Unidad 9

- Estimación.

Capítulo 8

ESTIMACIÓN

En los capítulos anteriores se han estudiado las nociones fundamentales de distribución de probabilidad y distribución muestral. Estamos ya en condiciones de tratar los métodos de inferencia estadística, los cuales comprenden los procedimientos para estimar parámetros de poblaciones y probar (contrastar) si una afirmación provisional sobre un parámetro poblacional se ve apoyada o desaprobada ante la evidencia de la muestra.

Hablando en general, hay dos tipos de inferencia: la deductiva y la inductiva. Una *inferencia deductiva* es un juicio o generalización que se basa en *axiomas, premisas* o *hipótesis*, de las cuales se deriva una *conclusión* mediante un razonamiento o proceso dialéctico a priori. Por ejemplo, si se supone que dos monedas están perfectamente equilibradas y que entonces la probabilidad de cada una de caer "cara" es 0.5 (premisa), la media o número esperado de "caras" en la jugada de las monedas debe ser 1 (conclusión). Si las premisas son ciertas, las conclusiones no pueden ser falsas.

Una *inferencia inductiva*, por otra parte, es un juicio o generalización derivado de observaciones empíricas o experimentales; la conclusión se obtiene mediante un razonamiento a posteriori. En el caso de la jugada de las monedas, por ejemplo, una inferencia inductiva es una conclusión sobre el número promedio de "caras" con base en los resultados de una muestra de prueba. Si los resultados de las pruebas son diferentes, la conclusión también será diferente. No se requiere una suposición a priori sobre la naturaleza de las monedas. La inferencia estadística es primordialmente de naturaleza inductiva y llega a generalizaciones respecto de las características de una población al valerse de observaciones empíricas de la muestra.

Es muy probable que una estadística muestral sea diferente del parámetro de la población y sólo por coincidencia sería el uno exactamente igual al otro. La diferencia entre el valor de una estadística muestral y el correspondiente parámetro de la población se suele llamar *error de estimación*. Sólo se sabría cuál es el error si se conociera el parámetro poblacional, pero éste por lo general se desconoce. La única

manera de tener alguna certeza al respecto es hacer todas las observaciones posibles del total de la población en la mayoría de las aplicaciones prácticas, lo cual, desde luego, es imposible o impracticable.

Y en efecto, la razón de ser de la inferencia estadística es la falta de conocimientos acerca de las características de la población. Pero que tales características se desconozcan no impide el que se actúe. Los hombres de ciencia, los empresarios, los políticos, los militares y las personas de toda condición deciden sobre las características de la población, características que a menudo se llaman "estados de la naturaleza", sin estar del todo ciertos de qué cosa son realmente. Lo que la mayoría hace, si bien inconscientemente, es "jugar con ventajas" que estiman lo mejor que pueden. Por ejemplo, todos conducen en las carreteras como si estuvieran completamente seguros de que no pueden morir en accidentes puesto que se sienten muy seguros de que las ventajas están abrumadoramente a su favor. Las estadísticas de la mortalidad en las carreteras son buenos indicadores de las posibilidades o probabilidades. A veces, unas cuantas personas hacen mal en confiar en la probabilidad que tienen de no ser víctimas de accidentes de tráfico. Pero esto se debe sobre todo a que tales accidentes no ocurren con frecuencia, para que sigan conduciendo con confianza en las carreteras.

De forma parecida, las inferencias estadísticas se hacen por posibilidades o probabilidades. De la media de una muestra se hacen inferencias sobre la media de la población. No se sabe exactamente cuál es la diferencia entre estas dos medias, ya que la última es desconocida en la mayoría de los casos. No obstante, sí se sabe que es más bien poca la probabilidad de que esta diferencia sea mayor que, por ejemplo, tres o aun dos errores estándares. A veces se yerra en actuar sobre esta probabilidad, y ese es el precio que debe pagarse por no llevar a cabo la tarea, a menudo imposible, de investigar toda la población en cuestión.

Los problemas que se tratan en la inferencia estadística se dividen generalmente en dos clases: los problemas de estimación y los de prueba de hipótesis. Como al estimar un parámetro poblacional desconocido se suele hacer una afirmación o juicio, este último ofrece solamente una *estimación*. Es un valor particular obtenido de observaciones de la muestra. No hay que confundir este concepto con el de *estimador*, que se refiere a la *regla o método* de estimar un parámetro poblacional. Por ejemplo, se dice que $\overline{X}$ es un estimador de μ porque la media muestral proporciona un método para estimar la media de la población. Un estimador es por naturaleza una estadística y como tal tiene una distribución. El procedimiento mediante el cual se llega a la obtención y se analizan los estimadores se llama *estimación estadística*, que a su vez se divide en *estimación puntual* y *estimación por intervalos*. En este capítulo se van a considerar ambas modalidades de estimación; en el capítulo 9 se tratará lo relacionado con pruebas de hipótesis estadísticas.

8.1 ESTIMACIÓN PUNTUAL

Si a partir de las observaciones de una muestra se calcula un solo valor como estimación de un parámetro de la población desconocido, el procedimiento se denomina *estimación puntual*, ya que se utiliza como estimación un solo punto del conjunto de todos los posibles valores. Supongamos, por ejemplo, que se desee estimar μ, la inteligencia promedio de todos los estudiantes universitarios del país. Sea X la variable aleatoria que indica el IQ de cada estudiante. Se toma una muestra aleatoria de n estudiantes y se denota $\overline{X}$ el IQ medio de la muestra. Si al tomar una muestra de 100 estudiantes se obtiene que el IQ promedio es 115 y este número se toma como un estimativo de

μ (IQ de toda la población universitaria del país), entonces decimos que 115 es una estimación puntual de μ. En general se tiene que:

Un estimador puntual T de un parámetro θ es cualquier estadística que nos permita a partir de los datos muestrales obtener valores aproximados del parámetro θ. Para indicar que T es un estimador del parámetro θ escribimos $\hat{\theta} = T$. Es decir, con una escritura tal queremos decir que estamos empleando o vamos a emplear la expresión o fórmula dada mediante T para obtener valores próximos al valor del parámetro θ.

Así por ejemplo, si decimos que $\overline{X}$ es un estimador de la media poblacional μ, entonces escribimos $\hat{\mu} = \overline{X}$, y con ello queremos indicar que se va a tomar la media aritmética de los datos como valor próximo a la media poblacional μ.

Para comprender mejor el proceso de estimación, es usual emplear una analogía como el acto de hacer disparos al blanco. En este símil el revólver corresponde al estimador; un impacto de la bala, una estimación y el centro del blanco corresponde al parámetro.

Además de tener $\overline{X}$ como estimador de la media poblacional μ se tiene a $S^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2 / (n - 1)$ como estimador de la varianza σ^2 y a $\overline{p}$ (proporción muestral) como estimador de la proporción poblacional π. Como ya se ha dicho antes, es muy probable que haya error cuando un parámetro es estimado, puesto que la muestra no es más que una parte de un conjunto mucho más grande dentro de las observaciones posibles. Por lo mismo es aventurado afirmar que el valor correspondiente a la población sea el mismo calculado por la muestra. Es cierto que si el número de observaciones al azar se hace suficientemente grande, éstas proporcionarían una media que casi sería semejante al parámetro; pero a menudo hay limitaciones de tiempo y de recursos y hay que decidir al disponer de sólo unas cuantas observaciones. Para poder utilizar la información que se tenga de la mejor forma posible, se necesita identificar las estadísticas que sean "buenos" estimadores. Hay cuatro criterios que se suelen aplicar para determinar si una estadística es un buen estimador: *Insesgamiento, eficiencia, consistencia* y *suficiencia.*

8.2 PROPIEDADES DE UN ESTIMADOR

Adicional a las propiedades que ya mencionamos para "un buen estimador", existe otra que en cierta forma comprende conjuntamente las propiedades de insesgamiento y eficiencia. Se trata del *error cuadrático medio.*

Sea T *un estimador del parámetro θ (desconocido). El error cuadrático medio de* T, *denotado ECM(T), se define como el valor esperado de $(T - \theta)^2$.* Esto es,

$$ECM\ (T) = E\ [(T - \theta)^2] \tag{8-1}$$

¿Cuál es la información que nos proporciona el error cuadrático medio?

Comencemos señalando el significado del valor esperado. Como ya fue explicado, el valor esperado de una variable es el número al cual tiende el promedio de las observaciones cuando repetimos el experimento indefinidamente y así estará relacionado con los valores que se darán con mayor frecuencia (probabilidad). Por tanto, al tomar $E\ [(T - \theta)^2]$ nos estamos refiriendo al promedio de los cuadrados de las observaciones. Si éste es un número pequeño, debemos aceptar que hay una tendencia para que los valores $(T - \theta)^2$ sean pequeños, y así lo será también la diferencia

$(T - \theta)$, lo que se traduce en que T tiende a producir respuestas (numéricas) próximas al parámetro θ. Esto es muy bueno, puesto que de lo que se trata es de "determinar" el valor de θ (desconocido por nosotros) mediante la fórmula T que sí se la podemos aplicar a los datos muestrales.

El "poder" que tenga T para "producir" valores próximos a θ depende de dos condiciones básicas. Una es la "fuerza" o intensidad con que tiende a dar esos valores (insesgamiento) y la otra es la "fuerza" que tenga para no permitir que se aparte del camino que lo conduce a θ (eficiencia). Estas dos condiciones matemáticamente quedan establecidas y precisadas en el teorema siguiente:

Teorema 8.1 Si T es un estimador del parámetro θ
$$\Longrightarrow ECM(T) = E[X^2] - [\theta - E(T)]^2 \tag{8-2}$$

La demostración de este teorema no presenta dificultad y es como sigue:

$$\begin{aligned} ECM(T) &= E[(T - \theta)^2] = E[T^2 - 2\theta T + \theta^2] = E(T^2) - E(2\theta T) + E(\theta^2) = \\ &= E(T^2) - 2\theta E(T) + E(\theta^2) = E(T^2) - [E(T)]^2 + [E(T)]^2 - 2\theta E(T) + \theta^2 = \\ &= (E(T^2) - [E(T)]^2) + ([E(T)]^2 - 2\theta E(T) + \theta^2) = V(T) + [\theta - E(T)]^2 \end{aligned}$$

Luego, $ECM(T) = V(T) + [\theta - E(T)]^2$

De (8-2) se tiene que el error cuadrático medio de T será pequeño en la medida en que su varianza también sea pequeña y lo mismo ocurra con $[\theta - E(T)]^2$, es decir, $\theta - E(T)$.

Un valor pequeño de $V(T)$ significa que T presenta poca variabilidad; y $\theta - E(T)$ pequeño quiere decir que $E(T)$ tiende al valor θ a medida que el experimento se repite, que por estar relacionado este hecho con los valores de mayor frecuencia, indicaría que T tiende a dar valores próximos a θ.

La diferencia $\theta - E(T)$ se llama sesgo del estimador.

Aunque el error cuadrático medio recoge dos ideas básicas que debe tener un buen estimador, puede resultar impreciso para tomar decisiones en algunos casos. En el ejercicio que desarrollaremos a continuación podremos comprobar lo que acabamos de comentar.

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria proveniente de una población de media desconocida μ y varianza $\sigma^2 = 81$. Considere a $T_1 = \overline{X}$ y $T_2 = \sum_{i=1}^{n} X_i/(n+1)$ como estimadores de μ, obtenga $ECM(T_1)$ y $ECM(T_2)$.

Utilizaremos el resultado del teorema 8.1 para obtener el error cuadrático medio

Para T_1 se tiene $ECM(T_1) = V(T_1) - [\mu - E(T_1)]^2$

$V(T_1) = V(\overline{X}) = \dfrac{\sigma^2}{n}$, $E(T_1) = E(\overline{X}) = \mu$, así que $ECM(T_1) = ECM(\overline{X}) = \dfrac{\sigma^2}{n} +$

$$+ [\mu - \mu]^2 = \frac{\sigma^2}{n} = \frac{81}{n}$$

Para T_2 se tiene $ECM(T_2) = V(T_2) + [\mu - E(T_2)]^2$

$$V(T_2) = V\left[\frac{1}{(n+1)}\sum_{i=1}^{n} X_i\right] = \frac{1}{(n+1)^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{(n+1)^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{(n+1)^2}(n\sigma^2) =$$

$$= \frac{n}{(n+1)^2}\sigma^2 = \frac{81n}{(n+1)^2}$$

$$E(T_2) = E[\frac{1}{(n+1)}\sum_{i=1}^{n} X_i] = \frac{1}{(n+1)}\sum_{i=1}^{n} E[X_i] = \frac{1}{(n+1)}\sum_{i=1}^{n}\mu = \frac{1}{(n+1)}(n\mu) = \frac{1}{(n+1)}\mu$$

$$ECM(T_2) = \frac{81n}{(n+1)^2} + [\mu - \frac{n\mu}{(n+1)}]^2 = \frac{81n}{(n+1)^2} + [\frac{\mu}{n+1}]^2 =$$

$$= \frac{81n}{(n+1)^2} + \frac{\mu^2}{(n+1)^2} = \frac{81n + \mu^2}{(n+1)^2}$$

En resumen, $ECM(T_2) = \dfrac{81n + \mu^2}{(n+1)^2}$

Supongamos que debemos escoger uno de los dos estimadores.
Como el único medio con que contamos para hacer la selección es el error cuadrático medio entonces, si se tiene en cuenta lo que éste representa, debemos tomar el estimador que tenga el menor error cuadrático medio. Por tanto, la solución al problema se encuentra en la respuesta del siguiente interrogante: ¿cuál de los dos estimadores tiene menor error?
Como en las dos expresiones se presenta el valor n, para precisar ideas, tomemos $n = 9$ y se obtiene así:

$$ECM(T_1) = \frac{81}{9} = 9, \cdot ECM(T_2) = \frac{729 + \mu^2}{100}$$

Como en la fórmula $ECM(T_2)$ se incluye μ, entonces $ECM(T_2)$ depende de μ y tenemos así que si

$$\mu < \sqrt{171} \Longrightarrow ECM(T_2) = \frac{729 + \mu^2}{100} < \frac{729 + 171}{100} = \frac{900}{100} = 9$$

y así $ECM(T_2) < ECM(T_1)$ y de esta forma debemos escoger T_2 como estimador de μ.

Por otra parte, si $\mu > \sqrt{171} \Longrightarrow ECM(T_2) = \frac{729 + \mu^2}{100} > \frac{729 + 171}{100} = \frac{900}{100} = 9$

y deberíamos escoger T_1.

Este sencillo ejercicio nos pone de presente lo que ya habíamos señalado acerca de lo que puede ocurrir con el error cuadrático medio. Observe que lo que pretendemos es contar con criterios que garanticen una buena selección del estimador, sin importar el valor particular que tenga el parámetro objeto de estudio.

Con el propósito de precisar esos criterios comenzaremos por considerar el error cuadrático medio en sus partes y así iniciamos el estudio de la diferencia $\theta - E(T)$.

Se dice que una estadística T es un estimador insesgado de θ, si se cumple que $E[T] = \theta$ para cualquier valor de θ.

De acuerdo con lo anterior y al tener en cuenta que $E(\bar{X}) = \mu$, entonces $\bar{X}$ es un estimador insesgado de μ.

En cambio $T_2 = \dfrac{1}{(n+1)}\sum_{i=1}^{n} X_i$ y al tener en cuenta que $E(T_2) = \dfrac{n\mu}{n+1}$ entonces T_2 no es un estimador insesgado de μ. En este caso decimos que T_2 es un estimador *sesgado* de μ.

También podemos decir que un estimador insesgado es aquel que tiene un sesgo igual a cero.

La proporción muestral $\bar{p}$ también es un estimador insesgado de la proporción poblacional π.

¿Es $S^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 / (n - 1)$ un estimador insesgado de la varianza poblacional σ^2?

Si S^2 fuera un estimador insesgado debería cumplirse que $E(S^2) = \sigma^2$. Veamos si ello ocurre.

Recordemos que en la sección 7.5 se estableció la identidad,

$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2 \Longrightarrow$ al dividir ambos miembros por $(n - 1)$ se tiene $\dfrac{\sum_{i=1}^{n}(X_i - \mu)^2}{n-1} = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} + \dfrac{n(\bar{X} - \mu)^2}{n-1} \Longrightarrow$ Al tomar el valor esperado a ambos miembros de la igualdad tenemos, $\dfrac{1}{(n-1)} \sum_{i=1}^{n} E[(X_i - \mu)^2 = E[S^2] + \dfrac{n}{n-1} E[(\bar{X} - \mu)^2] \Longrightarrow \dfrac{1}{n-1}(n\sigma^2) = E[S^2] + \dfrac{n}{n-1}(\sigma^2/n) \Longrightarrow \dfrac{n\sigma^2}{n-1} = E[S^2] + \dfrac{\sigma^2}{n-1} \Longrightarrow$ $E[S^2] = \dfrac{n\sigma^2}{n-1} - \dfrac{\sigma^2}{n-1} = \sigma^2$ y así $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ es un estimador insesgado de σ^2.[1]

Los dos resultados anteriores quedan resumidos en el siguiente teorema:

Teorema 8.2 Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de cierta distribución de media μ y varianza σ^2. Entonces, *i)* $T_1 = \bar{X}$ es un estimador insesgado de μ, *ii)* $T_2 = S^2$ es un estimador insesgado de σ^2.

Al tener presente el significado del valor esperado podemos explicarnos por qué la propiedad del insesgamiento es una buena condición para un estimador, ya que por medio de esta condición aseguramos que las estimaciones que hagamos mediante dicho estimador se encuentran alrededor de, o que tienen como centro el parámetro en cuestión y establecemos así la siguiente regla de procedimiento:

Regla 1. Si tiene un estimador T_1 y un estimador T_2 del parámetro θ y uno de ellos es insesgado, entonces escoja el insesgado.

En virtud de esta regla y por lo ya visto entre los dos estimadores $T_1 = \bar{X}$, $T_2 = \dfrac{1}{n+1}\sum_{i=1}^{n} X^2$ propuestos para μ, escogemos $T_1 = \bar{X}$.

Volviendo al símil del revólver para un estimador, uno insesgado sería el caso para el revólver cuyos disparos lograsen impactos alrededor del centro del blanco. Sin embargo, también podría darse el caso de un revólver cuyos disparos hiciesen impactos alrededor de un punto que no es el centro del blanco; este sería el caso de un estimador sesgado. En las figuras 8.1 y 8.2 se ilustra gráficamente el caso de un estimador insesgado (cuadro a la izquierda) y un estimador sesgado (cuadro a la derecha).

También puede ocurrir que ambos revólveres hagan impactos alrededor del centro del blanco, pero uno de ellos puede lograr impactos más concentrados alrededor del centro del blanco que el otro. *Véanse* figuras 8.3 y 8.4

[1] En el suplemento VII se demuestra que si se toma la varianza muestral como $\sum_{i=1}^{n}(X_i - \bar{X})^2/n$ entonces resulta un estimador sesgado. En el suplemento VIII se da una comprobación numérica del resultado.

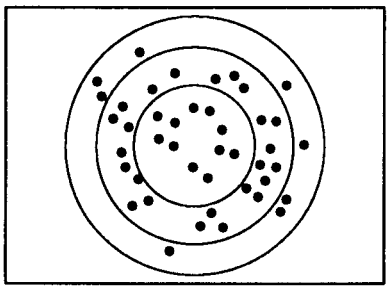

Figura 8.1 *Estimador insesgado.*

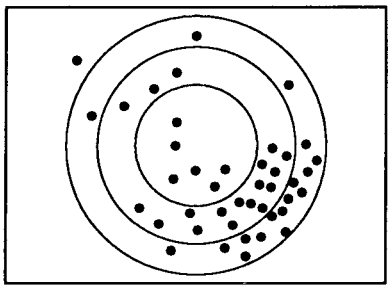

Figura 8.2 *Estimador sesgado.*

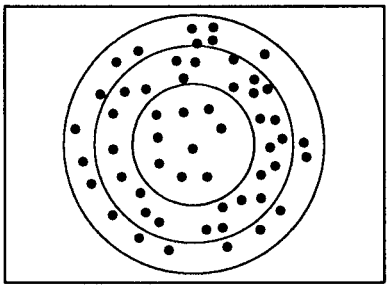

Figura 8.3 *Estimador concentrado.*

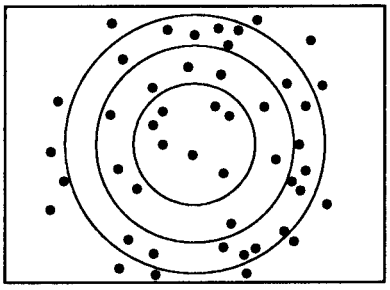

Figura 8.4 *Estimador disperso.*

Es claro que el revólver cuyo comportamiento se señala en el cuadro de la figura 8.3 sería preferible al de la figura 8.4. En esta consideración radica la noción de eficiencia.

Dados dos estimadores T_1 y T_2 del parámetro θ, ambos insesgados, decimos que T_1 es más eficiente que el estimador T_2 si la varianza de T_1 es menor que la varianza de T_2. Esto es, $V(T_1) < V(T_2)$.
La segunda regla de procedimiento que debemos tener presente es:

Regla 2. Si usted tiene dos estimadores T_1 y T_2 del parámetro θ, ambos insesgados, escoja el de menor varianza.

Las dos propiedades anteriores son las que podríamos llamar las de "primera mano" en cuanto a las buenas propiedades del estimador. Adicional a éstas se toman en cuenta otras características de los estimadores, pero cuyo significado intuitivo es más difuso que las que hemos citado; nos referimos a las propiedades de consistencia y suficiencia.

La *consistencia* se refiere al comportamiento de un estimador, particularmente cuando es sesgado a medida que la muestra se va tomando de un tamaño mayor y tenemos así:

Sea T un estimador del parámetro θ, decimos que T es un estimador consistente para θ, si se cumple que $P[|T - \theta| \leq \epsilon] \longrightarrow 1$, cuando $n \longrightarrow \infty$.

Con base en la anterior definición, el estimador consistente es aquel para el cual a medida que aumenta el tamaño de la muestra, la probabilidad de que se acerque al parámetro va siendo mayor.

En general no es tarea fácil demostrar que un estimador es consistente, puesto que su demostración incluye el cálculo de un límite y esta tarea puede requerir de fundamentos matemáticos de alguna consideración.

Como última característica de los buenos estimadores tenemos la *suficiencia.*

Un estimador T del parámetro θ, se dice suficiente cuando es capaz de sustraer de la muestra toda la información que ésta contenga acerca del parámetro.

Una explicación no sobra para precisar mejor la idea de suficiencia. Cuando se registran ciertos datos numéricos, cada valor observado lleva en sí parte de la media y de la variabilidad de la población de donde procede; pues bien, un estimador suficiente para la media es aquel que tiene la capacidad de sustraer de esos números (datos) todo lo que éstos tengan acerca de la media de la población de donde provienen. En este orden de ideas, un estimador de la media que sólo tome para promediar la primera y última observación no sería un estimador suficiente. Sea oportuna aquí la observación según la cual los estimadores de mayor uso como la media muestral, la varianza muestral y la proporción muestral son buenos estimadores.

Hasta ahora hemos presentado el concepto de estimador y reseñado las propiedades fundamentales para que un estimador sea considerado "bueno". Pero es posible que ya nos estemos haciendo la pregunta: ¿cómo se obtiene un estimador? En realidad se siguen varios métodos para obtener estimadores, unos más adecuados que otros en determinadas circunstancias. Los métodos más comunes son el de máxima verosimilitud, método de los momentos muestrales y método de los mínimos cuadrados.

En este texto explicaremos el *método de la máxima verosimilitud,* ya que es el procedimiento que proporciona por lo general buenos estimadores.

8.3 ESTIMACIÓN DE MÁXIMA VEROSIMILITUD

La estimación de máxima verosimilitud consiste en considerar todos los valores posibles del parámetro de la población y calcular la probabilidad de que se obtenga ese estimador particular, dados todos los valores posibles del parámetro. En el siguiente ejercicio se explica el fundamento de tal método:

Suponga que desea conocer la proporción de estudiantes de cierta universidad que están a favor de la propuesta de un cambio de sede de la universidad.

Suponga asimismo que se escogieron aleatoriamente 10 estudiantes y que 4 de ellos respondieron "sí". Esto es, $n = 10$, $x = 4$, en donde, $X =$ número de estudiantes que respondieron "sí" (del total de los 10).

Ahora vamos a determinar la probabilidad de obtener cuatro respuestas "sí" de acuerdo con la proporción verdadera que puede darse en la población universitaria considerada. Para una mayor sencillez en el desarrollo del ejercicio, admitimos que la población estudiantil del caso es tan grande que la proporción π de los que están a favor no se ve alterada si cada estudiante entrevistado queda descartado para una posterior entrevista sobre la citada encuesta. Además, que cada uno de ellos en su respuesta no influye ni está influenciado por la de cualquier otro estudiante. De esta manera, hemos creado las condiciones para considerar una distribución binomial para la variable que hemos definido y tenemos así,

$$P[X = x] = \binom{10}{x} \pi^x (1 - \pi)^{10 - x}; \qquad x = 0, 1, 2, \ldots, 10$$

Si tomamos $x = 4 \Longrightarrow P[X = 4] = \binom{10}{4} \pi^4 (1 - \pi)^6 = 210\pi^4 (1 - \pi)^6$

Ahora calcularemos las probabilidades respectivas para distintos valores de π.

Si $\pi = 0.1 \Longrightarrow P[X = 4] = 210(0.1)^4(0.9)^6 = 0.0112$
Si $\pi = 0.2 \Longrightarrow P[X = 4] = 210(0.2)^4(0.8)^6 = 0.0881$

Al continuar de esta manera obtenemos la tabla 8.1.

Tabla 8.1 $P[X = 4] = 210\pi^4(1 - \pi)^6$ *para distintos valores de* π.

Proporción de población ***p***	**Probabilidad** **P (*X* = 4)**
0.0	0.0000
0.1	0.0112
0.2	0.0881
0.3	0.2001
0.4	0.2508
0.5	0.2051
0.6	0.1115
0.7	0.0368
0.8	0.0055
0.9	0.0001
1.0	0.0000

En la tabla podemos notar que la máxima probabilidad se da para $\pi = 0.4$ que coincide precisamente con la proporción muestral $\frac{4}{10} = 0.4$. Esto es, el valor de la proporción muestral hace máxima a la función,

$$L(\pi) = 210\pi^4(1 - \pi)^6$$

También puede decirse que es más probable obtener una proporción muestral de 0.4 cuando la verdadera proporción es de 0.4 que si tiene otro valor.

Naturalmente, la proporción verdadera π puede ser distinta al valor muestral obtenido. Es decir, cuando entrevistemos a los estudiantes puede ocurrir que solamente tres respondan "sí"; y al seguir la metodología expuesta, el valor estimado para π en ese caso sería $\hat{\pi} = 0.3$, aunque el verdadero sea 0.4. Pero este es el riesgo al que nos vemos abocados al proceder por muestreo solamente y al no investigar toda la población. Más adelante en este mismo capítulo indicaremos cómo puede disminuirse la posibilidad de que tomemos decisiones erradas.

También es obvio que son posibles otros valores de π distintos a los que hemos indicado en la tabla 8.1 y de esta manera sería más preciso considerar la función $L(\pi)$ como una función continua definida en el intervalo $[0, 1]$. Esta función la llamaremos *función de verosimilitud* y formalmente queda definida como,

$$L(\pi) = P[X = 4 | \pi = p] = 210\pi^4(1 - \pi)^6; \qquad 0 \leq \pi \leq 1$$

De esta forma $L(\pi)$ es la probabilidad condicional de $X = 4$ dado $\pi = p$ y su gráfica es como se señala en la figura 8.5.

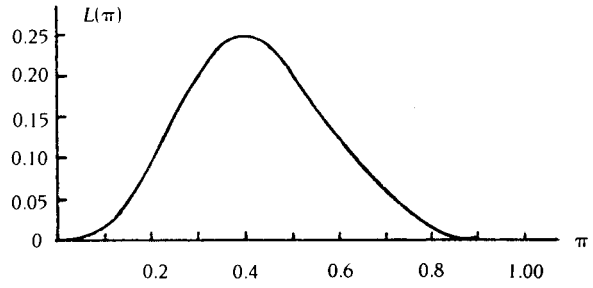

Figura 8.5 *Función de verosimilitud* $L(\pi) = P[X = 4 | \pi = p]$ *para* $n = 10$.

Aunque en la gráfica podemos apreciar que el máximo se halla en $\pi = 0.4$. Este resultado se puede obtener al recurrir a la herramienta matemática conocida como *criterio de la derivada para obtención de valores extremos* que se estudia en los cursos elementales de cálculo. Al aplicar este procedimiento a la citada función de verosimilitud, tenemos:

Según el criterio de la primera derivada, $L'(\pi) = 210\,[4\pi^3(1 - \pi)^6 - 6\pi^4(1 - \pi)^5] = 0 \Longrightarrow 210\pi^3(1 - \pi)^5\,[4(1 - \pi) - 6\pi] = 0 \Longrightarrow 4(1 - \pi) - 6\pi = 0 \Longrightarrow 4 - 4\pi - 6\pi = 0 \Longrightarrow 4 - 10\pi = 0 \Longrightarrow \pi = \frac{4}{10} = 0.4.$

Aplicamos el criterio de la segunda derivada y tenemos, $L''(\pi) = 630\pi^2(1 - \pi)^4\,[4 - 24\pi + 30\pi^2] \Longrightarrow L''(0.40) = 630\,(0.4)^2\,(0.6)^4\,[4 - 24\,(0.4) + 30\,(0.4)^2] = -10.45$ y así en $\pi = 0.4$ ocurre un máximo como ya lo habíamos previsto.

Las ideas que hemos expuesto hasta aquí acerca de la estimación de máxima verosimilitud pueden ser formalizadas para adecuarlas mejor a una concepción teórica. Es lo que nos proponemos desarrollar a continuación.

Sea X una variable aleatoria con función de densidad $f(x;\ \theta)$, determinada por el parámetro θ. Suponga que de la población se extrae una muestra de tamaño n que proporciona los datos $x_1, x_2, \ldots, x_n$. Con estos datos formamos el producto $f(x_1;\ \theta)\ f(x_2;\ \theta)\ \ldots\ f(x_n;\ \theta)$. Este producto se llama *función de verosimilitud*. Esto es, la función de verosimilitud de una muestra está dada por,

$$L(\theta) = f(x_1;\ \theta)\ f(x_2;\ \theta)\ \ldots\ f(x_n;\ \theta)$$

Una *estimación de máxima verosimilitud* para el parámetro θ, es aquel valor de θ donde la función de verosimilitud asuma su máximo.
La expresión de θ en términos de la muestra aleatoria $X_1, X_2, \ldots, X_n$ se llama *estimador de máxima verosimilitud de* θ.

Dos ejemplos serán suficientes para que apreciemos el proceso para obtener estimadores de máxima verosimilitud.

El primero que analizaremos será para el caso de una distribución de Poisson. Recordemos que se dice que una variable tiene distribución de Poisson cuando su función de densidad (valores de probabilidad) está dada por,

$$f(x;\ \lambda) = P\ [X = x] = \frac{e^{-\lambda}\lambda^{x}}{x!} \quad ;\ x = 0,\ 1,\ 2,\ 3,\ \ldots$$

Para hacer una estimación de máxima verosimilitud para λ, comenzamos por considerar que se tienen n observaciones $x_1, x_2, \ldots, x_n$. Con estas observaciones formamos la función de verosimilitud,

$$L\ (\lambda) = f(x_1;\ \lambda)\ f(x_2;\ \lambda)\ f(x_3;\ \lambda)\ \ldots\ f(x_n;\ \lambda) = \frac{e^{-\lambda}\lambda^{x_1}}{(x_1)!}\frac{e^{-\lambda}\lambda^{x_2}}{(x_2)!}\frac{e^{-\lambda}\lambda^{x_3}}{(x_3)!}\ \ldots\ \frac{e^{-\lambda}\lambda^{x_n}}{(x_n)!}$$

$$= \frac{e^{-n\lambda}\lambda^{\Sigma x_i}}{(x_1)!\ (x_2)!\ (x_3)!\ \ldots\ (x_n)!}$$. Esto es, la función de verosimilitud nos queda:

$$L(\lambda) = \frac{e^{-n\lambda}\lambda^{\Sigma x_i}}{(x_1)!\ (x_2)!\ (x_3)!\ \ldots\ (x_n)!}$$

Ahora nos proponemos maximizar $L(\lambda)$. En la práctica, la maximización de la función de verosimilitud requiere de algunos cálculos más o menos complejos y por ello lo usual es maximizar el logaritmo natural de la función de verosimilitud. Hecho esto, tenemos,

$$g(\lambda) = Ln\ [L(\lambda)] = -n\lambda + (\Sigma x_i)\ Ln(\lambda) - Ln\ [(x_1)!\ (x_2)!\ \ldots\ (x_n)!]$$

Al tomar la primera derivada de $g(\lambda)$ tenemos,

$$g'(\lambda) = -n + \frac{\Sigma x_i}{\lambda} = 0 \Longrightarrow \frac{-n\lambda + \Sigma x_i}{\lambda} = 0 \Longrightarrow -n\lambda + \Sigma x_i = 0 \Longrightarrow$$

$$\hat{\lambda} = \frac{\Sigma x_i}{n} = \bar{x} \text{ es el punto crítico}$$

Al tomar la segunda derivada, obtenemos $g''(\lambda) = \frac{-\Sigma x_i}{n}$, que por ser $x_1, x_2, \ldots, x_n$ números enteros no negativos $\Longrightarrow g''(\bar{x}) < 0$. Y así en $\hat{\lambda} = \bar{x}$ la función $L(\lambda)$ toma un máximo. Por tanto, $\hat{\lambda} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$ es el estimador de máxima verosimilitud de λ.

El segundo caso que tomamos en cuenta es el de una población normal con varianza conocida $\sigma^2 = 4$.

La función de densidad en este caso está dada por

$$f(x;\ \mu) = \frac{1}{(\sqrt{2\pi})\ (2)}\ e^{-(1/2)(x\ -\ \mu)^2/4}$$

La función de verosimilitud está dada por

$$L(\lambda) = \frac{1}{(\sqrt{2\pi})\ (2)} e^{-(1/2)(x_1\ -\ \mu)^2/4} \frac{1}{(\sqrt{2\pi})\ (2)}\ e^{-(1/2)(x_2\ -\ \mu)^2/4}\ \ldots \frac{1}{(\sqrt{2\pi})\ (2)} e^{-(1/2)(x_n\ -\ \mu)^2/4}$$

$$= \frac{1}{(2\pi)^{n/2}2^n} e^{-(1/2)\Sigma(x_i - \mu)^2/4} \Longrightarrow L(\mu) = \frac{1}{(2\pi)^{n/2}2^n} e^{-(1/2)\Sigma(x_i - \mu)^2/4}$$

Al tomar logaritmo se recibe,

$$g(\mu) = -(1/2)\ \Sigma(x_i - \mu)^2/4 - (n/2)\,Ln\,(2\pi) - nLn\,(2)$$

Al derivar e igualar a cero se tiene, $g'(\mu) = -(1/2)\ \Sigma(x_i - \mu)(-2) = 0 \Longrightarrow$

$$\Sigma(x_i - \mu) = 0 \Longrightarrow \Sigma x_i - \Sigma\mu = 0 \Longrightarrow$$

$$\Sigma x_i - n\mu = 0 \Longrightarrow n\mu = \Sigma x_i \Longrightarrow \hat{\mu} = \frac{\Sigma x_i}{n} = \bar{x} \text{ es el punto crítico.}$$

Pasamos ahora a la segunda derivada, $g''(\mu) = \Sigma(-1) = -n$, lo cual nos indica que la segunda derivada es negativa y así en $\hat{\mu} = \bar{x}$ ocurre un máximo, y por tanto, $\hat{\mu} = \bar{X}$ es el estimador de máxima verosimilitud de la media de una población normal.

A continuación damos los estimadores de mayor uso en los estudios estadísticos.

Tabla 8.2 *Estimadores de los parámetros más usuales.*

1. $\bar{X} = \dfrac{\sum_{i=1}^{n} X_i}{n}$, la media muestral. Este estimador se emplea para estimar μ_X y se escribe $\hat{\mu}_X = \bar{X}$.

2. $S^2 = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$, la varianza muestral. Este estimador se emplea para estimar σ^2_X y se escribe $\hat{\sigma}^2 = S^2$.

3. $S = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$, la desviación estándar muestral. Este estimador se emplea para estimar σ_X y se escribe $\hat{\sigma}_X = S$.

4. $\bar{p} = \dfrac{\text{No. de individuos que poseen la característica}}{\text{Total de individuos escogidos}}$, la proporción muestral. Este estimador se emplea para estimar π y se escribe $\hat{\pi} = \bar{p}$.

5. $T = N\bar{X}$, el total poblacional. Este estimador se emplea para estimar el total poblacional τ y se escribe $\hat{\tau} = N\bar{X}$.

6. $T = N\bar{p}$, el total poblacional. Este estimador se emplea para estimar el total poblacional τ de individuos que poseen determinada característica y se escribe $\hat{\tau} = N\bar{p}$.

Veamos algunos ejercicios.

1. De una población se escogieron al azar 10 personas y se les tomó la estatura. Los resultados fueron (en cm): $x_1 = 160$, $x_2 = 170$, $x_3 = 170$, $x_4 = 150$, $x_5 = 160$, $x_6 = 180$, $x_7 = 160$, $x_8 = 170$, $x_9 = 130$, $x_{10} = 150$

Estime μ_x y σ_X

De acuerdo con las fórmulas 1 y 3, $\hat{\mu}_X = \bar{x} = 160$, $\hat{\sigma}_X = S = 14$.

2. En cierta universidad se desea conocer la opinión de los estudiantes acerca de ciertas medidas que han tomado las directivas. De 120 estudiantes consultados 90 estuvieron a favor. Estime la proporción de estudiantes que están a favor de las medidas.

Aplicamos la fórmula 4, y tenemos así $\hat{\pi} = \frac{90}{120} = \frac{3}{4} = 0.75$. Esto es, hay una proporción estimada de los que están a favor del 75%.

3. Un conjunto residencial está formado por 200 apartamentos. Una selección aleatoria de 18 apartamentos condujo a que, en promedio, viven 4.5 personas por apartamento. Estime el total de personas que viven en el conjunto residencial.

Como se trata de estimar un total, aplicamos la fórmula 5 y tenemos $200 \times 4.5 = 900$. Por tanto, se estima que en el conjunto residencial vivan 900 personas.

4. De un lote de 1,000 licuadoras se escogen aleatoriamente 30 y se encontró que 2 de ellas presentaban algún tipo de maltrato; ¿cuántas licuadoras se estima que presenten algún tipo de maltrato?

En este caso aplicamos la fórmula 6, y tenemos $1{,}000 \times \frac{2}{30} = \frac{2{,}000}{30} = 66.67$. Se estima entonces que se encuentran 67 licuadoras maltratadas.

8.4 EL ERROR ESTÁNDAR

Como es de suponer, un mismo estimador ofrece distintos valores para distintas muestras del mismo tamaño extraídas de la misma población. Ante esta circunstancia sería de mucha utilidad tener una medida de la variabilidad del estimador respecto del parámetro que se trata de estimar. Esta variabilidad se mide en términos de la desviación estándar del estimador, la cual recibe el nombre de *error estándar.*

El error estándar de un estimador T *de un parámetro* θ *es la desviación estándar del estimador.*

Así por ejemplo, si tomamos $\bar{X}$ como estimador de μ, entonces el error estándar de $\bar{X}$ es $\sigma_{\bar{X}} = \frac{\sigma_x}{n}$

De acuerdo con lo anterior, si $\sigma_x = 4$ y se toma una muestra de tamaño 25 entonces el error estándar de $\bar{X}$ está dado por $\sigma_{\bar{x}} = \frac{4}{5} = 0.8$.

El valor absoluto de la diferencia entre una estimación particular y el valor del parámetro se llama *error de estimación.* Así por ejemplo, si el valor de la media es $\mu = 4.8$ y la media muestral es $\bar{x} = 4.5$, entonces el error de estimación es $|4.5 - 4.8| = 0.3$. En realidad por cada valor estimado del parámetro se tiene un error de estimación por lo general diferente. Sin embargo, es posible fijar un intervalo dentro del cual se encontrarán la mayoría de los valores de error de estimación para un estimador y parámetro dados.

En la tabla que sigue se dan las fórmulas de los errores de estimación para algunos estimadores. En esta tabla se encontrarán también los estimadores para tales errores; esto es debido a que por lo general los parámetros que se incluyen en las fórmulas de los errores de estimación son desconocidos y hay que estimarlos.

Tabla 8.3 *Errores estándares de algunos estimadores.*

Parámetro	Estimador	Error estándar	Estimador del error estándar
μ	$\bar{X}$	$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$	$\hat{\sigma}_{\bar{x}} = S_{\bar{x}} = \frac{S}{\sqrt{n}}$
π	$\bar{p}$	$\sigma_{\bar{p}} = \frac{\pi(1-\pi)}{n}$	$\hat{\sigma}_{\bar{p}} = S_{\bar{p}} = \frac{\bar{p}(1-\bar{p})}{n}$
τ	$N\bar{X}$	$\sigma_{N\bar{X}} = \frac{N\sigma_x}{\sqrt{n}}$	$\hat{\sigma}_{N\bar{X}} = S_{N\bar{X}} = \frac{NS}{\sqrt{n}}$

Una agencia de encuesta selecciona 900 familias y calcula la proporción de éstas en cuanto hacen uso de cierto detergente. Si la proporción estimada es de $\hat{\pi} = 0.35$. ¿Cuál es el error estándar estimado?
Como no se conoce el verdadero valor de π, debemos utilizar la proporción estimada y de acuerdo con la tabla 8.3, se tiene $\hat{\sigma}_{\bar{p}} = \sqrt{\frac{(0.35)(0.65)}{900}} = 0.016$.

En un estudio de cierta característica X de una población se sabe que la desviación estándar es $\sigma_x = 3$. Si se va a escoger una muestra de tamaño 100 de esta población para el estudio de la característica, halle el error estándar de $\bar{X}$.

En este caso es conocida la desviación estándar de la población, así que aplicaremos la fórmula $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$ y tenemos $\sigma_{\bar{x}} = \frac{3}{\sqrt{100}} = 0.3$. Observe que en este caso damos el valor verdadero del error estándar (no el estimado) de la media poblacional.

Se escogió al azar una muestra de diez clientes de un banco y se les preguntó el número de veces que habían utilizado el banco para llevar a cabo alguna transacción comercial. Los resultados fueron los siguientes: 0, 4, 2, 3, 2, 0, 3, 4, 1, 1. Estime el error estándar del número de transacciones promedio.
Para esta situación no se conoce la desviación estándar y por ello el error estándar será estimado por la fórmula $\hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{n}}$ y así tenemos, $\hat{\sigma}_{\bar{x}} = \frac{1.5}{\sqrt{10}} = 0.47$.

Ejercicios **8.1**

1. Los siguientes datos corresponden a los pesos (en kilogramos) de 15 hombres escogidos al azar y que laboran en una empresa: 72, 68, 63, 75, 84, 91, 66, 75, 86, 90, 62, 87, 77, 70, 69. Estime el peso promedio y la desviación estándar.

2. Entre los miembros de una comunidad se escogieron 150 personas al azar y se les preguntó si estaban de acuerdo con los programas que el gobierno estaba desarrollando para prevenir el consumo de drogas; la encuesta dio como resultado que 130 sí estaban de acuerdo. Estime la proporción de los que están de acuerdo y estime el error estándar.

3. De los 50 salones que tiene un edificio de una universidad se escogieron al azar 5 salones y se determinó el número de estudiantes que había en cada uno de ellos en la primera hora de clases. Estime el número de estudiantes que se encuentran en el edificio si todos los salones se hallan ocupados a esa hora, y si el número de estudiantes en cada uno de los salones inspeccionados fue: 24, 35, 16, 30, 28.

4. Con base en los datos del problema 1 del presente ejercicio, estime el error del peso promedio.

5. Para los datos del problema 3 del presente ejercicio, estime el error del número total de estudiantes.

8.5 ESTIMACIÓN POR INTERVALOS

En las anteriores secciones se trató el tema de la estimación puntual. Ahora nos proponemos considerar lo relativo a la estimación por intervalos, la cual consiste en determinar dos números entre los cuales se halla el parámetro estudiado con cierta certeza.

El procedimiento para obtener un intervalo (de confianza) para un parámetro, la media μ, por ejemplo, requiere de la determinación de un estimador del parámetro

y de la distribución del estimador. En la exposición que vamos a hacer se trata de obtener un intervalo de confianza para la media de una población normal.

Se sabe que si $x \sim N(\mu_x, \sigma^2_x)$ entonces $\overline{X} \sim N(\mu_x, \frac{\sigma^2_x}{n})$. Vamos a determinar números a y b tales que

$$P[a < \overline{X} < b] = 0.95 \tag{8-3}$$

Como ya lo hemos señalado, para determinar estos valores es necesario "estandarizar" X. Al hacer esto, tenemos:

$$P[\frac{a-\mu_x}{\sigma_x/\sqrt{n}} < \frac{\overline{X}-\mu_x}{\sigma_x/\sqrt{n}} < \frac{b-\mu_x}{\sigma_x/\sqrt{n}}] = P[\frac{\sqrt{n}(a-\mu_x)}{\sigma_x} < Z < \frac{\sqrt{n}(b-\mu_x)}{\sigma_x}] =$$

$$0.95 \Longrightarrow P[a < \overline{X} < b] = P[\frac{\sqrt{n}(a-\mu_x)}{\sigma_x} < Z < \frac{\sqrt{n}(b-\mu_x)}{\sigma_x}] = 0.95 \tag{8-4}$$

En realidad hay infinitos pares de números $\frac{\sqrt{n}(a-\mu_x)}{\sigma_x}$, $\frac{\sqrt{n}(b-\mu_x)}{\sigma_x}$ para los cuales se cumple la ecuación (8-4). De esas infinitas posibilidades vamos a escoger el par de números que se hallan situados simétricamente respecto de cero en la distribución normal. Así que, basados en la tabla II y al tener en cuenta que en el centro de la curva debe quedar el 95% del área o lo que es lo mismo 2.5% en cada cola, llegamos a que $\frac{\sqrt{n}(a-\mu_x)}{\sigma_x} = -1.96$, $\frac{\sqrt{n}(b-\mu_x)}{\sigma_x} = 1.96$

A partir de estas ecuaciones obtenemos, $a = \mu_x - 1.96 \frac{\sigma_x}{\sqrt{n}}$ y $b = \mu_x + 1.96 \frac{\sigma_x}{\sqrt{n}}$, que al volver a la expresión (8-3) nos queda

$$P[\mu_x - 1.96 \frac{\sigma_x}{\sqrt{n}} < \overline{X} < \mu_x + 1.96 \frac{\sigma_x}{\sqrt{n}}] = 0.95 \tag{8-5}$$

Ahora vamos a transformar la desigualdad que aparece en (8-5).

De $\mu_x - 1.96 \frac{\sigma_x}{\sqrt{n}} < \overline{X} < \mu_x + 1.96 \frac{\sigma_x}{\sqrt{n}}$ se obtiene

$-1.96 \frac{\sigma_x}{\sqrt{n}} < \overline{X} - \mu_x < 1.96 \frac{\sigma_x}{\sqrt{n}} \Longrightarrow -\overline{X} - 1.96 \frac{\sigma_x}{\sqrt{n}} < -\mu_x < -\overline{X} + 1.96 \frac{\sigma_x}{\sqrt{n}} \Longrightarrow$

$\overline{X} + 1.96 \frac{\sigma_x}{\sqrt{n}} > \mu_x > \overline{X} - 1.96 \frac{\sigma_x}{\sqrt{n}}$, que reordenado nos queda $\overline{X} - 1.96 \frac{\sigma_x}{\sqrt{n}} < \mu_x < \overline{X} + 1.96 \frac{\sigma_x}{\sqrt{n}}$. Así que, (8-5) puede ser sustituida por la expresión equivalente,

$$P[\overline{X} - 1.96 \frac{\sigma_x}{\sqrt{n}} < \mu_x < \overline{X} + 1.96 \frac{\sigma_x}{\sqrt{n}}] = 0.95 \tag{8-6}$$

El intervalo $[\overline{X} - 1.96 \frac{\sigma_x}{\sqrt{n}}, \overline{X} + 1.96 \frac{\sigma_x}{\sqrt{n}}]$ se llama *intervalo (aleatorio) de confianza para* μ_X.

A partir de los datos muestrales podemos determinar el valor de $\overline{X}$ y obtenemos así un intervalo numérico. Note que en el intervalo que hemos hallado 1.96 es un valor accidental en la medida en que ese valor aparece allí, puesto que se consideró al

inicio una probabilidad de 0.95. Si hubiésemos tomado otro valor de probabilidad aparecería el valor de la normal estándar correspondiente a tal probabilidad. Con esto pretendemos indicar que un intervalo de confianza para un caso como el presente tiene la forma $[\bar{X} - z\frac{\sigma_x}{\sqrt{n}}, \bar{X} + z\frac{\sigma_x}{\sqrt{n}}]$. Expresión esta que puede simplificarse mediante la escritura $\bar{X} \mp z\frac{\sigma_x}{\sqrt{n}}$. La diferencia $(\bar{X} + z\frac{\sigma_x}{\sqrt{n}}) - (\bar{X} - z\frac{\sigma_x}{\sqrt{n}}) = 2z\frac{\sigma_x}{\sqrt{n}}$, se llama *longitud del intervalo.* También es pertinente aclarar que la expresión dada en (8-6) tiene sentido mientras $\bar{X}$ conserve su carácter de variable aleatoria, porque si $\bar{X}$ es sustituida por un valor muestral para obtener un intervalo específico como, por ejemplo [150.2, 169.8], entonces sería incorrecto escribir $P[150.2 < \mu < 169.8] = 0.95$, puesto que μ representa un número fijo y así está o no está dentro del intervalo. En seguida precisamos los conceptos contenidos en la anterior presentación de un intervalo de confianza y que son válidos para cualquier intervalo de esta clase, trátese de media, varianza, proporción, diferencia de medias, cociente de varianzas o diferencia de proporciones.

Un intervalo de confianza para un parámetro θ es un intervalo construido alrededor del estimador del parámetro, de tal manera que podemos cifrar nuestra confianza de que el verdadero valor del parámetro quede incluido en dicho intervalo. Si los extremos son aleatorios el intervalo se dice *aleatorio.* Cuando el estimador es reemplazado por un número de acuerdo con los datos muestrales, el intervalo se dice *numérico.*

En caso de que el parámetro sea μ_x, el intervalo se construye alrededor de $\bar{X}$, como puede apreciarse en lo que hemos expuesto antes.

El nivel de confianza de un intervalo es una probabilidad (expresada en porcentaje) que representa nuestra seguridad de que el intervalo encierra el verdadero valor del parámetro θ. En el caso que hemos analizado, el nivel de confianza es del 95%.

En general el nivel de confianza se expresa en la forma $100(1 - \alpha)\%$. Para nuestro caso $(1 - \alpha) = 0.95$ y $100(1 - \alpha)\% = 95\%$. El valor α representa la probabilidad de que el parámetro quede por fuera del intervalo, y que para el caso presentado es de $\alpha = 0.05$. En la figura 8.6 se muestra gráficamente la situación de un intervalo de confianza del 95%.

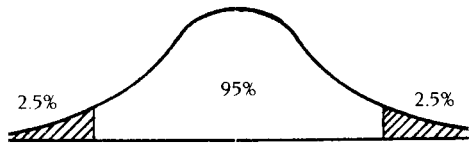

Figura 8.6 *Intervalo de confianza del 95% para μ_x en población normal.*

Para cada nivel de confianza existe un valor de tabla (normal, t, χ^2, F) asociado al nivel de confianza dado. Este valor se llama *coeficiente de confiabilidad* y se denota $z_{(1-\alpha/2)}$ para el caso de una normal; $t_{(k,\,1-\alpha/2)}$ para la distribución t; $\chi^2_{(k,\,1-\alpha/2)}$, $\chi^2_{(k,\,\alpha/2)}$ para una ji cuadrado; $F_{(1-\alpha/2;\,m,n)}$, $F_{(\alpha/2;\,m,n)}$ cuando se trata de la distribución F.

Observe que si deseamos un intervalo con un nivel de confianza de $100(1 - \alpha)\%$, en la tabla correspondiente debe buscarse un valor de variable para el cual el área de la cola superior (también inferior) sea del $100(\alpha/2)\%$. Esto se debe a que al tratarse de un intervalo del $100(1 - \alpha)\%$ de confianza, entonces la porción de área que no

será cubierta por el intervalo debe tener una medida de tamaño α y se toma como norma de procedimiento que este valor α se reparta en partes iguales entre las dos colas. En la figura 8.7 se muestra gráficamente la situación de un intervalo de confianza del $100(1 - \alpha)\%$ para la media en población normal.

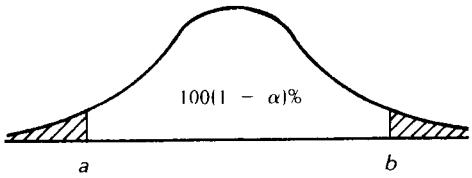

Figura 8.7 *Intervalo del $100(1 - \alpha)\%$ de confianza para μ en población normal.*

En el problema que hemos explicado, $z_{0.975} = 1.96$ es el coeficiente de confiabilidad. Los tres conceptos básicos que encierra un intervalo quedan resumidos en la expresión general para un intervalo de confianza.

$$\text{Estimador} \mp (\text{coef. de conf.})(\text{error estándar})$$

En el siguiente ejercicio se ilustra cómo se procede para hallar un intervalo de confianza para la media.

Suponga que se utiliza una variable aleatoria X para designar el peso de un pasajero de avión y que interesa conocer μ, el peso promedio de todos los pasajeros. Como hay limitaciones de tiempo y dinero para pesarlos a todos, se toma una muestra de 36 pasajeros de la cual se obtiene una media muestral $\bar{x} = 160$ libras. Suponga además que la distribución de los pasajeros tenga una distribución normal con desviación estándar $\sigma = 30$. Halle el intervalo con un nivel de confianza del 95%.

De acuerdo con lo expuesto el intervalo está dado por $\bar{X} \mp z_{(1-\alpha/2)}\dfrac{\sigma_x}{\sqrt{n}}$

$\bar{x} = 160$, $z_{0.975} = 1.96$, $\sigma_x = 30$, $n = 36$

Al reemplazar se tiene, $160 \mp (1.96)(30/6) = 160 \mp (1.96)(5) = 160 \mp 9.8$

– El intervalo pedido es [150.2, 169.8]

Para los mismos datos del problema anterior halle un intervalo del 90% de confianza para μ

$\bar{x} = 160$, $z_{0.95} = 1.645$, $\sigma_x = 30$, $n = 36$

Al reemplazar se tiene $160 \mp (1.645)(30/6) = 160 \mp (1.645)(5) = 160 \mp 8.23$

– El intervalo pedido es [151.78, 168.23]

Ahora calculemos un intervalo del 99%

$\bar{x} = 160$, $z_{0.995} = 2.575$, $\sigma_x = 30$, $n = 36$

Al reemplazar se tiene $160 \mp (2.575)(30/6) = 160 \mp 12.875$

– El intervalo pedido es [147.13, 172.88]

Estos tres intervalos para μ_x que hemos hallado para distintos niveles de confianza nos ponen de presente que a medida que se aumenta el nivel de confianza la longitud del intervalo también aumenta, como puede apreciarse en la figura 8.8.

Esta particularidad que hemos apreciado acerca de la longitud del intervalo cuando se cambia el nivel de confianza no es propia de este problema, sino que es una propiedad que siempre se da y tenemos así:

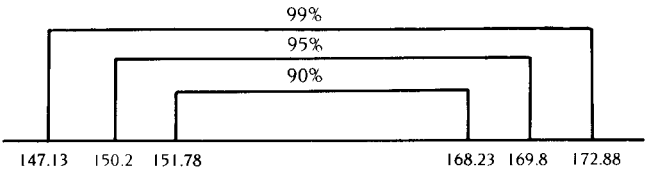

Figura 8.8 *Intervalos del 90, 95 y 99% de confianza para el peso promedio de los pasajeros del avión.*

Propiedad 1. Para un tamaño de muestra y una varianza dadas a medida que aumenta el nivel de confianza también lo hace la longitud del intervalo
Esta propiedad no debe sorprendernos si tenemos en cuenta que la longitud L de un intervalo de confianza está dada por $L = 2z\frac{\sigma_x}{\sqrt{n}}$

De la misma fórmula para la longitud de un intervalo se puede observar que a medida que n aumenta la longitud del intervalo disminuye y tenemos así la propiedad 2 acerca de la longitud de un intervalo de confianza:

Propiedad 2. Para un nivel de confianza y una varianza dadas cuando el tamaño de muestra aumenta la longitud del intervalo disminuye
Como se puede ver en la fórmula de la longitud de un intervalo, ésta se encuentra sujeta (supuesta la varianza fija) a la presencia de dos números cuyas acciones se contraponen en cuanto a la longitud; se trata del nivel de confianza (coeficiente de confiabilidad) y del tamaño de la muestra. De otra parte, mientras más pequeña sea la longitud del intervalo serán más cercanos los valores que encierra al parámetro que se dice contener con cierta certeza; por ello se afirma que la longitud del intervalo es una medida de la *precisión* de la estimación.

No sobra decir que para que un intervalo sea tomado en cuenta con algún interés, el nivel de confianza debe ser alto y por ello los niveles de confianza que se consideren para obtener intervalos de alguna utilidad son del 90% como mínimo. Una explicación a esta consideración que parece ser obvia la encontramos en las siguientes interpretaciones que se le dan a un intervalo de confianza para la media, pero que son válidas para otros parámetros.

Suelen presentarse dos interpretaciones para un intervalo de confianza, para la media en este caso. Una *probabilística* y otra *práctica.*

Desde el punto de vista de la probabilidad se dice: *"En el muestreo aleatorio simple de una población normal con media* μ_x *y varianza* σ^2_x *conocida el* $100(1-\alpha)\%$ *de todos los intervalos de la forma* $\bar{X} \mp z\frac{\sigma_x}{\sqrt{n}}$ *incluirá la media desconocida* μ_x*".*

A partir de esta interpretación podemos decir que, para el problema de los pesos de los pasajeros del avión, de 100 muestras de tamaño 36 que escojamos de estos pasajeros 95 de ellas (aproximadamente) producirán intervalos que contendrán el verdadero peso promedio μ_x. O lo que es lo mismo, de 100 intervalos obtenidos mediante la citada fórmula 95 de ellos contendrán el verdadero valor del parámetro. En este orden de ideas, ¿se considera usted tan desafortunado que si escoge uno de estos intervalos no contenga éste el verdadero valor del parámetro?

De la interpretación probabilística se desprende la práctica que se establece así: *"Si se realiza un muestreo aleatorio simple en una población normal con media* μ_x *y*

varianza conocida σ^2_x, se tiene el 100 $(1 - \alpha)$% de confianza de que el intervalo particular $\bar{x} \mp z \frac{\sigma_x}{\sqrt{n}}$ *contendrá el verdadero valor del parámetro desconocido* μ_x".

Así que podemos decir que tenemos una confianza o certeza del 95% de que el verdadero peso promedio de los pasajeros del avión está entre 150.2 y 169.8 libras; o que tenemos una certeza del 90% de que el verdadero peso promedio de los pasajeros se encuentra entre 151.8 y 168.2 libras.

Intuitivamente, se dice que construir un intervalo de confianza es como enlazar un novillo inmóvil. El parámetro que se desea estimar corresponde al "novillo" y el intervalo el "lazo" que el "vaquero" prepara. Cuando se extrae una muestra se construye un intervalo para un parámetro y se espera haberlo "enlazado". El nivel de confianza representa la posibilidad de "enlazarlo" o lo que es lo mismo de "cada 100 intentos, ¿cuántos de ellos terminarán "enlazando" el novillo?"

La metodología expuesta aquí para obtener un intervalo de confianza para la media de poblaciones normales con varianza conocida es la misma en términos generales que se utiliza para obtener el intervalo de confianza de cualquier parámetro. En lo que sigue presentaremos la fórmula para obtener el intervalo respectivo de acuerdo con el parámetro señalado y se acompañará de un ejercicio ilustrativo.

8.5.1 Intervalo de confianza para la media en poblaciones normales con varianza conocida

Es un intervalo construido de tal forma que podemos fijar de antemano el grado de certeza (confianza) de que el verdadero valor de la media μ_x quede incluido en él. Este intervalo se calcula mediante la fórmula simplificada.

$$\bar{X} \mp z_{(1-\alpha/2)} \frac{\sigma_x}{\sqrt{n}} \tag{8-7}$$

$z_{(1-\alpha/2)}$ = valor de la variable normal estándar que determina una cola superior de medida $\alpha/2$

Una muestra aleatoria de 36 cigarrillos de una marca determinada dio un contenido promedio de nicotina de 3.0 miligramos. Suponga que el contenido de nicotina de estos cigarrillos sigue una distribución normal con una desviación estándar $\sigma = 1.0$ miligramo. a) Obtenga e interprete un intervalo de confianza del 95% para el verdadero contenido promedio de nicotina en estos cigarrillos. b) El fabricante garantiza que el contenido promedio de nicotina es de 2.9 miligramos, ¿qué puede decirse de acuerdo con el intervalo hallado?

Sea μ_x = contenido promedio de nicotina.

Una vez que hemos precisado la fórmula a emplear, pasamos a determinar los valores de los elementos que la componen. En este caso, $\bar{x} = 3$, $z_{0.975} = 1.96$ (según la tabla II), $\sigma_x = 1$, $n = 36$.

Al reemplazar en (8-7) se tiene,

$$3 \mp (1.96)\left(\frac{1}{\sqrt{36}}\right) = 3 \mp (1.96)\left(\frac{1}{6}\right) = 3 \mp 0.33$$

– El intervalo pedido es [2.67, 3.33].

En cuanto a la interpretación se da en estos términos: "tenemos una certeza del 95% de que el verdadero contenido promedio de nicotina se halla entre 2.67 y 3.33 miligramos".

- Para responder el literal b), debemos tener en cuenta que si tenemos una certeza del 95% de que el verdadero valor del parámetro se halla entre 2.67 y 3.33, y al tener presente que 2.9 se halla entre este rango no podemos descartarlo como valor posible del parámetro.

En el mismo problema, supongamos que el fabricante garantiza que el promedio de nicotina está por debajo de 2.5 miligramos.
Si éste es el caso, tenemos que decir que de acuerdo con el intervalo hallado no se da la evidencia de que el promedio esté por debajo de 2.5 miligramos, puesto que corresponde a un valor que está por fuera del intervalo y la probabilidad de que se den valores de parámetro por fuera del intervalo es del 5%, con lo cual le corresponde a la región en donde se dice que está el parámetro una probabilidad del 0.025 de que ello ocurra. Esto es, hay una probabilidad de acuerdo con el intervalo hallado de 0.025 de que el verdadero valor del parámetro se halle en la región en donde está 2.5.

8.5.2 Intervalo de confianza para la media en poblaciones normales con varianza desconocida

Cuando este es el caso el intervalo está dado por la fórmula,

$$\bar{X} \mp t_{(n-1;\,1-\alpha/2)} \frac{S}{\sqrt{n}} \tag{8-8}$$

en donde $t_{(n-1;\,1-\alpha/2)}$ = valor de la variable con distribución t con $(n-1)$ grados de libertad que deja una cola superior de medida $\alpha/2$, S es la desviación estándar muestral.

Los siguientes son los registros de las mediciones del tiempo (en minutos) que tardaron 15 operarios para familiarizarse con el manejo de una máquina moderna recientemente adquirida por la empresa: 3.4, 2.8, 4.4, 2.5, 3.3, 4.0, 4.8, 2.9, 5.6, 5.2, 3.7, 3.0, 3.6, 2.8, 4.8. Suponga que los tiempos se distribuyen normalmente. a) Determine e interprete un intervalo del 95% de confianza para el verdadero tiempo promedio, b) el instructor considera que el tiempo promedio requerido por la población de los trabajadores que reciben instrucción sobre el manejo de esta máquina está por encima de los cinco minutos, ¿qué se puede decir de acuerdo con el intervalo hallado?
Como en la información no se da la varianza poblacional, utilizamos la fórmula (8-8). Calculamos la media y la desviación estándar muestral: $\bar{x} = 3.8$, $S = \sigma_{n-1} = 0.971$
El valor $t_{(14;\,0.975)}$ lo buscamos en la tabla IV y allí encontramos que es 2.145. Para la búsqueda de este número en la citada tabla debemos tomar en cuenta que el valor t buscado debe ser tal que deje una cola derecha de 2.5%.
Al reemplazar en la fórmula IV tenemos:

$$3.8 \mp (2.145)\frac{0.971}{\sqrt{15}} = 3.8 \mp (0.54)$$

- El intervalo pedido es [3.26, 4.34]
- Estamos 95% seguros de que el verdadero tiempo promedio que requieren los operarios para familiarizarse con la máquina está entre 3.26 y 4.34 minutos;
- De acuerdo con el intervalo hallado, no parece ser correcta la apreciación del instructor, puesto que el promedio 5 minutos está por fuera del intervalo hallado.

La fórmula (8-8) es particularmente empleada cuando se trata de poblaciones normales con varianza desconocida y el tamaño de muestra es máximo de 30. Cuando

se trata de muestras de tamaño superior a 30 se puede emplear la fórmula (8-7) de acuerdo con el teorema del límite central.

8.5.3 Intervalo de confianza para la diferencia de medias en poblaciones normales independientes

En esta situación hay que distinguir dos casos: Cuando las varianzas de las poblaciones involucradas son conocidas y cuando las varianzas de las dos poblaciones son desconocidas, pero se suponen iguales. Cuando se trata de poblaciones con varianzas conocidas emplearemos la fórmula,

$$(\bar{X} - \bar{Y}) \mp z_{(1 - \alpha/2)} \sqrt{\sigma^2_X / n_x + \sigma^2_Y / n_Y} \qquad (8\text{-}9)$$

para determinar un intervalo de confianza para la diferencia $\mu_X - \mu_Y$ (en este orden)

Nota. Usualmente la diferencia de los parámetros se toma en el orden tal que la diferencia muestral quede positiva.

Suponga que se desea medir la diferencia entre dos categorías de empleados en la actividad de seguros. Una está formada por personas con título superior y la otra por personas que sólo tienen estudios secundarios. Se toma una muestra de 45 empleados entre los primeros y la media de ventas resulta ser 32, en tanto que la media de una muestra de 60 empleados con estudios secundarios solamente, es 25. Suponga también que las ventas de los dos grupos se distribuyen normalmente con varianzas respectivas de 48 para los titulados y 56 para los que sólo tienen estudios secundarios. a) Calcule e interprete un intervalo del 90% de confianza para la verdadera diferencia de las medias, b) de acuerdo con el intervalo hallado, ¿hay evidencia de que las ventas medias de los grupos son iguales?

Definamos las variables, X = venta de un titulado, Y = venta de uno con sólo estudios secundarios.
Los parámetros a considerar son:
μ_X = venta promedio de los titulados, μ_Y = venta promedio de los que tienen sólo estudios secundarios.
Los valores muestrales promedios son respectivamente: $\bar{x} = 32$; $\bar{y} = 25$; así que construimos el intervalo por medio de la fórmula (8-9).
Los datos para la aplicación de esta fórmula son: $\bar{x} = 32, \bar{y} = 25, \sigma^2_x = 48, \sigma^2_Y = 56$, $n_x = 45$, $n_y = 60$, $z_{0.95} = 1.645$

Al reemplazar en (8-9) se tiene $(32 - 25) \mp (1.645) \sqrt{48/45 + 56/60} = 7 \mp (2.33)$

- El intervalo pedido es [4.67, 9.33]
- Tenemos una certeza del 90% de que la verdadera diferencia promedio de ventas se halla entre 4.67 y 9.33.

La condición de que las ventas medias son iguales se traduce por la condición $\mu_x = \mu_Y$ o lo que es lo mismo $(\mu_x - \mu_Y) = 0$, así que para que la igualdad entre las medias no pueda descartarse, el cero tiene que estar incluido en el intervalo. Como en el presente caso esto no sucede, entonces no hay evidencia de una igualdad entre las dos medias.

Se registraron los siguientes datos, en minutos, que tardan algunos hombres y mujeres en realizar cierta actividad en una empresa, los cuales fueron seleccionados aleatoriamente.

Hombres	**Mujeres**
$n_1 = 14$	$n_2 = 25$
$\bar{x}_1 = 17$	$\bar{x}_2 = 19$
$S^2_1 = 1.5$	$S^2_2 = 1.8$

Suponga que los tiempos para los dos grupos se distribuyen normalmente y que las varianzas son iguales, aunque desconocidas. a) Calcule e interprete un intervalo de confianza del 99% para la verdadera diferencia de medias. b) De acuerdo con el intervalo hallado, ¿hay evidencia de que los dos tiempos promedios son iguales?
Como puede observarse en este caso, no hay conocimiento de las varianzas poblacionales. Cuando esto ocurre, el intervalo para $\mu_x - \mu_Y$ se calcula mediante la fórmula,

$$(\bar{X} - \bar{Y}) \mp t_{(k,1-\alpha/2)} S_p \sqrt{1/n_x + 1/n_Y} \tag{8-10}$$

$t_{(k,1-\alpha/2)}$ = valor de la variable con distribución t con $k = n_x + n_Y - 2$ grados de libertad que determina un área superior de medida $\alpha/2$, $S^2_p = \frac{(n_x - 1)S^2_x + (n_Y - 1)S^2_Y}{n_x + n_Y - 2}$ es la varianza ponderada.

Como se indicó para el caso anterior, la diferencia de los parámetros se toma de tal forma que la diferencia muestral sea positiva.
En el presente problema tomamos,
X = tiempo que tarda una mujer en realizar la actividad, Y = tiempo que tarda un hombre en realizar la actividad.
Los datos para la fórmula (8-10) son: $\bar{x} = 19$, $\bar{y} = 17$, $n_x = 25$, $n_Y = 14$, $t_{(37;\,0.995)} = 2.704$ (de acuerdo con la tabla IV al aproximar a 40 grados de libertad).

La varianza ponderada está dada por $S^2_p = \dfrac{(24)(1.8) + (13)(1.5)}{14 + 25 - 2} = \dfrac{62.7}{37} = 1.69$

$\Longrightarrow S = 1.3$

Al reemplazar en (8-10) se tiene, $(19 - 17) \mp (2.704)(1.3)\sqrt{1/25 + 1/14} = 2 \mp 0.39$

- El intervalo pedido es [1.61, 2.39]
- Estamos 99% seguros de que la verdadera diferencia promedio de tiempo que gastan dichos hombres y mujeres en realizar la actividad se encuentra entre 1.61 y 2.39 minutos.

Como el 0 no está contenido en el intervalo, estos datos no evidencian una igualdad entre las dos medias. La utilización de la fórmula (8-10) requiere que las varianzas de las dos poblaciones aunque desconocidas, sean iguales. Cuando las varianzas son distintas la fórmula (8-10) sufre una modificación en cuanto a los grados de libertad, pero este tema no se discutirá en el presente texto.

8.5.4 Intervalo de confianza para la proporción y diferencia de proporciones

En algunos casos lo que interesa es determinar una proporción o diferencia de proporciones. Como por ejemplo, la proporción de personas que están a favor de determinado nuevo producto o si la proporción de artículos defectuosos que produce la máquina I es diferente a la proporción de defectuosos que produce la máquina II.

Cuando se trata de determinar un intervalo de confianza para una proporción se aplica la fórmula,

$$\bar{p} \mp z_{(1-\alpha/2)} \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \tag{8-11}$$

Como la fórmula (8-11) es una consecuencia del teorema del límite central, se recomienda para su aplicación tomar muestras de tamaño grande.

Una fábrica desea saber la proporción de amas de casa que preferirían una aspiradora "Central", dados la calidad y el precio. Se toma al azar una muestra de 100 amas de casa; 20 dicen que les gustaría la máquina. Calcule e interprete un intervalo del 95% de confianza para la verdadera proporción de amas de casa que preferirían la citada aspiradora.

Los datos para la aplicación de la fórmula (8-11) son $\bar{p} = \dfrac{20}{100} = 0.2$, $z_{0.975}$ $= 1.96$, $\sqrt{(0.2)(0.8)/100} = 0.04$

Al reemplazar se tiene: $0.2 \mp (1.96)(0.04) = 0.2 \mp 0.0784$

- El intervalo pedido es [0.122, 0.278]
- La verdadera proporción de amas de casa que preferirían la aspiradora está entre 12.2% y 27.8%

Se está considerando cambiar el procedimiento de manufactura de partes. Se toman muestras del procedimiento actual así como del nuevo para determinar si este último resulta mejor. Si 75 de 1,000 artículos del procedimiento actual presentaron defectos y lo mismo sucedió con 80 de 2,500 partes del nuevo, determine un intervalo de confianza del 90% para la verdadera diferencia de proporciones de partes defectuosas.

Cuando se trata de intervalos de confianza para la diferencia de proporciones empleamos la fórmula

$$\bar{p}_1 - \bar{p}_2 \mp z_{(1-\alpha/2)} \sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}} \tag{8-12}$$

en donde $\bar{p}_1$ y $\bar{p}_2$ son las proporciones muestrales de cada una de las características consideradas.

Sea π_1 = proporción de artículos defectuosos producidos por el procedimiento actual

π_2 = proporción de artículos defectuosos producidos por el procedimiento nuevo

$$\bar{p}_1 = \frac{75}{1{,}000} = 0.075, \qquad \bar{p}_2 = \frac{80}{2{,}500} = 0.032, \qquad z_{0.95} = 1.645$$

Al reemplazar en (8-12) se tiene

$$(0.075 - 0.032) \mp (1.645) \sqrt{\frac{(0.075)(0.925)}{1{,}000} + \frac{(0.032)(0.968)}{2{,}500}} = 0.043 \mp (0.0149)$$

- El intervalo pedido es [0.0281, 0.0579]
- Estamos 90% seguros de que la diferencia de proporciones está entre 0.0281 y 0.0579

8.5.5 Intervalo de confianza para varianza de poblaciones normales

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

$$\left[\frac{(n-1)S^2}{\chi^2_{(n-1;\,1-\alpha/2)}} \;,\; \frac{(n-1)S^2}{\chi^2_{(n-1;\,\alpha/2)}}\right] \tag{8-13}$$

Este intervalo es diferente a los estudiados anteriormente y puede expresarse así:

$$P\left(\frac{(n-1)S^2}{\chi^2_{(n-1;\,1-\alpha/2)}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{(n-1;\,\alpha/2)}}\right) = 1-\alpha^0$$

en donde $\chi^2_{(n-1;\,\alpha/2)}$ = valor de la variable con distribución ji cuadrado con $(n-1)$ grados de libertad que determina un área inferior de medida $\alpha/2$.
$\chi^2_{(n-1;\,1-\alpha/2)}$ = valor de la variable con distribución ji cuadrado con $(n-1)$ grados de libertad que determina un área superior de medida $\alpha/2$.

Un fabricante de baterías para automóvil asegura que las baterías que produce duran en promedio 2 años, con una desviación estándar de 0.5 años. Si 5 de estas baterías tienen duración 1.5, 2.5, 2.9, 3.2, 4.0 años, determine un intervalo de confianza del 95% para σ^2 e indique si es válida la afirmación del fabricante.
Como se trata de un intervalo de confianza para la varianza aplicamos la fórmula (8-13)
$n = 5$, $S^2 = 0.847$, $\chi^2_{(4;\,0.025)} = 0.484419$, $\chi^2_{(4;\,0.975)} = 11.1413$ (de acuerdo con la tabla III)

Al reemplazar se tiene:

$$\left[\frac{4(0.847)}{11.1413} \;,\; \frac{4(0.847)}{0.484419}\right] = [0.3,\ 7.0] \text{ como el intervalo pedido.}$$

Como el valor de varianza 0.25 está por fuera del intervalo, lo afirmado por el fabricante no está garantizado por los datos muestrales.

8.5.6 Intervalo de confianza para el cociente de varianzas

Si S^2_X y S^2_Y son las varianzas muestrales de muestras independientes de tamaños n_x y n_Y respectivamente de poblaciones normales, entonces un intervalo de confianza de $100(1-\alpha)\%$ para $\dfrac{\sigma^2_x}{\sigma^2_Y}$ es,

$$\left[\, a\,\frac{S^2_X}{S^2_Y},\ b\,\frac{S^2_X}{S^2_Y}\,\right] \tag{8-14}$$

donde $a = \dfrac{1}{F_{(1-\alpha/2;\,n_x-1,\,n_Y-1)}}$, $\quad b = F_{(1-\alpha/2;\,n_Y-1,\,n_x-1)}$

donde $F_{(1-\alpha/2;\,n_x-1,\,n_Y-1)}$ = valor de la distribución F con $(n_x-1,\ n_Y-1)$ grados de libertad que determina un área superior de medida $\alpha/2$

$F_{(1-\alpha/2;\,n_Y-1,\,n_x-1)}$ = valor de la distribución F con $(n_Y-1,\ n_x-1)$ grados de libertad que determina un área superior de medida $\alpha/2$

Determine un intervalo del 90% de confianza para el cociente $\frac{\sigma^2_X}{\sigma^2_Y}$ al tomar las variables definidas en el segundo ejercicio de la sección 8.5.3.

En ese ejercicio se tuvieron valores de varianzas muestrales

$$S^2_X = 1.8, \quad S^2_Y = 1.5 \quad a = \frac{1}{F_{(0.95;\ 13,\ 24)}} = \frac{1}{2.18} = 0.46$$

$$b = F_{(0.95;\ 24,\ 13)} = 2.42$$

Al reemplazar en (8-14) se tiene:

$$\left[0.46\left(\frac{1.8}{1.5}\right), 2.42\left(\frac{1.8}{1.5}\right)\right] = [0.552, 2.904] \text{ es el intervalo pedido}$$

Ejercicios 8.2

1. Una muestra aleatoria de 100 propietarios de automóvil en la ciudad de Bogotá indica que los automóviles recorren anualmente un promedio de 25,000 kilómetros con una desviación estándar de 4,000 kilómetros. Calcule e interprete un intervalo de confianza del 95% para el verdadero recorrido promedio anual.

2. Se administra un *test* estándar a una numerosa clase de estudiantes. La puntuación promedio de 100 estudiantes escogidos al azar fue de 75 puntos. Suponga que las puntuaciones tienen distribución normal con varianza $\sigma^2 = 2.500$ y determine un intervalo de confianza del 95% para la verdadera puntuación promedio. Interprete el intervalo hallado.

3. Sea X la vida útil en millas de cierta llanta, al ser X de distribución normal y con media desconocida. Suponga que se tomó una muestra aleatoria de tamaño 25 y se obtuvo una vida promedio $\bar{x} = 30{,}000$ millas y una desviación estándar $S = 4{,}000$. Calcule e interprete un intervalo de confianza del 95% para la verdadera vida promedio de estas llantas.

4. Con el propósito de estimar la proporción de estudiantes regulares que asistirán a los cursos intermedios, los profesores analizaron una muestra aleatoria de 200 estudiantes. Cuarenta y cinco de éstos indicaron que asistirían. Construya e interprete un intervalo de confianza del 90% para la verdadera proporción de los que asistirán a los cursos intermedios.

5. Una muestra aleatoria de tamaño $n_1 = 16$ que se tomó de una población con una desviación estándar $\sigma_1 = 5$ tiene una media $\bar{x}_1 = 80$. Una segunda muestra aleatoria de tamaño $n_2 = 25$ tomada de una población normal diferente con una desviación estándar $\sigma_2 = 3$, tiene media $\bar{x}_2 = 75$. Encuentre un intervalo de confianza del 95% para $\mu_1 - \mu_2$. De acuerdo con el intervalo hallado, ¿hay evidencia de que las dos medias son iguales?

6. Los siguientes datos corresponden a los pesos de 15 hombres escogidos al azar: 72, 68, 63, 75, 84, 91, 66, 75, 86, 90, 62, 87, 77, 70, 69. Obtenga e interprete un intervalo de confianza del 95% para el verdadero peso promedio.

7. En una encuesta tomada entre estudiantes universitarios, 300 de 500 que viven en el recinto universitario apoyan cierta proposición, mientras que de los 100 estudiantes que viven fuera del recinto universitario 64 la apoyan. Calcule e interprete un intervalo de confianza del 95% para la verdadera diferencia de proporciones de los que apoyan la proposición. De acuerdo con el intervalo hallado, ¿qué se puede decir sobre la igualdad entre las proporciones?

8. Se obtiene una muestra de 16 estudiantes con una media $\bar{x} = 68$ y una varianza de $S^2 = 9$ en un examen de estadística. Suponga que las calificaciones tienen distribución normal y determine un intervalo del 98% de confianza para σ^2.

9. El tiempo de recuperación fue observado para pacientes asignados al azar y sometidos a dos tipos distintos de procedimientos quirúrgicos. Los datos, en días, son los siguientes:

Procedimiento I	**Procedimiento II**
$n_1 = 21$	$n_2 = 23$
$\bar{x}_1 = 7.3$	$\bar{x}_2 = 8.9$
$S^2_1 = 1.23$	$S^2_2 = 1.49$

a) Calcule e interprete un intervalo del 90% de confianza para la verdadera diferencia entre las medias
b) ¿Evidencian estos datos una igualdad entre las medias?

10. Obtenga un intervalo del 95% de confianza para el cociente de las varianzas del problema 9.

11. Durante un periodo de 15 días se tomaron los tiempos gastados por dos estudiantes para transportarse de sus casas a la universidad. Las medias y varianzas fueron:

$\bar{x}_1 = 40.33$	$\bar{x}_2 = 42.54$
$S^2_1 = 1.53$	$S^2_2 = 2.96$

a) Calcule e interprete un intervalo de confianza del 95% para la diferencia de medias
b) De acuerdo con el intervalo hallado, ¿qué puede decirse de la igualdad de las medias?
c) Calcule e interprete un intervalo del 90% de confianza para el verdadero cociente de varianzas
d) De acuerdo con el intervalo hallado, ¿qué puede decirse de la igualdad entre las varianzas poblacionales?

12. Una muestra de 12 alumnas graduadas de una escuela de comercio mecanografió un promedio de 75.8 palabras por minuto con una desviación estándar de 7.5 palabras por minuto. Si se supone una distribución normal para la cantidad de palabras mecanografiadas, encuentre un intervalo de confianza del 99% para el número promedio de palabras para todas las graduadas en esta escuela.

13. Para los datos del problema 12, obtenga e interprete un intervalo de confianza del 95% para la varianza σ^2.

14. Una máquina produce arandelas. Se toma una muestra de 9 piezas y se les mide el diámetro interior, los resultados fueron: 0.99, 0.95, 1.01, 1.03, 0.97, 0.96, 0.97, 0.99, 1.01 cm. Encuentre e interprete un intervalo del 95% de confianza para el diámetro promedio.

15. Para los datos del problema 14 del presente ejercicio calcule e interprete un intervalo de confianza del 95% para la verdadera varianza.

16. Se selecciona una muestra aleatoria de 250 fumadores de cigarrillo y se encuentra que 75 prefieren la marca A. Encuentre e interprete un intervalo del 99% de confianza para la verdadera proporción. Si el fabricante de estos cigarrillos asegura que el 40% de los fumadores prefiere la marca A, ¿qué puede decirse según el intervalo hallado?

17. Suponga que se alimentan dos grupos de gallinas compuestos por 100 gallinas cada uno y que todos los factores, excepto los alimentos, son iguales para los dos grupos. La mortalidad para cada grupo fue la siguiente:

Gallinas	**Alimento A**	**Alimento B**
n	100	100
Número de muertes	20	15

Construya un intervalo de confianza del 98% para la verdadera diferencia entre las proporciones de las que mueren. ¿Evidencian estos datos una igualdad entre las dos proporciones?

18. Una compañía tiene dos departamentos que producen el mismo producto. Se tiene la sensación de que las producciones por hora son diferentes en los dos departamentos. Al tomar una muestra aleatoria de horas de producción en cada departamento se obtuvieron los datos siguientes:

	Departamento I	**Departamento II**
Tamaño de muestra	$n_1 = 64$	$n_2 = 49$
Media muestral	$\bar{x}_1 = 100$ unid.	$\bar{x}_2 = 90$ unid.

Se sabe que las varianzas de las producciones por hora son $\sigma^2_1 = 256$, $\sigma^2_2 = 196$ para los dos departamentos respectivamente. Obtenga e interprete un intervalo del 95% para la verdadera diferencia de la producción media. ¿Qué puede decirse de la sospecha que existe acerca de la diferencia entre la producción promedio?

8.6 TAMAÑO DE LA MUESTRA PARA ESTIMAR MEDIAS Y PROPORCIONES

El tamaño de la muestra que debemos escoger para hacer una estimación del parámetro con las características especificadas (de nivel de confianza y error de estimación) es un problema que tarde o temprano tenemos que resolver. La determinación del tamaño de la muestra es de importancia entre otras cosas porque:

1. Si se toma una muestra más grande de la indicada para alcanzar los resultados presupuestados, constituye un desperdicio de recursos (tiempo, dinero, etc.); mientras que una muestra demasiado pequeña conduce a menudo a resultados poco confiables.
2. Cuando elegimos una muestra de tamaño n sólo revisamos una fracción o parte de la población y con base en ella tomamos decisiones que afectan a toda la población. Es evidente que por este procedimiento se abre la posibilidad de que nos equivoquemos en nuestras decisiones, pero esta posibilidad depende en gran medida del tamaño de muestra o fracción de población que se haya escogido y por tanto analizado.

El tamaño que debe tener la muestra cuando estimamos media o proporción depende del nivel de confianza propuesto para el intervalo, así como del máximo error que estemos dispuestos a admitir entre el valor estimado y el valor real del parámetro que, como ya se dijo antes, corresponde al error de estimación.

A continuación indicamos cómo se determinaría el tamaño de la muestra a partir de la consideración del nivel de confianza y del error de estimación cuando hacemos muestreo con repetición o en poblaciones infinitas.

Supongamos que hemos fijado en d el error de estimación (precisión) y el nivel de confianza en $100(1 - \alpha/2)$ para la estimación de la media μ_x de una población normal con varianza desconocida, con estos datos formamos la ecuación,

$$d = z_{(1 - \alpha/2)} \frac{\sigma_x}{\sqrt{n}} \tag{8-15}$$

De la ecuación (8-15) se tiene $d^2 = z^2 \dfrac{\sigma^2}{n} \Longrightarrow nd^2 = z^2\sigma^2 \Longrightarrow n = \dfrac{z^2\sigma^2}{d^2}$ (8-16)

La ecuación (8-16) nos da la fórmula para obtener el tamaño de la muestra cuando tratamos de estimar un intervalo de confianza para la media con error de estimación y nivel de confianza dados.

De la ecuación (8-16) se desprende de manera clara que el tamaño de la muestra depende de dos elementos básicos (supuesta dada la varianza) que hay que sopesar cuando se va a tomar una decisión al respecto; se trata del nivel de confianza y del error de estimación y tenemos así:

1. **El tamaño de la muestra aumenta a medida que aumenta el nivel de confianza para un error de estimación y una varianza dados.**
2. **El tamaño de la muestra aumenta a medida que disminuye el error de estimación para un nivel de confianza y varianza dados.**

Un ingeniero trata de ajustar una máquina de refrescos de tal forma que el promedio del líquido dispensado se encuentre dentro de cierto rango. Sabe que la cantidad de líquido vertida por la máquina sigue una distribución normal con desviación estándar $\sigma = 0.15$ decilitros. También desea que el valor estimado que vaya a obtener de la media comparado con el verdadero no sea superior a 0.2 decilitros con una confianza del 95%. ¿De qué tamaño debe escoger la muestra, esto es, cuántas mediciones debe realizar para que se cumpla el plan propuesto?
Para aplicar (8-16), debemos determinar los elementos que la conforman; estos son:

$z = 1.96$, $d = 0.02$, $\sigma = 0.15$. Al reemplazar se tiene, $n = \frac{(1.96)^2(0.15)^2}{(0.02)^2} = 216.9$. Así que sería necesario la escogencia de una muestra de tamaño $n = 217$.

La fórmula que hemos presentado mediante la ecuación (8-16) es aplicable cuando el muestreo es con repetición o cuando se hace en población infinita. Si la población es finita y el tamaño de ésta debe ser tenido en cuenta, el tamaño de la muestra está dado por la ecuación (8-17),

$$n = \frac{Nz^2\sigma^2}{d^2(N-1) + z^2\sigma^2} \tag{8-17}$$

Para efectos de una planeación económica en cierta zona del país, es necesario estimar entre 10,000 establos lecheros, el número de vacas lecheras por establo con un error de estimación de 4 y un nivel de confianza del 95%. Si se sabe que $\sigma^2 = 1{,}000$. ¿Cuántos establos deben visitarse para satisfacer estos requerimientos?
Aplicamos (8-17) y para ello determinamos los elementos que la constituyen.

$N = 10{,}000$, $z = 1.96$, $\sigma^2 = 1{,}000$, $d = 4$. Al reemplazar se tiene

$n = \frac{10{,}000(1.96)^2(1{,}000)}{9{,}999(16) + (1.96)^2(1{,}000)} = 234.5 \Longrightarrow n = 235$ es el número de establos que deben escogerse para el estudio.

Si omitimos el tamaño de la población, caso en el cual volveríamos a la ecuación (8-16), entonces el tamaño de la muestra sería

$n = \frac{(1.96)^2(1{,}000)}{(4)^2} = 240$

En la práctica la fórmula (8-17) se emplea cuando el tamaño n de muestra presupuestado comparado con el tamaño N de la población es tal que $\frac{n}{N} > 0.05$

Como puede observarse en cualquiera de las dos ecuaciones (8-16) o (8-17), en ellas aparece el valor de la varianza; esto quiere decir que para poderlas aplicar es

necesario conocer dicho valor. Sin embargo, lo más frecuente es que sea desconocido; cuando esto sucede, la varianza es estimada por cualquiera de los medios siguientes:

1. Se toma una muestra preliminar llamada muestra piloto y la varianza se estima mediante el conocido estimador $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Este valor estimado se toma como valor de varianza y se reemplaza en la fórmula (8-16) o (8-17), según el caso, para estimar el tamaño de muestra posterior que ha de tomarse. En el caso de que el tamaño de la muestra piloto sea inferior a 30 se recomienda emplear el valor t en lugar del valor normal.
2. Se utilizan estimaciones previas que se hayan hecho acerca de la varianza en estudios anteriores.
3. Si existe evidencia de que la población estudiada tiene distribución normal, estimar σ mediante la fórmula de aproximación $\sigma = \frac{A}{4}$, siendo A la amplitud o rango de la población. Este método requiere del conocimiento del valor máximo y mínimo de la varianza investigada.

Una máquina llena cajas con cierto cereal. El supervisor desea conocer con un error de estimación de máximo 0.1 y un nivel de confianza del 90%, una media estimada del peso. Como la varianza era desconocida se procedió a escoger una muestra piloto para estimarla. Los resultados fueron los siguientes: 11.02, 11.14, 10.78, 11.59, 11.58, 11.19, 11.71, 11.27, 10.93, 10.94. ¿Cuántas cajas debe escoger para que se cumplan los requisitos propuestos?

Como el tamaño de la población es desconocido y la muestra piloto es de un tamaño inferior a 30, utilizaremos la fórmula $n = \frac{t^2S^2}{d^2}$.

Al hacer los cálculos se tiene $\bar{x} = 11.22$, $S^2 = 0.1$, $t_{(9,0.95)} = 1.8331$. Al reemplazar tenemos,

$n = \frac{(1.8331)^2(0.1)}{(0.1)^2} = 33.6$. Esto es, debe escoger 34 cajas.

Un transportador está interesado en conocer con un error de estimación de 0.02 y con un nivel de confianza del 99%, el peso promedio de una determinada clase de pescado que debe transportar. Experiencias pasadas le permiten suponer que el peso mínimo es de 1.48 libras y la máxima es de 2.47 libras por pescado. ¿De qué tamaño debe escoger la muestra? Suponga que los pesos de estos pescados se distribuyen normalmente.

Como el tamaño de la población es desconocido, lo mismo que la varianza, utilizamos la fórmula (8-16) previa estimación de σ como $\frac{2.47 - 1.48}{4} = 0.2475$. Al reemplazar en la citada fórmula se tiene,

$n = \frac{(2.575)^2(0.2475)^2}{(0.02)^2} = 1{,}015$ como tamaño de muestra indicado.

No se sorprenda por este valor si le parece alto. Tenga en cuenta el alto nivel de confianza y el bajo error de estimación.

En algunos casos el parámetro de interés se centra en la proporción poblacional. Por tanto, en estas circunstancias requerimos de la determinación del tamaño de la

muestra para proporciones. Cuando este es el caso, el tamaño de muestra está dado por:

$$n = \frac{z^2\bar{p}(1 - \bar{p})}{d^2} \tag{8-18}$$

en donde $\bar{p}$ corresponde a la proporción estimada, d al error de estimación, y la población se considera infinita.

Se está planeando una encuesta con el fin de determinar la proporción de familias que carecen de medios económicos para atender los problemas de salud. Existe la impresión de que esta proporción está próxima a 0.35. Se desea determinar un intervalo de confianza del 95% con un error de estimación de 0.05. ¿De qué tamaño debe tomarse la muestra?

En este caso tomamos $\bar{p} = 0.35$ (que es un valor estimado para π), $z = 1.96$. Al reemplazar en (8-18) se tiene $n = \frac{(1.96)^2(0.35)(0.65)}{(0.05)^2} = 350$. Por tanto, el tamaño de muestra indicado es $n = 350$ familias.

Cuando no se da estimación alguna para π, el cálculo de la muestra se hace mediante la fórmula (8-18) tomando $\bar{p} = 0.5$. Esto arroja por lo general una muestra mucho mayor de la indicada, pero es el precio que debemos pagar por no tener mayor información sobre el caso.

Un productor de semillas desea saber, con un error de estimación de 1% el porcentaje de las semillas que germinan en la granja de su principal competidor. ¿Qué tamaño de muestra debe tomarse para obtener un nivel de confianza del 95%.

Como no tenemos ninguna estimación de la proporción tomamos $\bar{p} = 0.5$ y al reemplazar en (8-18) se obtiene, $n = \frac{(1.96)^2(0.5)(0.5)}{(0.01)^2} = 9{,}604$. Así que debe escoger 9,604 semillas para probar.

Si el tamaño de la población debe ser tenido en cuenta el tamaño de muestra está dado por

$$n = \frac{Nz^2\bar{p}(1 - \bar{p})}{(N - 1)d^2 + z^2\bar{p}(1 - \bar{p})} \tag{8-19}$$

El decano de una facultad desea realizar una encuesta para determinar la proporción de estudiantes que está a favor del cambio de sede. Ya que entrevistar $N = 2{,}000$ estudiantes en un lapso razonable es casi imposible, determine el tamaño de muestra (número de estudiantes a entrevistar) necesario para estimar la proporción de estudiantes que están a favor, con un error de estimación de 0.05 y un nivel de confianza del 95%.

De nuevo observe que no tenemos información sobre algún estimador de la proporción.

Al aplicar la fórmula (8-19) se tiene, $n = \frac{2{,}000(1.96)^2(0.5)(0.5)}{(2{,}000 - 1)(0.05)^2 + (1.96)^2(0.5)(0.5)} = 322$. Por tanto, debe entrevistar a 322 estudiantes.

Ejercicios 8.3

1. Suponga que las estaturas de los hombres tienen distribución normal con desviación estándar de 2.5 pulgadas. ¿De qué tamaño se debe tomar la muestra si se desea determinar un intervalo de confianza del 95% para la media con un error de estimación 0.5?

2. Un químico ha preparado un producto diseñado para matar el 80% de un tipo particular de insectos; ¿de qué tamaño debe escoger la muestra para estimar la verdadera proporción si se requiere un intervalo del 95% y un error de estimación del 2%?

3. El rector de una universidad particular desea estimar el costo promedio de un año de estudios con un error de estimación menor de $20,000 y con una probabilidad de 0.95. Suponga que el administrador sólo sabe que el costo varía entre $20,000 y $30,000. ¿Cuántos estudiantes debe seleccionar?

4. Un técnico desea determinar el tiempo promedio que los operarios tardan en preparar sus equipos. ¿Qué tamaño debe tener la muestra si se necesita una confianza del 95% de que su media muestral estará dentro de 15 segundos del promedio real? Suponga que por estudios anteriores se sabe que $\sigma = 45$ segundos.

5. Se desea estimar el peso promedio de un lote de 500 naranjas. Para ello se va a escoger aleatoriamente cierto número de naranjas. Se desea que el error de estimación sea máximo de 2 onzas con un nivel de confianza del 90%. ¿Cuántas naranjas deben seleccionarse? Suponga que $\sigma = 5$.

6. Se desea estimar la proporción de estudiantes que están a favor de la legalización de las drogas prohibidas. El error de estimación se requiere del 1% y un nivel de confianza del 99%. ¿Cuántos estudiantes deben incluirse en la muestra?

7. Se desea estimar la fuerza promedio requerida para levantar un niño de seis años. Como no se tenía información sobre la varianza de esta población se procedió a tomar una muestra piloto para estimarla; los resultados fueron los siguientes: 2.24, 2.26, 2.47, 1.56, 1.72, 1.48, 2.40, 2.03, 1.72, 2.10, 1.74, 1.55. Si se desea estimar un intervalo del 95% de confianza con un error de estimación de 0.1. ¿De qué tamaño se debe escoger la muestra?

8. El jefe de personal de una empresa desea realizar una encuesta para determinar la proporción de trabajadores que está a favor de un cambio en el horario de trabajo. Como es imposible consultar a los $N = 500$ trabajadores en un lapso razonable, procede a escoger aleatoriamente cierto número de trabajadores para entrevistarlos; determine el número de trabajadores que debe entrevistarse si desea que la proporción estimada presente un error máximo del 5% y un nivel de confianza del 95%.