5. MUESTREO E INTERVALOS DE CONFIANZA

5.1. MÉTODOS DE MUESTREO

Muestreo es la actividad por la cual se toman ciertas muestras de una población de elementos de los cuales vamos a tomar ciertos criterios de decisión, el muestreo es importante porque a través de él podemos hacer análisis de situaciones de una empresa o de algún campo de la sociedad.

¿Y porque no se estudia la población completa? se preguntarían algunos, pero en ocaciones no es factible, veamos algunas razones para muestrar:

5.1.1. RAZONES DEL MUESTREO

1. La naturaleza destructiva de algunas pruebas.
Se quiere conocer la resistencia de los tornillos que se fabrica una planta, para conocerla es necesario destruír el producto, lógicamente no podemos probar toda la población porque nos quedaríamos sin productos.

2. La imposibilidad física de checar todos los elementos de la población.
Se quiere conocer el efecto de un nuevo insecticida en las moscas, como se puede comprender no es posible contactar a todas las moscas para realizar el estudio.

3. El costo de estudiar a toda la población es muy alto.
Se quiere conocer la opinión de la población sobre cierto personaje de la política, si en el país hay 100 millones de habitantes, se tendría que contratar mucho personal y equipo para realizar el estudio.

4. El tiempo para contactar a toda la población es inviable.
En ocasiones se necesita información rápida para tomar una decisión importante, tal vez estudiar a toda la población nos lleve más tiempo del que disponemos.

Por las razones anteriores, en muchos casos es conveniente el uso de muestras, pero para que podamos extraer conclusiones, es importante que elijamos bien las muestras para nuestros estudios.Hay cuestiones que debemos especificar a la hora de elegir una muestra:

1. El tipo de muestreo que se va a utilizar.

2. El tamaño de la muestra.

3. El nivel de confianza de las conclusiones que vamos a presentar.

5.1.1.1 CLASIFICACIÓN DE LOS MUESTREOS

Los métodos de muestreo pueden dividirse en dos grandes grupos: métodos de muestreo probabilísticos y métodos de muestreo no probabilísticos.

Muestreos no probabilísticos

No sirven para realizar generalizaciones, pues no se tiene certeza de que la muestra extraída sea representativa, ya que no todos los sujetos de la población tienen la misma probabilidad de se elegidos. En general se seleccionan a los sujetos siguiendo determinados criterios procurando que la muestra sea representativa.

Muestreo intencional u opinático: en el que la persona que selecciona la muestra es quien procura que sea representativa, dependiendo de su intención u opinión, siendo por tanto la representatividad subjetiva.

Muestreo sin norma: se toma la muestra sin norma alguna, la muestra podría ser representativa si la población es homogénea y no se producen sesgos de selección.

Muestreos probabilísticos

Los muestreos probabilísticos son aquellos en los que todos los individuos tienen la misma probabilidad de ser elegidos para formar parte de una muestra . Dentro de los métodos de muestreo probabilísticos encontramos los siguientes tipos:

1. Muestreo aleatorio simple

2. Muestreo sistemático

3. Muestreo estratificado

4. Muestreo por conglomerados

5.1.2. MUESTREO ALEATORIO SIMPLE

Priméramente se asigna un número a cada elemento de la población, después al azar (como una urna, tablas de números aleatorios, números aleatorios generados electrónicamente, etc) se eligen los elementos necesarios para la muestra.

La ventaja de este método es que es sencillo y de fácil comprensión. Sus desventajas es que requiere que se posea de antemano un listado completo de toda la población y que cuando se trabaja con muestras pequeñas es posible que no represente a la población adecuadamente.

Ejemplo

En una compañía con 150 trabajadores se quiere obtener una muestra aleatoria de 15 elementos para un chequeo médico. Se sigue el siguiente procedimiento:

1. Los trabajadores fueron numerados del 1 al 150

2. Mediante una tabla de números aleatorios se procede a seleccionarlos.

3. El punto de arranque en la tabla se fija mediante la hora en ese momento, 4:03, por lo tanto se inicia en la fila 4, columna 3.

4. Como los números de los trabajadores van desde 1 hasta 150 solo se toman en cuenta las primeras 3 cifras de cada número y se registran los números que se vayan encontrando en ese rango.

El primer número encontrado fue el 054 en la fila 4 columna 5, se siguen revisando los números horizontalmente, el siguiente seleccionado fue el 095 y así sucesivamente.

La muestra de 15 números fue la siguiente:

054	095	080	004	147
005	050	024	046	018
041	021	105	009	146

5.1.3. MUESTREO SISTEMÁTICO

Es necesario conocer el número de los elementos de la población (N) y el tamaño que deberá tener la muestra (n). Se define cada cuantos elementos de la población seleccionaremos uno para la muestra con k=N/n. Se comienza la selección eligiendo aleatoriamente el primer elemento entre 1 y k, luego se cuentan k elementos y se selecciona el segundo y así sucesívamente hasta completar la muestra.

Este método tiene las ventajas de ser fácil de aplicar, no es necesario tener un listado de toda la población y asegura una cobertura de unidades de todos los tipos.

Su desventaja es que si la constante de muestreo está asociada con el fenómeno de interés, las estimaciones obtenidas a partir de la muestra pueden contener un sesgo.

Ejemplo:

Suponga que la población de interés consiste de 2000 expedientes en un archivo. Para seleccionar una muestra de 100 con el método aleatorio simple primero se tendría que numerar todos los expedientes. En este método se selecciona el primer expediente de acuerdo al método aleatorio simple, luego como se quiere una muestra de 100, se divide 2000 / 100 = 20, y se selecciona un expediente cada 20.

5.1.4. MUESTREO ALEATORIO ESTRATIFICADO

En ciertas ocasiones resultará conveniente estratificar la muestra según ciertas variables de interés. Para ello debemos conocer la composición estratificada de la población objetivo a muestrear. Una vez calculado el tamaño muestral apropiado, este se reparte de manera proporcional entre los distintos estratos definidos en la población usando una simple regla de tres.

Entre sus ventajas, este método asegura que la muestra represente adecuadamente a la población en función de ciertas variables seleccionadas, además de obtener estimaciones más precisas

La desventaja es que se ha de conocer como se distribuye la población de acuerdo a las variables utilizadas para la estratificación.

Ejemplo:

Se quiere obtener una muestra de 50 estudiantes de la universidad. Se pretende que la muestra sea representativa en relación al lugar de origen de los estudiantes (si son de la localidad o son foráneos).Se sabe que en esta universidad el 30% de los estudiantes son foráneos.

Primero debemos identificar los estratos de la población y sus respectivas proporciones:

Estudiantes locales	0.70
Estudiantes foráneos	0.30

La muestra deberá mantener esas mismas proporciones, para lo cual es preciso multiplicar el tamaño de la muestra (n) por las proporciones de los estratos y obtenemos el número de elementos que serán seleccionados de cada estrato:

Estudiantes locales	(0.70)(50) = 35
Estudiantes foráneos	(0.30)(50) = 15

Ahora se procede a seleccionarlos por medio de alguno de los métodos anteriores.

5.1.5. MUESTREO ALEATORIO POR CONGLOMERADOS

El muestreo por conglomerados consiste en dividir la población en sectores o conglomerados, seleccionar una muestra aleatoria de esos sectores, y finalmente obtener una muestra aleatoria de cada uno de los sectores seleccionados.

Entre sus ventajas se encuentra que es muy eficiente cuando la población es muy grande y dispersa, además de que no es preciso tener un listado de toda la población, sólo de las unidades primarias de muestreo.

Su desventaja radica en que una muestra de conglomerados, usualmente produce un mayor error muestral (por lo tanto, da menor precisión de las estimaciones acerca de la población) que una muestra aleatoria simple del mismo tamaño.

Ejemplo:

Se quiere conocer la opinión de los padres de familia sobre los temas de educación sexual tratados en los libros de texto de primaria en la República Mexicana. Como la población está muy dispersa y es muy grande, es necesario hacer un muestreo por conglomerados en varias etapas.

Primero dividimos la República en sectores geográficos, que podrían ser los estados, y seleccionamos una muestra aleatoria de ellos. Luego en cada uno de ellos hacemos una selección aleatoria de escuelas primarias. Y por último en las escuelas seleccionadas obtenemos una muestra aleatoria de padres de familia.

5.2. DISTRIBUCIÓN MUESTRAL DE MEDIAS

DISTRIBUCION MUESTRAL DE MEDIAS

Si se extrae una muestra al azar de tamaño n, de una población infinita con media μ y una varianza s^2, entonces las observaciones de la muestra son variables aleatorias independientes e idénticamente distribuidas. La media de la muestra, calculada como

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \bullet\bullet\bullet + x_n)$$

Que es una combinación lineal de variables aleatorias dividida por una constante, que
También es una variable aleatoria normal, y el valor esperado y la varianza de la distribución por muestreo de $\bar{x}$ puede derivarse sencillamente. Primero, observamos que

$$E(\bar{x}) = E\left[\frac{1}{n}(x_1 + x_2 + \bullet\bullet\bullet\ x_n)\right]$$

$$= \frac{1}{n}\left[E(x_1) + E(x_2) + \bullet\bullet\bullet + E(x_n)\right]$$

$$= \frac{1}{n}(n\mu) = \mu$$

Es decir, esperanza de la media de la muestra es la media de la población.

Luego, puesto que se considera que las observaciones de la muestra son variables aleatorias independientes, la propiedad de aditividad se verifica para la varianza. Es decir, la varianza de la suma es la suma de las varianzas. Además, puesto que $V(x_i) = \sigma^2$ tenemos

$$V(\bar{x}) = V(\frac{1}{n}\sum x_i)$$

$$= \frac{1}{n^2}\left[V(x_1) + V(x_2) + \bullet\bullet\bullet V(x_n)\right]$$

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

$$= \frac{\sigma^2}{n}$$

En esta derivación hemos empleado el teorema de que la varianza de una constante multiplicado por una variable es igual al cuadrado de la constante multiplicado por la varianza de la variable.
El error estándar de la media, mide la variabilidad entre medias muestrales.

$$\sigma_{\bar{x}} = \sqrt{V(\bar{x})} = \frac{\sigma}{\sqrt{n}}$$

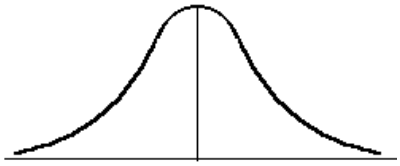
lo que revela que $\sigma_{\bar{x}}$ es menor que σ. Además, indica que cuando $n \to \infty,\ \sigma_{\bar{x}} \to 0$. Así, cuanto mayor es la muestra, tanto menor es la fluctuación entre medias muestrales extraídas de la misma población.
Si se toman muestras de una población finita, sin reposición, como en los casos anteriores, debe de introducirse un factor de corrección para población finitas para calcular el error estándar de la media. A saber:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Cuando la población progenitora es normal, la distribución de $\bar{x}$ por muestreo es también normal, por pequeña que sea la muestra. ¿Qué ocurre cuando no puede especificarse la distribución de probabilidad de la población a partir de la cual se obtiene la muestra? Para obtener una idea con respecto a la distribución de muestreo de $\bar{X}$ cuando el modelo de probabilidad de la población de interés no se especifica. Pasar al teorema del limite central.

-
-
-
-

Por lo tanto podemos decir que la



$$\mu_x$$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

De esta manera la ecuación para la transformación de cualquier media muestral en una variable normal estándar será:

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

Propiedades de la Media aritmética

Entre varias propiedades matemáticas importantes de la media aritmética para una distribución normal están:

Insesgamiento

Implica el hecho de que el promedio de todas las medias muestrales posibles (de un tamaño de muestra dado *n*) será igual a la media de población μ_x.

Eficiente

Se refiere a la precisión de la muestra de estadística como un estimador del parámetro de población.

Para distribuciones como la normal, la media aritmética se considera más estable de muestra a muestra que otras mediciones de tendencia central. Para

una muestra de tamaño n, la media de muestra se acercará más, en promedio, a la media de población que cualquier otro estimador.

Consistencia

Se refiere al efecto del tamaño de muestra sobre la utilidad de un estimador. Al incrementarse el tamaño de muestra, la variación de la media de muestra de la media de población se hace más pequeña, de manera que la media aritmética de muestra se vuelve una mejor estimación de la media de población.

5.3. ESTIMADORES PUNTUALES E INTERVALOS DE CONFIANZA

En la estadística tiene un papel destacado la noción de MUESTRA ALEATORIA.

Una muestra aleatoria de tamaño n es:

- Una colección de n variables aleatorias.
- Todas con la misma distribución.
- Todas independientes.

Esta definición idealiza la operación de repetir n veces la observación de la misma variable aleatoria, siendo las repeticiones independientes una de otra.

La colección de donde extraemos la muestra aleatoria, se denomina POBLACIÓN. Nuestra intención al tomar una muestra, es la de hacer INFERENCIA. Este término lo usamos en estadística para denotar al procedimiento con el que hacemos afirmaciones acerca de valores generales de la población mediante los números que observamos en la muestra.

Quizá un ejemplo aclare las ideas. Suponga que observamos el proceso de fabricación de las ``bolitas'' que se le ponen al envase de los desodorantes ``roll on''. No todas las bolitas van a tener el mismo diámetro, si escogemos, al azar una bolita, tendremos un valor para el diámetro que es una variable aleatoria. Podemos suponer que los diámetros tienen la distribución normal, debido a nuestra experiencia con el proceso, conocemos que la desviación estándar de la población es de 4 mm (aproximadamente). Pero, también por experiencia, sabemos que el diámetro promedio puede variar por desajuste de la maquinaria productora. De modo que tenemos:

- Una POBLACIÓN, que son todas las bolitas que se producen.
- Un PARÁMETRO de la población conocido (o casi) que es la desviación estándar.
- Otro PARÁMETRO cuyo valor es desconocido: la media .

Para tratar de conocer el valor del parámetro que desconocemos, tomamos una MUESTRA de la bolitas. Supongamos que son 100 bolitas en la muestra. Con un instrumento de precisión, y con mucho cuidado, medimos los diámetros de las 100 bolitas de la muestra y calculamos su promedio.

¿Qué nos dice el valor de la media de la muestra respecto a la media de la población?

- por una lado, definitivamente la media de la muestra NO va a ser igual a la de la población.
- por otra parte, no tenemos mejor información respecto a la media de la población que la que extraigamos de la muestra. Cualquier otra información no pasa de chisme.
- por último, sería muy extraño que si la población de bolitas tiene, por decir algo, un diámetro promedio de 45 mm, nos tocaran 100 bolitas en la muestra con un promedio de, digamos, 32 mm. Fíjese que no decimos imposible sino raro o extraño.
- además, si alguien nos preguntara ¿como cuánto es el diámetro promedio de la población de bolitas? Le contestaríamos diciendo el valor que hayamos visto en la muestra.
- a nuestra contestación debíamos agregarle alguna advertencia como: "mas o menos", o ``aproximadamente".

A un valor calculado con los datos de una muestra lo llamamos ESTADÍSTICA. Cuando usamos una estadística para jugar el papel de decir, aproximadamente, el valor de un parámetro de la población, le llamamos ESTIMADOR. Cuando andamos un poco pedantes le llamamos ESTIMADOR PUNTUAL (al decir ``puntual" queremos decir que para estimar el parámetro estamos usando un valor único).

Regresando a las bolitas del ``Roll on". Si la muestra de 100 bolitas arroja un valor del promedio de 43.5 mm, diríamos que ESTIMAMOS el promedio de la población en 43.5 mm.

Constrúyase Ud. mismo un ejemplo como el de las bolitas. En su ejemplo, describa

- una población.
- un parámetro para la población.
- una muestra.
- una estadística que le sirva como estimador.

Características probabilísticas de un estimador

Cuando se tiene una fórmula para estimar y se aplica a una muestra aleatoria, el resultado es aleatorio, es decir los estimadores son variables aleatorias.

Por ejemplo si se recibe un embarque de objetos que pueden

- estar listos para usarse ó
- defectuosos.

Podemos seleccionar, al azar, algunos de ellos para darnos una idea de la proporción de defectuosos en el embarque. El parámetro de interés es la

proporción de defectuosos en toda la población, pero lo que observamos es la proporción de defectuosos en la muestra. El valor de la proporción en la muestra es una variable aleatoria cuya distribución está emparentada directamente con la binomial (si se tratara del número de defectuosos, sería binomial).

Como cualquier variable aleatoria, el estimador tiene

- distribución de probabilidad.
- valor esperado.
- desviación estándar / varianza.

Valor esperado de un estimador y sesgo

El valor esperado de un estimador nos da un valor alrededor del cual es muy probable que se encuentre el valor del estimador. Para poner un ejemplo, si supieramos que el valor esperado de una estadística es 4, esto significaría que al tomar una muestra:

- No creemos que el valor de la estadística vaya a ser 4.
- Pero tampoco creemos que el valor de la estadística vaya a estar lejos de 4.

Ya que es muy probable que el valor del estimador esté cerca de su valor esperado, una propiedad muy deseable es que ese valor esperado del estimador coincida con el del parámetro que se pretende estimar. Al menos, quisiéramos que el valor esperado no difiera mucho del parámetro estimado.

Por esa razón es importante la cantidad que, técnicamente llamamos sesgo. El sesgo es la diferencia entre el valor esperado del estimador y el parámetro que estima.

Si el sesgo 0, se dice que el estimador es instigado y ésta es una característica buena para un estimador. Un estimador que es instigado tiene una alta probabilidad de tomar un valor cercano al valor del parámetro.

Varianza de un estimador

Otra propiedad importante de un estimador es su varianza (o su raíz cuadrada, la desviación estándar).

La importancia de la desviación estándar es que nos permite darle un sentido numérico a la cercanía del valor del estimador a su valor esperado.

Entre menor sea la desviación estándar (o la varianza) de un estimador, será más probable que su valor en una muestra específica se encuentre mas cerca del valor esperado. Para aclarar esto, considere dos estimadores T1 y T2, suponga que ambos son instigados y suponga que la varianza de T1 es menor que la de T2 ¿Qué quiere decir esto? Simplemente que en un entorno fijo del valor del parámetro, los valores de T1 son más probables que los de T2. O sea que vamos a encontrar a T1 más cerca del valor del parámetro que a T2. Esto hace que nuestras preferencias estén con T1.

Cuando un estimador tiene una varianza menor que otro decimos que el estimador es más eficiente.

En el pizarrón vemos algunos estimadores instigados:

- la proporción muestra como estimador de la proporción poblaciones.
- la media muestra como estimador del valor esperado poblaciones.
- la varianza de la muestra como estimador de la varianza de la población.

La distribución de probabilidad de una estadística

Quizá el resultado mas importante para la estadística es el Teorema del Límite Central. Este resultado nos indica que, para la estadística promedio de la muestra

- el valor esperado es la media de la población.
- la varianza es igual a la de la población dividida por el número de elementos de la muestra.
- la distribución de probabilidad es la normal.

Este teorema es muy importante porque permite calcular probabilidades acerca de dónde se encuentra el valor del promedio muestra. Es sólo cuestión de usar la tabla normal teniendo cuidado al estandarizar de usar la desviación estándar adecuada que es la de la población dividida por la raíz cuadrada del número de elementos de la muestra.

En el salón hacemos en forma detallada, ejemplos de estos cálculos.

Estimación del error de una medida directa

La estimación del error de una medida tiene siempre una componente subjetiva. En efecto, nadie mejor que un observador experimentado para saber con buena aproximación cuál es el grado de confianza que le merece la medida que acaba de tomar. No existe un conjunto de reglas bien fundadas e inalterables que permitan determinar el error de una medida en todos los casos imaginables. Muchas veces es tan importante consignar cómo se ha obtenido un error como su propio valor.

Sin embargo, la aplicación de algunos métodos estadísticos permite objetivar en gran medida la estimación de errores aleatorios. La estadística permite obtener los parámetros de una población (en este caso el conjunto de todas las medidas que es posible tomar de una magnitud), a partir de una muestra (el número limitado de medidas que podemos tomar).

Mejor valor de un conjunto de medidas

Supongamos que medimos una magnitud un número n de veces. Debido a la existencia de errores aleatorios, las n medidas serán en general diferentes

El método más razonable para determinar el mejor valor de estas medidas es tomar el valor medio. En efecto, si los errores son debidos al azar, tan probable es que ocurran por defecto como por exceso, y al hacer la media se compensarán, por lo menos parcialmente. El valor medio se define por:

y este es el valor que deberá darse como resultado de las medidas.

2. Tipos de estimación estadística

Estimación de parámetros:

Estimaciones sin sesgo:

Si la media de las dispersiones de muestreo con un estadístico es igual que la del correspondiente parámetro de la población, el estadístico se llamara estimador sin sesgo, del parámetro; si no, si no se llama estimador sesgado. Los correspondientes valores de tal estadístico se llaman estimación sin sesgo, y estimación con sesgo respectivamente.

Ejemplo 1: la media de las distribuciones de muestreo de medias e, media de la población. Por lo tanto, la media muestral es una estimación sin sesgo de la media de la población.

Ejemplo 2. Las medias de las distribuciones de muestreo de las variables es:

Encontramos, de manera que es una estimación sin sesgo de. Sin embargo, s es una estimación sesgada de. En términos de esperanza podríamos decir que un estadístico es instigado porque

Estimación Eficiente:

Si las distribuciones de muestreo de dos estadísticos tienen la misma media(o esperanza), el de menor varianza se llama un estimador eficiente de la media, mientras que el otro se llama un estimador ineficiente, respectivamente.

Si consideramos todos los posibles estadísticos cuyas distribuciones de muestreo tiene la misma media, aquel de varianza mínima se llama aveces, el estimador de máxima eficiencia, ósea el mejor estimador.

Ejemplo:

Las distribuciones de muestreo de media y mediana tienen ambas la misma media, a saber, la media de la población. Sin embargo, la varianza de la distribución de muestreo de medias es menor que la varianza de la distribución de muestreo de medianas. Por tanto, la media muestral da una estimación eficiente de la media de la población, mientras la mediana de la muestra da una estimación ineficiente de ella.

De todos los estadísticos que estiman la media de la población, la media muestral proporciona la mejor(la más eficiente) estimación.

En la practica, estimaciones ineficientes se usan con frecuencia a causa de la relativa sencillez con que se obtienen algunas de ellas.

Estimaciones de punto y estimaciones de intervalo, su fiabilidad:

Una estimación de un parámetro de la población dada por un solo número se llama una estimación de punto del parámetro. Una estimación de un parámetro de la población dada por dos puntos, entre los cuales se pueden considerar encajado al parámetro, se llama una estimación del intervalo del parámetro.

Las estimaciones de intervalo que indican la precisión de una estimación y son por tanto preferibles a las estimaciones de punto

Ejemplo:

Si decimos que una distancia sé a medido como 5.28 metros (m), estamos dando una estimación de punto. Por otra parte, si decimos que la distancia es 5.28 ± 0.03 m, (ósea, que esta entre 5.25 y 5.31 m), estamos dando una estimación de intervalo

El margen de error o la percepción de una estimación nos informa su fiabilidad.

Estimaciones De Intervalos De Confianza Para Parámetros De Población:

Sean y la media y la desviación típica (error típico) de la distribución de muestreo de un estadístico S. Entonces, si la distribución de muestreo de s es aproximadamente normal (que como hemos visto es cierto para muchos estadísticos si el tamaño de la muestra es N³30), podemos esperar hallar un estadisco muestral real S que este en los intervalos alrededor del 68.27 %, 95.45% y 99.7 % del tiempo restante, respectivamente.

La tabla 1. Corresponde a los niveles de confianza usados en la practica. Para niveles de confianza que no aparecen en la tabla, los valores Z_c se pueden encontrar gracias a las tablas de áreas bajo la curva normal.

Nivel de confianza	99.7 % 99% 98% 96% 95.45% 95% 90% 80% 6827% 50%
Z_c	3.00 2.58 2.33 2.05 2.00 1.96 1.645 1.28 1.00 0.6745

Intervalos de confianza para la media:

Si el estadístico s de la media de la muestra, entonces los limites de confianza respectivamente. Mas en general los limites de confianza para estimar la media de la población m viene dado por usando los valores de

Si el muestreo de la población es infinita por lo tanto viene dado por:

Si el muestro es sin reposición de una población de tamaño Np.

Ejemplo

Halar laos limites de confianza de 98% y 90%.para los diámetros de una bolsa

Solución:

Sea Z =Zc tal que al área bajo la curva normal a la derecha sea 1% . Entonces , por simetría el área del lado izquierdo de Z=-Zc . como el área total bajo la curva es 1, Zc= 0.49 por lo tanto, Zc=2.33. luego el limite de confianza es 98% son X= ±2.33s¤ÖN=0.824± 2.33(0.042/ Ö200)=0.824 ±0.069 cm.

Generalmente, la desviación típica de la población no es conocida. Así pues , para obtener los limites usamos la estimación s o S es satisfactorio si N>=30, si a aproximación es pobre y debe de empleare la teoría de pequeñas muestras.

Cálculo del tamaño de la muestra

A la hora de determinar el tamaño que debe alcanzar una muestra hay que tomar en cuenta varios factores: el tipo de muestreo, el parámetro a estimar, el error muestral admisible, la varianza poblacional y el nivel de confianza. Por ello antes de presentar algunos casos sencillos de cálculo del tamaño muestral delimitemos estos factores.

Parámetro. Son las medidas o datos que se obtienen sobre la población.

Estadístico. Los datos o medidas que se obtienen sobre una muestra y por lo tanto una estimación de los parámetros.

Error Muestral, de estimación o standard. Es la diferencia entre un estadístico y su parámetro correspondiente. Es una medida de la variabilidad de las estimaciones de muestras repetidas en torno al valor de la población, nos da una noción clara de hasta dónde y con qué probabilidad una estimación basada en una muestra se aleja del valor que se hubiera obtenido por medio de un censo completo. Siempre se comete un error, pero la naturaleza de la investigación nos indicará hasta qué medida podemos cometerlo (los resultados se someten a error muestral e intervalos de confianza que varían muestra a muestra). Varía según se calcule al principio o al final. Un estadístico será más preciso en cuanto y tanto su error es más pequeño. Podríamos decir que es la desviación de la distribución muestral de un estadístico y su fiabilidad.

Nivel de Confianza. Probabilidad de que la estimación efectuada se ajuste a la realidad. Cualquier información que queremos recoger está distribuida según una ley de probabilidad (Gauss o Student), así llamamos nivel de confianza a la probabilidad de que el intervalo construido en torno a un estadístico capte el verdadero valor del parámetro.

Varianza Poblacional. Cuando una población es más homogénea la varianza es menor y el número de entrevistas necesarias para construir un modelo reducido del universo, o de la población, será más pequeño. Generalmente es un valor desconocido y hay que estimarlo a partir de datos de estudios previos.

Tamaño de muestra para estimar la media de la población

Veamos los pasos necesarios para determinar el tamaño de una muestra empleando el muestreo aleatorio simple. Para ello es necesario partir de dos supuestos: en primer lugar el nivel de confianza al que queremos trabajar; en segundo lugar, cual es el error máximo que estamos dispuestos a admitir en nuestra estimación. Así pues los pasos a seguir son:

Veamos los pasos necesarios para determinar el tamaño de una muestra empleando el muestreo aleatorio simple. Para ello es necesario partir de dos supuestos: en primer lugar el nivel de confianza al que queremos trabajar; en segundo lugar, cual es el error máximo que estamos dispuestos a admitir en nuestra estimación. Así pues los pasos a seguir son:
1.- Obtener el tamaño muestral imaginando que N->a

Donde:

z correspondiente al nivel de confianza elegido
: varianza poblacional
e: error máximo

2.- Comprobar si se cumple

Si esta condición se cumple el proceso termina aquí, y ese es el tamaño adecuado que debemos muestrear.

Si no se cumple, pasamos a una tercera fase:

3.- Obtener el tamaño de la muestra según la siguie
n te fórmula:

Veamos un ejemplo: La Consejería de Trabajo planea un estudio con el interés de conocer el promedio de horas semanales trabajadas por las mujeres del servicio doméstico. La muestra será extraída de una población de 10000 mujeres que figuran en los registros de la Seguridad Social y de las cuales se conoce a través de un estudio piloto que su varianza es de 9.648. Trabajando con un nivel de confianza de 0.95 y estando dispuestos a admitir un error máximo de 0,1, ¿cuál debe ser el tamaño muestral que Empleemos?.

Buscamos en las tablas de la curva normal el valor de que corresponde con el

nivel de confianza elegido: = ±1.96 y seguimos los pasos propuestos arriba.

Tamaño de muestra para estimar la proporción de la población

Para calcular el tamaño de muestra para la estimación de proporciones poblaciones hemos de tener en cuenta los mismos factores que en el caso de la media. La fórmula que nos permitirá determinar el tamaño muestral es la siguiente:

: z correspondiente al nivel de confianza elegido
P: proporción de una categoría de la variable
e: error máximo
N: tamaño de la población

Siguiendo con el estudio planteado en el punto anterior, supongamos que tratamos de estimar la proporción de mujeres que trabajan diariamente 10 horas o más. De un estudio piloto se dedujo que P=0.30, fijamos el nivel de confianza en 0.95 y el error máximo 0.02.

5.3.1. INTERVALO DE CONFIANZA PARA UNA PROPORCIÓN

Sean $X_1, \ldots, X_n \rightsquigarrow \mathbf{Ber}(p)$. Si queremos estimar el parámetro *p*, la manera más natural de hacerlo consiste en definir la suma de estas --lo que nos proporciona una distribución Binomial :

$$X = X_1 + \cdots + X_n \rightsquigarrow \mathbf{B}(n, p)$$

y tomar como estimador suyo la v.a.

$$\hat{p} = \frac{X}{n}.$$

Es decir, tomamos como estimación de *p* la proporción de éxitos obtenidos en las *n* pruebas[8.1], $\hat{p}$.

La distribución del número de éxitos es binomial, y puede ser aproximada a la normal cuando el tamaño de la muestra *n* es grande, y *p* no es una cantidad muy cercana a cero o uno:

$$X \rightsquigarrow \mathbf{B}(n, p) \Rightarrow X \overset{\approx}{\rightsquigarrow} \mathbf{N}(np, npq)$$

El estimador $\hat{p}$ no es más que un cambio de escala de *X*, por tanto

$$\hat{p} = \frac{X}{n} \overset{\approx}{\rightsquigarrow} \mathbf{N}\left(p, \frac{pq}{n}\right) \qquad \Longrightarrow \qquad \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \approx Z \rightsquigarrow \mathbf{N}(0, 1)$$

Esta expresión presenta dificultades para el cálculo, siendo más cómodo sustituirla por la siguiente aproximación:

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}\hat{q}}{n}}} \approx Z \rightsquigarrow \mathbf{N}(0, 1)$$

Para encontrar el intervalo de confianza al nivel de significación α para *p* se considera el intervalo que hace que la distribución de $Z \rightsquigarrow \mathbf{N}(0,1)$ deje la probabilidad α fuera del mismo. Es decir, se considera el intervalo cuyos extremos son los cuantiles $\alpha/2$ y $1-\alpha/2$. Así se puede afirmar con una confianza de $1-\alpha$ que:

$$\underbrace{z_{\alpha/2}}_{-z_{1-\alpha/2}} \leq Z \leq z_{1-\alpha/2} \Leftrightarrow |Z| \leq z_{1-\alpha/2}$$

$$\Leftrightarrow \frac{|\hat{p}-p|}{\sqrt{\frac{\hat{p}\hat{q}}{n}}} \leq z_{1-\alpha/2}$$

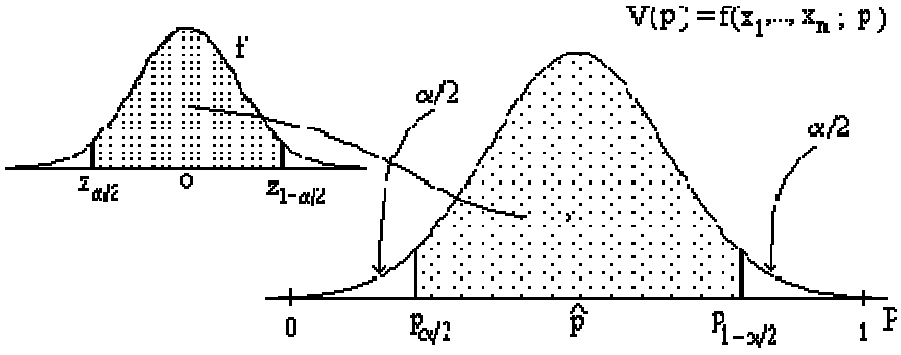
$$\Leftrightarrow |\hat{p}-p| \leq z_{1-\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Esto se resume en la siguiente expresión:

$$p = \hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

con una confianza de $1-\alpha$

Figura: Intervalo de confianza para una proporción.



8.6.2.1 Ejemplo

Se quiere estimar el resultado de un referéndum mediante un sondeo. Para ello se realiza un muestreo aleatorio simple con *n*=100 personas y se obtienen 35% que votarán a favor y 65% que votarán en contra (suponemos que no hay

indecisos para simplificar el problema a una variable dicotómica). Con un nivel de significación del 5%, calcule un intervalo de confianza para el verdadero resultado de las elecciones.

Solución: Dada una persona cualquiera (*i*) de la población, el resultado de su voto es una variable dicotómica:

$$X_i \rightsquigarrow \mathbf{Ber}(p)$$

El parámetro a estimar en un intervalo de confianza con $\alpha = 0,05$ es *p*, y tenemos sobre una muestra de tamaño *n*=100, la siguiente estimación puntual de *p*:

$$\hat{p} = \frac{35}{100} = 0,35 \Longrightarrow \hat{q} = 0,65$$

Sabemos que

$$\frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \overset{\approx}{\rightsquigarrow} \mathbf{N}(0,1)$$

En la práctica el error que se comete no es muy grande si tomamos algo más simple como

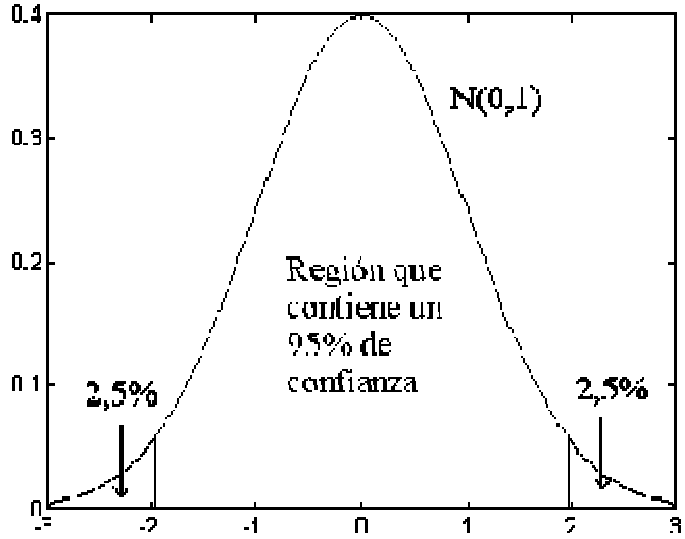
$$Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}\hat{q}}{n}}} \overset{\approx}{\rightsquigarrow} \mathbf{N}(0,1)$$

Así el intervalo de confianza buscado lo calculamos como se indica en la :

$$|Z| \leq z_{1-\alpha/2} \iff \frac{|0,35 - p|}{\sqrt{\frac{0,35 \times 0,65}{100}}} \leq z_{0,975} = 1,96$$
$$\iff p = 0,35 \pm 0,0935$$

Por tanto, tenemos con esa muestra un error aproximado de 9,3 puntos al nivel de confianza del 95%.

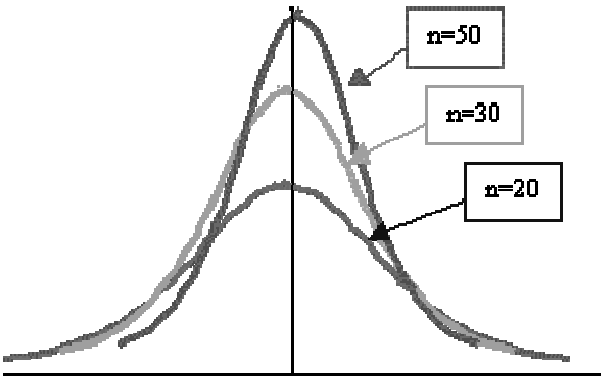
Figura: Región a partir de la cual se realiza una estimación confidencial para una proporción, con una confianza del 95%.



5.3.2. FACTOR DE CORRECCIÓN PARA UNA POBLACIÓN FINITA

Teorema del límite central

Si se seleccionan muestras aleatorias de *n* observaciones de una población con media μ y desviación estándar σ, entonces, cuando *n* es grande, la distribución muestral de medias tendrá aproximadamente una distribución normal con una media igual a μ y una desviación estándar de $\sigma/\sqrt{n}$. La aproximación será cada vez más exacta a medida de que *n* sea cada vez mayor.



Ejemplo

Para la dsitribución muestral de medias del ejercicio pasado, encuentre:

a. El error muestral de cada media
b. La media de los errores muestrales
c. La desviación estándar de los errores muestrales.

Solución:

a. En la tabla siguiente se ven las muestras, las medias de las muestras y los errores muestrales:

Muestra	**x**	**Error muestral, e=x-** μ
(0,0)	0	0 - 3 = -3
(0,2)	1	1 - 3 = -2
(0,4)	2	2 - 3 = -1
(0,6)	3	3 – 3 = 0
(2,0)	1	1 – 3 = -2
(2,2)	2	2 – 3 = -1
(2,4)	3	3 – 3 = 0
(2,6)	4	4 – 3 = 1
(4,0)	2	2 – 3 = -1
(4,2)	3	3 – 3 = 0
(4,4)	4	4 – 3 = 1
(4,6)	5	5 – 3 = 2
(6,0)	3	3 – 3 = 0
(6,2)	4	4 – 3 = 1
(6,4)	5	5 – 3 = 2
(6,6)	6	6 – 3 = 3

b. La media de los errores muestrales es μ_e, es:

$$\mu_e = \frac{(-3) + (-2) + (-1) + 0 + \ldots + 2 + 3}{16} = 0$$

c. La desviación estándar de la distribución de los errores muestrales $\sigma_{e,}$ es entonces:

$$\sigma_e = \sqrt{\frac{\Sigma\left[(e-\mu_e)^2 f\right]}{N}} = \sqrt{\frac{(-3-0)^2 1+(-2-0)^2 2+(-1-0)^2 3+(0-0)^2 4+(1-0)^2 3+(2-0)^2 2+(3-0)^2 1}{16}} = 1.58$$

La desviación estándar de la distribución muestral de un estadístico se conoce como ***error estándar del estadístico***. Para el ejercicio anterior el error estándar de la media denotado por σ_x, es 1.58. Con esto se puede demostrar que si de una población se eligen muestras de tamaño *n* ***con reemplazo***, entonces el error estándar de la media es igual a la desviación estándar de la distribución de los errores muestrales.

En general se tiene: $\sigma_x = \sigma_e$

Cuando las muestras se toman de una población pequeña y sin reemplazo, se puede usar la formula siguiente para encontrar σ_x .

$$\sigma_x = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

donde σ es la desviación estándar de la población de donde se toman las muestras, *n* es el tamaño de la muestra y *N* el de la población.

Como rfegla de cálculo, si el muestreo se hace sin reemplazo y el tamaño de la población es al menos 20 veces el tamaño de la muestra ($N \geq 20$), entonces se puede usar la fórmula.

El factor $\sqrt{\frac{N-n}{N-1}}$ se denomina ***factor de corrección*** para una población finita.

Ejemplo:

Suponga que la tabla siguiente muestra la antiguedad en años en el trabajo de tres maestros universitarios de matemáticas:

Maestro de matemáticas	Antiguedad
A	6
B	4
C	2

Suponga además que se seleccionan muestras aleatorias de tamaño 2 sin reemplazo. Calcule la antigüedad media para cada muestra, la media de la distribución muestral y el error estándar, o la desviación estándar de la distribución muestral.

Solución:

Se pueden tener ${}_3C_2$ =3 muestras posibles. La tabla lista todas las muestras posibles de tamaño 2, con sus respectivas medias muestrales.

Muestras	Antigüedad	Media Muestral
A,B	(6,4)	5
A,C	(6,2)	4
B,C	(4,2)	3

La media poblacional es: $\mu = \frac{2+4+6}{3} = 4$

La media de la distribución muestral es: $\mu_{\bar{x}} = \frac{5+4+3}{3} = 4$

La desviación estándar de la población es:

$$\sigma = \sqrt{\frac{(6-4)^2 + (4-4)^2 + (2-4)^2}{3}} = 1.63$$

El error estándar o la desviación estándar de la distribución muestral es:

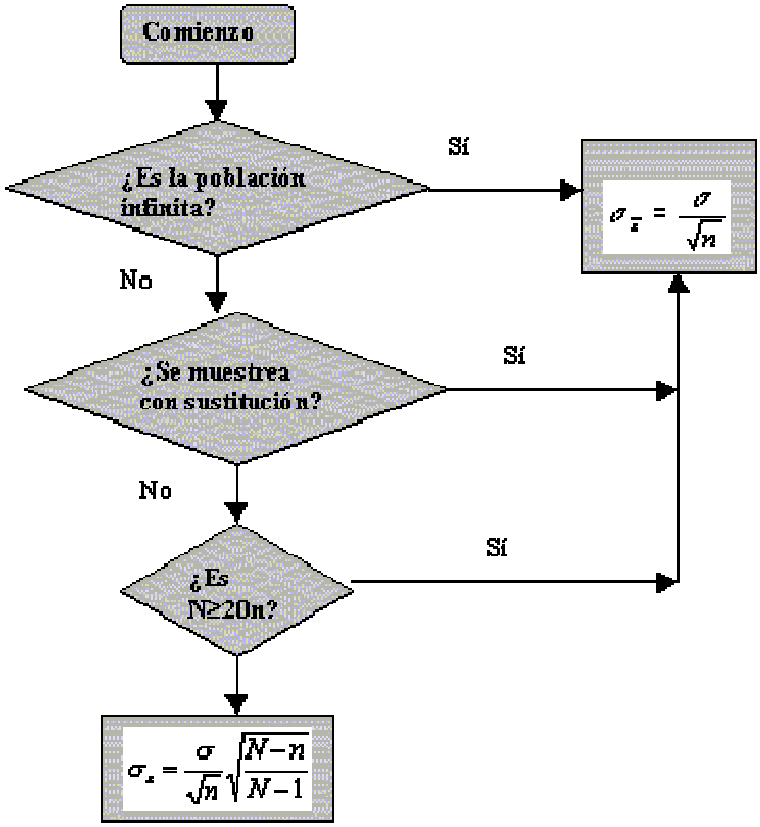
$$\sigma_{\bar{x}} = \sqrt{\frac{(5-4)^2 + (4-4)^2 + (3-4)^2}{3}} = 0.816$$

Si utilizamos la fórmula del error estándar sin el factor de correción tendriamos que: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1.63}{\sqrt{2}} = 1.152$

Por lo que observamos que este valor no es el verdadero. Agregando el factor de corrección obtendremos el valor correcto:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{1.63}{\sqrt{2}} \sqrt{\frac{3-2}{3-1}} = 0.816$$

El diagrama de flujo resume las decisiones que deben tomarse cuando se calcula el valor del error estándar:



Distribución Muestral de Medias

Si recordamos a la distribución normal, esta es una distribución continua, en forma de campana en donde la media, la mediana y la moda tienen un mismo valor y es simétrica.

Con esta distribución podíamos calcular la probabilidad de algún evento relacionado con la variable aleatoria, mediante la siguiente fórmula:

$$z = \frac{x - \mu}{\sigma}$$

En donde z es una variable estandarizada con media igual a cero y varianza igual a uno. Con esta fórmula se pueden a hacer los cálculos de probabilidad para cualquier ejercicio, utilizando la tabla de la distribución z.

Sabemos que cuando se extraen muestras de tamaño mayor a 30 o bien de cualquier tamaño de una población normal, la distribución muestral de medias tiene un comportamiento aproximadamente normal, por lo que se puede utilizar la formula de la distribución normal con $\mu = \mu_{\bar{x}}$ y $\sigma = \sigma_{\bar{x}}$, entonces la fórmula para calcular la probabilidad del comportamiento del estadístico, en este caso la media de la muestra , quedaría de la siguiente manera:

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

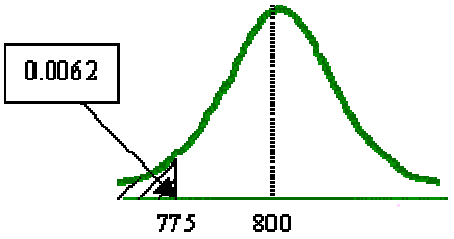
y para poblaciones finitas y muestro con reemplazo:

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n} \sqrt{\frac{N-n}{N-1}}}$$

Ejemplo:

Una empresa eléctrica fabrica focos que tienen una duración que se distribuye aproximadamente en forma normal, con media de 800 horas y desviación estándar de 40 horas. Encuentre la probabilidad de que una muestra aleatoria de 16 focos tenga una vida promedio de menos de 775 horas.

Solución:



$$z = \frac{775 - 800}{40 / \sqrt{16}} = -2.5$$

Este valor se busca en la tabla de z

$$P(\bar{x} \leq 775) = P(z \leq -2.5) = 0.0062$$

La interpretación sería que la probabilidad de que la media de la muestra de 16 focos sea menor a 775 horas es de 0.0062.

Ejemplo:

Las estaturas de 1000 estudiantes están distribuidas aproximadamente en forma normal con una media de 174.5 centímetros y una desviación estándar de 6.9 centímetros. Si se extraen 200 muestras aleatorias de tamaño 25 sin reemplazo de esta población, determine:

a. El número de las medias muestrales que caen entre 172.5 y 175.8 centímetros.
b. El número de medias muestrales que caen por debajo de 172 centímetros.

Solución:

Como se puede observar en este ejercicio se cuenta con una población finita y un muestreo sin reemplazo, por lo que se tendrá que agregar el factor de corrección. Se procederá a calcular el denominador de Z para sólo sustituirlo en cada inciso.

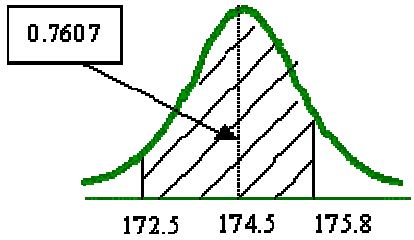
$$\sigma/\sqrt{n}\sqrt{\frac{N-n}{N-1}} = 6.9/\sqrt{25}\sqrt{\frac{1000-25}{1000-1}} = 1.36$$

a.

$$z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}\sqrt{\frac{N-n}{N-1}}} = \frac{172.5-174.5}{1.36} = -1.47$$

$$z = \frac{175.8-174.5}{1.36} = 0.96$$

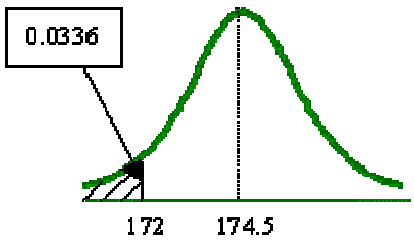
$$p(172.5 \leq \bar{x} \leq 175.8) = 0.7607$$



(0.7607)(200)=152 medias muestrales

b. $$z = \frac{172-174.5}{1.36} = -1.83$$

$$p(\bar{x} \leq 172) = 0.00336$$



(0.0336)(200)= 7 medias muestrales

5.3.3. ELECCIÓN TAMAÑO APROPIADO DE UNA MUESTRA

A la hora de determinar el tamaño que debe alcanzar una muestra hay que tomar en cuenta varios factores: el tipo de muestreo, el parámetro a estimar, el error muestral admisible, la varianza poblacional y el nivel de confianza. Por ello antes de presentar algunos casos sencillos de cálculo del tamaño muestral delimitemos estos factores.

Parámetro. Son las medidas o datos que se obtienen sobre la población.

Estadístico. Los datos o medidas que se obtienen sobre una muestra y por lo tanto una estimación de los parámetros.

Error Muestral, de estimación o standard. Es la diferencia entre un estadístico y su parámetro correspondiente. Es una medida de la variabilidad de las estimaciones de muestras repetidas en torno al valor de la población, nos da una noción clara de hasta dónde y con qué probabilidad una estimación basada en una muestra se aleja del valor que se hubiera obtenido por medio de un censo completo. Siempre se comete un error, pero la naturaleza de la investigación nos indicará hasta qué medida podemos cometerlo (los resultados se someten a error muestral e intervalos de confianza que varían muestra a muestra). Varía según se calcule al principio o al final. Un estadístico será más preciso en cuanto y tanto su error es más pequeño. Podríamos decir que es la desviación de la distribución muestral[(1)] de un estadístico y su fiabilidad.

Nivel de Confianza. Probabilidad de que la estimación efectuada se ajuste a la realidad. Cualquier información que queremos recoger está distribuida según una ley de probabilidad (Gauss o Student), así llamamos nivel de confianza a la probabilidad de que el intervalo construido en torno a un estadístico capte el verdadero valor del parámetro.

Varianza Poblacional. Cuando una población es más homogénea la varianza es menor y el número de entrevistas necesarias para construir un modelo reducido del universo, o de la población, será más pequeño. Generalmente es un valor desconocido y hay que estimarlo a partir de datos de estudios previos.

3.1.- Tamaño de muestra para estimar la media de la población

Veamos los pasos necesarios para determinar el tamaño de una muestra empleando el muestreo aleatorio simple. Para ello es necesario partir de dos supuestos: en primer lugar el nivel de confianza al que queremos trabajar; en segundo lugar, cual es el error máximo que estamos dispuestos a admitir en nuestra estimación. Así pues los pasos a seguir son:

1.- Obtener el tamaño muestral imaginando que $N \to \infty$:

$$n_o = \frac{Z^2_{\alpha/2}\sigma^2}{e^2}$$

donde:

$Z_{\alpha/2}$: z correspondiente al nivel de confianza elegido

σ^2: varianza poblacional

e: error máximo

2.- Comprobar si se cumple $N > n_o(n_o - 1)$

si esta condición se cumple el proceso termina aquí, y ese es el tamaño adecuado que debemos muestrear.

Si no se cumple, pasamos a una tercera fase:

3.- Obtener el tamaño de la muestra según la siguiente fórmula:

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

Veamos un ejemplo: La Consejería de Trabajo planea un estudio con el interés de conocer el promedio de horas semanales trabajadas por las mujeres del servicio doméstico. La muestra será extraída de una población de 10000 mujeres que figuran en los registros de la Seguridad Social y de las cuales se conoce a través de un estudio piloto que su varianza es de 9.648. Trabajando con un nivel de confianza de 0.95 y estando dispuestos a admitir un error máximo de 0,1, ¿cuál debe ser el tamaño muestral que empleemos?.

Buscamos en las tablas de la curva normal el valor de $Z_{\alpha/2}$ que corresponde con el nivel de confianza elegido: $Z_{\alpha/2} = \pm 1.96$ y seguimos los pasos propuestos arriba.

1.-

$$n_o = \frac{1.96^2 \cdot 9.648}{0.1^2} = 3706$$

2.- Comprobamos que no se cumple $N > n_o(n_o - 1)$, pues en este caso

10000 < 3706 (3706 - 1); 10000 < 13730730

3.-

$$n = \frac{3706}{1 + \frac{3706}{10000}} = 2704$$