Capítulo 12

LA INFERENCIA ESTADÍSTICA

Sumario

Objetivos específicos

Al finalizar el estudio del capítulo, el estudiante será capaz de:

1. Explicar por qué se requiere trabajar con muestras y en qué consiste la diferencia estadística, en particular, el proceso de estimación de los valores poblacionales, y qué requisitos deben estar presentes para hacer inferencias válidas.
2. Explicar qué son los valores poblacionales y los estimadores, y desde qué punto de vista son variables aleatorias.
3. Realizar la estimación de la media poblacional μ y calcular e interpretar los límites de confianza para μ.
4. Explicar cuáles son los factores que determinan el tamaño n de la muestra necesaria para estimar μ, y realizar el cálculo bajo ciertas especificaciones.
5. Realizar la estimación del valor poblacional P y calcular e interpretar los límites de confianza para P.
6. Calcular el tamaño requerido de la muestra para estimar P.

Resumen

En este tema se señalan las razones que justifican la utilización de muestras, se introduce y explica el concepto de inferencia estadística, se discuten e ilustran las técnicas empleadas para realizar estimaciones de los valores poblacionales y fijar el tamaño de muestra; además, se dan algunas ideas acerca de la utilidad de estas técnicas y sus limitaciones.

12.1. POBLACIÓN, MUESTRA E INFERENCIA ESTADÍSTICA

Los conceptos de población, muestra e inferencia estadística ya fueron discutidos con cierta amplitud; no es necesario, entonces, volver a tratarlos con todo detalle, únicamente se hará un resumen de las ideas básicas.[1]

En todo estudio o investigación existe un campo de referencia, universo o población al cual se desea generalizar los resultados o conclusiones que se obtengan. Por población o universo se entiende el total de personas, objetos o mediciones con una característica común. Más precisamente, la población la constituyen los valores numéricos asociados con las personas, objetos o mediciones; por lo tanto, una manera de definir la población es como "el total de todas las observaciones correspondientes a una cierta característica".

Las poblaciones pueden ser **finitas** o **infinitas**. En términos generales si se va a estudiar una situación existente en un momento dado, la población quizás puede ser muy grande, pero necesariamente finita. En cambio, si se analiza un proceso, el cual teóricamente puede repetirse indefinidamente bajo las mismas condiciones, la población es infinita. También, se utilizan los términos **población real** y **población hipotética** o **conceptual**, según sea que tenga existencia real o se trate simplemente de algo concebido por el investigador. Para efectos prácticos, por razones de conveniencia, en muchos casos se trabaja con poblaciones finitas como si se tratara de poblaciones infinitas. Esto facilita las aplicaciones de las técnicas y no introduce errores de consideración en los resultados.

Las investigaciones y experimentos se llevan a cabo con el propósito de llegar a conclusiones o leyes de carácter general, es decir, buscando resultados que se puedan generalizar

1. Leer cuidadosamente de las secciones 1.5 a 1.8, y la 1.10 del capítulo 1, "La Naturaleza de la Estadística", donde se discuten con mayor detalle los conceptos tratados en esta sección.

a todos los elementos de la población estudiada. Para este propósito, lo apropiado es estudiar, observar o "censar" todos los elementos que constituyen la población. En la mayoría de las situaciones, sin embargo, no se hace, más bien se toma una muestra y sus resultados se generalizan a toda la población.

¿Por qué se utilizan muestras en lugar de estudiar toda la población? Por varias razones: la población es infinita o muy grande y es imposible estudiarla toda, resulta muy costoso o tomaría mucho tiempo abarcarla en su totalidad, los elementos se transforman o destruyen al ser estudiados. Sin embargo, la razón más importante es el hecho de que, salvo situaciones especiales, no es necesario estudiar a toda la población, ya que los resultados de una muestra bien seleccionada, de tamaño razonable, serán suficientemente precisos para alcanzar los fines prácticos, perseguidos con la recolección de los datos.

Se entiende por **inferencia estadística** el proceso mediante el cual se generalizan los resultados observados, en una **muestra aleatoria**, a la población de la cual se extrajo y se evalúa el error asociado con esa generalización. La inferencia estadística supone que la muestra es probabilística o aleatoria, es decir, ha sido seleccionada utilizando un mecanismo que da, a cada elemento de la población, una probabilidad conocida de ser incluido en la muestra. Al ser las muestras aleatorias, su composición la determina el azar y los errores de muestreo –diferencia entre el valor estimado con la muestra y el valor verdadero o poblacional– son aleatorios, tienen una distribución probabilística y es posible aplicar la teoría de las probabilidades para medir la confiabilidad de las inferencias.[2]

Si la muestra es seleccionada en forma intencional o de juicio, los errores no tendrán un comportamiento aleatorio, sino que dependerán de los prejuicios y tendencias conscientes o inconscientes de quien hizo la selección, es decir, serán sesgados, y por ello no es posible aplicar la teoría de las probabilidades para evaluar las inferencias o generalizaciones; por lo tanto, la inferencia estadística no puede ser aplicada a las muestras no aleatorias.[3]

La preferencia por el uso de muestras aleatorias para la realización de inferencias obedece a las siguientes tres razones básicas:

- La selección aleatoria elimina los sesgos de selección.
- Produce errores aleatorios que son medibles utilizando modelos probabilísticos.
- El error de muestreo puede hacerse tan pequeño como se quiera, aumentando el tamaño de la muestra.

2. La muestras probabilísticas o aleatorias son **medibles**, es decir, son diseñadas de forma tal que es posible medir la confiabilidad de las inferencias a partir de los propios datos muestrales.
3. Obviamente, la forma más simple de determinar el error de muestreo es comparando el valor muestral con el valor poblacional, pero si este último se conociera, no tendría sentido realizar el estudio ni utilizar una muestra.

La selección aleatoria no implica igual probabilidad, hay esquemas de muestreo en el cual todos los elementos reciben la misma probabilidad de entrar en la muestra (métodos de selección con igual probabilidad) y hay otros en los que la selección se hace con probabilidades desiguales para los diferentes elementos de la población. Seleccionar con probabilidades desiguales no hace que la muestra deje de ser aleatoria, solo implica que, al calcular las estimaciones, se debe tener el cuidado de usar "ponderaciones" que compensen la desigual probabilidad utilizada en la selección.

Así, por ejemplo, si en una ciudad hay 12 000 familias con teléfono de línea fija y 8000 sin teléfono, y se seleccionan al azar 200 familias con teléfono y 200 sin teléfono, la muestra total de 400 es aleatoria, pero no de igual probabilidad, ya que las familias con teléfono tuvieron una probabilidad de $\frac{200}{12\,000} = \frac{1}{60}$ y aquellas sin teléfono una de $\frac{200}{8000} = \frac{1}{40}$. Aunque cualquier estimación realizada con la muestra total de 400 familias requiere una ponderación adecuada para tomar en cuenta esa desigual probabilidad, la muestra es aleatoria y permite efectuar inferencias estadísticas.

Dentro de la inferencia estadística se distinguen dos tipos de problemas: uno en el cual interesa **estimar** características o propiedades de la población (su promedio, su variancia, la proporción de elementos que tienen una cierta característica) y otro en el que es importante someter a **prueba hipótesis** que se tienen acerca de esas características de la población. Este capítulo se concentrará en el primer problema, o sea, en el de la estimación, la cual, puede ser puntual o por intervalo como se verá oportunamente.

$$\text{Inferencia estadística} \begin{cases} \text{Estimulación} \begin{cases} \text{Puntual} \\ \text{Por intervalo} \end{cases} \\ \text{Prueba de hipótesis} \end{cases}$$

12.2. VALORES DE LA POBLACIÓN Y ESTIMADORES

Las medidas que representan características o propiedades de la población se denominan valores de la población,[4] como ejemplos pueden citarse:

μ = Media aritmética o promedio de la población.

σ^2 = Variancia de la población.

R = Recorrido de la población.

Me = Mediana de la población.

P = Proporción de los elementos de la población que tienen una cierta característica.

n = Coeficiente poblacional de correlación lineal entre dos variables.

Usualmente, los valores poblacionales no son conocidos y una de las tareas básicas de la inferencia estadística es su **estimación**, la cual se entiende como el cálculo, a partir de los datos de la muestra, de un valor numérico que será considerado, para efectos prácticos, como una aproximación o estimador del valor de la población.

Un estimador entonces es una función de los valores de la muestra utilizada para aproximar o estimar el valor poblacional; por estimación se entiende tanto el valor numérico producido por una muestra específica como el proceso de cálculo de ese valor y su respectiva generalización a la población.

Con frecuencia, un valor de la población puede ser estimado de varias maneras. Por ejemplo μ, la media poblacional, podría ser estimada por:

i. La media de la muestra $\overline{X}$

ii. El promedio simple del valor mayor y el menor en la muestra

iii. La mediana de la muestra (me)

Sin embargo, los estadísticos procuran seleccionar estimadores que reúnan la mayoría de una serie de requisitos considerados deseables, algunos de los cuales se comentan seguidamente.

Insesgado: uno de los requisitos deseables de un estimador es la calidad de insesgado. Por esto, se quiere indicar que el estimador debe ser tal que si se calcula para todas las muestras posibles de un cierto tamaño y luego se hace un promedio de los valores resultantes, este será igual al valor poblacional.

4. Un término que también se utiliza, con cierta frecuencia, es el de "parámetros". Aquí se ha preferido "valores de la población" por considerar que transmite una mejor idea del concepto.

Un ejemplo de un estimador insesgado es la media muestral $\overline{X}$. Si de una población se toman todas las muestras posibles de un tamaño dado n, y se calcula para cada una de ellas $\overline{X}$, puede comprobarse que el promedio de todos los $\overline{X}$ es igual a μ. Esto sucede porque el promedio de la muestra es un estimador insesgado de μ.

La expresión

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2,$$

que intuitivamente, por su estructura, parece un estimador insesgado de σ^2, no lo es. Como ya se mencionó,[5] el estimador insesgado de σ^2 es s^2, cuya expresión algebraica es:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2.$$

Note que debe usarse $n-1$ en lugar de n para tener una estimación insesgada de la variancia de la población σ^2, a partir de una muestra de tamaño n.

Consistente: otra condición deseable en un estimador es la consistencia, o sea, la propiedad de que la estimación se aproxime al valor de la población con probabilidad que tiende a 1, conforme aumenta el tamaño de la muestra. Si un estimador es consistente, aumentando apropiadamente el tamaño de la muestra, se puede estar casi seguro de que el valor estimado está muy cerca del de la población.

Por ejemplo, en el caso de $\overline{X}$, a medida que aumenta el tamaño de la muestra, $\overline{X}$ está cada vez más cerca de μ, es decir, la diferencia $\overline{X} - \mu$ tiende a cero al aumentar n.

Variancia mínima: una propiedad también importante es la de variancia mínima. La cual hace preferir, entre dos o más estimadores **insesgados** y **consistentes**, el que varíe menos de muestra a muestra, o sea, al más estable. Por ejemplo, en la distribución normal, por ser simétrica, μ = mediana; entonces, una forma de estimar μ es empleando el promedio $\overline{X}$ o la mediana de la muestra; sin embargo, corrientemente se prefiere utilizar $\overline{X}$ porque tiene mayor estabilidad en muestras repetidas.[6]

No todos los estimadores reúnen las tres condiciones citadas, algunos son insesgados, pero no tienen variancia mínima y viceversa. En la práctica, se trata de escoger aquellos que reúnan las mejores propiedades. Sin embargo, en ciertos casos, razones prácticas o de costo, pueden llevar a utilizar estimadores que no tengan todos los requisitos antes citados.

5. Ver sección 9.3.
6. En muestras muy pequeñas, la variabilidad de ambos estimadores es muy parecida, pero al aumentar n, la eficiencia de $\overline{X}$ sobre la mediana es clara. En muestras grandes, la variabilidad de la mediana de muestra a muestra es un 57% más alta que la de $\overline{X}$; por lo tanto, para estimar con igual precisión μ usando la mediana, se requiere una n un 57% mayor. En otras palabras, si la muestra requerida para estimar μ, bajo un cierto nivel de precisión utilizando $\overline{X}$ es 100, la muestra equivalente, si se quisiera usar la mediana como estimador, sería de alrededor de 157 elementos.

12.3. LOS ESTIMADORES COMO VARIABLES ALEATORIAS

Los estimadores son funciones de los valores de la muestra. Por ello, cuando se seleccionan diferentes muestras al azar de una población, es natural esperar que las estimaciones varíen de muestra a muestra. Por ejemplo, si se tiene la población formada por los estudiantes de una universidad y se empiezan a tomar muestras al azar de 25 estudiantes y para cada una de ellas se calcula el peso promedio, es natural que los promedios de esas muestras no sean iguales, sino que presenten variaciones. Es decir, $\overline{X}$ variará de muestra a muestra. Igual cosa sucederá con s^2 y con cualquier otro estimador. Además, como las muestras se seleccionan al azar, la variación será aleatoria.

Si los valores del estimador varían de muestra a muestra, es decir, si el estimador es una variable, es lógico que tenga una distribución y esta, a su vez, tenga su media aritmética, su mediana, su variancia, etcétera.

Suponga que de una población se seleccionan muestras de tamaño n y para cada una se calcula $\overline{X}$. Se plantean varias preguntas acerca de la variable $\overline{X}$, entre ellas:

a) ¿Cuál es la distribución de $\overline{X}$?

b) ¿Cuál es la media de la distribución de $\overline{X}$? $E(\bar{X}) = ?$

c) ¿Cuál es la variancia de la distribución de $\overline{X}$? $E[(\overline{X} - E(\overline{X}))^2] = ?$

Para responder a estas preguntas, es necesario tomar en cuenta si la población es finita o infinita y si el muestreo al azar es con reemplazo o sin reemplazo.

En el muestreo **con reemplazo,** los elementos de la muestra son seleccionados de uno en uno, pero el seleccionado es devuelto a la población (reemplazado) antes de la siguiente selección y, por lo tanto, participa en ella. En el muestreo **sin reemplazo,** por el contrario, el elemento seleccionado no es devuelto a la población (permanece afuera) y no participa en las siguientes selecciones.

Aunque el muestreo con reemplazo no presenta dificultades especiales, en la práctica se utiliza, casi exclusivamente, el muestreo sin reemplazo. Se procede así porque no hay ninguna ventaja en tener dos o más veces el mismo elemento en la muestra y, porque al trabajar sin reemplazo, la muestra incluye el máximo de elementos diferentes y resulta, por ello, más representativa. Realmente, la utilidad del concepto de muestreo con reemplazo es eminentemente teórica y cumple, en la teoría de muestreo, una función similar al de la competencia pura y perfecta en la teoría económica.

Considere, como ilustración, una población formada por cinco escuelas de una zona rural: A, B, C, D, E, con los siguientes números de maestros:

$$\text{A: 2, B: 3, C: 4, D: 5, E: 6, } \mu = 4, \ \sigma_X^2 = 2.$$

Suponga que se toman muestras sin reemplazo de tamaño dos y, para cada una, se calcula el número promedio de maestros. ¿Cuántas muestras diferentes son posibles? Las diez siguientes con sus promedios:

Muestras: (A,B) (A,C) (A,D) (A,E) (B,C) (B,D) (B,E) (C,D) (C,E) (D,E)

Composición: (2,3); (2,4); (2,5); (2,6); (3,4); (3,5); (3,6); (4,5); (4,6); (5,6)

Promedios: 2,5; 3,0; 3,5; 4,0; 3,5; 4,0; 4,5; 4,5; 5,0; 5,5

Puede apreciarse que los promedios varían de muestra en muestra entre 2,5 y 5,5 con la siguiente distribución:

2,5			
3,0			
3,5	3,5		
4,0	4,0	$\mu = 4,0$	$\sigma^2_{\bar{X}} = 0,75$
4,5	4,5		
5,0			
5,5			

Y si se toman muestras de $n = 3$, ¿cuántas muestras se obtienen y cómo variarán sus promedios? Las diez siguientes con los promedios y la distribución indicada:

(A,B,C)	(A,B,D)	(A,B,E)	(A,C,D)	(A,C,E)	(A,D,E)	(B,C,D)	(B,C,E)	(B,D,E)	(C,D,E)
3,0	3,3	3,7	3,7	4,0	4,3	4,0	4,3	4,7	5,0
3,0									
3,3									
3,7	3,7								
4,0	4,0				$\mu = 4$			$\sigma^{2\bar{X}} = 0,33$	
4,3	4,3								
4,7									
5,0									

Los ejemplos anteriores comprueban que $\overline{X}$ es una variable, la cual tiene su distribución con su media y su variancia. Indican, además, al ser la muestra más grande, la variación de los promedios dentro de un intervalo más pequeño, es decir, están más concentrados. Esto se refleja en la variancia σ^2, la cual se reduce de 0,75 (con muestras de tamaño 2) a 0,33 (con muestras de tamaño 3). La μ, por el contrario, no varía y se mantiene igual a la media de la población (esto no debe sorprender porque $\overline{X}$ es un estimador insesgado (μ).

¿Ahora bien, cuáles son la media y la varianza de la distribución de $\overline{X}$?

Valor esperado de $\overline{X}$: cualquiera que sea la situación: muestreo con reemplazo o sin reemplazo, población finita o infinita, el valor esperado $E(\bar{X})$ de la media muestral, o sea, es igual a la media de la población:

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

Variancia de $\overline{X}$: la variancia de la variable $\overline{X}$ designada por $\sigma^2_{\bar{X}}$, sí es diferente según la situación que se considere en cuanto a la población y el tipo de muestreo.

Muestreo simple al azar sin reemplazo de una población finita:

$$\sigma^2_{\bar{X}} = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n},$$

donde,

N = Tamaño de la población

σ^2 = Variancia de la población

n = Tamaño de la muestra

El factor $\frac{N-n}{N-1}$ se denomina **factor de corrección para poblaciones finitas.**

Muestreo simple al azar con reemplazo de una población finita o muestreo simple al azar de una población infinita:

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}.$$

Como puede notarse, cuando se trabaja con muestreo simple al azar con reemplazo en una población finita o cuando la población es infinita, el factor de corrección se omite y la variancia de $\overline{X}$ se reduce a $\frac{\sigma^2}{n}$.[7]

12.4. DISTRIBUCIÓN DE LA MEDIA MUESTRAL ($\overline{X}$)

Conocida la media y la variancia de $\overline{X}$, solo queda por determinar su distribución. Esta es especificada en el teorema del límite central.

7. Otro caso en el cual puede omitirse el factor de corrección y usar la expresión más simple $\frac{\sigma^2}{n}$ es cuando n es pequeña en relación con N, ya que, entonces, el valor de $\frac{N-n}{N-1}$ es muy cercano a 1 y su omisión no produce un error de importancia.

12.4.1. Teorema del límite central

Si una población cualquiera (continua o discreta) tiene media μ y variancia finita σ^2, la media muestral $\overline{X}$ poseerá una distribución que se aproxima a la distribución normal, con media μ y variancia $\frac{\sigma^2}{n}$ al aumentar n el tamaño de la muestra.

En realidad hay dos situaciones:

Población muestreada, es decir, la población de donde se saca la muestra es normal o muy cercana a la normal. En este caso, la distribución de $\overline{X}$ será normal prácticamente para cualquier tamaño de muestra.

La población muestreada tiene una distribución que no es normal. En este caso, todo es cuestión de tomar una muestra suficientemente grande para lograr la aproximación.

De acuerdo con el teorema, cualquiera que sea la forma de la distribución de la población (la distribución "madre"), la de la media muestral $\overline{X}$ tenderá a la normal al aumentar el tamaño de la muestra. Este resultado es sumamente importante porque capacita para utilizar la distribución normal al hacer inferencias en una gran cantidad de situaciones, con solo tener una muestra suficientemente grande.

En la práctica, una gran cantidad de fenómenos sigue una distribución muy cercana a la normal o a una distribución unimodal con bastante acumulación en la parte central, de manera que, generalmente, es correcto usar la normal para realizar inferencias acerca del promedio aunque la muestra sea moderadamente pequeña; esta es una de las principales razones por las cuales la distribución normal es tan importante en estadística. Además, en muestras grandes, la distribución de $\overline{X}$ es prácticamente normal, aún cuando la población de la cual fue tomada la muestra fuera bastante asimétrica.

Finalmente, aunque una buena aproximación en la zona central, alrededor del promedio, se logra por lo general con un tamaño de muestra relativamente pequeño, no sucede lo mismo en las colas, donde se requiere un tamaño de muestra mucho mayor para obtener suficiente cercanía entre la distribución de $\overline{X}$ y la curva normal.

12.5. EL ERROR ESTÁNDAR DEL PROMEDIO

La desviación estándar de $\overline{X}$, es decir, $\frac{\sigma}{\sqrt{n}}$, se denomina error estándar del promedio y se indica con $\sigma_{\overline{X}}$.

El error estándar mide la variabilidad de $\overline{X}$ de muestra a muestra, y por ello puede emplearse para medir la confianza que merece $\overline{X}$ como estimador de μ. Evidentemente, cuanto mayor sea el error estándar, mayor es la probabilidad de que los valores de $\overline{X}$

se alejen de μ, y menor será la confianza que puede tenerse en que el $\overline{X}$ dado por una muestra específica esté cerca de μ.

Dos son los factores que determinan el valor de $\sigma_{\overline{X}}$: la variancia de la población (σ^2) y el tamaño de la muestra (n). En la práctica, el investigador puede reducir σ^2 utilizando un diseño experimental o un diseño muestral apropiado, pero una vez planteado su estudio o investigación, el único camino abierto para reducir el error estándar del promedio es aumentar el tamaño de la muestra.

Como puede apreciarse en el gráfico de la figura 12.1, al aumentar n, $\sigma_{\overline{X}}$ disminuye, al principio sensiblemente, pero luego cada vez más lento. Llega un momento en el cual solo es posible lograr una reducción de alguna importancia en el error estándar, con un aumento tan grande en la muestra que es mejor no intentarlo por el gasto y esfuerzo adicional requerido.

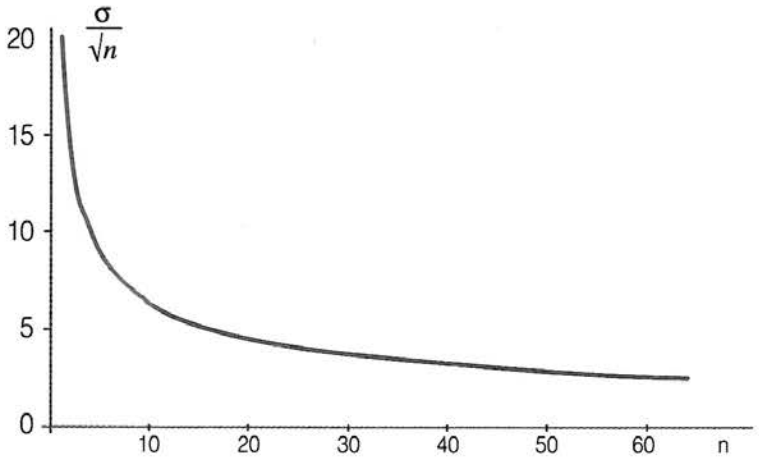


Figura 12.1. Gráfico de la evolución del error estándar al aumentar el tamaño de la muestra (n) para una población con $\sigma = 20$

12.6. ESTIMACIÓN DE LA MEDIDA μ

Para estimar μ puede tomarse simplemente $\overline{X}$. Esta sería una estimación puntual de μ. Por ejemplo, si se quiere estimar el peso promedio de los estudiantes de la Universidad Estatal a Distancia, puede tomarse una muestra de 25 alumnos, calcular su peso promedio, y tomar ese valor como el promedio de la población.

Es evidente, sin embargo, la utilidad que tendría contar con una idea acerca del error asociado con la estimación hecha. La estimación puntual es insuficiente si no se posee alguna medida de la confianza que merece. El problema se resuelve utilizando la estimación por intervalos, o sea, construyendo un intervalo, a partir de la estimación puntual,

dentro de cuyos límites se espera, con un determinado grado de confianza, en el cual esté contenido el verdadero valor poblacional.

Suponga que se toman 25 alumnos al azar y se obtiene $\overline{X} = 145$ libras. La estimación 145 puede estar muy alejada de μ o poco alejada, esto no se puede saber. Para superar esta dificultad, incertidumbre, puede usarse la estimación por intervalo. Si se afirma que el peso promedio de los estudiantes universitarios está entre 0 y 400 libras, no hay posibilidad real de error, el promedio no puede ser inferior a 0 y es prácticamente imposible que sea mayor que 400 libras. La probabilidad de que la afirmación esté equivocada es 0.

En cambio, si se afirma que el promedio está entre 141 y 149 libras, entonces existe probabilidad de error (μ podría ser, por ejemplo, igual a 139). Si se dice que μ está entre 144 y 146, existirá aún mayor probabilidad de fallar, pues el intervalo es mucho más estrecho.

Evidentemente, entre más amplio sea el intervalo, mayor será la confianza de que contenga a μ y, por el contrario, entre más estrecho se defina, menor será la probabilidad de que lo abarque y mayor la de equivocarse al decir que contiene a μ.

Lógicamente, para que un intervalo diera completa confianza de contener el valor poblacional debería ser demasiado amplio y carecería de utilidad práctica; por ello, se define un intervalo razonablemente estrecho para fines prácticos y se acepta un cierto grado de probabilidad de que no contenga el valor de la población. En seguida, se ilustra la construcción de un intervalo de confianza para μ.

12.7. LÍMITES DE CONFIANZA PARA μ[8]

Suponga que se tiene una población normal con media μ y variancia σ^2, y de ella se toman muestras al azar de tamaño n.

¿Cuál es la distribución de la media muestral $\overline{X}$? Por el teorema del límite central, se sabe que $\overline{X}$ se distribuye normalmente con media μ y variancia $\frac{\sigma^2}{n}$

8. Antes de comenzar el estudio de este subtema, se recomienda revisar las secciones del capítulo 11 que se refieren a la curva normal.

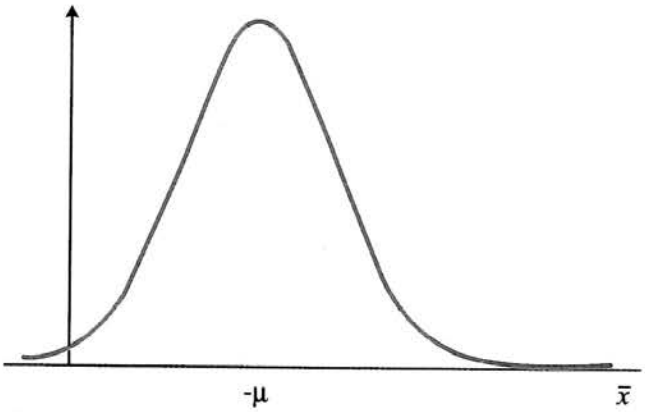


Figura 12.2. Gráfico que muestra la distribución de la media muestral con $\sigma_{\bar{x}}^2 = \sigma^2/n$

Otra pregunta que puede plantearse es: ¿dentro de qué rango se encontrará el 95% de los valores de $\overline{X}$?

Se sabe, por la teoría estadística, que en una curva normal estándar el 95% central del área bajo la curva está entre–1,96 y +1,96; por lo tanto, en la distribución de $\overline{X}$, el 95% de los valores se encontrará entre $\mu - 1,96\sigma_{\bar{x}}$ y $\mu + 1,96\sigma_{\bar{x}}$.

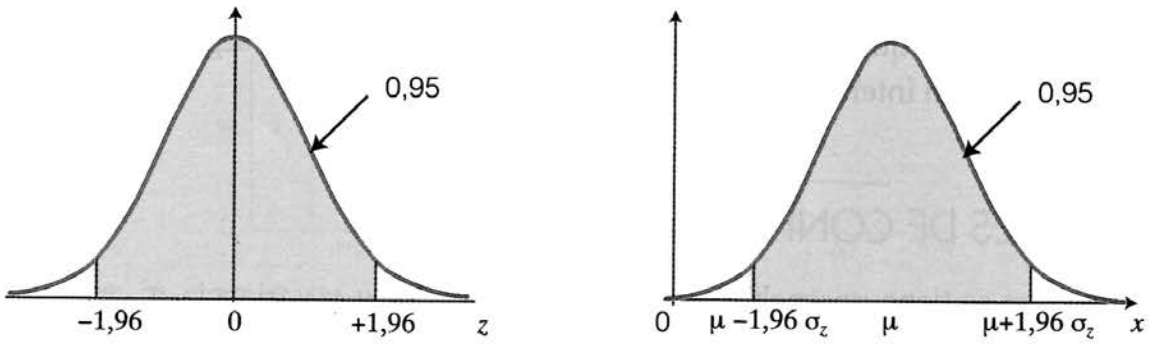


Figura 12.3. Gráfico de la curva normal estándar

La probabilidad de que $\overline{X}$ asuma un valor en el intervalo $\mu \pm 1,96\sigma_{\bar{x}}$ es de 0,95. Esto puede expresarse simbólicamente en la siguiente forma:

$$P[\mu - 1,96\sigma_{\overline{x}} \leq \overline{X} \leq \mu + 1,96\sigma_{\overline{x}}] = 0,95.$$

De la expresión anterior puede derivarse una fórmula para el cálculo de un intervalo de confianza del 95% para μ.

Restando μ en los componentes de la desigualdad:

$$P[-1,96\sigma_{\overline{x}} \leq \overline{X} - \mu \leq +1,96\sigma_{\overline{x}}] = 0,95.$$

Restando $\overline{X}$:

$$P[-\overline{X} - 1{,}96\sigma_{\overline{X}} \leq -\mu \leq \overline{X} + 1{,}96\sigma_{\overline{X}}] = 0{,}95.$$

Multiplicando por –1 la desigualdad:

$$P[\overline{X} + 1{,}96\sigma_{\overline{X}} \geq -\mu \geq \overline{X} - 1{,}96\sigma_{\overline{X}}] = 0{,}95.$$

Reordenando la desigualdad:

$$P[\overline{X} - 1{,}96\sigma_{\overline{X}} \leq -\mu \leq \overline{X} - 1{,}96\sigma_{\overline{X}}] = 0{,}95.$$

$$P[L_1 \leq \mu \leq L_2] = 0{,}95,$$

donde:

L_1 = límite inferior = $\bar{X} - 1{,}96\sigma_{\overline{X}}$.

L_2 = límite superior = $\bar{X} + 1{,}96\sigma_{\overline{X}}$.

Los límites anteriores son para un nivel de confianza de 95%. También pueden derivarse, en igual forma, para otros niveles como 99%, 90%, etc. El único cambio que debe introducirse en la expresión es sustituir el valor de la normal estándar $z = 1{,}96$, correspondiente a la probabilidad de 95%, por el valor de z del nivel de confianza deseado. A continuación, se presenta la fórmula general para el cálculo de los límites de confianza:

$$L_1 = \overline{X} \pm z_1 \frac{\sigma}{\sqrt{n}}.$$

$$L_1 = \overline{X} - z_1 \frac{\sigma}{\sqrt{n}}. \qquad P[L_1 \leq \mu \leq L_2] = 1 - a.$$

$$L_2 = \overline{X} + z_1 \frac{\sigma}{\sqrt{n}}.$$

Donde:

$1 - \alpha$: representa el grado de confianza de que el intervalo contiene el valor de la población μ.

z_t: es un valor de Z en la normal estándar, tal que, entre $-z_t$ y $+z_t$, queda comprendida $(1 - \alpha)\%$ del área bajo la curva.

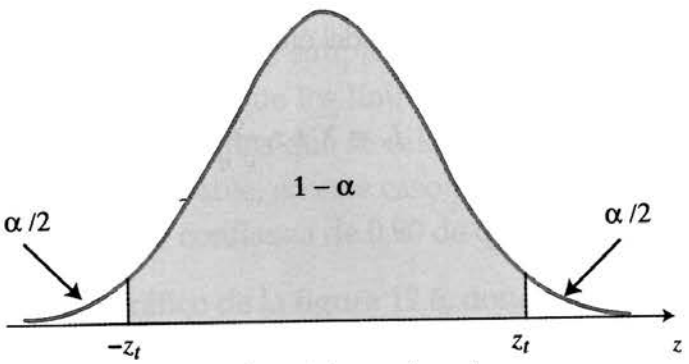


Figura 12.4. Gráfico de $(1 - \alpha)\%$ del área bajo la curva entre $-z_t$ y $+z_t$

Para calcular z_t, se determina $\frac{\alpha}{2}$, luego $1 - \frac{\alpha}{2}$ y, finalmente, se obtiene el valor de z_t consultando la tabla acumulada de la normal estándar y buscando el valor de z para el cual se encuentra acumulada un $\left(1 - \frac{\alpha}{2}\right)$ del área bajo la curva.

Ejemplo 1

Como ilustración, considere el caso antes mencionado del peso de los estudiantes en el cual, al tomarse una muestra al azar de 25, se obtuvo un peso promedio de 145 libras. Suponga que $\sigma^2 = 225$ y calcule un intervalo de confianza del 90%.

En este caso, $\overline{X} = 145$ y $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = 15 = \frac{15}{\sqrt{25}} = \frac{15}{5} = 3$.

$$1 - \alpha = 0,90$$

$$\alpha = 0,10$$

$$\frac{\alpha}{2} = 0,05$$

$$1 - \frac{\alpha}{2} = 0,95.$$

Si se consulta la tabla de la normal estándar, se encuentra que para $z = 1,645$ el área acumulada de la curva es 0,95 y, dada la simetría de la curva, hasta −1,645 hay un 5%; por ello, entre −1,645 y 1,645 se tiene un 90% del área bajo la curva.[9]

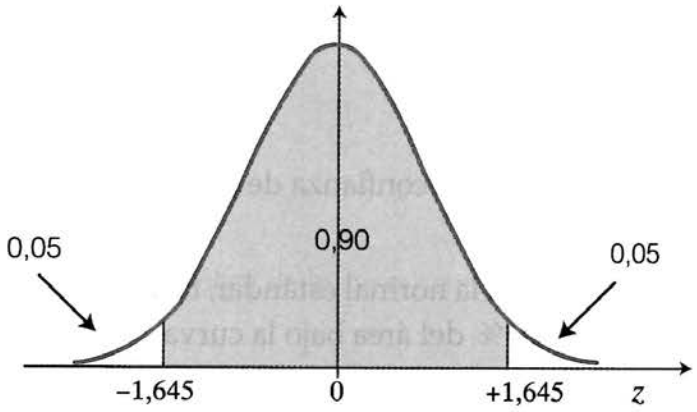


Figura 12.5. Gráfico que muestra el 90% del área bajo la curva entre −1,645 y 1,645

$$L_i = \overline{X} \pm z_t \frac{\sigma}{\sqrt{n}}$$

9. El valor de z también puede ser obtenido recurriendo directamente a la función DISTR.NORM. ESTAND:INV de la hoja de cálculo Excel.

$$L_1 = 145 - 1,645 \cdot 3 = 155 - 4,935 = 140,06$$
$$L_2 = 145 + 1,645 \cdot 3 = 155 + 4,935 = 149,94$$
$$P[140,06 \leq \mu \leq 149,94] = 0,90.$$

¿Cuál es el significado de esta expresión? Se ve a continuación.

12.7.1. Interpretación del intervalo de confianza

Como ha sido indicado, $\overline{X}$ varía de muestra a muestra y, en consecuencia, aunque para un problema dado z_t, $\sigma_{\bar{x}}$ y n son fijas, los límites también varían de muestra a muestra.

Resulta entonces que el intervalo $\overline{X} - 4,935$ a $\overline{X} + 4,935$ es aleatorio, es decir, cambia de muestra a muestra al variar $\overline{X}$. Por esto, 0,90 debe interpretarse en el sentido de que si se sacan muestras de tamaño 25 y para cada una de ellas se calcula un intervalo de confianza de 90%, alrededor de 90 de cada 100 de esos límites contendrán el valor poblacional μ.

En el caso particular del ejemplo considerado, en la expresión:

$$P[140,06 \leq \mu \leq 149,94] = 0,90,$$

parece que μ es la variable y el valor 0,90 significa la probabilidad de que μ esté entre 140,06 y 149,94 es de 0,90. Sin embargo, esto no tiene sentido, pues μ no es una variable sino un valor fijo, la media de la población de donde se ha sacado la muestra. La media poblacional μ está comprendida en el intervalo o no lo está, por lo tanto, solamente pueden hacerse las siguientes dos afirmaciones correctas:

$$P[140,06 \leq \mu \leq 149,94] = 0$$
o
$$P[140,06 \leq \mu \leq 149,94] = 1.$$

El valor 0,90 no es, en rigor, una probabilidad, pues μ está dentro del intervalo 140,06 a 149,94 o está fuera de él; 0,90 indica, simplemente, el grado de confianza que puede tenerse, en este caso específico, de que los límites calculados contengan a μ. En otras palabras, como de cada 100 intervalos que se calculan 90 contienen a μ (frecuencia de 90%), se considera bastante probable, en este caso particular, que μ esté entre 140,06 y 149,94, y se asigna un grado de confianza de 0,90 de que así sea.

Lo anterior se ilustra en el gráfico de la figura 12.6, donde se ha trazado una línea horizontal que representa a μ y se han dibujado líneas verticales correspondientes a límites

de confianza dados por muestras sucesivas. Los intervalos varían de posición, unos cubren a μ y otros no. La proporción del total de intervalos que, en un número grande de muestras, cubren a μ se representa por $(1-\alpha)$. En el caso específico antes considerado, se espera que un 90% de los intervalos contengan a μ y el resto no.

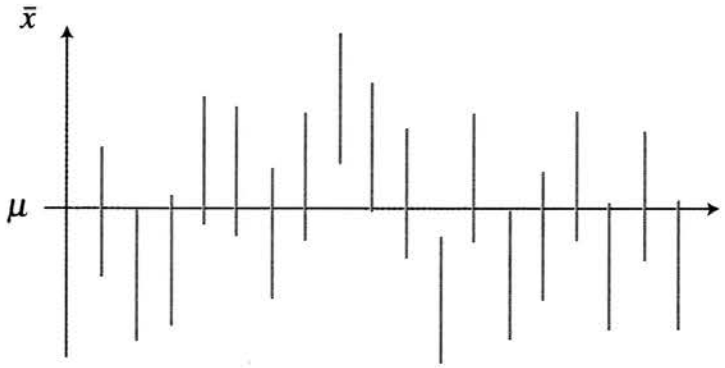


Figura 12.6. Gráfico que muestra el intervalo de confianza

Ejemplo 2

Se sabe, por experiencia, que el tiempo para completar con éxito una prueba de destreza en un curso de mecánica automotriz tiene $\sigma = 10$ minutos. Una muestra de 30 estudiantes de una universidad técnica, debidamente adiestrados, son sometidos a la prueba se obtiene $\overline{X} = 91$ minutos.

Compruebe que los límites de confianza del 97% para μ son 87 y 95.

Escriba una interpretación clara del significado del valor 0,97 en los límites calculados en la parte a).

Solución:

a) $1-\alpha = 0,97 \Rightarrow \alpha = 0,03$

$$\Rightarrow \frac{\alpha}{2} = 0,015$$

$$\Rightarrow z_{0,985} = 2,17$$

$$Li = \overline{X} \pm z_i \frac{\sigma}{\sqrt{n}} = 91 \pm (2,17) \cdot \frac{10}{\sqrt{30}}$$

$$= 91 \pm (2,17) \cdot \frac{10}{5,48}$$

$$= 91 \pm (2,17) \cdot (1,825) = 91 \pm 3,96 \approx 91 \div 4$$

$$P[87 \leq \mu \leq 95] = 0,97.$$

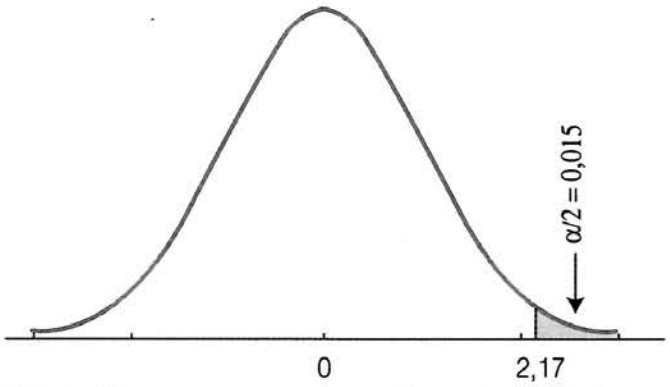


Figura 12.7. Gráfico que muestra la confianza de 97% de que μ es un valor comprendido entre 87 y 95

b) Si se toman muchas muestras de tamaño 30 y para cada una se calcula un intervalo de confianza del 97%, aproximadamente 97 de cada 100 intervalos contendrán μ.

Una forma "práctica" de interpretarlo es decir que se tiene una confianza de 97% de que μ es un valor comprendido entre 87 y 95.

Para este ejemplo específico, la interpretación sería: se tiene una confianza del 97% de que el tiempo promedio para completar exitosamente la prueba de destreza está entre 87 y 95 minutos.

12.8. EL TAMAÑO DE LA MUESTRA

La importancia del tamaño de la muestra en las investigaciones estadísticas se ha mencionado varias veces en este libro. Se ha señalado, también, que el tamaño de la muestra está determinado por factores como: variabilidad de la población, precisión deseada en las estimaciones, recursos disponibles para el estudio, etc. Ahora se está en capacidad de profundizar un poco más sobre el tema, considerando el problema para el caso de muestreo simple al azar.

Suponga que se quiere estimar el ingreso promedio mensual por familia en una gran ciudad. Por estudios anteriores, se estima que $\sigma = 300$ mil colones.

¿De qué tamaño debe ser la muestra para obtener una probabilidad de 95% de que la discrepancia entre $\overline{X}$ y μ (error de estimación) no será mayor de 20 mil colones?

Este problema se puede plantear simbólicamente en la siguiente forma:

$$P[|\overline{X} - \mu| \leq 20] = 0,95$$
$$P[-20 \leq \overline{X} - \mu \leq 20] = 0,95.$$

Por el teorema del límite central, se sabe que $\overline{X}$ tiene distribución normal con media μ, y variancia $\frac{\sigma^2}{n}$. Además, se conoce que $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = z$ es una variable normal estándar, por lo tanto, el problema se puede resolver dividiendo en la desigualdad anterior todos los miembros por $\frac{\sigma}{\sqrt{n}}$ para lograr una variable z con distribución normal estándar:

$$P\left[\frac{-20}{300/\sqrt{n}} \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq \frac{20}{300/\sqrt{n}}\right] = 0,95$$

$$P\left[\frac{-20}{300/\sqrt{n}} \leq z \leq \frac{20}{300/\sqrt{n}}\right] = 0,95$$

En la normal estándar se sabe que

$$P[-1,96 \leq z \leq 1,96] = 0,95.$$

De lo anterior, para satisfacer la condición deseada, se deduce que debe cumplirse la relación $\frac{20}{300/\sqrt{n}} = 1,96$ y, por tanto,

$$\sqrt{n} - \frac{1,96 - 300}{20}$$

$$n = \left[\frac{1,96 - 300}{20}\right]^2 \quad (*)$$

$$n = (1,96 - 15)^2 = (29,4)^2 = 864,36 = 865.$$

Los cálculos señalan que debe tomarse como mínimo una muestra de 865 familias.

Analizando los componentes de (*) puede establecerse una fórmula general para el cálculo del tamaño de muestra, cuando se utiliza muestreo simple al azar. Esta se incluye seguidamente:

$$n = \left[\frac{z_t \sigma}{d}\right]^2.$$

Donde:

σ = desviación estándar de la población.

$d = \overline{X} - \mu$ = discrepancia permisible o error máximo de estimación permitido.

z_t = valor de z, normal estándar, para un nivel de confianza de $(1 - \alpha)\%$.

Ejemplo 3

Para el mismo caso del ingreso de las familias, suponga ahora que se quiere tener una confianza de 90% de que la diferencia entre $\overline{X}$ y μ no será mayor de 20 mil colones.

En este caso,

$\sigma = 300$ $\quad d = 20$ $\quad 1 - \alpha = 0,90$ $\quad \frac{\alpha}{2} = 0,05$

$\alpha = 0,10$ $\quad z_{0,95} = 1,645$

$$n = \left(\frac{1,645 \cdot 300}{20}\right)^2 = (24,675)^2 \approx 609.$$

La muestra debe ser de 609 familias. El valor calculado resulta menor que el obtenido en la página anterior porque se redujo el nivel de confianza de 95% a 90%.

La fórmula antes incluida para el tamaño de la muestra puede utilizarse para resolver otro tipo de problemas, despejando la variable que interese.

Ejemplo 4

Para el mismo caso del ingreso de las familias de la ciudad, suponga que se toma una muestra de 350 familias. ¿Cuál es la probabilidad de que el error de estimación sea 30 mil colones o menos?

Ya se vio que $n = \left(\frac{z_t \sigma}{d}\right)^2$. Despejando z_t, se tiene:

$$z_t = d \cdot \frac{\sqrt{n}}{\sigma} = (\bar{X} - \mu)\frac{\sqrt{n}}{\sigma}.$$

Sustituyendo por los valores conocidos:

$$z_t = 30 \cdot \frac{\sqrt{350}}{300} = 1,87.$$

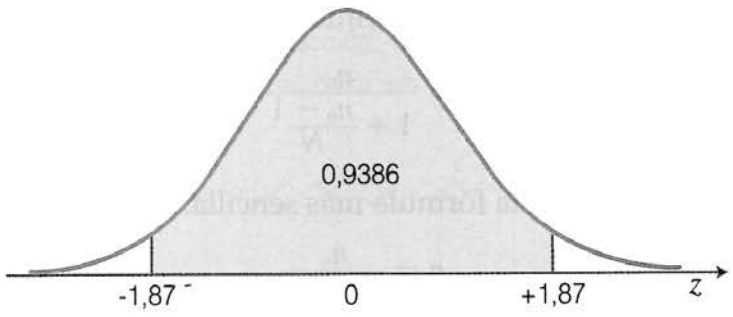


Figura 12.8. Gráfico que muestra el ingreso promedio mensual por familia

¿Cómo debe interpretarse esta probabilidad? Si se toman muestras de tamaño 350, se espera que, en aproximadamente 94 de cada 100 de ellas, la diferencia entre $\overline{X}$ y μ, sea menor de 30 mil colones. Por lo tanto, se tiene una confianza de 94% de que el error de estimación será de 30 mil colones o menor.

12.9. EL TAMAÑO DE LA MUESTRA EN POBLACIONES FINITAS

La fórmula para el tamaño de la muestra, discutida y utilizada anteriormente, se aplica a poblaciones infinitas o a poblaciones finitas cuando se trabaja con reemplazo o cuando n representa una fracción despreciable de N, es decir, siempre que el factor de corrección para poblaciones finitas $\frac{N-n}{N-1}$ sea cercano a 1.

Si no sucede esto, la fórmula debe modificarse para tomar en cuenta el tamaño de la población. Así la expresión:

$$d = z_t \frac{\sigma}{\sqrt{n}}.$$

Debe cambiarse por la expresión:

$$d = z_t \sqrt{\frac{N-n}{N-1}} \frac{\sigma}{\sqrt{n}}.$$

Despejando n de la relación anterior se tiene:

$$n = \frac{\left(\frac{z_t \sigma}{d}\right)^2}{1 + \frac{1}{N}\left[\left(\frac{z_t \sigma}{d}\right)^2 - 1\right]}.$$

Como puede notarse, la expresión cuadrática es la fórmula para n en poblaciones infinitas. Si se define $n_0 = \left(\frac{z_t \sigma}{d}\right)^2$, se tiene la fórmula general para el tamaño de la muestra:

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}}.$$

La cual puede reemplazarse por la fórmula más sencilla:

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

que da un valor prácticamente igual al anterior.

El tamaño de la muestra puede calcularse, entonces, en dos pasos, obteniéndose primero n_0 y luego aplicando la fórmula anterior.

El siguiente ejemplo permite apreciar el efecto que tiene el factor de corrección en el caso de poblaciones finitas.

Ejemplo 5

Como parte de un estudio internacional, se quiere estimar el gasto promedio mensual en medicinas de las personas de 70 años y más de una ciudad. Se sabe que el número total de personas mayores de 70 años es aproximadamente de 8000; además, por una investigación similar llevada a cabo en otra ciudad, se presume que $\sigma^2 = 360\,000$.

¿De qué tamaño debe ser la muestra para tener una probabilidad de 99% de que la discrepancia entre $\overline{X}$ y μ no sea mayor de 100 dólares?

$$\sigma = 600 \qquad z_t = 2,576$$

$$d = 100 \qquad N = 8000$$

$$n_0 = \left(\frac{z_t \sigma}{d}\right) = \left(\frac{2,576 \cdot 600}{100}\right)^2 = (15,46)^2 \approx 239$$

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{239}{1 + \frac{239}{8000}} = \frac{239}{1 + 0,030} = \frac{239}{1,030} \approx 232.$$

Si se ignora el hecho de que la población es finita, se tomaría una muestra de 239 y, por lo tanto, mayor en 7 elementos de la realmente necesaria. En este caso, como puede apreciarse, carece de importancia tomar en cuenta el tamaño de la población.

¿Qué sucedería si se desea que la discrepancia permisible (d) sea de 50 dólares?

$$n_0 = \left(\frac{2,576 \cdot 600}{50}\right) = (30,912)^2 = 955,55 \approx 956$$

$$n = \frac{956}{1 + \frac{956}{8000}} = \frac{956}{1 + 0,1195} = \frac{956}{1,12} = 853,57 \approx 854.$$

En este caso sí tiene importancia la reducción. Si se calcula el tamaño de la muestra ignorando N, se tomarían 102 elementos más, es decir, un 12% más de los necesarios para la precisión deseada.

12.10. FACTORES DETERMINANTES DEL TAMAÑO DE LA MUESTRA

Como acaba de verse, la fórmula general para el tamaño de la muestra, en muestreo simple al azar, es:

$$n = \frac{\left(\frac{z_t \sigma}{d}\right)^2}{1 + \frac{1}{N}\left(\frac{z_t \sigma}{d}\right)^2}.$$

Esta fórmula revela que el tamaño de la muestra depende de la variabilidad de la población (σ^2), de la precisión deseada en las estimaciones (d), de la confianza que se desea tener $(1 - \alpha)$ y del tamaño de la población (N).

La importancia de estos factores no es igual en todos los casos. En primer lugar, cuando la población es infinita o es grande, el cociente n/N es pequeño, el factor de corrección $N - 1$ es muy cercano a 1 y puede ignorarse. En este caso, la fórmula se simplifica y adopta la expresión muy conocida:

$$n_0 = \left(\frac{z_t \sigma}{d}\right)^2$$

la cual no depende de N.[10]

Una conclusión interesante es, al contrario de lo que generalmente se cree, es que el tamaño de la población solo tiene importancia cuando es pequeña o cuando la muestra que se obtendrá representa una proporción elevada de población, o sea, cuando el factor de corrección $\frac{N-n}{N-1}$ es significativamente inferior a 1.

Lo anterior explica por qué, en contraposición a lo sugerido por el sentido común, la estimación, por ejemplo, de las intenciones de voto a nivel nacional en una elección presidencial se realiza con muestras de tamaño similar en Costa Rica, México y Estados Unidos, a pesar de la gran diferencia que existe entre el número de votantes de estos países.[11]

Cabe señalar, finalmente, que un factor muy importante en la fijación del tamaño de la muestra es el costo. Este no aparece en la fórmula antes mencionada porque se trata de un elemento no estadístico; sin embargo, juega un papel fundamental y, en muchos casos, este factor decide el tamaño de la muestra. La dificultad del muestreo es realmente un problema de lograr un balance entre precisión estadística y costo; así, los muestristas tratan de lograr la mayor precisión para un costo dado o el mínimo costo para una

10. Cuando la población es muy grande o infinita la expresión $1 + n_0/N$ muy cercana a 1 o igual a 1, y la corrección del tamaño de muestra máximo no para tomar en cuenta el tamaño de la población no tiene ningún efecto. A eso se debe que el tamaño de la muestra final sea igual a no.
11. Muestras de tamaño entre 1200 y 1500 son muy comunes.

determinada precisión. El problema del costo es fácil de resolver en muestreo simple al azar, pero si se trata de un diseño más complejo, el balance entre costo y precisión reviste especial importancia.

En resumen, los factores determinantes del tamaño de la muestra son:

a) Variabilidad de población (σ^2)

b) Recursos económicos disponibles (costo)

c) Precisión deseada (d)

d) Confianza que se quiere tener $(1 - \alpha)$

e) Tamaño de la población (N)

El tamaño de la muestra es directamente proporcional a la variabilidad de la población, a su tamaño y a la confianza con que se desean las estimaciones; e inversamente proporcional a la magnitud del error que se está dispuesto a aceptar y al costo.

12.11. ESTIMACIÓN EN EL CASO DE PROPORCIONES

En muchas oportunidades no interesa realizar inferencias acerca de un promedio poblacional μ, sino respecto a la proporción de elementos en la población que tienen una cierta característica, o sea, acerca de una proporción P.

Suponga que al director de un colegio le interesa la proporción de sus $N = 2000$ alumnos con computadora disponible en su hogar (N_1). El valor poblacional de interés es $P = \frac{N_1}{N}$ y, para estimarlo, podría tomarse una muestra de n estudiantes, determinar cuántos de ellos tienen computadora en el hogar (n_1) y luego calcular el valor muestral $p = \frac{n_1}{n}$. Con este valor p, se estima el valor poblacional P.[12]

Obviamente, resulta de interés tener un procedimiento para evaluar la confianza que merece p como estimador de P, el cual permita calcular su error estándar, construir límites de confianza y hacer otro tipo de inferencias respecto a P.

Al respecto, es importante señalar que todos los procedimientos vistos en este capítulo para hacer inferencias para la media μ; si se cumplen ciertas condiciones, pueden aplicarse al caso de las proporciones con pequeños ajustes.

En primer término, una proporción es, en realidad, la media de una variable {0, 1}, o sea, de una variable X_i que solo puede asumir esos dos valores. Suponga, por ejemplo, que hay 20 colegios en una región escolar e interesa la variable $X_i =$ **disponibilidad**

12. Note que se usa P (mayúscula) para el valor poblacional y p (minúscula) para el valor muestral.

de gimnasio dentro del colegio. Llame a cada colegio, averigüe si hay o no gimnasio, y anote un 1 cuando lo hay y un 0 cuando no lo hay. Como resultado de este ejercicio, termine con la siguiente información:

$$0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1,$$

la cual señala que, del conjunto de todos de colegios ($N = 20$), un total de $N_1 = 8$ tienen gimnasio. Esto da $P = \frac{N_1}{N} = \frac{8}{20} = 0,40$, e indica que un 40% de los colegios tiene gimnasio.

Ahora bien, si se define $X_i =$ **tenencia de gimnasio**, se tiene:

$$X_1 = 0,\ X_2 = 1,\ X_3 = 0,\ X_4 = 1,...,X_{19} = 1,\ X_{20} = 1$$

y el total de colegios con gimnasio sería: $X = \sum X_i = N_1 = 8$ y, a su vez, la media de la variable X_i:

$$\mu = \frac{X}{N} = \frac{\sum X_i}{N} = \frac{N_1}{N} = P = \frac{8}{20} = 0,40.$$

La media de X_i es igual a P. Esto señala que en una variable $\{0,1\}$, la media equivale a la proporción de unos, es decir, a la proporción de elementos con la característica de interés.

De igual forma, si se tiene una muestra de n elementos e interesa una variable $\{0,1\}$, la media de la variable es igual a p:

$$\bar{X} = \frac{x}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{n_1}{n} = p.$$

Ahora bien, si una p es realmente un promedio, los principios de inferencia vistos para la media muestral $\overline{X}$ le son aplicables; la única duda por aclarar es respecto a la distribución probabilística de la media muestral p, la cual se requiere para evaluar la confianza de las inferencias.

Por la teoría estadística, se sabe que cuando n es moderadamente grande y el valor de P no es muy pequeño, la distribución del valor muestral p se aproxima por una distribución normal con media P y variancia $\sigma^2 = \frac{p(1-p)}{n}$. Esta propiedad de p, el valor muestral, permite usar la curva normal, que ya se conoce, para realizar inferencias acerca de la proporción poblacional P.[13]

13. La condición básica para utilizar con confianza la normal, en la evaluación de las inferencias con proporciones, es que n sea moderadamente grande y que $np > 5$. Si la muestra n es pequeña, el procedimiento más apropiado para calcular la confianza de las inferencias es usar la distribución binomial; y si p es muy pequeña, debe recurrirse a la distribución de Poisson. El desarrollo de estos puntos, sin embargo, están fuera del alcance de este libro, pero pueden consultarse en un texto de métodos estadísticos.

Ejemplo 6

En una encuesta de opinión telefónica, realizada en agosto del 2009, se entrevistó una muestra nacional de 350 personas de 18 años y más. Una de las preguntas fue si la persona tenía celular. Un total de 210 respondió afirmativamente.

a) Estime la proporción de personas de 18 años y más, residentes en viviendas con teléfono de línea fija, que tienen celular.

b) Calcule un intervalo de confianza del 90% para el valor P poblacional correspondiente. Interprete el valor obtenido.

Cálculo de p:

$$p = \frac{n_1}{n} = \frac{210}{350} = 00,60.$$

Un 60% de los entrevistados tiene celular.

12.11.1. Intervalo de confianza para P

Como se quiere una confianza de 90%, la curva normal indica que debe usarse el valor $z = 1,645$.

Haciendo los ajustes en las fórmulas vistas para el caso de μ, se obtiene la correspondiente para el caso de P:

$$L_i = p \pm z_t \sqrt{\frac{p(1-p)}{n}} = 0,60 \pm 1,645 \sqrt{0,60 \cdot \frac{0,40}{350}}$$

$$= 0,60 \pm 0,043$$

$$L_1 = 0,60 - 0,04 = 0,56 \qquad L_2 = 0,60 + 0,04 = 0,64.$$

Los valores 0,56 y 0,64 permiten afirmar, con una confianza de 90%, que la proporción de la población adulta residente en viviendas con teléfono de línea fija, que posee celular está entre 56 y 64%.

$$P[0,56 \leq P \leq 0,64] = 0,90.$$

12.11.2. Tamaño de la muestra para estimar P

Una pequeña universidad desea estimar la proporción de sus estudiantes que podrían estar interesados en un sistema especial de préstamos para la compra de computadora. Como se está en época de vacaciones y se desea iniciar el programa en el próximo año lectivo, y además se cuenta con una base de datos donde aparecen los números

telefónicos de los estudiantes, se decide llevar a cabo una encuesta telefónica. El total de estudiantes es de 5000.

¿Qué tamaño de muestra debe usarse si se quiere estimar la proporción de estudiantes interesados, con un error no mayor de 5 puntos porcentuales y con una confianza de 95%?

Para resolver este problema, debemos partir de la fórmula de n vista en la sección 12.8, pero adaptándola para el caso de las proporciones; en estas, la desviación estándar poblacional es $\sigma = \sqrt{P(1-P)} = \sqrt{P \cdot Q}$. Haciendo la sustitución correspondiente, se obtiene la fórmula:

$$n = \left(\frac{z_1 \sigma}{d}\right)^2 = \left(\frac{z\sqrt{P \cdot Q}}{d}\right)^2.$$

Es importante notar que, para calcular el tamaño de la muestra, se necesita conocer P, el valor que se quiere estimar. En la práctica, este inconveniente se supera a partir de información previa o haciendo una conjetura razonable acerca del posible nivel de P, e introduciendo ese valor en la fórmula de cálculo. Cuando no se tiene ninguna idea de cuál puede ser, se utiliza el valor $P = 0,5$ que hace máxima σ y, por lo tanto, produce el valor máximo de n requerido para cumplir las condiciones especificadas para la muestra. Como esta es la situación en el ejemplo considerado, se usa $P = 0,50$.

$$n = \left(\frac{z_t \sqrt{P(1-P)}}{d}\right)^2 = \left(1,96\sqrt{\frac{0,50 \cdot 0,50}{0,50}}\right)^2 = (19,6)^2 = 384,16 \approx 385.$$

La muestra máxima que debe usarse para estimar el número de estudiantes, con un error no mayor de 5%, es de 385. Ahora bien, como se trata de una población finita no muy grande, 5000 estudiantes, se hace la corrección para población finita muestreada sin reemplazo:

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{384}{1 + \frac{384}{5000}} = \frac{384}{1,077} \approx 357.$$

El ajuste señala que el tamaño debe ser 357, o sea, 28 unidades menos que el indicado por el valor sin ajustar. Se trata de una diferencia de poca significación práctica, el cual perfectamente podría haberse ignorado, especialmente en un estudio telefónico donde la entrevista tomará muy poco tiempo y resulta poco costosa.

Conviene señalar que se ha estimado un tamaño de muestra para las condiciones especificadas. Obviamente, la confianza que merezca la estimación dependerá al final de los resultados de la encuesta. Así, por ejemplo, si al realizar la encuesta con una muestra de 357 estudiantes la proporción interesada en el programa de financiamiento de microcomputadoras resulta ser de 20%, el error muestral de estimación sería de:

$$d = z_t \sqrt{\frac{p(1-p)}{n}} = 1,96 \cdot \sqrt{\frac{0,20 \cdot 0,80}{357}} = 1,96\sqrt{0,000448} = 0,0415 = 4,2\%,$$

valor menor al 5% originalmente fijado. Esto sucede porque la $p = 0,20$ implica una desviación estándar más pequeña que la supuesta al calcular el tamaño de la muestra.

12.11.3 Error estándar de p

La proporción P de la población se estima con el valor muestral $p = \frac{x}{n}$, donde n es el tamaño de la muestra y $x = \sum x_i$ representa el número de elementos en la muestra que tienen la característica de interés.

Siendo p un promedio (de una variable $\{0,1\}$), su variancia puede obtenerse usando la fórmula general para la variancia de un promedio:

$$\sigma_{\bar{x}}^2 = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}$$

$$\sigma_p^2 = \frac{N-n}{N-1} \cdot \frac{P(1-P)}{n} = \frac{N-n}{N-1} \cdot \frac{P \cdot Q}{n}.$$

Ahora bien, en la mayoría de los casos, P no se conoce y la variancia de la población $P \cdot Q = P(1-P)$ debe aproximarse con la variancia de la muestra que es $p \cdot q$, esto produce la expresión

$$\sigma_p^2 = \frac{N-n}{N-1} \cdot \frac{p(1-p)}{n}.$$

Además, si la población es infinita o si el cociente $\frac{n}{N}$ es pequeño, entonces el factor de corrección $\frac{N-n}{N-1}$ es cercano a 1 y puede despreciarse, lo cual produce la expresión:

$$\sigma_p^2 = \frac{p(1-p)}{n} = \frac{p \cdot p}{n} \qquad (q = 1-p).$$

Bajo estas condiciones, el error estándar de una proporción p, basado en una muestra de tamaño n, es:

$$\sigma_p = \sqrt{\frac{p \cdot q}{n}}.$$

La posibilidad de aplicar la teoría del muestreo al caso de proporciones es sumamente importante, ya que permite enfrentar un número grande de problemas prácticos utilizando la misma metodología discutida antes para el caso de variables continuas. Seguidamente se presentan algunas ilustraciones.

Ejemplo 7

Se ha tomado una muestra al azar de 50 personas de 65 años y más y 16 de ellos califican como diabéticos. ¿Dentro de qué límite puede esperarse que esté contenida, con un 90% de confianza, la verdadera proporción de personas de 50 años y más en la población que califica como diabéticos?

$$p = \frac{16}{80} = 0,20 \quad \sigma_p = \sqrt{\frac{p \cdot q}{n}} = \sqrt{\frac{0,20 \cdot 0,80}{80}} = \sqrt{0,002} = 0,044721.$$

Como $\alpha = 0,10$, entonces, $z_{1-\frac{\alpha}{2}} = z_{0,95} = 1,645$, de donde

$$p - 1,645\sqrt{\frac{pq}{n}} < P < p + 1,645\sqrt{\frac{pq}{n}}$$
$$0,20 - 1,645 \cdot 0,044721 < P < 0,20 + 1,645 \cdot 0,044721$$
$$P[0,13 < P < 0,27] = 0,90.$$

Con una confianza de 90%, se espera que la verdadera proporción de personas de 50 años y más diabéticos esté entre 13 y 27%.

Ejemplo 8

Un fabricante desea conocer la proporción de estudiantes de secundaria, de una ciudad, que usa los "tenis" que produce. Él cree que esa proporción no es mayor del 30%, y desea estimarla con un error no mayor de 4 puntos $(p \pm 0,04)$ y con una confianza del 95%. ¿Qué tamaño de muestra le recomendaría usted utilizar?

La fórmula del tamaño de la muestra es: $n = \left(\frac{z_t \sigma}{d}\right)^2$.

Además, en este caso, se tiene que $z_{\frac{\alpha}{2}} = z_{0,975} = 1,96d = 0,04$

$$\sigma = \sqrt{0,30 \cdot 0,70} = \sqrt{0,21}.$$

El tamaño de muestra resulta:

$$n = \left(\frac{1,96 \cdot \sqrt{0,21}}{0,04}\right)^2 = \frac{(1,96)^2 \cdot 0,21}{0,0016} = 504,21 \approx 505.$$

Se le recomendaría tomar una muestra de 505 estudiantes. Note que el resultado se ha redondeado hacia arriba siguiendo una práctica común en el caso de tamaños de muestra.

Ejemplo 9

En la sección 12.10, se señaló que el tamaño de la muestra n requerido para un cierto tipo de estudio no debe aumentar proporcionalmente al aumentar el tamaño de la población. Para mostrar este punto considere un ejemplo simple.

En una elección muy cerrada, un partido político desea conocer, por diversas razones, la proporción de electores que piensan votar por él en: a) una ciudad mediana con alrededor de 100 000 votantes, b) en la capital y su zona metropolitana con cerca de un millón de votantes y c) en todo el país con un número de votantes cercano a los 10 millones. Para llenar estas necesidades planea realizar una encuesta estadística por muestreo utilizando la entrevista cara a cara.

Si se quiere un margen de error máximo de 0,03 y una confianza de 95%, ¿qué tamaño de muestra debe usarse para cada una de estas encuestas?

Solución

En este caso, como el nivel de confianza es 95%, $z = 1,96$; suponga, además, por tratarse de una elección cerrada, que $\sigma = 0,50$.[14] Aplicando la fórmula se obtiene:

$$n_0 = \left(z \cdot \frac{\sigma}{d}\right)^2 = \left(1,96 \cdot \frac{\sqrt{0,5 \cdot 0,5}}{0,03}\right)^2 \approx 1068.$$

Luego, se procede a hacer el ajuste, por tratarse de una población finita que se muestrea sin reemplazo, utilizando la fórmula general:

$$n = \frac{n_0}{1 + \frac{n_0}{N}}.$$

$N = 100\,000$ (caso ciudad mediana)

$$n = \frac{1068}{1 + \frac{1068}{100\,000}} = \frac{1068}{1,01068} = 1057.$$

$N = 1\,000\,000$ (caso capital zona metropolitana)

$$n = \frac{1068}{1 + \frac{1068}{1\,000\,000}} = \frac{1068}{1,00107} = 1067.$$

$N = 10\,000\,000$ (caso todo el país)

$$n = \frac{1068}{1 + \frac{1068}{10\,000\,000}} = \frac{1068}{1,00011} = 1068.$$

Como puede notarse, en la ciudad mediana el ajuste por población finita reduce la muestra de 1068 a 1057, o sea en 11 casos, una variación muy pequeña, sin ninguna importancia práctica. En las otras dos situaciones casi no hay cambio. Esto sucede porque la relación entre la muestra máxima n y la población N, es decir, $\frac{n}{N}$ es un valor sumamente pequeño: 0,01068,

14. La desviación estándar poblacional máxima en el caso de proporciones es:
$\sigma = \sqrt{PQ} = \sqrt{0,50 \cdot 0,50} = 0,50$.

en el primer caso; 0,00107, en el segundo y 0,00011, en el tercero; en consecuencia, la división por $1 + \frac{n_0}{N}$ tiene muy poco efecto en el tamaño de la muestra.

Por ello, para fines prácticos, en las tres encuestas antes mencionadas se utilizará posiblemente una muestra de alrededor de 1100.[15]

Ahora bien, ¿cuál sería el tamaño si se quisiera hacer la estimación en una ciudad pequeña con unos 2000 votantes? Observe:

$$n = \frac{1068}{1 + \frac{1068}{2000}} = \frac{1068}{1 + 0,534} = \frac{1068}{1,534} = 697.$$

Aquí sí hay un cambio radical, con fuertes implicaciones sobre el costo y el tiempo para realizar la encuesta. La muestra máxima de 1068 se reduce a 697, una disminución de 371 casos. Esto sucede porque la población es pequeña y la muestra máxima n representa un 53,4% de ella. En la práctica, posiblemente se use una de 700 casos.

15. Las empresas que realizan estudios de opinión y de mercado, y los mismos muestristas, cuando se trata de muestras de cierto tamaño, prefieren, por razones prácticas y de negociación con sus clientes, redondear el tamaño de la muestra generalmente hacia arriba a múltiplos de 100. Cuando son más pequeñas, el redondeo es a múltiplos de 50.

EJERCICIOS DE AUTOEVAUACIÓN

1. Un hospital donde hay 40 médicos, 30 hombres y 10 mujeres, es invitado a enviar una delegación de 2 doctores y 2 doctoras a una celebración que se realizará en una ciudad europea. El director decide que la selección sea por rifa y para ello asigna un número de 1 a 30 a cada uno de los doctores y de 31 a 40 a cada una de las doctoras. Luego, en una bolsa se colocan 30 bolitas de igual tamaño, numeradas de 1 a 30 y la conserje procede a sacar una bola, se le da al director y luego otra; los dos doctores con esos números son escogidos para integrar la delegación. Después, se ponen 10 bolas de 31 a 40 y se procede en la misma forma para escoger las dos doctoras que completarán la delegación.
   a) ¿Cuál fue la probabilidad de un doctor de ser incluido en la delegación? ¿Cuál la de una doctora?
   b) En su opinión, ¿la delegación de 4 médicos es una muestra probabilística de los 40 médicos del hospital? ¿Por qué?
   c) ¿Qué nombre recibe el sistema de muestreo utilizado para seleccionar la muestra de 4 doctores y doctoras?
2. Para una población de $N = 16$ elementos, con $\sigma = 18$, calcular la variancia de $\overline{X}$ para muestras de tamaño $n = 4$, suponiendo que:
   a) La selección se hizo al azar simple sin reemplazo.
   b) La selección se hizo al azar simple con reemplazo.
3. Para una población formada por los elementos
   $A = 2, B = 6, C = 8, D = 10, E = 10, F = 12$
   a) Verifique $\mu = 8$ y $\sigma^2 = 10,67$.
   b) Enumere todas las muestras posibles de tamaño dos (con reemplazo); compruebe que son 36.
   c) Verifique que si se escoge un elemento al azar y se deja afuera, y luego el otro, el número de muestras distintas posibles de tamaño 2 es de 15.
   d) Para cada una de las muestras enumeradas en el punto c, calcule el promedio y luego indique en cuántas muestras difiere del promedio de la población a lo más en una unidad; en cuántas difiere a lo más en dos unidades; etcétera.
   e) ¿Cuál diría usted que es la probabilidad de que un promedio difiera a lo más en dos unidades del promedio de la población?

4. En una empresa con un número elevado de empleados, se desea estimar su peso promedio (μ) con una muestra. Se estima que $\sigma^2 = 49$.

   a) Haga un gráfico en el cual se observe cómo varía el error estándar del promedio al aumentar n, tamaño de la muestra. Escriba un pequeño comentario sobre lo que muestra el gráfico.

   b) Si se quiere un error estándar de un kilogramo en el promedio, ¿de qué tamaño debe ser la muestra de trabajadores?

   c) ¿A qué tamaño debe elevarse el tamaño de la muestra si se quisiera un error estándar de ½ kl?

   d) Si se quisiera tener una confianza de 95% de que la diferencia entre $\overline{X}$ y μ no sea mayor de un kl, ¿de qué tamaño debe ser la muestra?

5. El Ministerio de Educación de un país quiere estimar el gasto promedio realizado, al inicio de clases, por las familias que tienen hijos en el sistema público de educación secundaria, para equiparlos adecuadamente para el curso. Como desea hacer un cálculo confiable, pero dispone de recursos humanos y económicos limitados, decide utilizar una muestra estadística de hogares con estudiantes matriculados en colegios públicos. Un pequeño estudio piloto en el cual se probó el cuestionario, sugiere que el valor de σ puede estimarse en 30 000 colones. ¿Qué tamaño de muestra sería necesario si el Ministerio quiere tener una confianza de 95% de que el valor estimado del gasto promedio por familia no diferirá del promedio poblacional en más de 5000 colones?

6. Suponga que, en el ejemplo anterior, la muestra utilizada finalmente es de 650 familias y el gasto promedio de cada una resulta ser 118 500 colones.

   a) Obtenga un intervalo de confianza de 95% para la media poblacional.

   b) Escriba la interpretación del intervalo de confianza.

7. Conociendo que en el país hay una total de 400 000 familias con hijos en el sistema público de educación secundaria, obtenga:

   a) El gasto total realizado por las familias con hijos en el sistema escolar al inicio del curso.

   b) Un intervalo de confianza del 95% para la estimación anterior.

8. En una región costera de un país hay un total de 350 escuelas públicas. La Secretaría de Instrucción Pública desea estimar el número de pupitres por reemplazar por su mal estado, pero temiendo que los interesados exageren sus necesidades, decide hacer un estudio directo con personal calificado en una muestra de escuelas. Por una investigación realizada varios años antes, se estima que $a = 55$. ¿Qué tamaño

de muestra sugeriría utilizar usted, sabiendo que interesa un error de 10 pupitres y un 90% de estimación?

9. Suponga que el estudio mencionado en el punto 8 es realizado finalmente con una muestra de 100 escuelas y arroja una media de 24 pupitres por escuela que deben ser reemplazados.

   a) Calcule un intervalo de confianza del 90% para la media poblacional.

   b) Interprete los límites obtenidos.

   c) Estime el total de pupitres que deben ser reemplazados en toda la región costera.

10. El presidente de un club social desea conocer el grado de apoyo que tendría dentro de los socios una propuesta para incrementar la cuota de mantenimiento. Para ello, decide hacer un sondeo a una muestra de socios.

    a) Estime el tamaño de muestra requerido, si se desea estimar la proporción que está de acuerdo con la propuesta, con una confianza de 90%, y con un error no mayor de 5 puntos porcentuales.

    b) Suponga que, finalmente, la encuesta se lleva a cabo con un total de 600 socios y 186 se pronuncian a favor de la propuesta. Calcule un intervalo de confianza del 95% para el valor poblacional.

    c) Interprete el intervalo calculado en b.

    d) Cambiarían los límites calculados en b si se sabe que el total de socios es de 10 000? Haga los cálculos pertinentes.

RESPUESTA A LOS EJERCICIOS DE AUTOEVALUACIÓN

1. a) Probabilidad de un doctor: $\frac{2}{30} = \frac{1}{15}$. Probabilidad de una doctora: $\frac{2}{10} = \frac{1}{5}$.

   b) Sí, es una muestra probabilística de los médicos del hospital, porque cada uno (hombre o mujer) tuvo una probabilidad conocida de ser seleccionado. Sin embargo, la muestra no es de igual probabilidad: las doctoras tuvieron una probabilidad tres veces mayor de ser seleccionadas que los doctores.

2. $N = 16$ y $\sigma = 18$.

   a) Cálculo de $\sigma_{\bar{X}}^2$, suponiendo muestreo al azar simple sin reemplazo y $n = 4$.

$$\sigma_{\bar{X}}^2 = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n} = \frac{16-4}{15} \cdot \frac{(18)^2}{4} = \frac{12}{15} \cdot \frac{324}{4} = 64{,}8.$$

   b) Cálculo de $\sigma_{\bar{X}}^2$, suponiendo muestreo simple al azar sin reemplazo

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} = \frac{(18)^2}{4} = 81.$$

3. $A = 2, B = 6, C = 8, D = 10, E = 10, F = 12, \quad N = 6.$

   a) Cálculo de μ y σ^2:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{48}{6} = 8.$$

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} = \frac{(-6)^2 + (-2)^2 + 0 + (2)^2 + (2)^2 + (4)^2}{6} = \frac{64}{6} = 10{,}67.$$

   b) Enumeración de todas las muestras posibles de tamaño dos con reemplazo:

   AA BB CC DD EE FF

   AB AC AD AE AF BC BD BE BF CD CE CF DE DF EF

   BA CA DA EA FA CB DB EB FB DC EC FC ED FD FE

   Puede apreciarse que son 36. La otra forma de verlo es notando que, al escoger la primera, se pueden dar 6 resultados diferentes y, como hay reposición, la selección de la segunda también tiene 6 resultados posibles; por lo tanto, el total de muestras posibles es $6 \cdot 6 = 36$.

c) Si se toma un elemento y se deja afuera y luego se toma otro (muestreo sin reemplazo), no pueden darse repeticiones de un mismo elemento y quedan fuera las 6 muestras anotadas en la primera línea del punto b. El número de 30 puede obtenerse si se nota que hay 6 posibilidades, al seleccionar el primer elemento, pero para el segundo solo hay 5 posibilidades porque el seleccionado queda fuera, por lo tanto, el número total de muestras posibles es de $6 \cdot 5 = 30$.

Sin embargo, las 30 están formadas por 15 grupos de dos que tienen la misma composición y solo difieren en el orden de obtención, líneas segunda y tercera en b; por lo tanto, el número de muestras diferentes es de 15.

Muestra	Composición	Promedio	$\overline{X} - \mu$
AB	(2,6)	4	–4
AC	(2,8)	5	–3
AD	(2,10)	6	–2
AE	(2,10)	6	–2
AF	(2,12)	7	–1
BC	(6,8)	7	–1
BD	(6,10)	8	0
BE	(6,10)	8	0
BF	(6,12)	9	1
CD	(8,10)	9	1
CE	(8,10)	9	1
CF	(8,12)	10	2
DE	(10,10)	10	2
DF	(10,12)	11	3
EF	(10,12)	11	3

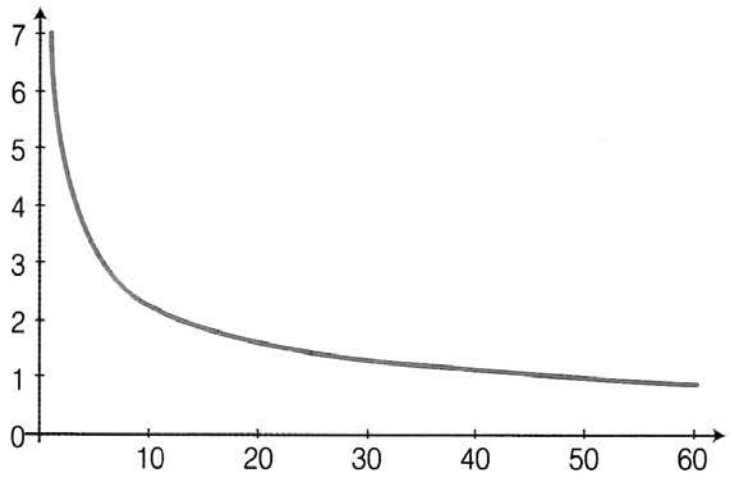
$\overline{X} - \mu$		frecuencia
0	.....	2
±1	.....	5
±2	.....	4
±3	.....	3
±4	.....	1

Difieren a lo más en una unidad: 7; a lo más en dos: 11; a lo más en tres: 14; y a lo más en cuatro todas las 15 muestras.

e) La probabilidad de que el promedio de una muestra al azar de tamaño 4 difiera a lo más en dos unidades del promedio de la población es: $\frac{11}{15} = 0,73$.

4. Se sabe que $\sigma^2 = 49 \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{7}{\sqrt{n}}$.

a) Variación del error estándar al aumentar n.



Conforme la muestra aumenta, el error estándar del promedio se reduce en proporción a $\sqrt{n}$ cada vez que la muestra se cuadruplica, el error disminuye la mitad. Así, por ejemplo, para $n = 1$ es 7, para $n = 4$ es 3,5, para $n = 16$ es 1,75, etcétera.

b) $\frac{\sigma}{\sqrt{n}} = 1 \Rightarrow \sqrt{n} = \sigma$ y $n = \sigma^2 = 7^2 = 49$.

c) Si se quiere un error estándar de $\frac{1}{2}$ la muestra debe cuadruplicarse, o sea $n = 196$.

d) Si se quiere $P(|\bar{X} - \mu|1) = 0,95$ ¿cuál debe ser el tamaño de n?

Se estandariza la expresión anterior:

$$P(|\overline{X} - \mu| < 1) = 0,95 \Rightarrow P\left(\frac{|\overline{X} - \mu|}{\frac{\sigma}{\sqrt{n}}} < \frac{1}{\frac{\sigma}{\sqrt{n}}}\right) = 0,95 \Rightarrow P\left(z < \frac{1}{\frac{\sigma}{\sqrt{n}}}\right) = 0,95.$$

Se sabe que $P(|z|1,96) = 0,95$. Entonces,

$$\frac{1}{\frac{\sigma}{\sqrt{n}}} = 1,96 \Rightarrow \sqrt{n} = \sigma \cdot 1,96$$

$$\Rightarrow n = \sigma^2 \cdot (1,96)^2$$
$$= 49 \cdot (1,96)^2$$
$$= 49 \cdot 3,8416 = 188,23$$
$$n \approx 189.$$

5. $\sigma = 30\,000$, $d = \bar{X} - \mu < 5000$. Valor curva normal para 95% es $z = 1{,}96$.

Como se trata de una población grande:

$$n = \left(z \cdot \frac{\sigma}{d}\right)^2 = \left(1{,}96 \cdot \frac{30\,000}{5000}\right)^2 = (11{,}76)^2 = 138{,}29 \approx 139.$$

6. a) $n = 650$, $\bar{X} = 118\,500$, $z = 1{,}96$.

$$L_i = \bar{X} \pm z\frac{\sigma}{\sqrt{n}} = 118\,500 \pm 1{,}96 \cdot \frac{30\,000}{\sqrt{650}} = 118\,500 \pm 2307$$

$P(116\,193 < \mu < 120\,807)$.

b) Se tiene una confianza del 95% que el gasto promedio realizado al inicio del curso por las familias, cuyos hijos están en el sistema estatal de educación secundaria, está entre 116 193 y 120 807 colones.

7. $N = 400\,000$ (familias con hijos en el sistema público de educación secundaria)

El gasto total es igual a $N\bar{X} = 400\,000 \cdot 118\,500 = 47\,400$ millones de colones.

El intervalo de confianza se obtiene multiplicando cada uno de los límites para el promedio, calculados en el punto 6, por $N = 400\,000$. El resultado se incluye seguidamente en millones:

$P(46\,477{,}2 < N\mu < 48\,322{,}8) = 0{,}95$.

8. $N = 350$ escuelas públicas. $\sigma = 55$, $d = 10$

$$n = \left(\frac{1{,}645 \cdot 55}{10}\right)^2 = (9{,}0475)^2 = 81{,}857 \approx 82.$$

Pero como la población es pequeña, se debe hacer el ajuste por población finita:

$$n = \frac{n_0}{1 + \frac{n_0}{n}} = \frac{81{,}86}{1 + \frac{81{,}86}{350}} = \frac{81{,}86}{1{,}2339} = 66{,}34 \approx 67.$$

Se propone una muestra de 67 escuelas.

9. $n = 100$, $\overline{X} = 24$.

a) $$L_i = \bar{X} \pm z\sqrt{\frac{N-n}{N-1}} \cdot \frac{\sigma}{\sqrt{n}} = 24 \pm 1{,}645 \cdot \sqrt{\frac{250}{349}} \cdot \frac{55}{\sqrt{100}} = 24 \pm 7{,}66$$

$P(16{,}34 < \mu < 31{,}66) = 0{,}90$.

b) Puede afirmarse, con una confianza del 90%, que el número promedio de pupitres por escuela que debe ser reemplazado está entre 16 y 32.

c) Total de pupitres que deben reemplazarse: $N\overline{X} = 350 \cdot 24 = 8400$.

10. a) $P = Q = 0,50;\ 1 - \alpha = 0,90;\ z = 1,645;\ d = p - P = 0,05$

$$n = \left(z \cdot \frac{\sqrt{P \cdot Q}}{d}\right)^2 = \left(\frac{1,645 \cdot 0,5}{0,05}\right)^2 = (16,45)^2 = 270,6 \approx 271$$

b) $n = 600, \quad n_1 = 186$ (de acuerdo con los cambios propuestos)

$$p = \frac{186}{600} = 0,31;\ z = 1,96$$

$$L_i = \overline{X} \pm z\sqrt{\frac{P \cdot Q}{n}} = 0,31 \pm 1,96 \cdot \sqrt{\frac{0,31 \cdot 0,69}{600}}$$
$$= 0,31 \pm 1,96 \cdot 0,018881 = 0,31 \pm 0,037.$$

$P(0,273 < P < 0,347) = 0,95.$

c) Puede afirmarse, con una confianza del 95%, que la proporción de socios que apoya la propuesta está entre un 27% y un 35%.

d) Como se trata de muestreo sin reemplazo de una población finita, el error estándar es:

$$\sqrt{\frac{N-n}{N-1}} \cdot \sqrt{\frac{P \cdot Q}{n}} = \sqrt{\frac{9600}{9999}} \cdot \sqrt{\frac{0,31 \cdot 0,69}{600}}$$
$$= 0.979845 \cdot 0,018881 = 0,0185,$$

y el intervalo de confianza vendría dado por:

$$L_i = \overline{X} \pm z\sqrt{\frac{N-n}{N-1}} \cdot \sqrt{\frac{P \cdot Q}{n}} = 0,31 \pm 1,96 \cdot 0,0185 = 0,31 \pm 0,036.$$

El cambio en los límites, al considerar el tamaño de la población, es insignificante. Esto ocurre porque el tamaño de la muestra es relativamente pequeño con respecto al de la población y, entonces, el ajuste por población finita no tiene un efecto significativo.