VI. MUESTREO

Importancia: Una vez que definimos, explicamos e ilustramos el concepto de **probabilidad,** vimos que constituía el eje rector para hacer análisis económico ampliado, sobre la base de la estadística descriptiva, a partir de la inferencia estadística (que se basa en el análisis de una muestra para inferir las características de la población de la que proviene). Lo anterior fue muy valioso porque a partir de la **naturaleza** y número de resultados posibles que se generan en un experimento, pudimos describir cómo se forman y que características tienen las distribuciones probabilísticas discretas y continuas.

Así, continuando con su aplicación, ahora veremos como se usa para la obtención de muestras probabilísticas, que obtendremos de poblaciones finitas e infinitas. Motivo por el cual es conveniente introducir de manera formal la definición de los siguientes conceptos:

VI.1 CONCEPTO DE UNIVERSO Y MUESTRA:

UNIVERSO O POBLACIÓN: Se define como la suma de las unidades elementales.

Si el número de unidades elementales es igual al número de observaciones; se dice que la población es la suma de las observaciones.

Por ejemplo: Si hay 600 personas e interesa la variable X y el peso en Kgs. de los personas, cada persona es una unidad elemental y por lo tanto la población son las 600 personas.

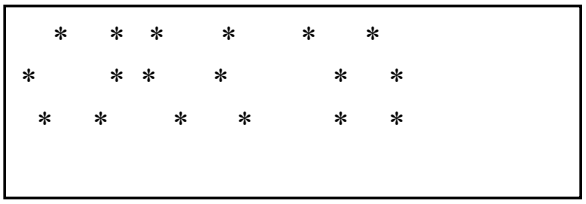
El tamaño de una población se representa generalmente por N. Luego, una población en sentido estadístico es un conjunto de elementos -generalmente definida- que puede conocerse por medio de un análisis completo y exhaustivo.

La población puede ser: **FINITA** o **INFINITA**.

El ejemplo de las 600 personas ilustra una población FINITA, una población INFINITA podrá referirse por ejemplo al número de moscos que hay en el mundo entero. Cada una de las unidades elementales, tiene varias características identificables y numerables; es decir que cada característica puede representarse por un número.

Ejemplo: Si la población es de animales, sus características pueden ser:
-Su peso;
-La dieta a que están sujetos;
-Su producción (según su clase: vacas, gallinas, etc.).

En la teoría de la probabilidad moderna, una población se representa gráficamente en la siguiente forma:



(R, Q, P)

Donde R representa el conjunto-Universo o población;
Q es una σ álgebra de Boole;
P es la medida de la probabilidad dentro de la población.
Una muestra es un conjunto de n observaciones-unidades elementales-extraídas de la población. Esta n es el tamaño de la muestra.

Si en el ejemplo anterior se seleccionan 20 personas de la población de 600, se ha tomado una muestra de tamaño: n = 20.

El tipo de muestra y representatividad que se obtiene con ella depende de la forma en que haya sido extraída la muestra de la población. Así se habla de procedimients o métodos como el muestreo simple aleatorio, de muestreo sistemático, de muestreo estratificado, por conglomerados, etcétera. Todos ellos tienen en común el hecho de que seleccionan la muestra al azar, que se conoce como muestra probabilística y tiene características importantes que más adelante describiremos.

METODOLOGIA DEL MUESTREO ESTADÍSTICO. (11)

VI.2 MÉTODOS DE MUESTREO:

Existen: el muestreo empírico y el probabilista. El primero, suele usarse cuando se tiene un amplio conocimiento del fenómeno que se investigará y cuando existen estudios previos al respecto; tal que el estadístico tiene antecedentes y el costo para la investigación es reducido. Este tipo de muestreo se recomienda cuando no se desea un análisis profundo y preciso sobre las características del universo que se estudia. Este método resulta en ocasiones

bueno, ya que capta con relativa facilidad las características de la población en estudio. Como podrá notarse, no es del todo científico y no permite por sí mismo llegar a estimaciones precisas, resultando difícil realizar inferencias en la estimación.

El método científico -por lo contrario- proporciona una medida de la magnitud del error y de la confianza con que se puede tomar los resultados. Generalmente suele ser más costoso y quizás tome un poco más de tiempo el realizarlo, en especial cuando hay problema de información sobre el número de unidades que integran el universo y algunas otras características que permiten el cálculo rápido del tamaño de la muestra, teniendo además que gastarse cierto número de horas-hombre en la recavación de la información requerida.

Es recomendable, sin embargo, usar siempre el método científico para dotar a los estudios de seguridad matemática, aún cuando se tengan que hacer esfuerzos extraordinarios para conseguir los recursos monetarios necesarios.

En otras palabras, estos términos no son otra cosa más que sinónimos de una selección aleatoria y una selección arbitraria respectivamente.

Un muestreo probabilístico es aquel cuyo error de muestreo es calculado, condición que existe solo cuando se usa la selección aleatoria.

La palabra "aleatoria" se refiere al método de seleccionar una muestra, más bien que a la muestra particular elegida. Cualquier muestra posible puede ser al azar o aleatoria, por muy poco representativa que pueda ser de la población, con tal que haya sido obtenida siguiendo la regla de dar una probabilidad igual a cada una de las muestras posibles.

Por otra parte, una muestra arbitraria o a criterio, es aquella cuyo error no es determinado ni asignada ninguna probabilidad de selección a los elementos o unidades que la componen.

VI.2.1 ERRORES DE MUESTREO Y DE NO MUESTREO.

La exactitud o confiabilidad de una muestra, depende de dos tipos básicos de errores: errores de muestreo, que se reflejan en estimaciones matemáticas de la precisión de estimadores provenientes de muestras particulares, y se manifiestan en diferentes formas clasificadas bajo la notación de sesgos o distorsiones.

Los errores de muestreo se miden a través de las llamadas fórmulas de error estándar. De acuerdo con estas fórmulas, se hacen estimaciones de la precisión de estimadores muéstrales particulares y siguiendo el procedimiento apropiado; estas mismas fórmulas sirven de base para determinar el tamaño de la muestra requerida, de acuerdo con una precisión especificada

previamente. Las fórmulas del error estándar han sido desarrolladas para una gran variedad de diseños muéstrales y en la actualidad es una cuestión rutinaria su aplicación a cada uno de los casos.

Los errores de muestreo surgen de la variación en los estimadores provenientes de distintas muestras del mismo tamaño.

El valor de los errores determina la precisión con que los valores muéstrales estiman a los parámetros poblacionales.

La probabilidad de que cualquier estimador caiga dentro de un cierto rango del parámetro poblacional, se obtiene por medio de la teoría de la probabilidad para distintos diseños muéstrales.

Así, en base a esta teoría, el margen de error -o error de muestreo- que se puede esperar con un diseño de muestreo y tamaño de muestra determinados, se puede calcular a diferentes niveles de precisión bajo el supuesto de una selección aleatoria, la cual requiere que cada miembro de la población tenga la misma probabilidad de ser seleccionado. Luego, una vez que se conocen el error estándar y la precisión buscada, se pueden calcular: el tamaño de la muestra y los recursos necesarios para la investigación.

Contrariamente, el tema de los errores de no muestreo es a la fecha un tema que requiere una vasta experiencia y la cual es ajena a la disciplina matemática.

Incluidas en el concepto de errores no de muestreo, están las innumerables influencias que tienden a distorsionar o sesgar los estimadores provenientes de la muestra, la selección arbitraria de los miembros de la muestra, fraseo perjudicial en las preguntas, actitudes preconcebidas por el entrevistador y muchos otros factores pueden producir valores muestrales que no representaran a los valores de los parámetros de la población, no importa que tan grande sea la muestra.

Distintos a los errores de muestreo, éste tipo de sesgo es independiente del tamaño de la muestra.

VI.2.2 SELECCIÓN DE LA UNIDAD DE MUESTREO.

La aplicación de los métodos de muestreo estadístico tiene por objeto, seleccionar algunos elementos del universo que se trata de estudiar, para poder hacer inferencias sobre sus características. La selección de estas unidades se hace a partir de una lista, mapas, croquis, directorios-o una combinación de

estos elementos informativos-, los que deben contener todas las unidades de interés y permitir determinar la probabilidad de su inclusión; así mismo, que en el momento de levantar la encuesta, la identificación de cada unidad en la muestra sea hecha sin ninguna ambigüedad.

Al conjunto de todos los elementos se le llama: **MARCO MUESTRAL**.

De acuerdo a la forma de seleccionar estas unidades se pueden dar las siguientes maneras de hacerla:

Reemplazo:

Las selecciones sucesivas de una muestra probabilística pueden hacerse con o sin reemplazo de las unidades obtenidas en las selecciones previas; por ello al primer procedimiento se le llama muestreo con reemplazo y al segundo sin reemplazo.

En el muestreo con reemplazo, al hacer las estimaciones, cada unidad de la muestra debe considerarse en un número igual al de veces que haya salido en la muestra.

Etapas:

Las unidades que tengan que investigarse a través del cuestionario, posiblemente convenga agruparlas y estos grupos a su vez se vuelvan a agrupar y así sucesivamente. Dependiendo del número de agrupamientos de las unidades de interés -o últimas unidades de muestreo-, es el nombre que se le da. Si el marco muestral no presentó agrupamientos, el muestreo se llamará monoetápico -selección directa de las unidades de interés-; si el marco muestral presenta agrupamientos de un sólo orden se llamará bietápico, o lo que es lo mismo se seleccionarán primero los grupos de unidades-de primera etapa-y finalmente se seleccionarán los de interés o de segunda etapa, y así sucesivamente se tendrá el muestreo trietápico, tetraetápico, etc.

Probabilidad:

Si las unidades de muestreo en cada etapa son seleccionadas con la misma probabilidad, el muestreo se llamará equiprobable; en el caso contrario se dice que es de probabilidades variables de selección en la ó las etapas que correspondan.

Estratos:

La precisión al hacerse las estimaciones básicamente dependen de dos factores:

a) Del tamaño de la muestra; y
b) De la variabilidad o heterogeneidad de la población.

Es evidente que mientras más grande sea la muestra, representara más fielmente a la población, tal que se pueden mejorar las estimaciones aumentando el tamaño de la muestra.

En cuanto al segundo factor para aumentar la precisión, puede dividirse el marco muestral, -si es que se dispone de los medios necesarios-en clases homogéneas llamados estratos y seleccionar separadamente en cada estrato una muestra, garantizando con esta forma cualquier representación deseada de todos los estratos de la población. La denominación de un modelo de muestreo se forma indicando estos conceptos-etapa, probabilidad y con ó sin reemplazo.

VI.2.3 MANEJO DE LAS TABLAS DE NÚMEROS ALEATORIOS

La selección de las unidades de muestreo debe hacerse basándose en las leyes del azar; esto es, debe asignarse a cada unidad del marco muestral una probabilidad de inclusión en la muestra. Con este método la muestra se obtiene en selecciones sucesivas de una unidad, cada una con una probabilidad asignada de antemano, según sea el modelo de muestreo que se utilice, hasta completar el número de unidades que deben incluirse en la muestra para cada etapa. Un procedimiento práctico para seleccionar las unidades, es utilizando una tabla de números aleatorios como la que aparece en el apéndice N de la sección de tablas estadísticas.

CONSTRUCCIÓN DE LAS TABLAS DE NÚMEROS ALEATORIOS

Estas tablas consisten de dígitos puestos de manera tal que cada uno de ellos reciba igual probabilidad de ser seleccionados. Estas tablas se construyen de diferentes maneras:

- Usando la computadora de manera similar al proceso de la ruleta.
- Usando ciertas funciones matemáticas; ó
- Usando instrumentos mecánicos basados esencialmente en el principio de la ruleta.

El uso de las tablas de números aleatorios puede ilustrarse con el siguiente ejemplo, relativo a la **selección aleatoria** de la muestra.

Supóngase que se van a seleccionar tres Escuelas: de Medicina, Veterinaria y Zootecnia para ser consideradas como muestra de las 18 escuelas de Medicina Veterinaria y Zootecnia existentes en el país:

Si n = 3 y N = 18. Decimos que el universo está constituido por dos dígitos; si N fuera 4327, diriamos que está constituido por cuatro dígitos; El número de dígitos del universo es el límite máximo para trabajar dichas tablas. Así, en nuestro ejemplo, se hace la relación o numeración de las escuelas que integrel universo: a cada uno de las 18 Escuelas se le asigna un número de dos dígitos: 01, 02, 03, ..., 18.

En seguida se seleccionan pares de números de la tabla de manera consistente.

Por ejemplo: la selección podría empezar en la parte superior de la tabla, - primera columna -, la siguiente columna, etc.. Esto produce los siguientes pares de dígitos: 01, 04, 06.

Estos dígitos identifican la escuela en la población que será considerada como elemento de la muestra.

Si el número par al azar excede el número de unidades posibles de muestreo (N = 18) como el número 31, el número es ignorado y se selecciona el siguiente número, 16 -por ejemplo- y al seguir seleccionando para completar el tamaño de la muestra y ésta vuelve a aparecer, en este caso también se ignora y se continúa buscando un número distinto a 16 y no mayor que 18.

De esta manera se obtienen las tres escuelas que formarán la muestra. Esta no es la única manera para seleccionar pares de dígitos en la tabla de manera horizontal, diagonal, en zig-zag, etc. Lo importante es que el procedimiento sea consistente.

El segundo medio de selección probabilística, el **sistemático,** es en esencia una simple variante del procedimiento anterior. Implica la selección de las unidades de la muestra de manera sistemática empezando con uno de los dígitos.

Esto es, si hay N unidades muestrales en la población, y se desean n para la muestra, cada N/n unidad es seleccionada, empezando con un número aleatorio.

Así usando el ejemplo anterior cada sexta unidad será seleccionada: (N/n = 18 / 3 = 6) empezando con un número aleatorio entre 01 y 06 inclusive. Este número aleatorio se puede obtener también de la tabla de números aleatorios.

MODELOS DE MUESTREO

Los modelos de muestreo tienen por objeto indicar el número de unidades que deben incluirse en la muestra, dependiendo de la forma que estas se seleccionan, de la confianza que se requiera al hacer las inferencias del error de muestreo que se pueda permitir y del fondo disponible para la realización de la encuesta.

VI.2.4 MUESTREO SIMPLE ALEATORIO

Recordando que por muestreo probabilista se entiende un método de muestreo en el que cada miembro de la población tiene una probabilidad conocida de ser incluida en la muestra. Cuando todos los miembros de la población tienen la misma probabilidad de ser seleccionados se denomina muestreo simple aleatorio.

Si una caja contiene seis pedacitos de papel numerados del 1 al 6; si se desea elegir una muestra de la caja de tamaño 3, sin reemplazo, el muestreo simple aleatorio indica que la probabilidad de cada uno de los 6 papelitos es 1/6. Al extraer el segundo, la probabilidad de cada uno es 1/5 y así sucesivamente. En este caso cada número dentro de la caja tiene la misma probabilidad de ser seleccionado.

Esto es, la probabilidad del número 4 es igual a:
P (A o B o C) = 1/6 + 5/6 x 1/5 + 5/5 x 4/5 x 1/4 = 3/6.

En general, se puede decir que si el tamaño de la muestra es **n** y el de la población **N**, en el muestreo simple aleatorio, cada miembro de la población tiene una probabilidad de encontrarse en la muestra de **n/N** .

Por ejemplo: Si de entre 120 estudiantes se seleccionan 10 al azar y todos tienen la misma probabilidad de ser elegidos, cada uno de los 120 estudiantes, tiene una probabilidad de 10/120 de estar en la muestra.

Ahora ¿cuál es la probabilidad de seleccionar una muestra de tamaño n a partir de una población de tamaño N?
Suponiendo de N = 6 y n = 3:

$$\begin{bmatrix} N \\ n \end{bmatrix} = \begin{bmatrix} 6 \\ 3 \end{bmatrix} = \frac{6!}{3!(6-3)!} = \frac{6!}{3!*3!} = 20 \text{ muestras posibles}$$

Cuando se adopta el muestreo aleatorio simple cada muestra tiene igual probabilidad de ser seleccionada y es de 1/20.

En general, se dice que cuando se selecciona una muestra de tamaño n, a partir de una población de tamaño N por muestreo simple aleatorio la probabilidad de que se seleccione una cualquiera de las $\begin{bmatrix} N \\ n \end{bmatrix}$ muestras posibles será: $\frac{1}{\begin{bmatrix} N \\ n \end{bmatrix}}$

Lo anterior se refiere a los casos en que el muestreo se realizo sin reemplazo. Lo mismo sucede cuando se realiza con reemplazo, aunque en la práctica se utiliza generalmente el muestreo sin reemplazo.

VI.2.5 MUESTREO ESTRATIFICADO (11)

De acuerdo con este método, la población se divide en estratos basados en características consideradas relevantes para el sujeto bajo estudio, y se seleccionan las unidades de muestreo de cada uno de los estratos.

Por ejemplo: investigando tiendas al menudeo en la ciudad de Cuernavaca; las tiendas en la ciudad podrán clasificarse primero por tipo de tienda (abarrotes, farmacias, etc.) y luego por tamaño de tienda. Para cada estrato, tipo o tamaño de tienda, se puede estimar el número de tiendas y calcularse cuántas de estas tiendas -unidades de muestreo- deben incluirse en la muestra. Es común en tales casos, seleccionar la mayoría de las unidades de muestreo de los estratos conteniendo las tiendas grandes y sólo una pequeña proporción de unidades de muestreo de los estratos que contienen relativamente pocas tiendas.

Para que sea útil el muestreo estratificado se deben reunir las siguientes tres condiciones:

1) Deben conocerse ciertas características relevantes que influencian fuertemente el fenómeno bajo estudio:

2) Que la población sea susceptible de dividirse de acuerdo con las características relevantes:

3) La división relativa de la población debe conocerse con cierto grado de precisión.

Una muestra estratificada puede obtenerse aún cuando no se pudieran identificar los elementos del estrato, siempre y cuando se conozca después de haberse seleccionado la muestra. El problema sin embargo, es que los errores de muestreo de las estimaciones resultan mayores que si se hubiera estratificado antes.

Si el número de unidades de muestreo seleccionadas de cada estrato es proporcional al tamaño relativo del estrato en la población, el resultado es una muestra estratificada proporcional, lo contrario es una muestra estratificada no proporcional. Esto último es preferible si los diversos estratos no son homogéneos con respecto a la característica bajo estudio.

El error de muestreo de una muestra estratificada puede considerase menor que el de una muestra simple aleatoria del mismo tamaño. Lo anterior se debe a que el diseño de estratificaciones hace uso de información adicional, considerando la división de la población de acuerdo con las características relevantes y sirve para reducir el margen de error de muestreo.

El problema con este método, es que aún cuando se conocen las características relevantes y en base a ellas se estratifica, el tamaño relativo de los estratos en la población no siempre se conoce con gran exactitud.

Debido a esta escasez de información, las ventajas obtenidas con la estratificación se pierden con las variaciones introducidas por la información incorrecta referente al tamaño de los estratos en la población, elemento que desafortunadamente se subestima frecuentemente.

Los diseños de estratificación se pueden combinar con otras como por ejemplo:

- Muestreo por área; y
- Los esquemas de muestreo por conglomerados.

Ejemplo de la situación anterior podría ser el siguiente: digamos que México podría subdividirse en estratos regionales, tales como:

- Norte;
- Sur;
- Este: y
- Oeste.

Con áreas seleccionadas dentro de cada uno de estos estratos o regiones y con miembros de la muestra seleccionados al interior de cada una de estas áreas, en grupos o "racimos". Similarmente, la selección de los miembros de una muestra estratificada podría realizarse, ya sea usando procedimientos aleatorios o arbitrarios.

VI.2.6 MUESTREO POLIETAPICO

Este método requiere la selección de las unidades de muestreo en diferentes etapas, existiendo unidades de primera, segunda, etc.., etapa en un diseño muestral.

Por ejemplo: si el interés es conocer la opinión de los médicos veterinarios zootecnistas sobre los programas de estudio de las diferentes escuelas y facultades de Medicina Veterinaria y Zootecnia y si para ello se decide realizar la investigación en la ciudad de México, entonces la clasificación de la ciudad en distritos permite obtener la unidad de primera etapa; la clasificación en colonias es la unidad de la segunda etapa; la selección de las manzanas a muestrear es la unidad de tercera etapa; y la selección aleatoria de los médicos residentes en las manzanas previamente seleccionadas, constituyen la unidad de cuarta etapa.

VI.2.7 MUESTREO POR ÁREAS

Cuando la población se distribuye sobre un área muy grande, la selección de los elementos de la muestra de toda el área puede resultar un procedimiento ineficiente y costoso. Estos es particularmente cierto, si a las personas que entrevistan se les paga por hora y la mayor parte del tiempo se va en viajar. El muestreo por áreas fue diseñado para resolver este problema. Se basa en una subdivisión apriori de la población en áreas; la selección de algunas de estas áreas con la ayuda de los métodos de muestreo aleatorio y la restricción a la selección de las unidades que integrarán la muestra, solamente en esas áreas.

La restricción geográfica sirve para concentrar los esfuerzos de trabajo en ciertas regiones, provocando reducciones sustanciales en el costo del trabajo de campo en comparación a una muestra del mismo tamaño proveniente de un diseño distinto al de áreas.

Esta técnica de muestreo puede usarse para trabajar con muestras irrestrictas y estratificadas. De hecho en investigaciones de gran escala la técnica de estratificar áreas es generalmente la regla, porque asegura la representatividad de todos los segmentos relevantes de la población a costos bajos.

En cada investigación el diseño de áreas se realiza en varias etapas; cada etapa sirve para restringir el área geográfica de la cual se seleccionarán las unidades de la muestra.

VI.3 APLICACIONES

VI.3.1 APLICACIÓN DEL MUESTREO SIMPLE ALEATORIO

Aún cuando este método es el más simple de los clasificados como probabilisticos, su sencillez no deja de ser útil para ilustrar las ventajas que se derivan de la aplicación de esta metodología al análisis de fenómenos económicos, al igual que los demás métodos de muestreo estadístico, se caracteriza por proporcionar estimación de los caracteres de la población.

Se asigna igual probabilidad de selección a cada unidad perteneciente a la población. Si N es el número de unidades, la probabilidad de selección de cualesquiera de ellas es: 1/N .

El número de muestras distintas de tamaño n, sacadas de las N unidades de la población está dado por:

$$\begin{bmatrix} N \\ n \end{bmatrix} = \frac{N!}{n!(N-n)!}$$

Los estimadores obtenidos serán insesgados cuando su esperanza matemática sea igual al parámetro poblacional:

$$E(\bar{y}) = \bar{Y}$$

Demostración : $\bar{y} = \frac{1}{n}\sum y_i$

$E(\bar{y}) = \frac{\sum (y_1 + y_2 + ... + y_n)}{n} = \frac{E(y_1) + E(y_2) + ... + E(y_n)}{n} = E(\bar{y}) = \frac{n\bar{Y}}{n}$ por lo tanto $E(\bar{y}) = \bar{Y}$

El estimador del total de la población definido por $\hat{Y} = N\bar{y}$ es insesgado porque:

$E(\hat{Y}) = E(N\bar{y}) = NE(\bar{y}) = NY = Y$

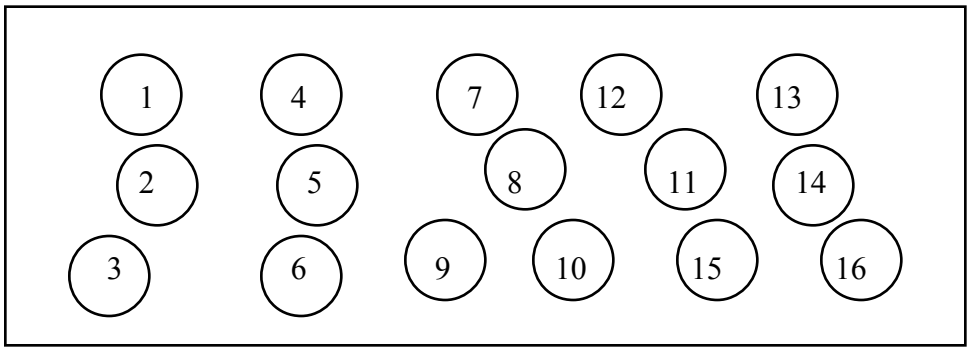
Aplicaciones: Para ello se supone que se conoce el tamaño de la muestra requerida, el cual se estudiará posteriormente en detalle.

Objetivo: Se desea estimar el total de familias en la localidad "γ" con una muestra simple aleatoria cuyo tamaño está dado por cuatro manzanas.

Notación:

F = n/N= Fracción de muestreo.
N = Número de manzanas en la localidad.
$\hat{Y}$ = Población total estimada
$\bar{y}$ = Promedio de familias por manzana en la muestra
m = Promedio de personas por familia.

El mapa de la localidad revela la siguiente distribución de la manzanas.



Las manzanas se numeran siguiendo un orden determinado: ascendente o descendente en este caso, resultaron ser 16 en total.

Conociendo N = 16 y n = 4 se seleccionará la muestra con la tabla de "números aleatorios". Suponiendo que las manzanas seleccionadas son:
los número 16, 3, 9 y 11.

En seguida, se hace un listado de las manzanas seleccionadas registrando el número de familias que existen en cada una de ellas. Los resultados son:

La manzana 16 tiene 4 familias
" 3 " 9 "
" 9 " 9 "
" 11 " 10 "

Recordando que el total de familias se estima por:

$$\hat{Y} = N\bar{y}; \text{si N=16 y } \bar{y} = \frac{1}{n}\sum y_i = \frac{1}{4}(4+9+9+10) = \frac{32}{4} = 8$$

Tendremos que $\hat{Y}$ =16(8); $\hat{Y}$ =128 familias en la localidad.

Se puede verificar que el cálculo del total de las familias en la localidad tenga un 95% de probabilidad de haber caído en el intervalo de confianza con la siguiente fórmula:

$$N\bar{y} - \frac{tNs}{\sqrt{n}} * \sqrt{1-F} \leq \hat{Y} \leq N\bar{y} + \frac{tNs}{\sqrt{n}} * \sqrt{1-F}$$

Donde t es el valor de la normal desviada correspondiente a la confianza de probabilidad deseada cuando n es menor que 30 y **s** es la varianza muestral.

Como se recordará:

Con α = 5% y un número infinito de grados de libertad se halla en tablas t_α = 1.96; sabemos que:

$$S^2 = \frac{\sum (y_i - \bar{y})^2}{n} = \frac{\sum y_i^2}{n} - (\bar{y})^2 = \frac{278}{4} - 8 = 5.5$$

como $S = \sqrt{s^2} = \sqrt{5.5} = 2.3$ y $t_\alpha = \pm 1.96$, tendremos que

Limites de confianza $= \frac{16(32)}{4} \pm \frac{(1.96)(16)(2.3)}{\sqrt{4}} * \sqrt{1 - \frac{4}{16}} =$ 125 a 131

El total estimado de familias (128) se halla entre 125 y 131 con una seguridad o confianza del 95%.

El número total de habitantes se puede saber multiplicando el total estimado $(\hat{Y})$ por el promedio de personas por familia (m)

Si m= 5.4; $\hat{Y}$ = 128

m $\hat{Y}$ = 5.4(128) = 691 habitantes en la localidad "gama"

VI.3.2 MUESTREO POR ÁREAS, COMBINADO CON EL SIMPLE ALEATORIO Y EL ESTRATIFICADO.

Por ejemplo: Considérese el siguiente diseño muestral hecho para captar las características del gasto familiar en consumo en 2002 y 2003.

Se diseñó una muestra probabílistica multietápica del país que fue dividido en áreas. En un muestreo multietápico, cada persona (y familia) en el universo bajo estudio, tiene una probabilidad de ser incluida en la muestra, la cual esta asociada con las probabilidades de selección de la unidad de muestreo en la cual se localiza la persona, en cada una de las etapas.

Lo primero que se hizo fue seleccionar con números aleatorios a las unidades de muestreo de la primera etapa que eran de dos tipos; áreas urbanas y áreas rurales. En la segunda etapa, con números aleatorios se seleccionaron áreas más pequeñas o manzanas dentro de las unidades de la primera etapa, seleccionadas previamente. La tercera etapa consistió en la división de las manzanas en áreas más pequeñas llamadas segmentos; con números aleatorios se seleccionaron los segmentos donde el entrevistador debía tener la información de cada una de las familias que lo integraban. Finalmente dentro de cada familia todos los adultos más uno de cada tres adolescentes seleccionados aleatoriamente, contestaron el cuestionario.

En este caso particular el modelo muestral comprendió tres etapas. La **estratificación** en el muestreo por áreas se hace generalmente en la primera etapa (es decir, laas áreas se integran en estratos), ya que a partir de ella la población debe dividirse en forma tal, que se asegure la representatividad de los estratos. En el ejemplo que se ilustra, todas las unidades de muestreo de la primera etapa, áreas urbanas y rurales, fueron agrupadas en estratos de acuerdo con ciertos criterios para minimizar la variabilidad dentro de los estratos. Los criterios usados fueron flexibles ya que el propósito principal era obtener hasta donde fuera posible homogeneidad en las unidades de muestreo en la primera etapa de cada una de los estratos, así como la integración de estos últimos con un número aproximadamente igual de familias. Se seleccionaron automáticamente 14 áreas urbanas, porque contenían un número de familias mayor que el establecido por estrato.

Del resto de las áreas urbanas, se seleccionó una de cada estrato, con probabilidad proporcional a su tamaño. Similarmente en los estratos rurales, un pueblo o área fue seleccionado con probabilidad proporcional a su tamaño.

En total, se seleccionaron 103 unidades de la primera etapa, conteniendo 191 poblaciones. De las 103 unidades de la primera etapa; 49 eran urbanas y 54 rurales.

Una vez que se han diseñado las áreas y agrupado en estratos, en cada estrato se seleccionan ciertas áreas usando algún criterio, generalmente se aplica el llamado "probabilidad proporcional al tamaño", con el cual cada área tiene una probabilidad (proporcional) de ser seleccionada de acuerdo a su tamaño o significación dentro del estrato. Por ejemplo: Supongamos que deseamos seleccionar con probabilidad proporcional a su tamaño una de las siguientes cinco ciudades que integran un estrato:

Ciudad	Población	Población Acumulada (en miles)	Dígitos Aleatorios	Probabilidad
A	100,000	100	01 - 10	10 ÷ 35
B	40,000	140	11 - 14	4 ÷ 35
C	60,000	200	15 - 20	6 ÷ 35
D	70,000	270	21 - 27	7 ÷ 35
E	80,000	350	28 - 35	8 ÷ 35
Total estrato 350,000				35 ÷ 35

Un procedimiento es la selección de un número aleatorio formado por dos dígitos de cualquier tabla de números aleatorios, y luego seleccionar la ciudad cuyo rango de dígitos incluye los números aleatorio. Si el número aleatorio es mayor que 35, nuevamente se seleccionan otros números hasta obtener uno que sea igual a 35 o menos.

Por ejemplo: Si el número aleatorio es el número 22 se selecciona la ciudad D como la muestra del estrato, por que de acuerdo con la penúltima columna del cuadro anterior, el 22 es uno de los siete dígito que representan la ciudad D: Si fuera 06, la muestra contendría la ciudad A.

En esencia, se sigue el mismo procedimiento para seleccionar las manzanas de la segunda y las familias de la tercera etapa del muestreo por áreas, ya que por lo general no se requieren estratificaciones adicionales. Así, si la ciudad A es seleccionada en la muestra podría dividirse en manzanas y seleccionarse con probabilidad proporcional unas cuantas de estas con la ayuda de la tabla de los números aleatorios.

Una vez seleccionadas las manzanas, las familias se listarán en cada manzana y el número requerido de ellas se obtendría usando una vez más la tabla de números aleatorios.

Obsérvese que en poblaciones grandes y dispersas este procedimiento resulta ventajoso no sólo en la fase de la entrevista, sino también en la fase de preparación del marco muestral, ya que las definiciones y listados de las familias solo se hacen para las unidades de la primera etapa que caen en la muestra y los listados de familias se requieren solamente de aquéllas manzanas consideradas en la muestra.

VI.3.3 MUESTREO POR RACIMOS O CONGLOMERADOS

Este método, que es en esencia una extensión del muestreo por áreas, consiste en la aplicación de las últimas unidades del muestreo en localidades adyacentes en lugar de permitir su dispersión en todas las áreas que comprenden la muestra.

Por ejemplo: Una muestra de 300 familias podría obtenerse seleccionando 60 grupos de 5 manzanas en lugar de seleccionar individualmente a 300 familias.

Esta concentración de las unidades de muestreo reduce considerablemente el tiempo y dinero estimados para el llenado del cuestionario, por lo que se aconseja cuando el entrevistador tenga que cubrir una gran área como en el caso del muestreo en áreas rurales. Sin embargo con este se pierde cierta eficiencia en la muestra.

Esta pérdida se deriva de la tendencia que tienen por vivir como vecinos las personas con iguales características, actitudes o aún hábitos de consumo. Así, una persona de altos ingresos es más probable que este al lado de otra de igual nivel; y no de una de bajos ingresos, lo que ocasiona que las unidades de muestreo en lugar de ser independientes estén correlacionadas. Mientras más alta sea la correlación positiva, menor será la eficiencia del modelo por racimos; en donde la ineficiencia resulta de la reducción en la precisión de los estimadores muestrales.

VI.3.4 MUESTREO REPLICADO

Hasta el momento, se han ilustrado métodos que requieren la selección de una sola muestra de la población. Un procedimiento alternativo es dividir la muestra en un número igual de sub-muestras y seleccionar cada una de las sub-muestras de la población como si cada una de ellas fuera la única muestra a seleccionar.

La muestra total, consiste en un número de sub-muestras replicadas, cada una de ellas tratando de proporcionar en su área de influencia una imagen completa del universo. Si se desean entrevistar 400 personas en un área de 10 000 personas, cada: 25, ... (10 000 entre 400) sería entrevistado comenzando con un número aleatorio entre 01 y 25.

Si se decide seleccionar 5 en lugar de una muestra cuyo tamaño total sea de 400 personas, cada una de las cinco sub-muestras deberá contener 80 unidades de muestreo. Para ello se puede dividir a la población en 125, (10 000 entre 80 = 125). Son así iguales cada una conteniendo 80 unidades de muestreo; luego se seleccionan 5 números aleatorios entre 01 y 125 que se consideran, cada uno como punto de arranque o primer unidad de muestreo que faltan en cada sub-muestra, se seleccionan progresivamente cada 125 familias. El resultado, son 5 sub-muestras replicadas o interpenetrantes con 80 unidades cada una, que agregadas suman una muestra con 400 unidades de muestreo.

VI.4 DEFINICIONES BÁSICAS [14]

Error de muestreo:

Sea μ el valor de un parámetro de la población que se estudia mediante el muestreo, y $\bar{x}$ una función definida mediante la muestra, que estima el valor de μ.

Error de muestreo $= |\mu - \bar{x}|$ que debe ser menor o igual al máximo error de variación permitido $\varepsilon|\mu|$; es decir $\varepsilon|\mu| \geq |\mu - \bar{x}|$.; como no conocemos μ

decimos: $\varepsilon|\bar{x}| \geq |\mu - \bar{x}|$

VI.4.1 LÍMITES DE CONFIANZA:

Lo antes dicho se basa en el hecho de que cuando no se conocen los parámetros (μ y σ) de la población se pueden estimar recurriendo a muestras que permiten calcular intervalos dentro de los cuales puede estar contenido el valor de los parámetros. Estos intervalos se llaman intervalos de confianza y sus extremos se llaman límites de confianza.

El grado de confianza de que el parámetro está contenido en el intervalo se determina por el número de errores estándar a los cuales les corresponde un área bajo la curva que se denomina "coeficiente de confianza" (ε épsilon). Al riesgo de que el valor estimado de μ no se encuentre dentro del intervalo de confianza construido alrededor de la media de la muestra, se le llama nivel de significación (α) y es el área o probabilidad complementaria del coeficiente de confianza.

Así $\varepsilon = 1 - \alpha$ ó $\varepsilon + \alpha = 1$ = área bajo la curva..

De esta manera el intervalo de confianza se determina con:

Límites de confianza $= \bar{x} \pm Z\alpha\sigma_{\bar{x}}$

donde: $\bar{x}$ = Media muestral;

$Z\alpha$ = Valor especifico de Z en la tabla, asociado con determinado valor de α y ε;

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

Error estándar para una población infinita.

n = Tamaño de la muestra;

σ = Desviación estándar de la población.

VI.4.2 DISTRIBUCIÓN DE MUESTREO[(10)]

Distribución de Muestras (de medias y proporciones)

Por analogía, la distribución de muestreo que se deriva del universo, con determinado tamaño de muestra n y $\sigma_{\bar{x}}$, tendrá

$\mu\bar{X} = E(\bar{X})$ y una varianza $(\bar{X}) = \frac{\sigma^2}{n}$ para una población infinita y Var $(\bar{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$ para una población finita donde

σ^2 = Varianza del universo. La varianza de $\bar{X}$ se representa con $\sigma^2_{\bar{X}}$, cuya raíz cuadrada $\sigma_{\bar{x}}$ se denomina ERROR ESTÁNDAR para distinguirla de σ = Desviación estándar del universo o raíz cuadrada de σ^2. Luego en una distribución de muestreo

$\mu\bar{X} = E(\bar{X})$ y $\sigma\bar{x} = \frac{\sigma}{\sqrt{n}}$

Ejemplo: Supóngase la población N =3 con los términos

(xi) : 1, 2 y 3.

Su; $\mu = \frac{1+2+3}{3} = \frac{\sum x_i}{N} = 2$

Su; $\sigma = \sqrt{\frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3}} = \sqrt{\frac{\sum (xi-\mu)^2}{N}} = \sqrt{\frac{2}{3}} = 0.81$

CUYOS VALORES SON FIJOS

Si tomamos muestras de tamaño 2, esto es n = 2 de N = 3 sin reemplazo, habrá

$\begin{bmatrix} N \\ n \end{bmatrix} = \frac{N!}{(N-n)!n!} = \frac{3*2*1}{(3-2)!2!} = \frac{3*2*1}{1!(2*1)} = \frac{6}{2} = 3$

Interpretación: Hay tres muestras de tamaño 2, cuya composición de cada una es: 1, 2; 1, 3; y 3,2.

Estandarizando la nueva variable aleatoria $\bar{X}$, tendremos:

Muestra	$\bar{X}_i$	$\bar{X}_i - \mu$	$Z_i = \frac{\bar{X}_i - \mu}{\sigma_{\bar{X}}}$	Ordenada Yi	Área bajo la curva
1, 2	1.5	-0.5	-1.25	0.18265	0.394
1, 3	2.0	0.0	0.0	0.39894	0
2,3	2.5	0.5	+1.25	0.18265	0.394
		0	0		

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{0.81}{\sqrt{2}}\sqrt{\frac{3-2}{3-1}} = \frac{0.81}{1.41}\sqrt{\frac{1}{2}}$$

$\sigma_{\bar{X}}$=0.56(0.7) = 0.4

De lo anterior puede decir que:

No. Desviaciones	-0.125 (0.4)	ó	+0.125(0.4)
o errores estándar	-0.05	ó	+0.05

que nos sirve para graficar los valores estandarizados de las tres $\bar{X}$ 15: 1.5, 2.0 y 2.5, obteniéndose:

Observese que aún cuando N = 3, es demasiado pequeña, esta distribución tiende a la normal por el teorema del límite central. Donde:

Xi : valores originales.

Zi : valores originales ahora expresados en unidades de desviación estándar

μ : Media del universo.

$E(\bar{X})$: Esperanza matemática de las $\bar{X}$

Luego usando la distribución de muestreo vemos que hay tres medias muéstrales (1.5, 2.0 y 2.5) llamadas "ESTADÍSTICAS", que cada una de ellas puede estimar el valor verdadero del parámetro μx que generalmente se desconoce su valor en la vida real, el cual podemos estimar que está en el rango $|\mu - \bar{X}|$ = error de muestreo, con cierto grado de confianza.

El ***error de muestreo*** o precisión en la estimación se mide y se calcula con las fórmulas del ***error estándar*** (en términos de probabilidad) de la media o de la proporción según sea el caso.

Así supongamos que deseamos estimar el valor de μx, para ello supongamos que seleccionamos aleatoriamente la muestra A, que está compuesta por las unidades de muestreo 1 y 2 y por consiguiente tiene una media aritmética ($\bar{x}$) = 1.5 y una desviación estándar de (s) = 0.5.

Muestra	Composición	Media de la Muestra $\bar{X}_y$	Desviación estándar de la muestra (**s**)
A	1, 2	1.5	$\sqrt{\frac{(1-1.5)^2+(2-1.5)^2}{2}}=\sqrt{\frac{0.5}{2}}=0.5$
B	1, 3	2.0	$\sqrt{\frac{(1-2)^2+(3-2)^2}{2}}=\sqrt{\frac{2}{2}}=1.0$
C	2,3	2.5	$\sqrt{\frac{(2-2.5)^2+(3-2.5)^2}{2}}=\sqrt{\frac{0.5}{2}}=0.5$

La media ($\bar{X}$) y desviación estándar (**S**) de las muestras difieren según la muestra elegida, pero;

$$E(\bar{X}_i)=\frac{6}{3}=2=\mu x=\mu\bar{x}=\frac{\sum \bar{X}_i}{N}$$

Se pueden generar distintas distribuciones a partir del cálculo de la muestra con o sin reemplazo.

Cuando la selección es con reemplazo se usa la fórmula $N^n= 3^2 = 9$

Interpretación: hay 9 muestras de tamaño 2, cuya composición es:

Muestra	Composición	Media de la muestra (xi)	P ($\bar{x}$i)
A	1,1	1.0	1 ÷ 9
B	1,2	1.5	1 ÷ 9
C	1,3	2.0	1 ÷ 9
D	2,1	1.5	1 ÷ 9
E	2,2	2.0	1 ÷ 9
F	2,3	2.5	1 ÷ 9
G	3,1	2.0	1 ÷ 9
H	3,2	2.5	1 ÷ 9
I	3,3	3.0	1 ÷ 9
		18.0	9 ÷ 9

$$\mu\bar{x}=\frac{\sum \bar{x}_i}{N}=\frac{18}{9}=2=E(\bar{x})$$

$$\mu\bar{x} = E(\bar{x})P(\bar{x}) = \frac{1/9 + 1.5/9 + ... + 2.5/9 + 3/9}{9} = \frac{18}{9} = 2$$

Las distribuciones de muestras más importantes son:

a) De muestreo: Medias, que se obtienen con: Teorema de Límite Central y Ley de los Grandes Números.
b) De proporciones. que también se obtienen con el Teorema de Límite Central y la Ley de los Grandes Números.

Por otra parte cuando el tamaño de la muestra (n) y la población (N) son grandes, el número de muestras es muy grande; por lo que se recomienda simplificar el procedimiento de obtención de muestras probabilisticas.

Con este objeto, se usa el teorema del Límite Central para demostrar que se puede utilizar la media de la muestra para representar la media de la población.

El teorema del Límite Central, establece que si una población es normal, con media y desviación estándar, μx y σx, entonces si tomamos muestras de tamaño n y a estas les calculamos sus medias aritméticas, la nueva distribución constituida por las medias de las muestras, es una distribución muestral, normal con:

$\mu\bar{x} = E(\bar{x})$ y $\sigma_{\bar{x}} = \frac{\sigma x}{\sqrt{n}}$ para una población infinita

La ley de los Grandes Números establece que si una población tiene μx y σx independientemente de que sea o no normal; si el tamaño de la muestra n crece, entonces la distribución que resulta de las medias muestrales se aproximan a la normal con $\mu\bar{x}$ y $\sigma_{\bar{x}}$. Luego:

Medias de las muestras ($\bar{X}$)	P($\bar{X}$)
1.5	1 ÷ 3
2.0	1 ÷ 3
2.5	1 ÷ 3

$$E(\bar{x}) = \frac{1.5}{3} + \frac{2.0}{3} + \frac{2.5}{3} = \frac{6}{3} = 2 = \mu\bar{x} = \mu$$

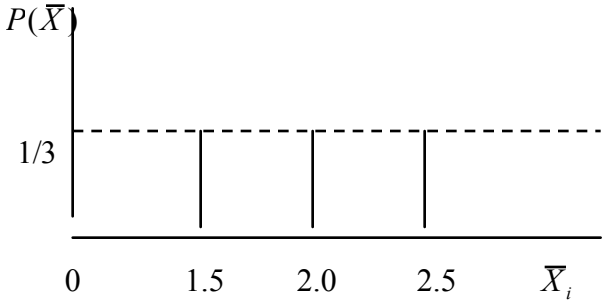
$$\sigma\bar{x} = \sqrt{\frac{(1.5-2)^2 + (2-2)^2 + (2.5-2)^2}{3}} = \sqrt{\frac{0.50}{3}} = \sqrt{0.17} = 0.4$$

también $\sigma\bar{x}$ se obtiene con $\sigma\bar{x} = \frac{\sigma x}{\sqrt{n}}$ cuando n es muy grande

$$\sigma\bar{x} = \frac{\sigma x}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{0.81}{\sqrt{2}}\sqrt{\frac{3-2}{3-1}} = \frac{0.81}{1.41}\sqrt{\frac{1}{2}} = \frac{0.81}{1.41}(0.7) = (0.57)(0.7) = 0.4$$

$$\sigma\bar{x} = 0.4$$

Gráfica de la nueva distribución de muestreo con $\mu\bar{x}$ = 2 y $\sigma\bar{x}$ =0.4



valores originales, sin estandarizaciòn

Si se desea calcular el intervalo de confianza dentro del cual se halle contenido el valor de μ, para calcular el intervalo de confianza el investigador determina el grado de confianza, (ξ) deseado en la estimación. El grado de confianza lo determina el número de errores estándar deseado, que en términos de probabilida, a su vez determina el error de muestreo.

Sabemos que n = 2

$\bar{x}$ = 1.5

s = 0.5

con ξ = 95% probabilidad (área bajo la curva) de que μx se halle en el intervalo $\bar{X} \pm Z\alpha\sigma\bar{x}$; donde α = 5% = probabilidad de que no sea así, se denomina nivel de significación.

Derivado de lo anterior diremos que a un ξ = 95% le corresponden 1.96 errores estándar = 1.96 $\sigma\bar{x} = Z\alpha\sigma\bar{x}$.

Así $\bar{X} \pm Z\alpha\sigma\bar{x}$ y como; $\sigma\bar{x} = \frac{s}{\sqrt{n}} = \frac{0.5}{\sqrt{2}} = \frac{0.5}{1.41} = 0.35$

por lo tanto 1.5 ± 1.96 (0.35)

1.5 ± 0.70

luego el límite inferior del intervalo es 0.80 = 1.50 - 70 y el límite superior del intervalo es 2.20 = 1.50 + 0.70.

Interpretación: Hay una probabilidad del 95% que el valor μx se halle en el intervalo de 0.80 a 2.20.

Generalizando, si la muestra seleccionado hubiera sido la B o la C, tendríamos:

B	C
$\bar{X}$ = 2	$\bar{X}$ = 2.5
s = 1.0	s = 0.5

Para B $\sigma_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{1}{\sqrt{2}} = \frac{1}{1.41} 0.70$; para C, $\sigma_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{0.5}{\sqrt{2}} = \frac{0.5}{1.41} 0.35$

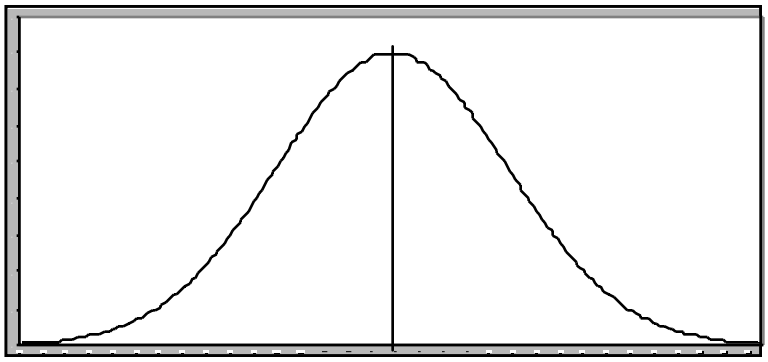
n = 2 n = 2

$\overline{X} \pm Z\alpha\sigma\overline{X}$	$\overline{X} \pm Z\alpha\sigma\overline{X}$
$2 \pm 1.96(0.70)$	$2.5 \pm 1.96(0.35)$
2 ± 1.37	2.5 ± 0.70

Intervalo: De 0.63 a 3.37 Intervalo: De 1.80 a 3.20

Conclusión: En los tres casos el valor de μx = 2 se halla contenido con una seguridad, probabilidad o confianza del 95% y con un riesgo de α = 5% de que no sea así, en los intervalos antes calculados.

Gráficamente.:



$\bar{x} - Z_{\alpha}\sigma_x$ μ_x $\bar{x} + Z_{\alpha}\sigma_x$

A: 0.80=1.5-0.70	1.5	1.5+0.70=2.20
B: 0.63=2.0-1.37	2.0	2.0+1.37=3.35
C: 1.80=2.5-0.70	2.5	2.5+0.70=3.25

Si conectamos estos resultados con la definición básica de que el error de muestreo $(\bar{x} - \mu)$ se determina con el error estándar de la media , en términos de probabilidad, $\sigma_{\bar{x}}$, y con la situación ideal de que siempre esperamos que el error de muestreo sea igual o menor al error permitido (e), observamos que:

1.- con la muestra 1: $e|\mu_{\bar{x}}| \geq |\bar{x} - \mu|$ ya que $e=0.70 \geq |1.5 - 2.0|$

2.- con la muestra 2: tenemos $e= 1.37 \geq |2 - 2|$ y

3.- con la muestra 3: vemos que $e= 0.70 \geq |2.5 - 2.0|$,

en los tres casos es satisfactorio ver que el error de muestreo es inferior al error permitido.

Nuevo ejemplo; ahora supongamos que ε = 50% y Zα = 0.68.

luego α = 50%

Muestra	$\bar{X}$	s	$\sigma\bar{X}$	Zα	Zα$\sigma\bar{X}$	Limites		Contiene a μx
						Inferior	Superior	
A	1.5	0.5	0.35	0.68	0.238	1.262	1.738	No
B	2.0	1.0	0.70	0.68	0.476	1.524	2.476	Si
C	2.5	0.5	0.35	0.68	0.238	2.262	2.738	No

La muestra A y C no contienen a μx porque el grado de confianza ξ es muy bajo; es decir hay menos área sobre la curva que ocasiona una Zα muy baja que al ser combinada en Zα$\sigma\bar{X}$ originan un intervalo más pequeño en torno a $\bar{X}$, en la fórmula $\bar{X} \pm Z\alpha\sigma\bar{X}$, con lo que aumentan la probabilidad α, de que $\bar{X}$ no represente a μx.Estos resultados se corroboran con el suguiente análisis:

Con la muestra 1: e=$0.238 \leq |1.5 - 2.0|$, por eso el intervalo de confianza no contiene a la media poblacional;

Con la muestra 2: e= $0.476 \geq |2.0 - 2.0|$, por eso contiene a la media poblacional y con la muestra 3: e= $0.238 \leq |2.5 - 2.0|$, por eso no contiene a la media poblacional.

VI.4.3 ERROR PERMITIDO Y ERROR DE MUESTREO.

De lo anterior podemos decir que e = error permitido = Zα$\sigma\bar{X}$.

Se dice que es el error permitido; α y n condicionan los valores de Zα y de $\sigma\bar{X}$.

Así, como: e = Zα $\sigma\bar{x} = \dfrac{\bar{x} - \mu\bar{x}}{\sigma\bar{x}} * \dfrac{\sigma x}{\sqrt{n}} = \dfrac{\bar{x} - \mu\bar{x}}{\sigma\bar{x}}\sigma\bar{x} = \bar{x} - \mu x$

$e = |\bar{x} - \mu x| =$ error de muestreo; también: $e|\mu x| =$ error permitido.

Idealmente siempre queremos que $e|\mu x| \geq |\bar{x} - \mu x|$. Observe que ambos requieren del error estándar $(\sigma\bar{x})$ para su cálculo.

Por otra parte mostrando los valores de mayor uso de Zα, ξ y α, de la ecuación (1) tenemos: $(\sigma\bar{x})$ para su cálculo.

$\bar{X}$ - Zα $\sigma\bar{x}$; límite inferior del intervalo

$\bar{X} + Z\alpha\ \sigma\bar{x}$; límite superior del intervalo

Zα	1.00	1.96	2.00	3.00
ξ	0.68	0.95	0.955	0.997
α	0.32	0.05	0.045	0.003

Ejemplo: Se desea conocer el ingreso medio de los trabajadores de la Cía. PEPSI COLA con el fin de estudiar las condiciones de trabajo y en su caso pedir mejorías en la revisión del Contrato Colectivo de Trabajo. Para ello seleccionamos una muestra aleatoria de 49 trabajadores cuyo ingreso medio es de $ 5,500 /mes.

Estudios previos realizados por la Facultad de Contaduría y Administración revelan que la σ del universo es de $ 700/ mes. Con α = 5%, estimar el ingreso medio de los trabajadores.

n = 49

σ = 700/mes

$\bar{X}$ = 5500/mes

α = 5 %

Zα = ± 1.96

$\bar{X} \pm Z\alpha\sigma\bar{X}$

sustituyendo

$5500 \pm 1.96(100)$

5500 ± 196

$\sigma\bar{x} = \frac{\sigma x}{\sqrt{n}} = \frac{700}{\sqrt{49}}$; ; por lo tanto $\sigma\bar{x}$ = 100

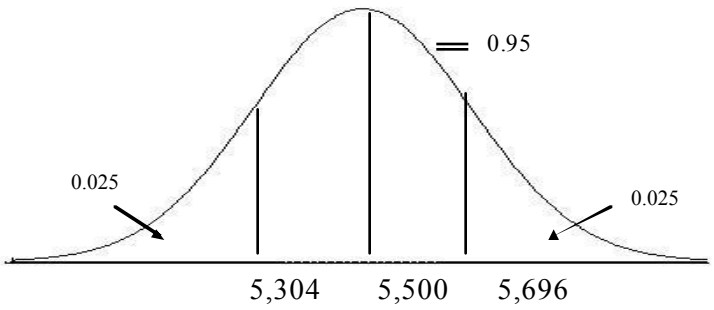
Limites de confianza = 5,500 ± 196

Intervalo de confianza : 5,304 a 5,696

donde el límite inferior = 5,304

el límite superior = 5,696

INTERPRETACIÓN: El ingreso medio μx de los trabajadores de la PEPSI se halla entre los $5,304 y $ 5,696 con una probabilidad o seguridad del 95%.



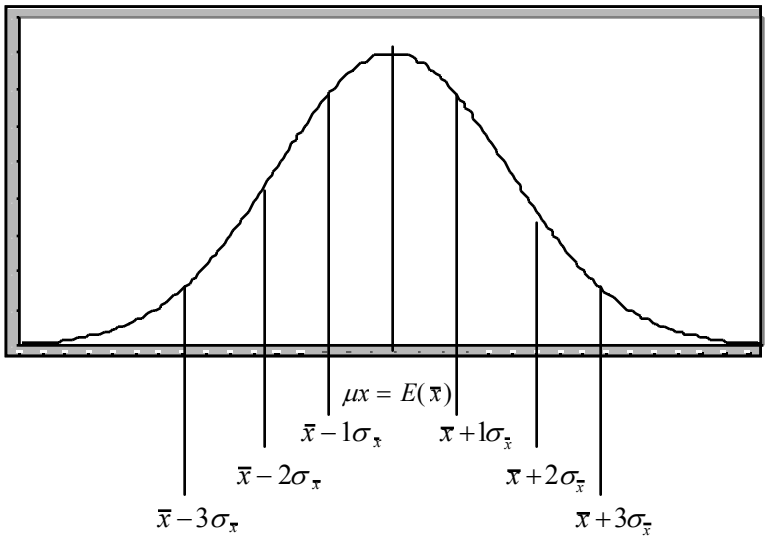
$$\bar{x} - Z_\alpha \sigma_{\bar{x}} \qquad \bar{x} \qquad \bar{x} + Z_\alpha \sigma_{\bar{x}}$$

En este caso se estima μx con la variable aleatoria asociada x mediante $\bar{X}$ proveniente de n = 49 con α = 5% y un ξ = 95% que les corresponde una Zα = 1.96 = Número de desviaciones estándar y $\sigma\bar{x}$ = 100,
tal que:
$P(\bar{x} - Z\alpha\sigma\bar{x} \le \mu x \le \bar{x} + Z\alpha\sigma\bar{x}) = 1 - \alpha = 95\%$; en otras palabras digamos que $e|\mu x| = 6\%$ con $|\mu x|$ tenemos:

$$P|\mu x - \bar{x}| \ge e|\mu x| = 0.06|\mu\chi| = 1 - \xi = 1 - 0.95 = 5\% = \alpha$$

Ello significa que el error en la estimación del valor de μx en valores absolutos es:
| error en la estimación de μx | = Zx $\sigma\bar{x}$, por lo que
error máximo permitido = error en la estimación de $e|\mu\chi|$
Derivado de lo anterior se puede escribir e = Zα $\sigma\bar{x}$

Gráficamente dichas relaciones se ven así:



donde $\sigma\bar{x} = \dfrac{\sigma}{\sqrt{n}}$ para una población infinita

$\sigma\bar{x} = \dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}}$ para una población finita

VI.5 DETERMINACIÓN DEL TAMAÑO DE LA MUESTRA (n)
Sabemos que:

$e = Z\alpha\sigma\bar{x}$ como $\sigma\bar{x} = \frac{\sigma x}{\sqrt{n}}$ para una población infinita

$$e = Z\alpha\frac{\sigma x}{\sqrt{n}}$$

$$\sqrt{n} = \frac{\sigma x}{e}Z\alpha$$

$$n = \left[\frac{\sigma x}{e}Z\alpha\right]^2 = \frac{Z\alpha^2\sigma^2 x}{e^2}$$ para una población infinita.

$$n = \frac{Z^2\alpha\sigma_x^2 N}{e^2 N - e^2 + Z\alpha^2\sigma_x^2}$$ para una población finita.

Ejemplo: En una población infinita ¿qué, tamaño de muestra será necesario para producir un intervalo de confianza del 90% en que está la media de la población verdadera, con un error permitido de 1.0 en cualquier sentido si la desviación estándar de la población es 10.00?

Solución: Sabemos que σ_x = 10.0

e = 1.0

Zα = ± 1.65

para α = 10 %

$$n = \left[\frac{\sigma x}{e}Z\alpha\right]^2 = \left[1.65\frac{10.0}{1.0}\right]^2 = (1.65)^2 = 272.2$$

Con este tamaño de muestra de 272.25 aseguramos que el error de muestreo = $|\bar{x} - \mu_x| \leq error\ permitido = Z_\alpha\ \sigma_{\bar{x}}$ donde $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$

Así, con Z_α =1.65 y $\sigma_{\bar{x}} = \frac{10}{\sqrt{272.25}}$

$e = 1.65(10/\sqrt{272.25}) = \frac{16.5}{16.5} = 1 = error\ \ permitido = errordemuestreo$, lo cual es muy aceptable.

También, si sabemos que el error estándar = 10/16.5= 0.606

Luego aplicandolo al error permitido (e) en términos probabilisticos tendremos que e = 1.65(0.606)= 1, se comprueba que el error de muestreo se mide con el error estándar en términos probabilisticos.

CONSIDERACIONES:

1. Hay ocasiones en que conocemos N, en ese caso $n = \frac{N}{Ne^2 + 1}$

Ejemplo: Con N= 603 y e= 5%

Tenemos $n = \frac{603}{1 + 603(0.05)^2} = \frac{603}{2.5075} = 240.47$

2. Cuando no conocemos nada $n = \frac{1}{e^2}$ digamos si e = 5 %

$n = \frac{1}{(0.05)^2} = \frac{1}{0.0025} = 400$

3. Trabajando con proporciones o atributos diremos que en el muestreo simple aleatorio: cada elemento tiene la misma probabilidad de ser seleccionado y, por ejemplo con n = 300 y α = 5%, ξ = 95% $Z\alpha$ = 1.96, el error permitido (e) o margen de error permitido para p = 0.5 = q será igual a:

$e = \sqrt{\frac{pq}{n} * Z\alpha} = \sqrt{\frac{(0.5)(0.5)}{300}} * 1.96$

$e = \sigma p * Z\alpha = 5\%$

VI.5.1 EVALUACIÓN DEL TAMAÑO DE LA MUESTRA [1]

Partiendo de $n = \frac{Z^2\sigma^2}{e^2}$ donde e: es el error mínimo permitido, que lo determina el investigador, ya que sólo él está en condiciones de fijarlo para aceptar su resultado muestral. Por ejemplo puede especificar que si la media obtenida de la muestra es $ 6 mayor o menor que la media verdadera (poblacional), considerará que el estimador $\bar{x}$ obtenido mediante la muestra es satisfactorio. Por lo tanto e = $6, y el intervalo de confianza es $\bar{x} \pm$ $ 6.

$Z\alpha$ se establece mediante el nivel de confianza del intervalo; por ejemplo si el investigador desea que el resultado de la estimación sea ξ = 99.73% prácticamente seguro, ξ= 99.73%, de que la media estimada de la población con base en la muestra esté dentro del recorrido de la verdadera media de la población ± $ 6 ó $\mu x \pm$ $6, el valor de $Z\alpha$ es 3.

Así, una vez que tenemos el tamaño de la muestra, el resultado de la muestra debe ser evaluado. Esto puede ser hecho encontrando el ERROR

[1] Tomado del Profesor Stephen P. Shao.

ESTÁNDAR DE LA MEDIA $S\bar{x}$, de acuerdo con la desviación estándar de la muestra $\hat{s}$.
45

Si el producto de $Z\alpha$ $S\bar{x}$ es menor que el error de muestreo especificado, la estimación de la muestra es considerada satisfactoria. Si el producto es mayor, el tamaño de la muestra deberá ser revisado e incrementado.

Ejemplo: El Gerente de una estación de servicio desea muestrear las notas de venta a fin de encontrar la cantidad (media) promedia por venta durante un período dado.

Para ello indica que 1) el máximo error muestral no deberá ser mayor que 20 ¢ por arriba o por abajo de la verdadera media; 2) el nivel de confianza deberá ser ξ 99.73%; y 3) la desviación estándar de la población basada en su experiencia, es estimado en 80 %. Encontrar el tamaño de la muestra adecuada con estas especificaciones.

SOLUCION:

1.-El intervalo de confianza es $\mu x \pm$ \$ 0.20 luego e = \$ 0.20

2.-Para ξ= 99.73 tenemos $Z\alpha = 3$

3.- $n = \left[\frac{Z\alpha\sigma x}{e}\right]^2 = \left[\frac{3(0.80)}{0.20}\right]^2 = 12^2 = 144$ tamaño de la muestra

Ahora suponga que trabajando con esa muestra seleccionada aleatoriamente se aplica y encontramos:

$\bar{x}$ = \$ 2.70

$\hat{s}$ = \$ 0.72

luego $S\bar{x} = \frac{\hat{S}}{\sqrt{n}} = \frac{0.72}{\sqrt{144}} = \$0.06$

Construimos el intervalo de confianza :

$\bar{x} \pm Z\alpha S\bar{x} = 2.70 \pm 3(0.06) = 2.70 \pm 0.18 = 2.52$ a 2.88

Puesto que $Z\alpha$ $S\bar{x}$ = 0.18 = error de muestreo es menor que el error permitido e = 0.20, se acepta el tamaño de la muestra.

Sin embargo ahora supóngase que con; n = 144 $\hat{s}$ = \$ 0.84, entonces

$S\bar{x} = \frac{\hat{S}}{\sqrt{n}} = \frac{0.84}{\sqrt{144}} = 0.07$ luego:

$\bar{x} \pm Z\alpha S\bar{x} = 2.70 \pm 3(0.07) = 2.70 \pm 0.21$

Como el error de muestreo calculado 0.21 es mayor que el error permitido (e = 0.20), el tamaño de la muestra se revisa como sigue, partiendo de una población infinita:

$$n = \left[\frac{Z\alpha Sx}{e}\right]^2 = \left[\frac{3(0.84)}{0.20}\right]^2 = 158.76 = 159$$

Por lo tanto el tamaño de la muestra aumenta a 159.
Ahora bien; con Sx = 0.80
¿Cuál es el tamaño de la muestra si ξ = 95.45 % y $Z\alpha$ = 2 ?.

$$n = \left[\frac{Z\alpha Sx}{e}\right]^2 = \left[\frac{2(0.80)}{0.20}\right]^2 = 8^2 = 64$$

De este ejemplo numérico se deduce que el tamaño de la muestra depende significativamente de los valores que tomen e, $Z\alpha$, σ_x. En poblaciones finitas, N, es determinante.

Una vez establecida las "definiciones básicas" a continuación empezamos a aplicarlas en temas fundamentales que constituye la ESTADISTICA INDUCTIVA moderna.

Aún cuando la exposición y composición de estos temas no es fácil, yo espero que el esfuerzo didáctico que adopte le permita al lector su fácil entendimiento y manejo continuo en la solución de problemas de su empresa, principalmente, en las áreas de ventas, compras, producción, organización y finanzas.

VI.6 PRECISIÓN - ERRORES DE MUESTREO

Como se indicó, la confiabilidad en las estimaciones se mide por medio de los errores de muestreo, que a su vez, se determinan con las fórmulas de los errores estándar, en términos de probabilidad, es decir: $Z_{\alpha\sigma_{\bar{x}}}$. Con ese propósito a continuación ilustramos las fórmulas de los ERRORES ESTÁNDARES de los principales diseños muestrales, las cuales son muy importantes ya que a partir de ellas se calculan:

a) Tamaño de la muestra,
b) Límites de confianza,
c) Errores de muestreo y
d) Se prueban hipótesis.

Muestreo simple aleatorio

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N*n}};$$ con proporciones: $$\sigma_p = \sqrt{p*q\frac{N-n}{N*n}}$$

Muestreo estratificado.

$$\sigma_{\bar{x}} = \sqrt{\sum_{i=1}^{k} w_i^2 s_i^2 \frac{N_i - n_i}{N_i * n_i}}$$; con proporciones: $$\sigma p = \sqrt{\sum_{i=1}^{k} w_i^2 pq \frac{N_i - n}{N_i * n}}$$

$Si^2 = pq$

donde:

i: estratos: 1,2,3,4,5,.....,K.

W_i : Proporción del estrato en la población = $\frac{N_i}{\sum N_i}$

$P_i = \frac{n_i}{n}$; n= tamaño de la muestra; n_i= muestra en el estrato i-ésimo, N_i = estrato i-ésimo.

Muestreo replicado

$$\sigma_{\bar{x}} = \left|\frac{\bar{X}max - \bar{X}min}{K}\right|\sqrt{\frac{K(Z-K)}{Z(K-1)}}$$

donde:

$\bar{X}$ max: La media mayor en la muestra replicada.

$\bar{X}$ min: La media menor en la muestra replicada.

Z : Tamaño de cada zona

K : Número de replicaciones.

Ejemplos aplicando los errores estándar en la determinación de la precisión.

Se desea estimar con un 95 % de confianza, la proporción verdadera de familias que tienen encendida su T.V. entre las 7 y 10 de la noche. En otras palabras, se busca el intervalo alrededor de la proporción muestral.

Con N = 10,000 familias

Con n = 400 familias con televisión.

VI.6.1 MUESTREO SIMPLE ALEATORIO[11]

Se selecciona una muestra aleatoria y se encuentra que 280 de las 400 televisores están encendidos una o más veces en el tiempo señalado, luego el porcentaje muestral es igual a:

$n_i / n = 70\% = 280/400$

$$\sigma_p = \sqrt{p * q \frac{N - n}{N * n}} =$$

$$\sigma_p = \sqrt{0.70(0.30)\frac{10{,}000 - 400}{10{,}000(400)}} = 2.3\%$$

Por motivos prácticos decimos que en una muestra grande, dos errores estándar proporcionan el intervalo de confianza del 95.45 %, para la proporción verdadera de TV encendidas entre los 7 y 10 de la noche; la estimación del intervalo será:

70 % ± 2 (2.3) ó entre 65.4 % y 74.6 %.

VI.6.2 ESTRATIFICADO:

Estrato	N_i	Nº de entrevistas n	Nº de T.V. encendidas entre 7 y 10 hrs. n_i	$P_i = n_i / n$
1	7,000	200	160	160 ÷ 200 = 80 %

2	1,000	100	40	40 ÷ 100 = 40 %
3	2,000	100	60	60 ÷ 100 = 60 %
	10,000	400	260	

$$\sigma_p = \sqrt{(0.70)^2(0.8)(0.2)\frac{7{,}000-200}{7{,}000*200} + (0.10)^2(0.4)(0.6)\frac{1{,}000-100}{1{,}000*100} + (0.20)^2(0.6)(0.4)\frac{2{,}000-100}{2{,}000*100}} =$$

$$\sqrt{0.0003807 + 0.0000216 + 0.0000912} = \sqrt{0.0004935}\,;\ \ \sigma_p = 0.022 \text{ ó } 2.2\%$$

En este caso, el intervalo es $65\% \pm 2(2.2\%)$ ó entre 60.6% y 69.4%.