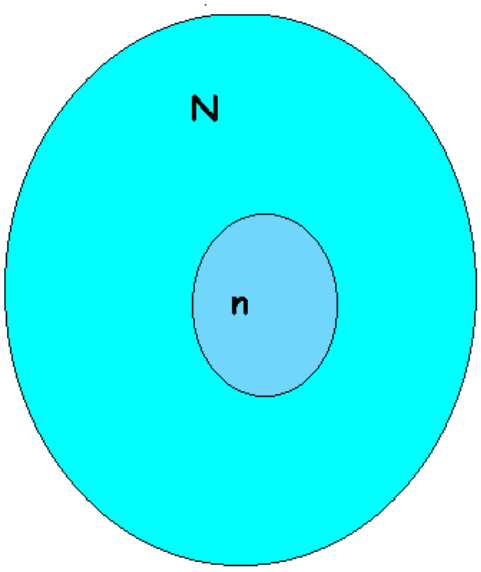
2 MUESTREO SIMPLE ALEATORIO PARA MEDIAS Y TOTALES.

2.1 Definición

El muestreo aleatorio simple no solamente es el más sencillo de aplicar, sino que constituye la unidad elemental de diseño a partir de la cual se suelen plantear muestras complejas. También es el que se apoya en el menor número de supuestos y en esa sencillez reside su flexibilidad y capacidad de aplicación a todo tipo de poblaciones.

Suponga que se tiene una población con las siguientes características:



a) El tamaño de la población es N.

a) El tamaño de la muestra es n.

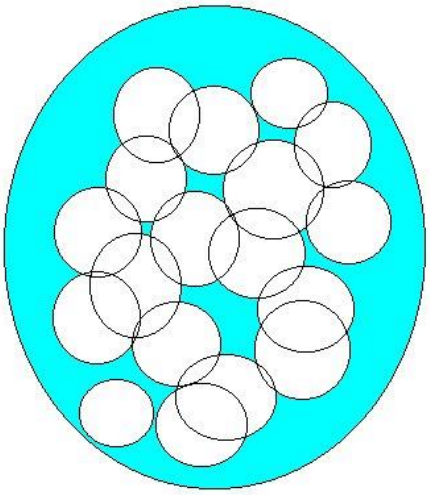
b) Las unidades se seleccionan sin reemplazo, lo que equivale a selecciones sucesivas con probabilidades asociadas a las unidades no seleccionadas en cada extracción iguales a

$$\frac{1}{N-i} \qquad \text{para i= 0,1,2,3,.....,n-1}$$

c) Las muestras que tengan las mismas unidades aunque el orden de extracción sea distinto se consideran iguales y por tanto una muestra es diferente de otra, cuando al menos existe una unidad diferente.

Puesto que se seleccionan sin reemplazo (b) y el orden no importa (c), el número total de muestras está dado por todas las formas posibles de seleccionar n unidades de N en total. Este número de formas corresponde a las combinaciones de los N elementos de la población tomados n a la vez:

$$\binom{N}{n} = \frac{N!}{(N-n)!n!}$$



2.2 Notación

La notación que se empleará en el muestreo aleatorio simple (M.A.S.) será la siguiente:

N	Tamaño de la población
n	Tamaño de la muestra
y_i	El valor de la variable estudiada en la i-ésima unidad de la muestra o de la población.
f	Fracción de Muestreo $f = \frac{n}{N}$
Y	Total de la población $Y = \sum_{i=1}^{N} y_i$
$\bar{Y}$	Media de la población $\bar{Y} = \frac{\sum_{i=1}^{N} y_i}{N}$

$\bar{y}$ Media de la Muestra

$$y = \frac{\sum_{i=1}^{n} y_i}{n}$$

$\hat{\bar{Y}}$ Estimador de la Media

$\hat{Y}$ Estimador del Total

$$\sigma^2 = \frac{\sum_{i=1}^{N} (y_i - \bar{Y})^2}{N}$$ Varianza poblacional

$$S^2 = \frac{\sum_{i=1}^{N} (y_i - \bar{Y})^2}{N-1}$$ Cuasivarianza poblacional

2.3 Números Aleatorios

Para el proceso de selección de la muestra se han elaborado, con la finalidad de obtener las ventajas de la aleatorización y eliminar posibles sesgos, las llamadas Tablas de Números Aleatorios. Estas vinieron a sustituir algunos dispositivos físicos como las urnas.

La primera tabla de números aleatorios de la que se tiene noticia fue "Random Sampling numbers"; Tracts for Computers editada por la Universidad de Cambridge.

El procedimiento de elaboración consistió en tomar números a partir de resultados censales, con ellos se integró una tabla de 41,600 dígitos. Otras tablas conocidas son las de Fisher y Yates, quienes en 1943 construyeron su tabla de 100,000 dígitos (Statistical Tables for use in Biological Agricultural and Medical Research).

Una de las más extensas, pues comprende 1,000,000 de dígitos, es la de la Rand Corporation, elaborada en 1955.

Las tablas se suelen presentar en columnas de 3,4 ó 5 dígitos. Para el empleo correcto de éstas, se deben seguir unas sencillas reglas:

a) Conocer previamente el tamaño de la población N y de la muestra n

b) Se toma una página de las tablas y se parte de cualquier posición tomando el número de dígitos que convenga. El arranque puede darse por coordenadas aleatorias de acuerdo al número de columnas y renglones de la página.

c) Se procede a tomar consecutivamente números en columna o renglón, conservando aquellos menores o iguales a N y suprimiendo los mayores o repetidos en caso de muestreo sin reemplazo hasta completar n.

Generadores de Números Aleatorios

Actualmente, las computadoras cuentan con la función Random que genera números con comportamiento aleatorio basado en algoritmos de congruencias, y aunque los dígitos generados no son estrictamente aleatorios, tienen las propiedades de éstos, lo cual se verifica con diversas pruebas estadísticas, como de uniformidad, rachas, etc. Esta función se incluye en hojas de cálculo y diversos modelos calculadoras de bolsillo.

Las funciones de generación de números pseudoaleatorios usualmente devuelven un número con distribución uniforme en el intervalo (0,1). El argumento puede ser falso o corresponder a una semilla de arranque para la secuencia. Por ejemplo, Excel cuenta con la función ALEATORIO(), la cual se puede utilizar de la siguiente fórmula para generar una muestra un valor entre 1 y N=500. La fórmula asegura la misma probabilidad de aparición para todos los números enteros en el intervalo.

A=ENTERO(ALEATORIO()*500)+1)

En otra plataforma de cálculo se utilizaría una instrucción equivalente.

2.4 Número de Muestras y Probabilidad estar en la Muestra

La probabilidad de disponer de una muestra particular está dada por:

$$\frac{1}{\binom{N}{n}}$$

Una forma sencilla de verificar esto es la siguiente:

Si las unidades de una muestra particular toman los valores $y_1,y_2,...,y_n$; la probabilidad de obtenerlas en ese orden procediendo sin reemplazo, está dada por:

$$\frac{1}{N}\cdot\frac{1}{N-1}\cdot\frac{1}{N-2}\cdots\frac{1}{N-n+1}=\frac{(N-n)!}{N!}$$

Como el orden no importa, entonces se multiplica por todas las posibles formas de ordenar o permutar **n** elementos tomados todos a la vez, es decir **n!**

$$\frac{(N-n)!n!}{N!}=\frac{1}{\binom{N}{n}}$$

Si cada unidad se toma con reemplazo, entonces la probabilidad de una muestra particular está dada por

$$\underbrace{\frac{1}{N}\cdot\frac{1}{N}\cdots\frac{1}{N}}_{n}=\frac{1}{N^n}$$

En un muestreo aleatorio simple sin reemplazo, la probabilidad de que una unidad, en particular con valor y_o, sea elemento de la muestra, está dada por la probabilidad de seleccionar dicho elemento en la primera extracción, esto es 1/N. En la segunda, su probabilidad está condicionada a extraer cualquiera de las N-1 restantes y enseguida extraer la que interesa con probabilidad 1/(N-1). En todos los casos se concluye que la probabilidad de cada extracción es 1/N. A continuación se expone esta secuencia:

1ª Extracción	$= \frac{1}{N}$
2ª Extracción	$\frac{N-1}{N}\frac{1}{N-1} = \frac{1}{N}$
3ª Extracción	$\frac{N-1}{N}\frac{N-2}{N-1}\frac{1}{N-2} = \frac{1}{N}$
……	……
nª Extracción	$\frac{N-1}{N}\frac{N-2}{N-1}\ldots\ldots\frac{1}{N-(n-1)} = \frac{1}{N}$

Como son eventos mutuamente excluyentes, la probabilidad de la unión está dada por la suma de las probabilidades, es decir, la probabilidad de observar la unidad en la 1ª, 2ª, ó nésima extracción estará dada por

$$\underbrace{\frac{1}{N}+\frac{1}{N}+\ldots+\frac{1}{N}}_{n} = \frac{n}{N}$$

Por lo tanto, la probabilidad de que cualquier elemento pertenezca a la muestra es el cociente

$$\frac{n}{N}$$

2.5 Estimadores para Medias y Totales

El estimador usual de la media poblacional $\overline{Y}$ es la media muestral.

$$\boxed{\hat{\overline{Y}} = \overline{y}}$$

El estimador del total **Y** se obtiene de la siguiente forma:

Recuerde que el total de una población se puede expresar con la fórmula:

$Y = \sum_{i=1}^{N} y_i$ Si esta expresión se multiplica y se divide con N, la relación no se altera:

$\overline{Y} = \frac{N\sum_{i=1}^{N} y_i}{N}$ y, por definición de $\overline{Y}$, se tiene: $Y = N\overline{Y}$

Para estimar el total se adopta una forma lógica, basta conocer N y una estimación de la media $\overline{Y}$:

$$\hat{\overline{Y}} = N\hat{\overline{Y}}$$

Como la media muestral $\bar{y}$ es el estimador adoptado de $\overline{Y}$ tendremos como estimador del total:

$$\hat{Y} = Ny$$

2.6 Esperanza y Varianza de los Estimadores.

Para obtener expresiones para la esperanza y varianza del estimador $\bar{y}$ se recurrirá a un modelo de aleatorización, conocido como método de Cornfield.

Sea X_i una variable aleatoria dicotómica tal que:

$$X_i = \begin{cases} 1 & \text{Si la observación } y_i \in \text{en la muestra} \\ 0 & \text{Si la observación } y_i \notin \text{en la muestra} \end{cases}$$

Por la forma como se definió Xi, se trata de una variable aleatoria que se distribuye Bernoulli, de modo que: $\Pr(X_i = 1) = \frac{n}{N}$ $\Pr(X_i = 0) = 1 - \frac{n}{N}$

También el hecho de que X se distribuye Bernoulli permite expresar fácilmente su esperanza y varianza:

$$E(X_i) = P = \frac{n}{N} \qquad V(X_i) = PQ = \frac{n}{N}\left(1 - \frac{n}{N}\right)$$

También involucraremos a la covarianza, la cual se puede expresar como la siguiente diferencia:

$$COV(X_i X_j) = E(X_i X_j) - E(X_i)E(X_j)$$

Se debe obtener una expresión para $E(X_i X_j)$

$$X_i X_j = \begin{cases} 1 \text{ Si } y_i\ y_j \in \text{en la muestra con probabilidad } \frac{n}{N}\frac{n-1}{N-1} \\ 0 \text{ Si } y_i\ y_j \notin \text{en la muestra con probabilidad asociada a tres casos} 1 - \frac{n}{N}\frac{n-1}{N-1} \end{cases}$$

De aquí se obtiene: $E(X_i X_j) = (1)\left(\frac{n}{N}\frac{n-1}{N-1}\right) + (0)\left(1 - \frac{n}{N}\frac{n-1}{N-1}\right)$

$$\boxed{E\left(X_i X_j\right) = \frac{n}{N}\frac{n-1}{N-1}}$$

Ahora la esperanza del producto se sustituye en la expresión de la covarianza

$$\begin{aligned} COV\left(X_i X_j\right) &= E\left(X_i X_j\right) - E\left(X_i\right)E\left(X_j\right) \\ &= \frac{\mathrm{n}}{\mathrm{N}}\frac{n-1}{N-1} - \left(\frac{n}{N}\right)^2 \\ &= \frac{\mathrm{n}}{\mathrm{N(N-1)}}\left(n-1-\frac{n}{N}(N-1)\right) \\ &= \frac{\mathrm{n}}{\mathrm{N(N-1)}}\left(n-1-\frac{nN}{N}+\frac{n}{N}\right) \\ &= \frac{\mathrm{n}}{\mathrm{N(N-1)}}\left(-1+\frac{n}{N}\right) \end{aligned}$$

$$\boxed{COV\left(X_i X_j\right) = -\frac{n}{N(N-1)}\left(1-\frac{n}{N}\right)}$$

Ahora bien, de acuerdo a la definición de X_i, podemos expresar a la media muestral como una suma de todos los valores de la población multiplicados por una variable indicadora que adopta solamente los valores (0,1) y que por tanto apunta solamente a los valores correspondientes a las unidades en muestra.

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{i=1}^{N} X_i y_i}{n}$$

Se verifica a continuación que la media muestral es un estimador insesgado.

$$E(y) = E\left(\frac{\sum_{i=1}^{N} X_i y_i}{n}\right)$$

$$= \frac{\sum_{i=1}^{N} E\left(X_i\right) y_i}{n}$$

$$= \frac{\sum_{i=1}^{N} \frac{n}{N} y_i}{n}$$

$$= \frac{\sum_{i=1}^{N} y_i}{N} = \overline{Y}$$

Como resultado inmediato, el estimador del total también es un estimador insesgado.

$$E(\hat{Y}) = E(Ny) = NE(y) = N\overline{Y} = Y$$

Para abordar el problema de la varianza del estimador se definen a continuación dos estadísticas que involucran a toda la población. La varianza y la cuasivarianza parametrales.

$$\sigma^2 = \frac{\sum_{i=1}^{N}(y_i - \overline{Y})^2}{N}, \qquad S^2 = \frac{\sum_{i=1}^{N}(y_i - \overline{Y})^2}{N-1}$$

Los desarrollos algebraicos se suelen simplificar con el empleo de S^2, sobre todo bajo el enfoque de análisis de varianza, de ahí su presencia más frecuente en todo tipo de desarrollos.

Se aplica el modelo de aleatorización para obtener la varianza del estimador.

Entonces su varianza se expresa: $V(y) = V\left(\frac{\sum_{i=1}^{N} X_i y_i}{n}\right) = \frac{1}{n^2} V\left(\sum_{i=1}^{N} X_i y_i\right)$

Por otra parte, recuérdese que la varianza de una suma de variables aleatorias es:

$$\left\| V(\sum x) = \sum V(x) + 2\sum_{i<}\sum_{j} COV(x_i x_j) \right\|$$

Con los resultados anteriores se obtiene la varianza de $\bar{y}$

$$\mathrm{V(y)} = \frac{1}{\mathrm{n}^2}\left(\sum_{i=1}^{N} V(X_i y_i) + 2\sum_{i<}^{N-1}\sum_{j}^{N} COV(X_i y_i, X_j y_j)\right)$$

$$= \frac{1}{\mathrm{n}^2}\left[\sum_{i=1}^{N} \mathrm{y}_i^2 V(X_i) + 2\sum_{i<}\sum_{j} y_i y_j COV(X_i X_j)\right]$$

$$= \frac{1}{\mathrm{n}^2}\left[\sum_{i=1}^{N} \mathrm{y}_i^2 \frac{n}{N}\left(1 - \frac{n}{N}\right) - 2\sum_{i<}\sum_{j} y_i y_j \frac{n}{N(N-1)}\left(1 - \frac{n}{N}\right)\right]$$

$$= \frac{1}{\mathrm{nN}}\left(1-\frac{n}{N}\right)\left[\sum_{i=1}^{N}\mathrm{y}_{\mathrm{i}}^{2}-\frac{2}{(N-1)}\sum_{i<}\sum_{j}y_i y_j\right]$$

$$= \frac{1\text{-}\mathrm{f}}{\mathrm{Nn}}\left[\frac{N-1}{N-1}\sum_{i=1}^{N}\mathrm{y}_{\mathrm{i}}^{2}-\frac{2}{(N-1)}\sum_{i<}\sum_{j}y_i y_j\right]$$

$$= \frac{1\text{-}\mathrm{f}}{\mathrm{Nn}}\left[\frac{N}{N-1}\sum_{i=1}^{N}\mathrm{y}_{\mathrm{i}}^{2}-\frac{1}{(N-1)}\left\{\sum_{i=1}^{N}y_i^2+2\sum_{i<}^{N-1}\sum_{j}^{N}y_i y_j\right\}\right]$$

$$= \frac{1\text{-}\mathrm{f}}{\mathrm{Nn}}\left[\frac{N}{N-1}\sum_{i=1}^{N}\mathrm{y}_{\mathrm{i}}^{2}-\frac{1}{N-1}\left(\sum_{i=1}^{N}y_i\right)^2\right]$$

$$= \frac{1\text{-}\mathrm{f}}{\mathrm{Nn}}\frac{N}{N-1}\left[\sum_{i=1}^{N}\mathrm{y}_{\mathrm{i}}^{2}-\frac{Y^2}{N}\right]$$

$$= \frac{1\text{-}\mathrm{f}}{\mathrm{Nn}}\frac{N}{N-1}\left[\sum_{i=1}^{N}\left(y_i^2-\bar{Y}^2\right)\right]$$

$$\boxed{V(y)=(1-f)\frac{S^2}{n}}$$

De donde se surge fácilmente la varianza del estimador del total $\hat{Y}=N\bar{y}$, la cual tiene la siguiente expresión:

$$V\left(\hat{Y}\right)=N^2\left(1-\frac{n}{N}\right)\frac{S^2}{n}$$

Su verificación Se sabe que si K es una constante y X es una variable aleatoria:

$$\left\|V(KX)=K^2V(X)\right\|$$

Aplicando el resultado a nuestro caso:

$$V\left(\hat{Y}\right)=V\left(N\bar{y}\right)$$
$$= N^2V\left(\bar{y}\right)$$
$$= N^2\left(1-\frac{n}{N}\right)\frac{S^2}{n}$$

$$\boxed{V\left(\hat{Y}\right)=N^2\left(1-\frac{n}{N}\right)\frac{S^2}{n}}$$

Tanto la varianza de $\bar{y}$ como la de $\hat{Y}$ se expresan en función de S^2, parámetro generalmente desconocido. En la práctica se procede con estimaciones de las varianzas de $\bar{y}$ y $\bar{Y}$ calculadas en base al estimador de S^2 el cual será s^2:

$$s^2 = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}{n-1}$$

El estimador s^2 de la S^2 poblacional es un estimador insesgado.

$$E\left(s^2\right) = S^2$$

Recuérdense las siguientes expresiones y sus equivalencias:

$$\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2 = \sum_{i=1}^{N} X_i^2 - N\bar{X}^2$$

$$V(X) = E\left[\left(X - \bar{X}\right)^2\right] = E\left(X^2\right) - \bar{X}^2$$

Debido a que $\bar{y}$ es una variable aleatoria se tendrá: $E\left(\bar{y}^2\right) = V\left(\bar{y}\right) + \bar{Y}^2 = \left(1 - \frac{n}{N}\right)\frac{S^2}{n} + \bar{Y}^2$

Por lo tanto si se recurre nuevamente al modelo de aleatorización se verifica su insesgamiento.

$$E\left(s^2\right) = E\left\{\frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}{n-1}\right\}$$

$$= \frac{1}{n-1}E\left\{\sum_{i=1}^{n} y^2{}_i - n\bar{y}^2\right\}$$

$$= \frac{1}{n-1}E\left\{\sum_{i=1}^{N}(X_i y_i)^2 - n\bar{y}^2\right\}$$

$$= \frac{1}{n-1}E\left\{\sum_{i=1}^{N} X_i^2 y_i^2 - n\bar{y}^2\right\}$$

$$= \frac{1}{n-1}\left\{\sum_{i=1}^{N} E(X_i^2) y_i^2 - nE\left(\bar{y}^2\right)\right\}$$

$$= \frac{1}{n-1}\left\{\sum_{i=1}^{N}\frac{n}{N}y_i^2 - n\left[\left(1-\frac{n}{N}\right)\frac{S^2}{n}+\bar{Y}^2\right]\right\}$$

$$= \frac{n}{n-1}\left\{\frac{1}{N}\sum_{i=1}^{N}y_i^2 - \left(1-\frac{n}{N}\right)\frac{S^2}{n}-\bar{Y}^2\right\}$$

$$= \frac{n}{n-1}\left\{\frac{1}{N}\sum_{i=1}^{N}y_i^2 - \frac{N\bar{Y}^2}{N}-\left(1-\frac{n}{N}\right)\frac{S^2}{n}\right\}$$

$$= \frac{n}{(n-1)}\left\{\frac{1}{N}\left(\sum_{i=1}^{N}y_i^2 - N\bar{Y}^2\right)-\left(1-\frac{n}{N}\right)\frac{S^2}{n}\right\}$$

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

$$= \frac{nS^2}{N(n-1)}\left\{(N-1)-(N-n)\frac{1}{n}\right\}$$

$$= S^2$$

Por tanto se trata de un estimador insesgado.

$$\boxed{E(s^2) = S^2}$$

Al aplicar los resultados previos se tendrán los estimadores de las varianzas insesgados de $\bar{y}$ y $\hat{Y}$ dados por:

$$\boxed{\hat{V}(\bar{y}) = \left(1-\frac{n}{N}\right)\frac{s^2}{n}} \qquad \boxed{\hat{V}(\hat{Y}) = N^2\left(1-\frac{n}{N}\right)\frac{s^2}{n}}$$

2.7 Muestreo con reemplazo

El muestreo aleatorio simple supone selección aleatoria sin reemplazo, pero ¿qué ventaja ofrece seleccionar la muestra sin reemplazo? Se analizan a continuación las consecuencias de seleccionar una muestra con reemplazo. Seleccionar la muestra con reemplazo es equivalente a disponer de una serie de N casillas vacías colocadas en línea y arrojar n bolas. Habrá casillas en las que caigan cero bolas y otras que podrán tener 1,2,….n bolas. La distribución asociada es una multinomial con los siguientes parámetros: $E(X_i) = nP_i = n(1/N)$, $V(X_i) = nP_i(1-P_i) = n(1/N)(1-1/N)$ y $Cov(X_i, X_j) = -nP_iP_j = -n/N^2$. El estimador de la media con reemplazo es insesgado y la diferencia está en su varianza.

$$V(\bar{y}_R) = \frac{1}{n^2}V\left(\sum_{i=1}^{N}y_iX_i\right)$$

$$= \frac{1}{n^2}\left[\sum_{i=1}^{N}y_i^2V(X_i) + 2\sum_{i<}^{N-1}\sum_{j}^{N}y_iy_jCov(X_i, X_j)\right]$$

$$= \frac{1}{n^2}\left[\sum_{i=1}^{N} y_i^2 V(X_i) + 2\sum_{i<}^{N-1}\sum_{j}^{N} y_i y_j Cov(X_i, X_j)\right]$$

$$= \frac{1}{n^2}\left[\sum_{i=1}^{N} y_i^2 \frac{n}{N}\left(\frac{N-1}{N}\right) - 2\sum_{i<}^{N-1}\sum_{j}^{N} y_i y_j \frac{n}{N^2}\right]$$

$$= \frac{1}{n^2}\frac{n}{N^2}\left[\sum_{i=1}^{N} y_i^2 (N-1) - 2\sum_{i<}^{N-1}\sum_{j}^{N} y_i y_j\right]$$

$$= \frac{1}{n^2}\frac{n}{N^2}\left[N\sum_{i=1}^{N} y_i^2 - \left[\sum_{i=1}^{N} y_i^2 + 2\sum_{i<}^{N-1}\sum_{j}^{N} y_i y_j\right]\right]$$

$$= \frac{1}{n^2}\frac{n}{N^2}\left[N\sum_{i=1}^{N} y_i^2 - \frac{N^2}{N^2}\left(\sum_{i=1}^{N} y_i\right)^2\right]$$

$$= \frac{1}{n^2}\frac{nN}{N^2}\left[\sum_{i=1}^{N} y_i^2 - N\left(\frac{1}{N}\sum_{i=1}^{N} y_i\right)^2\right]$$

$$= \frac{1}{nN}\left[\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^{2^2}\right]$$

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

Expresión conocida en el caso de la varianza de la media para poblaciones infinitas. Es evidente que el estimador calculado a partir de una muestra sin reemplazo tiene una menor varianza que el calculado a partir de una muestra con reemplazo.

2.8 Intervalos de Confianza para Medias y Totales

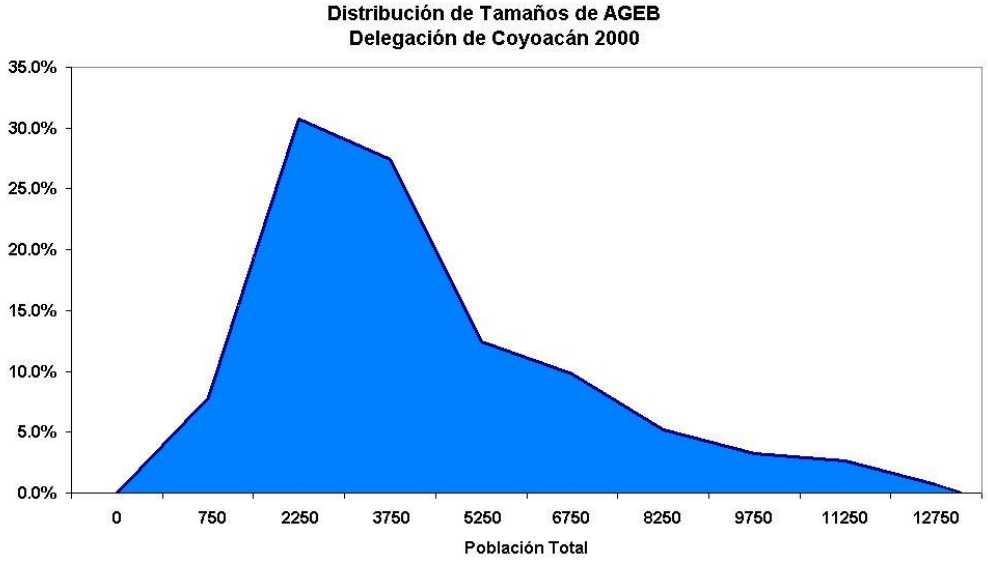
Generalmente se supone que los estimadores de la media $\bar{Y}$ y el total Y se distribuyen en forma normal en torno a los parámetros. Esta suposición se basa en ciertos resultados análogos al Teorema Central del Límite, el cual es válido para poblaciones infinitas. Hájek encontró que la condición necesaria y suficiente para que se considere que la distribución de $\bar{y}$ tiende a la normalidad es:

$$\underset{v \to \infty}{Lím} \frac{\sum_{i=1}^{n_v}\left(y_{vi} - \bar{Y}_v\right)^2}{\left(n_V - 1\right)S_v^2} = 0$$

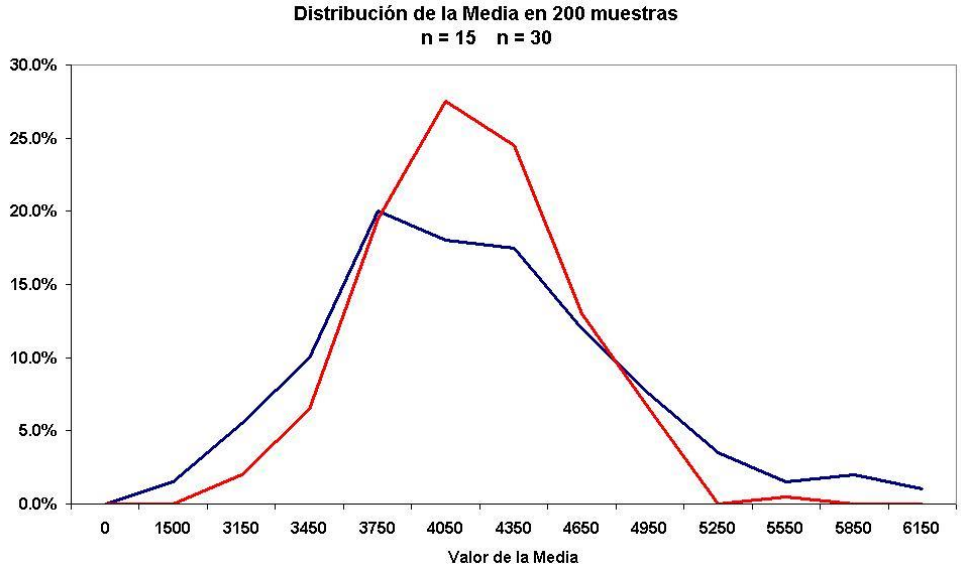
Sin embargo, influye de manera definitiva el conocimiento previo que se tenga de la variable, ya que variables con un comportamiento francamente asimétrico, como son: los tamaños de las ciudades, de empresas ó tiendas, el ingreso de la población, etc.; requieren tamaños mayores de muestra para su convergencia a la normalidad que los requeridos para variables de comportamiento simétrico, como son las medidas antropométricas y sus equivalentes en cualquier tipo de organismos.

Las muestras relativamente pequeñas de poblaciones asimétricas suelen conservar parcialmente esa asimetría en la distribución de sus correspondientes medias.

Considere como ejemplo la distribución de tamaños de población de 153 Areas Geoestadísticas Básicas de la Delegación de Coyoacán según el censo de población y vivienda del año 2000. La distribución de sus tamaños tiene un comportamiento claramente asimétrico. El tamaño promedio de las 153 AGEBs es de 4,185.8 personas.



Mediante simulación de Montecarlo se extrajeron 200 muestras de tamaño 15 y 200 muestras de tamaño 30. En la siguiente gráfica se presentan las distribuciones empíricas de ambas simulaciojnes. Puede observarse que en la muestras de tamaño 15 hay claros rastros de asimetría. En la distribución de las muestras de tamaño 30 la presencia de la asimetría es menor y desde luego con una menor varianza en torno al valor promedio y aproximación a la normalidad. El error estándar calculada empíricamente para n =15 en base a las 200 muestras fue de 628.2 y el correspondiente a n = 30 fue de 419.4. Ambos valores se aproximan a los valores poblacionales de 637.5 y 417.2 respectivamente.



Si se supone que $\bar{y} \approx N\left(\bar{Y}, \left(1-\frac{n}{N}\right)\frac{S^2}{n}\right)$ y por otro lado recordamos que para una variable aleatoria Z→N (0,1) un intervalo del 100 (1-α)% de confianza se obtiene de la siguiente forma:

$$P\left[-Z_{(1-\alpha/2)} < Z < Z_{(1-\alpha/2)}\right] = 1-\alpha$$

Se estandariza la media $Z = \frac{\bar{y}-\bar{Y}}{\sqrt{\left(1-\frac{n}{N}\right)\frac{S^2}{n}}}$ y se obtienen los límites del intervalo.

$\therefore$

$$P\left[-Z_{(1-\alpha/2)} < \frac{\bar{y}-\bar{Y}}{\sqrt{\left(1-\frac{n}{N}\right)\frac{S^2}{n}}} < Z_{(1-\alpha/2)}\right] = 1-\alpha$$

$\therefore$

$$P\left[\bar{y} - Z_{(1-\alpha/2)}\sqrt{\left(1-\frac{n}{N}\right)\frac{S^2}{n}} < \bar{Y} < \bar{y} + Z_{(1-\alpha/2)}\sqrt{\left(1-\frac{n}{N}\right)\frac{S^2}{n}}\right] = 1-\alpha$$

Así, los límites del intervalo buscado serán:

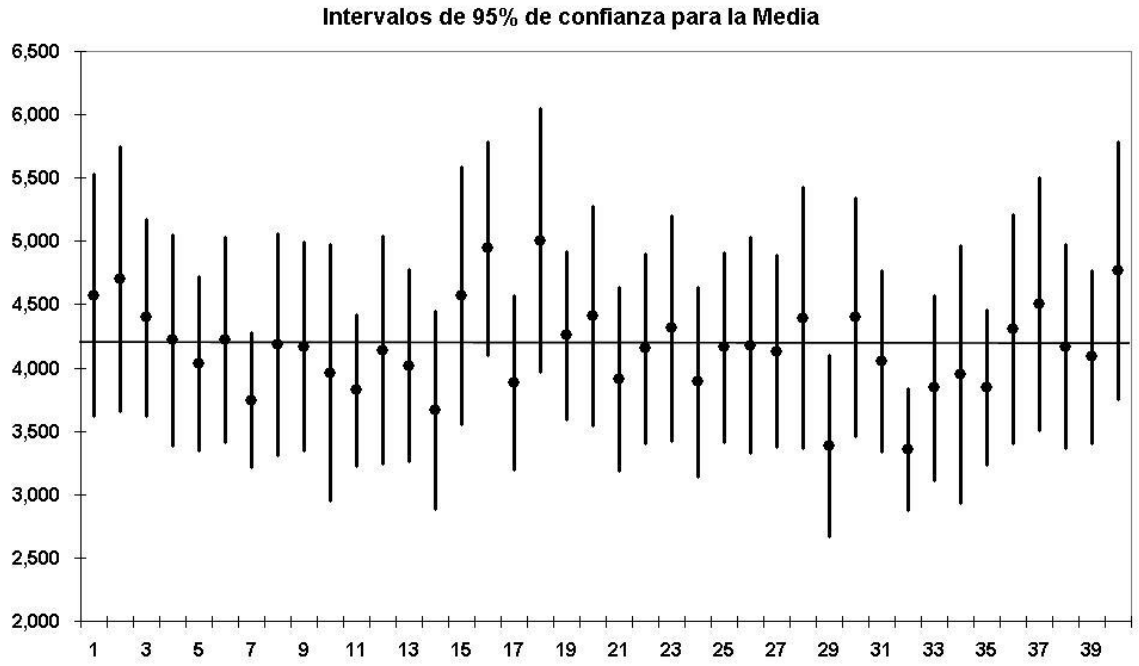
$$\boxed{\bar{y} \pm Z_{(1-\alpha/2)}\sqrt{\left(1-\frac{n}{N}\right)\frac{S^2}{n}}}$$

Debido a que $\hat{\bar{Y}} = N\bar{y}$, se tiene como corolario que los límites para un intervalo de 100(1-α) para el total Y, serán:

$$\boxed{\hat{Y} \pm NZ_{(1-\alpha/2)}\sqrt{\left(1-\frac{n}{N}\right)\frac{S^2}{n}}}$$

Al desconocer S^2 se puede utilizar su estimador s^2. En sentido estricto la distribución a utilizar sería la t de Student con n-1 grados de libertad, pero si n>50 resulta indistinto para efectos prácticos utilizar valores percentilares de la normal estándar.

En la siguiente gráfica se presenta una serie de intervalos de 95% de confianza para la media calculados a partir de las primeras 40 muestras de tamaño n = 30 de las AGEBs de Coyoacán. Los intervalos de las posiciones 29 y 32 no cubren al parámetro. Las amplitudes varían debido a los diferentes valores de la estimación de S^2.



Ejemplo 2.1

En una biblioteca se han puesto los libros en 130 anaqueles de tamaño semejante. El número de libros de 15 estantes seleccionados al azar fue registrado en la siguiente forma:

28,23,25,33,31,18,22,29,30,22,26,20,21,28,25

Estime el total de libros en la biblioteca y calcule un intervalo de confianza de 95% para el total.N = 130

$$\bar{y} = \frac{\sum_{i=1}^{15} y_i}{15} = 25.4$$

n = 15

$$s^2 = \frac{\sum_{i=1}^{15} \left(y_i - \bar{y}\right)^2}{14} = 19.257143$$

Como n es relativamente pequeña, se utiliza el valor percentilar de t para 97.5% y 14 grados de libertad.

$t_{97.5\%,\ 14\ gl}$ = 2.145 s=4.3882961 $\hat{Y} = N\bar{Y}$ $\hat{Y}$ = 130 (25.4) = 3302

Intervalo de confianza:

$$N\bar{y} \pm NZ_{(1-\alpha/2)}\sqrt{\left(1-\frac{n}{N}\right)\frac{S^2}{n}} \qquad 3302 \pm (130)(2.145)\sqrt{\left(1-\frac{15}{130}\right)\frac{19.257143}{15}}$$

$$3302 \pm 290.165$$

El intervalo solicitado de 95% de confianza para el total Y es (3005,3599)