 0.2 + 0.6375\right)$$

Finalmente,

$$\mu_1 - \mu_2 \in \left(-0.4375\ ,\ 0.8375\right)$$

En conclusión, con un nivel de confianza del 95 % se infiere que la diferencia de medias poblacionales referidas al contenido de fibra en el jugo de las marcas A y B está entre -0.4375 y 0.8375 g. Lo anterior indica que, de cada 100 muestras seleccionadas, en 95 se encuentra el parámetro $\mu_1 - \mu_2$.

También en este ejemplo se infieren las tres situaciones siguientes, considerando igualdad de varianzas poblacionales:

i) $\mu_1 - \mu_2 = 0 \ \Rightarrow\ \mu_1 = \mu_2$;

ii) $\mu_1 - \mu_2 > 0 \ \Rightarrow\ \mu_1 > \mu_2$;

iii) $\mu_1 - \mu_2 < 0 \ \Rightarrow\ \mu_1 < \mu_2$

Por otro lado, se supone que las varianzas poblacionales son distintas. En este caso se utiliza la expresión 3.11 para determinar el intervalo de confianza, pero inicialmente se calculan los grados de libertad, como se indica a continuación:

$$g = \frac{\left(\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}\right)^2}{\frac{\left(\frac{\hat{S}_1^2}{n_1}\right)^2}{n_1+1} + \frac{\left(\frac{\hat{S}_2^2}{n_2}\right)^2}{n_2+1}} - 2 = \frac{\left(\frac{(1.2)^2}{22} + \frac{(0.9695)^2}{25}\right)^2}{\frac{\left(\frac{(1.2)^2}{22}\right)^2}{22+1} + \frac{\left(\frac{(0.9695)^2}{25}\right)^2}{25+1}} - 2 = \frac{0.0106}{0.0001863 + 0.000054} - 2 \cong 42.11$$

En la Figura 3.15 se observa el valor aproximado de *t* obtenido al leer una tabla *t-student* con $g = 42$ grados de libertad, para un 0.975 = 97.5 % de probabilidad.

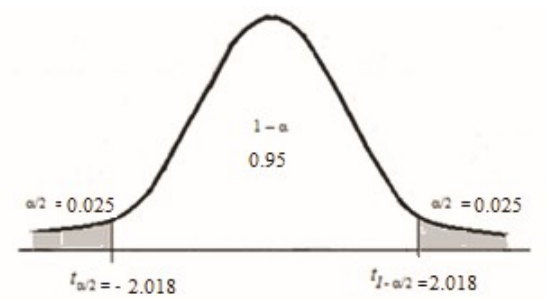

Figura 3.15 Intervalo de confianza sobre la curva t-student

Fuente: los autores con la ayuda del *software* libre R.

$$t_{1-\frac{\alpha}{2}} = t_{0.975,42} = 2.018$$

Luego, se usa la expresión 3.11

$$\mu_1 - \mu_2 \in \left((\bar{X}_1 - \bar{X}_2) - t_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}} \ , \ (\bar{X}_1 - \bar{X}_2) + t_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}} \right)$$

Al reemplazar los correspondientes valores resulta:

$$\mu_1 - \mu_2 \in \left((10-9.8) - 2.018 \sqrt{\frac{(1.2)^2}{22} + \frac{(0.9695)^2}{25}} \ , \ (10-9.8) + 2.018 \sqrt{\frac{(1.2)^2}{22} + \frac{(0.9695)^2}{25}} \right)$$

Efectuando las operaciones de tipo aritmético se obtiene:

$$\mu_1 - \mu_2 \in \left((10-9.8) - 2.018(0.3210) \ , \ (10-9.8) + 2.018(0.3210) \right)$$

Luego

$$\mu_1 - \mu_2 \in \left(0.2 - 0.6477 \ , \ 0.2 + 0.6477 \right)$$

Por lo tanto,

$$\mu_1 - \mu_2 \in \left(-0.4477 \ , \ 0.8477 \right)$$

En conclusión, con un nivel de confianza del 95 % se infiere que la diferencia de medias poblacionales referidas al contenido de fibra en el jugo de las marcas A y B está entre -0.4477 y 0.8477 g, considerando varianzas desiguales (heterocedasticidad). Lo anterior indica que, de cada 100 muestras seleccionadas, 95 tienen el parámetro $\mu_1 - \mu_2$.

Al considerar varianzas poblacionales diferentes, también se infieren las tres situaciones siguientes:

i) $\mu_1 - \mu_2 = 0 \Rightarrow \mu_1 = \mu_2$;

ii) $\mu_1 - \mu_2 > 0 \Rightarrow \mu_1 > \mu_2$;

iii) $\mu_1 - \mu_2 < 0 \Rightarrow \mu_1 < \mu_2$

3.2.5 Intervalos de confianza para estimar la varianza poblacional

En este apartado se busca obtener un intervalo de la forma *(a,b)* que contenga al parámetro σ^2 con un nivel de confianza de $(1 - \alpha)$. Se busca determinar los valores *a* y *b* tales que:

$$P(a \leq \sigma^2 \leq b) = 1 - \alpha$$

Una representación gráfica de esta situación se visualiza en la Figura 3.16. Es necesario señalar que se trata de una curva asimétrica.

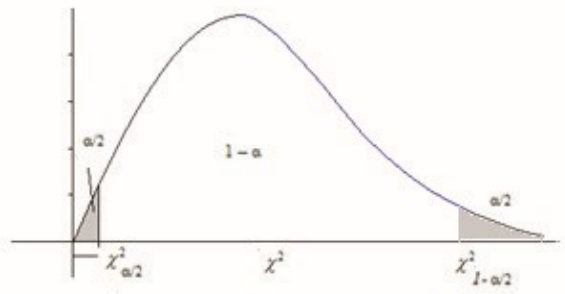

Figura 3.16 Intervalo de confianza para la varianza poblacional

Fuente: los autores con la ayuda del *software* libre R.

La siguiente variable tiene distribución chi-cuadrado con *n-1* grados de libertad:

$$\chi^2 = \frac{(n-1)\hat{S}^2}{\sigma^2}$$

Luego, en concordancia con su correspondiente distribución de probabilidad, se tiene:

$$P(\chi^2_{\frac{\alpha}{2}} \leq \chi^2 \leq \chi^2_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

En concordancia con la Figura 3.16, se escribe:

$$P(\chi^2_{\frac{\alpha}{2}} \leq \frac{(n-1)\hat{S}^2}{\sigma^2} \leq \chi^2_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

Aplicando propiedades de las desigualdades de números reales positivos en cada uno de los miembros de la desigualdad del evento involucrado en el cálculo de la probabilidad, resulta:

$$P\left(\frac{1}{\chi^2_{\frac{\alpha}{2}}} \geq \frac{1}{\frac{(n-1)\hat{S}^2}{\sigma^2}} \geq \frac{1}{\chi^2_{1-\frac{\alpha}{2}}}\right) = 1-\alpha$$

De la anterior expresión se obtiene:

$$P\left(\frac{(n-1)\hat{S}^2}{\chi^2_{\frac{\alpha}{2}}} \geq \sigma^2 \geq \frac{(n-1)\hat{S}^2}{\chi^2_{1-\frac{\alpha}{2}}}\right) = 1-\alpha$$

La desigualdad anterior se expresa de la siguiente manera:

$$P\left(\frac{(n-1)\hat{S}^2}{\chi^2_{1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1)\hat{S}^2}{\chi^2_{\frac{\alpha}{2}}}\right) = 1-\alpha$$

Por lo tanto, el intervalo de confianza para estimar la varianza poblacional está dado por la expresión 3.13.

$$\sigma^2 \in \left(\frac{(n-1)\hat{S}^2}{\chi^2_{1-\frac{\alpha}{2}}}\ ,\ \frac{(n-1)\hat{S}^2}{\chi^2_{\frac{\alpha}{2}}}\right) \tag{3.13}$$

Puesto que también es factible calcular la siguiente probabilidad:

$$P\left(\sqrt{\frac{(n-1)\hat{S}^2}{\chi^2_{1-\frac{\alpha}{2}}}} \leq \sigma \leq \sqrt{\frac{(n-1)\hat{S}^2}{\chi^2_{\frac{\alpha}{2}}}}\right) = 1-\alpha$$

entonces el intervalo de confianza para estimar la desviación estándar poblacional se obtiene mediante la expresión 3.14:

$$\sigma \in \left(\sqrt{\frac{(n-1)\hat{S}^2}{\chi^2_{1-\frac{\alpha}{2}}}}\ ,\ \sqrt{\frac{(n-1)\hat{S}^2}{\chi^2_{\frac{\alpha}{2}}}}\right) \tag{3.14}$$

Ejemplo 3.17. El siguiente ejemplo ha sido adaptado de Freund *et al.* (2000). Una clase de motor experimental es sometida a 16 pruebas para evaluar su consumo

de combustible. En esa muestra aleatoria se obtiene una desviación estándar corregida de 2.2 litros. Bajo el supuesto de normalidad para los datos, construir un intervalo de confianza del 99 % para estimar la varianza poblacional como indicador de la verdadera variación del consumo de combustible; asimismo, determinar un intervalo de confianza para estimar la desviación estándar poblacional.

Los requerimientos son los siguientes:

$$\hat{S} = 2.2$$
$$n = 16$$
$$1 - \alpha = 0.99 \rightarrow \alpha = 0.01$$
$$\frac{\alpha}{2} = 0.005$$
$$1 - \frac{\alpha}{2} = 0.995$$

En la Figura 3.17 se observan los valores de los cuantiles obtenidos al leer una tabla *chi-cuadrado n -1*=16 – 1=15 grados de libertad, para un 0.005=0.5 % y un 0.995=99.5 % de probabilidad acumulada, respectivamente.

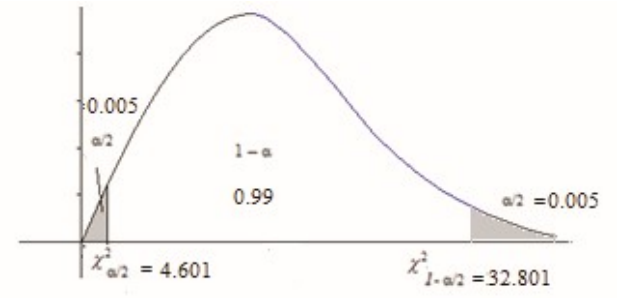

Figura 3.17 Intervalo de confianza sobre la curva chi-cuadrado
Fuente: los autores con la ayuda del *software* libre R.

$$\chi^2_{1-\frac{\alpha}{2}} = \chi^2_{0.995,15} = 32.801$$
$$\chi^2_{\frac{\alpha}{2}} = \chi^2_{0.005,15} = 4.601$$

Luego, usando la expresión 3.13:

$$\sigma^2 \in \left(\frac{(n-1)\hat{S}^2}{\chi^2_{1-\frac{\alpha}{2}}} , \frac{(n-1)\hat{S}^2}{\chi^2_{\frac{\alpha}{2}}} \right)$$

y reemplazando los valores, resulta:

$$\sigma^2 \in \left(\frac{(16-1)(2.2)^2}{32.801} \ , \ \frac{(16-1)(2.2)^2}{4.601} \right)$$

Efectuando las operaciones de tipo aritmético se obtiene:

$$\sigma^2 \in \left(\frac{72.6}{32.801} \ , \ \frac{72.6}{4.601} \right)$$

Por consiguiente,

$$\sigma^2 \in (2.213 \ , \ 15.779)$$

En conclusión, con un nivel de confianza del 99 % se infiere que la varianza poblacional en referencia al consumo de combustible está entre 2.213 y 15.779. Lo anterior indica que, de cada 100 muestras seleccionadas, 99 tienen el parámetro σ^2.

Ahora, para estimar la desviación estándar poblacional se usa la expresión 3.14:

$$\sigma \in \left(\sqrt{\frac{(n-1)\hat{S}^2}{\chi^2_{1-\frac{\alpha}{2}}}} \ , \ \sqrt{\frac{(n-1)\hat{S}^2}{\chi^2_{\frac{\alpha}{2}}}} \right)$$

y al sustituir los valores pertinentes se obtiene:

$$\sigma \in \left(\sqrt{\frac{(16-1)(2.2)^2}{32.801}} \ , \ \sqrt{\frac{(16-1)(2.2)^2}{4.601}} \right)$$

Así, entonces,

$$\sigma \in \left(\sqrt{2.213} \ , \ \sqrt{15.779} \right)$$

Por lo tanto,

$$\sigma \in (1.4876 \ , \ 3.9722)$$

En conclusión, con un nivel de confianza del 99 % se infiere que la desviación estándar poblacional en referencia al consumo de combustible está entre 1.4876 y 3.9722 litros. Lo anterior indica que, de cada 100 muestras seleccionadas, 99 tienen el parámetro σ.

3.2.6 Intervalos de confianza para estimar el cociente de varianzas poblacionales

En este apartado se desea obtener un intervalo de la forma *(a,b)* que contenga el parámetro $\frac{\sigma_1^2}{\sigma_2^2}$, con un nivel de confianza de $(1-\alpha)$. Se busca determinar los valores *a* y *b* tales que:

$$P(a \leq \frac{\sigma_1^2}{\sigma_2^2} \leq b) = 1-\alpha$$

Una representación gráfica de esta situación se observa en la Figura 3.18. Es necesario señalar que se trata también de una curva asimétrica para la derecha.

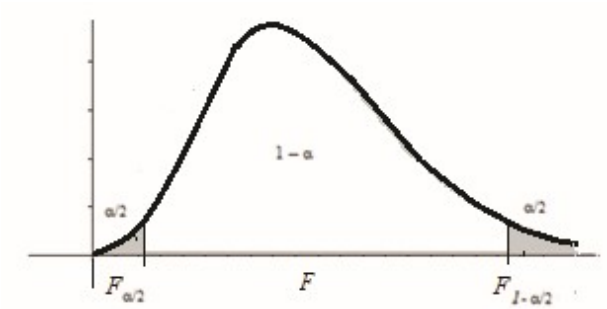

Figura 3.18 Intervalo de confianza para el cociente de varianzas poblacionales
Fuente: los autores con la ayuda del *software* libre R.

En consonancia con la expresión 2.11, la variable *F* tiene distribución de Fischer con n_1-1 grados de libertad para el numerador y n_2-1 grados de libertad para el denominador,

$$F = \frac{\hat{S}_1^2 \sigma_2^2}{\hat{S}_2^2 \sigma_1^2}$$

Luego, en concordancia con la distribución de probabilidad de *Fisher*, se tiene:

$$P(F_{\frac{\alpha}{2}} \leq F \leq F_{1-\frac{\alpha}{2}}) = 1-\alpha$$

De acuerdo con la Figura 3.18, se escribe:

$$P(F_{\frac{\alpha}{2}} \leq \frac{\hat{S}_1^2 \sigma_2^2}{\hat{S}_2^2 \sigma_1^2} \leq F_{1-\frac{\alpha}{2}}) = 1-\alpha$$

Aplicando propiedades de las desigualdades de números reales positivos en cada uno de los miembros de la desigualdad del evento al cual se le calcula la probabilidad, resulta:

$$P\left(\frac{1}{F_{\frac{\alpha}{2}}} \geq \frac{1}{\frac{\hat{S}_1^2\sigma_2^2}{\hat{S}_2^2\sigma_1^2}} \geq \frac{1}{F_{1-\frac{\alpha}{2}}}\right) = 1-\sigma$$

De la anterior expresión se obtiene:

$$P\left(\frac{\hat{S}_1^2}{\hat{S}_2^2 F_{\frac{\alpha}{2}}} \geq \frac{\sigma_1^2}{\sigma_2^2} \geq \frac{\hat{S}_1^2}{\hat{S}_2^2 F_{1-\frac{\alpha}{2}}}\right) = 1-\sigma$$

La desigualdad anterior se expresa de la siguiente forma:

$$P\left(\frac{\hat{S}_1^2}{\hat{S}_2^2 F_{1-\frac{\alpha}{2}}} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{\hat{S}_1^2}{\hat{S}_2^2 F_{\frac{\alpha}{2}}}\right) = 1-\sigma$$

Por lo tanto, el intervalo de confianza para estimar el cociente de varianzas poblacionales está dado por la expresión 3.15.

$$\frac{\sigma_1^2}{\sigma_2^2} \in \left(\frac{\hat{S}_1^2}{\hat{S}_2^2 F_{1-\frac{\alpha}{2}}} \ , \ \frac{\hat{S}_1^2}{\hat{S}_2^2 F_{\frac{\alpha}{2}}}\right) \qquad (3.15)$$

Ejemplo 3.18. Se hizo un estudio para comparar los contenidos de nicotina de dos marcas de cigarrillos. En una muestra aleatoria de 10 cigarrillos de la marca A se encuentra un promedio de 3.5 mg de nicotina, con una desviación estándar corregida de 0.5 mg, mientras que en una muestra aleatoria de 8 cigarrillos de la marca B se obtiene un promedio de 2.9 mg de nicotina, con una desviación estándar de 0.7 mg; se supone que los dos conjuntos de datos provienen de muestras independientes seleccionadas de poblaciones normales. Construir un intervalo de confianza del 98 % para estimar el cociente de varianzas poblacionales.

En este caso se tiene:

Marca A	Marca B
$n_1 = 10$	$n_2 = 8$
$\overline{X}_1 = 3.5$	$\overline{X}_2 = 2.9$
$\hat{S}_1 = 0.5$	$\hat{S}_2 = 0.7483$

El valor de la desviación estándar corregida para la marca B se obtiene de la siguiente manera:

$$\hat{S}_2^2 = \frac{n_2}{n_2 - 1} S_2^2 = \frac{8}{7}(0.7)^2 = 0.56 \rightarrow \hat{S}_2 = 0.7483$$

Además,

$$1 - \alpha = 0.98 \rightarrow \alpha = 0.02$$

$$\frac{\alpha}{2} = 0.01$$

En la Figura 3.19 se observan los valores de los cuantiles obtenidos al leer una tabla *F* de Fisher con n_1-*1*=10-1=9 grados de libertad para el numerador, y n_2-*1*=8-1=7 grados de libertad para el denominador; estos corresponden a un 0.01=1 % y 0.99=99 % de probabilidad.

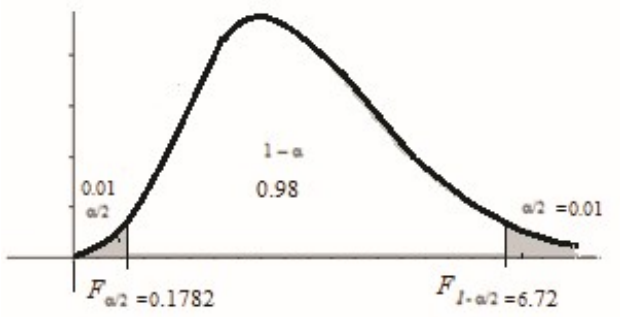

Figura 3.19 Intervalo de confianza sobre la curva de Fisher

Fuente: los autores con la ayuda del *software* libre R.

$$F_{1-\frac{\alpha}{2}} = F_{0.99,9,7} = 6.72$$

$$F_{\frac{\alpha}{2}} = F_{0.01,9,7} = \frac{1}{F_{0.99,7,9}} = \frac{1}{5.61} = 0.1782$$

Luego, usando la expresión 3.15:

$$\frac{\sigma_1^2}{\sigma_2^2} \in \left(\frac{\hat{S}_1^2}{\hat{S}_2^2 F_{1-\frac{\alpha}{2}}} \ , \ \frac{\hat{S}_1^2}{\hat{S}_2^2 F_{\frac{\alpha}{2}}} \right)$$

al reemplazar los respectivos valores resulta:

$$\frac{\sigma_1^2}{\sigma_2^2} \in \left(\frac{(0.5)^2}{(0.7483)^2(6.72)} \ , \ \frac{(0.5)^2}{(0.7483)^2(0.1782)} \right)$$

Efectuando las operaciones de tipo aritmético se obtiene:

$$\frac{\sigma_1^2}{\sigma_2^2} \in \left(\frac{0.25}{3.7628} \ , \ \frac{0.25}{0.09978} \right)$$

Por consiguiente,

$$\frac{\sigma_1^2}{\sigma_2^2} \in (0.0664 \ , \ 2.5055)$$

En conclusión, con un nivel de confianza del 98 % se infiere que el cociente de las varianzas poblacionales en referencia a la cantidad de nicotina de las dos marcas de cigarrillos está entre 0.0664 y 2.5055. Lo anterior indica que, de cada 100 muestras seleccionadas, en 98 se encuentra el parámetro $\frac{\sigma_1^2}{\sigma_2^2}$.

Actividades para el estudio independiente Capítulo 3

3.1 Para la variable aleatoria *X,* con distribución exponencial, determinar el estimador máximo verosímil. Recuerde que la función de densidad de probabilidad está dada por:

$$f(x,\lambda)=\begin{cases}\lambda e^{-\lambda x} & \text{si } x\geq 0 \quad \text{con } \lambda>0\\ 0 & \text{si } x<0\end{cases}$$

3.2 En la empresa azucarera AA, la cantidad de azúcar depositada por la máquina M en cada uno de los paquetes se distribuye normalmente con desviación estándar de 2 g; de un lote de 500 paquetes se toma una muestra aleatoria de 25 paquetes (bolsas) empacados por tal máquina, y se encuentra un contenido promedio de 2500 g. Construir un intervalo de confianza del 98 % para estimar la verdadera media de empacado en las bolsas en el lote que se produzca a través de la máquina M.

3.3 De un lote de 200 unidades del producto LM se ha seleccionado una muestra aleatoria de 10 unidades, obteniéndose un peso neto promedio de 1000 g, con una desviación estándar corregida de 5 g; si se asume que los datos provienen de una distribución normal, determinar un intervalo de confianza del 95 % para estimar el verdadero peso promedio de las unidades del producto en la población.

3.4 En el departamento de Boyacá, Colombia, se tomó una muestra aleatoria de 500 ciudadanos y se les preguntó si pertenecen o no a la población económicamente activa de este departamento; 350 de los encuestados respondieron que sí pertenecen a esta población. Construir un intervalo de confianza del 99 % para estimar la verdadera proporción de ciudadanos que pertenecen a la población económicamente activa de este departamento.

3.5 En un centro de distribución de computadores se ofrecen computadores de dos marcas diferentes en un periodo de tiempo específico; se selecciona aleatoriamente un mes y se encuentra que se venden 350 computadores, de un total de 500, de la marca A, y 333, de un total de 450, de la marca B. Determinar un intervalo de confianza del 98 % para estimar la diferencia entre las verdaderas proporciones de las marcas A y B de computadores que se venden en todo el mercado en ese mes.

3.6 Se analiza el contenido de oro presente en una aleación; en una muestra especial de 40 circuitos integrados se encontró un contenido medio de 5.8 u.i

de oro, con una desviación estándar de 0.6 u.i de oro; asimismo, se inspecciona el contenido de oro en otra muestra aleatoria de 50 circuitos integrados corrientes, detectándose un contenido promedio de 5 u.i, con una desviación estándar de 0.8 u.i; se supone que las muestras provienen de poblaciones normales. Construir un intervalo de confianza del 95 % para estimar la diferencia de contenidos medios de oro de la primera clase de circuito con respecto a la segunda.

3.7 El siguiente ejemplo ha sido adaptado de Canavos (1988). Un determinado procedimiento produce cierta clase de cojinetes de bola cuyo diámetro interior es de 5 cm; se selecciona una muestra aleatoria de 12 de esos cojinetes, y al medir sus diámetros internos se obtiene una desviación estándar corregida 0.03 centímetros. Bajo el supuesto de normalidad para los datos, construir un intervalo de confianza del 99 % para estimar la varianza poblacional; asimismo, determinar un intervalo de confianza para estimar la desviación estándar poblacional.

3.8. Se tiene la creencia de que los egresados de la titulación de Administración de Empresas obtienen un salario promedio mayor que el de los egresados de la titulación de Economía; además, se quiere saber si la variación de sus correspondientes salarios difiere. Para comprobarlo se ha tomado una muestra aleatoria de 10 administradores, obteniéndose una media muestral de 2 600 000 pesos por mes, con una varianza corregida de 1 200 000 pesos, mientras que en una muestra aleatoria de 13 economistas se ha obtenido un promedio de 2 400 000 pesos por mes, con una varianza de 1 300 000; se supone que los dos conjuntos de datos provienen de muestras independientes seleccionadas de poblaciones normales. Construir un intervalo de confianza del 98 % para estimar el cociente de varianzas poblacionales.

Ejercicios para el capítulo 3

3.1 Indagar sobre la forma como se construye el intervalo de confianza para muestras pareadas, también denominadas muestras relacionadas o emparejadas; además, proporcionar un ejemplo de aplicación.

3.2 Se supone que la cantidad de cemento que una máquina empaca en cada bulto es una variable aleatoria con desviación típica poblacional de 1 kg. Se toma una muestra aleatoria de tamaño 20 y se obtiene una media de 50 kg. Obtener un intervalo de confianza del 95 % para estimar la media poblacional.

3.3 Un jugador de fútbol anota 120 goles en 500 lanzamientos (cobros) desde el punto penal, en los entrenamientos. Construir un intervalo de confianza del 99 % para estimar la verdadera proporción de la efectividad del jugador.

3.4 Una determinada clase de estufa de gas se somete a 11 pruebas para evaluar su consumo. En esa muestra se obtiene una desviación estándar corregida de 1.8 litros. Construir un intervalo de confianza del 95 % para estimar la varianza poblacional como indicador de la verdadera variación del consumo de gas; asimismo, construir un intervalo de confianza para la desviación estándar poblacional.

3.5 En un centro de distribución de monitores para computador de mesa se distribuyen monitores de dos marcas diferentes en un periodo de tiempo determinado; en una semana se venden 200 monitores, de un total de 350, de la marca A, y 180, de un total de 300, de la marca B. Hallar un intervalo de confianza del 96 % para estimar la verdadera diferencia entre las proporciones de las marcas A y B que se venden en todo el mercado.

3.6 En una muestra de 20 unidades de un producto se obtuvo un peso neto promedio de 250 g, con una desviación estándar corregida de 5. Encontrar un intervalo de confianza del 98 % para estimar el verdadero peso promedio de las unidades del producto.