UNIVERSIDAD COOPERATIVA DE COLOMBIA
FACULTAD DE CIENCIAS ECONÓMICAS Y ADMINISTRATIVAS

CURSO DE MUESTREO

Prof. Dagoberto Salgado Horta

Ibagué, Julio 2009

INDICE

INTRODUCCIÓN

En toda investigación estadística existe un conjunto de elementos sobre los que se toma información. Este conjunto de elementos es lo que se denota con el nombre de población o universo estadístico. Cuando se toma información de todos y cada uno de los elementos de dicha población, decimos que se realiza un censo. Sin embargo, esto no siempre es posible, ya sea porque es costoso, requiere mucho tiempo, o bien porque la toma de información lleve consigo la destrucción de los elementos en cuestión, o que la población tenga infinitos elementos. Este problema hace que el investigador tome la información de una parte de la población, proceso que recibe el nombre de **muestreo**.

Toda sociedad requiere INFORMACIÓN $\rightarrow$ toma de decisiones. Ya que la información cuesta dinero, el investigador debe determinar que tanta información debe comprar. Demasiado poca información le impide realizar buenas estimaciones, mientras que mucha información ocasiona un despilfarro de dinero.

CARACTERÍSTICAS DE LA INFORMACIÓN

1. Calidad { Suficiente / Confiable
2. Oportuna (tiempo)
3. Bajo costo.

El objetivo de la mayoría de las investigaciones estadísticas consiste en hacer generalizaciones válidas, con información muestral, acerca de poblaciones de las cuales provienen las muestras.

Estadística moderna es una teoría de la información con la inferencia como su objetivo. El medio para la inferencia es la MUESTRA.

Método de búsqueda de información < Censo / Muestreo

¿Qué es el muestreo?. Es una técnica inductiva para estimar totales o promedios. La estimación puede ser tan exacta como queramos al incrementar el tamaño de la muestra. Puede ser acompañada por un límite de error de estimación o bien expresada como un intervalo de confianza.

¿Qué se pretende con el curso muestreo?

Objetivos del Curso:

1. Diseñar los instrumentos o cuestionarios de la encuesta.
2. Economía de adquirir una cantidad específica de información.
3. Tipos de muestreo $\rightarrow$ para científicos, sociales, comercio, administración, economía, ciencias forestales. Los físicos realizan EXPERIMENTOS, los otros realizan ENCUESTAS.
4. Determinación de tamaños de muestra con error especificado. El muestreo abarca 3 grandes áreas.
   4.1. Diseño de la muestra.
   4.2. Determinación de tamaño de muestra.
   4.3. Inferencias.

Con la investigación por muestreo se persigue encontrar perfiles de la población (medidas, resumen, comportamientos del colectivo). No se persigue encontrar características personales.

Justificación

Se presenta de forma intuitiva una introducción general a los métodos de muestreo y gradualmente, se va haciendo especial énfasis en los aspectos conceptuales y analíticos.

Este curso es de particular interés para los alumnos de la Licenciatura de Estadística, sobre todo requisito fundamental para las asignaturas de Análisis de Datos y Análisis Multivariante.

TEMA 1. ORGANIZACIÓN DE UNA INVESTIGACIÓN POR MUESTREO DE ENCUESTA

La encuesta por muestreo es una metodología que abarca más allá del muestreo propiamente dicho, el cual consiste en el método de selección de la muestra, determinación del tamaño de muestra y la inferencia estadística. La finalidad de una encuesta por muestreo es obtener información para satisfacer una necesidad definida. La necesidad de recopilar datos surge en todo campo de la actividad humana.

Ejemplo:
- Población.
- Mano de obra.
- Agricultura.
- Industria.
- Comercio Interno.

Una investigación por muestreo se puede dividir en 3 etapas básicas:

1. Planificación.
2. Recolección de la Información.
3. Análisis de los resultados.

VENTAJAS MUESTREO Vs. CENSO.

a. Costo reducido (Los gastos son menores que los que se realizarían si le lleva a cabo un censo).
b. Mayor rapidez (El muestreo emplea menos tiempo en recopilar y procesar los datos que el censo).
c. Mayor exactitud. Se espera que una encuesta bien empleada produzca resultados más exactos que el censo. En el censo surgen más errores por la complejidad y magnitud del trabajo. El muestreo emplea personas de mayor calibre, es posible capacitarlos mejor y supervisar su trabajo.
d. Estimar validamente el margen de error y decidir si los resultados son suficientemente exactos. Un censo completo no revela el margen de incertidumbre al cual está sometido. En poblaciones pequeñas $\rightarrow$ censo.

TIPOS DE ENCUESTA POR MUESTREO

Según el objetivo que se persiga en la investigación por muestreo, las encuestas se clasifican en:

1. Descriptivas.
2. Analíticas.
3. Exploratorias.

OBJETIVOS

Descriptivas: Permiten describir el comportamiento del fenómeno en estudio, es decir, con ellas se puede conocer cierta información sobre grandes grupos. Ejemplo: número de hombres que ven televisión.

Analíticas: aquellas que permiten hacer comparaciones entre subgrupos de una población para averiguar si existen ciertas diferencias entre ellos y formular o verificar hipótesis sobre sus causas. Se emplean técnicas multivariantes.

Exploratorias: proporcionan un mecanismo de búsqueda cuando se está comenzando a indagar sobre un tema particular. Sirven de base para estudios posteriores y requieren un análisis descriptivo.

DISEÑO DE ENCUESTAS

Formulación del problema de investigación.

Se refiere al planteamiento del problema a investigar, es decir, definir el qué, por qué, para qué y cómo.

Definición de Objetivos.

La primera tarea es fijar en términos concretos los objetivos de la encuesta.

Cobertura: población que se cubrirá. Los objetivos de la encuesta deben definir la población que se quiere cubrir.

VARIABLES Y ATRIBUTOS INVOLUCRADOS EN LAS HIPÓTESIS.

Variable: función real valorada. Característica que puede tomar diferentes valores.

Atributos: De acuerdo a los atributos la variable se clasifica en:

Tipos de Variables
- Cuantitativas
  - Discretas
  - Continuas
- Cualitativas o Categóricas
  - Nominal
  - Ordinal

Escalas de Medición.

- Nominal.
- Ordinal.
- Intervalo.
- Razón o proporción.

DISEÑO DE CUESTIONARIOS

El cuestionario es una parte muy importante de la encuesta por muestreo. Habiendo decidido cuál es la información que se desea obtener, el problema de su presentación requiere considerable habilidad. Las preguntas deben ser claras, sin ambigüedades y al punto. Las preguntas vagas no proporcionan respuestas claras. Deben evitarse las preguntas que orientes respuestas. Como éstas podrían depender en alguna medida del

ORDEN en que se hacen la preguntas, debe considerarse también su orden. Una pequeña prueba previa siempre es útil para decidir sobre un método efectivo para plantear las preguntas. Todos términos técnicos que se utilizan deberán ser definidos adecuadamente. (Des Raj, 39).

Tipos de cuestionarios

- Autoadministrado
- Entrevistas
- Por teléfono, correo, personal —> inspección

Tipos de Preguntas

- Abiertas
- Cerradas
- Selección Múltiple
- Selección dicotómica

Redacción de las Preguntas

Debe ser clara, sin ambigüedades. Es importante cuidar el lenguaje en relación al público que está dirigido la encuesta.

Debe ser en positivo la redacción de la pregunta. Se recomienda que no contenga más de 20 palabras. El número de preguntas esta asociado inversamente a la tasa de respuesta.

Preguntas referentes a temas sensibles.

Se debe evitar preguntas que induzcan respuestas y también, se debe tener cuidado si el orden de la preguntas induce a las respuestas.

Prueba Piloto.

El objetivo fundamental de realizar una encuesta piloto es probar la validez, confiabilidad y precisión del cuestionario o instrumento de medición. También, tiene como propósito, determinar costos y tamaño de muestra de la encuesta. Una vez probado el instrumento, se determina la composición final del cuestionario.

CONCEPTUALIZACION Y DISEÑO DEL INSTRUMENTO

En la investigación por muestreo, esos conceptos deben ser convertidos en preguntas en un cuestionario que permite la recolección de los datos empíricos relevantes para analizar.

Lógica de Conceptualización.
Ejemplo: Estatus social puede ser definido por varios elementos: ingreso, prestigio ocupacional, educación, riqueza, poder, estatus familiar y valores morales.

Para permitir rigurosa investigación, sin embargo, tales conceptos generales deben ser especificados, esto es, deben ser reducidos para especificar, indicadores empíricos.

Operacionalización.
Los conceptos son codificados generales de la experiencia y observaciones.

En ciencias tales conceptos toman la forma de variables que traen una colección de atributos relacionados.

- Operacionalización es el proceso mediante el cual los investigadores especifican observaciones empíricas que pueden ser tomados como atributos contenidos dentro de un concepto dado.

CALIDAD DE LA MEDICIÓN

Los elementos siguientes deben ser considerados en el diseño de la encuesta, a fin de garantizar la adquisición de información de calidad.

- Precisión
- Confiabilidad
- Validez:
  - ♦ Validez de Contenido
  - ♦ Validez de Constructo

FORMATO DE PRESENTACIÓN DEL CUESTIONARIO

El formato del cuestionario debe ser tan importante como la naturaleza y redacción de las preguntas. Una inapropiada presentación del cuestionario puede conducir a respuestas erróneas.

Se debe evitar:

- ♦ Cuestionarios demasiados largos, ya que el N° de preguntas está asociado inversamente a la tasa de repuesta.
- ♦ Varias preguntas en una sola línea.
- ♦ Preguntas abreviadas.
- ♦ Demasiadas páginas del cuestionario —> que el entrevistado sienta que gasta poco tiempo en responder el cuestionario.
- ♦ Cuestionario muy comprimido en espacio son desastrosos.

Formatos para respuestas

[ ] Si	( ) Si	1. Si
[ ] No	( ) No	2. No
[ ] No sabe	( ) No sabe	3. No sabe

Preguntas Contingencia

A menudo en una encuesta, ciertas preguntas serán claramente relevantes solo para un subconjunto de respondientes.

Pregunta de Contingencia: significa que la segunda pregunta es un contingente, cuya respuesta depende de la primera.

El uso apropiado de estas preguntas puede facilitarle al respondiente la tarea de responder el cuestionario y también puede mejorar la calidad de los datos producidos.

La segunda pregunta se debe evitar que comience ¿ Si..... condicionalmente porque puede inducir a respuesta.

Estas segundas preguntas deben ser indentadas sobre el cuestionario, encerradas en cajas y conectadas con la pregunta base a través de flechas.

Preguntas Matriz.

Típico caso es el de escala Likert.
A menudo, Ud. deseará preguntar varias cuestiones que tengan el mismo conjunto de categorías de respuesta.

Ejemplo:

17. Al lado de cada afirmación que se presenta más abajo, indique si Ud. está completamente de Acuerdo (CA), Acuerdo (A), en Desacuerdo (D), Completamente en Desacuerdo (CD) o Indeciso (I).

		CA	**A**	**D**	**CD**	**I**
a.	Este país necesita más leyes y orden	[ ]	[ ]	[ ]	[ ]	[ ]
b.	La política debe ser el desarme.	[ ]	[ ]	[ ]	[ ]	[ ]
c.	Durante los disturbios se deben disparar perdigones a los saqueadores.	[ ]	[ ]	[ ]	[ ]	[ ]

Existen algunos peligros inherentes al uso de este formato como: Los respondientes pueden desarrollar algún patrón de respuesta.

Preguntas referentes a temas sensibles.

Se deben evitar las preguntas directas que comprometan la integridad física, emocional, moral o espiritual del encuestado.

SECUENCIA Y ORDENAMIENTO DE LAS PREGUNTAS

El orden en el cual las preguntas son presentadas pueden afectar las respuestas, así como toda la actividad de recolección. Por ejemplo, la presencia de una pregunta puede afectar las respuestas dadas en las siguientes preguntas.

Algunos investigadores intentan este efecto "aleatorizando" el orden de las preguntas.

La solución más segura es la sensibilidad del problema. Se debe construir más de una versión del cuestionario que contenga diferentes ordenamientos de las preguntas.

El orden de las preguntas depende el tipo de cuestionario, si es autoadministrado o entrevista. En el primer caso, usualmente es mejor comenzar el cuestionario con el conjunto de preguntas más interesantes. Las preguntas iniciales no deben ser amenazantes. Las preguntas de identificación se deben dejar por el final de la encuesta.

REPRODUCCIÓN DEL CUESTIONARIO

El método de reproducción del cuestionario es importante para el logro de éxito del estudio, un cuestionario nítidamente reproducido indicará a una alta tasa de respuesta y así, proporcionara mejores datos.

Varios métodos están disponibles, y los cuales dependerán de los recursos disponibles, facilidades locales y tiempo.

INSTRUCCIONES

Cada cuestionario, si es autoadministrado o si es administrado por el encuestador, debe contener instrucciones claras y comentarios introductorios donde sean apropiados.

Instrucciones Generales
Cada cuestionario autoadministrado debe comenzar con instrucciones básicas para seguir su completación.

Introducciones
Si el cuestionario esta organizado de acuerdo al contenido de subsecciones, es útil introducir cada sección oraciones cortas relacionadas con el contenido y propósito.

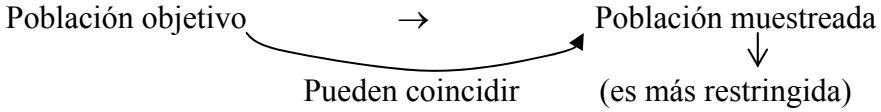
Instrucciones Específicas.
Algunas preguntas pueden requerir instrucciones específicas para facilitar la respuesta apropiada. Caso de respuestas múltiples.

Instrucciones para el entrevistador
Proporcionar instrucciones claras en el lugar apropiado para los entrevistadores.

ETAPAS PRINCIPALES INVESTIGACIÓN POR MUESTREO

- **PLANEACIÓN Y EJECUCIÓN DE UNA ENCUESTA**

a. Objetivos de la encuesta. Fijar en términos concretos los objetivos de la encuesta. No aclarar la finalidad de la encuesta disminuirá su valor en última instancia.
b. Población bajo muestreo: los objetivos de la encuesta deben definir la población que se quiere cubrir. La palabra población se emplea para denominar el conjunto del que se elige la muestra. Implica la definición de lo que es población y de sus elementos. Evitar ambigüedades. El encuestador debe ser capaz de decidir en el campo sin demasiados titubeos si un caso dudosos pertenece o no a la población.
La población que se procura cubrir será por lo general diferente de la que en realidad es objeto de muestreo. Los resultados que se obtengan serán aplicados a la población muestreada.



c. El marco. Lista, mapa, que sirve como guía al universo que se cubrirá, debe examinarse que esté libre de defecto y actualizada.
d. Unidad de muestreo. Para los propósitos de la selección de la muestra debe ser posible dividir a la población en unidades de muestreo.
e. Selección de la muestra → objetivo del curso.
f. Información que se recopilará. Qué información se busca obtener debe ser considerada en una de la primeras etapas de planeación de a encuesta. Sólo deben de tenerse datos de interés para los propósitos de la encuesta. Un cuestionario demasiado largo produce una baja general en la calidad de los resultados. Lo práctico es preparar BOSQUEJOS de los cuadros que debe producir la encuesta, a sí se eliminará información no pertinente.
g. Grado de precisión deseado.

Resultados $\rightarrow$ incertidumbre $\rightarrow$ { Muestra
Errores en las mediciones deseadas.

La falta de certeza se reduce al tomar muestras grandes y emplear mejores dispositivos. Implica costos y tiempo. Es mejor especificar el grado de precisión deseado.

h. Método de obtener la información.
   - Encuesta que emplea un cuestionario autoadministrado.
   - Entrevistas.
   - Encuestas con preguntas abiertas y/o cerradas.
   - Encuestas por teléfono, correo o visitas personales.

i. Referencia de tiempo y período de referencia.
   - Referencia de tiempo (período al que pertenecerán los resultados de la encuesta).
   - Período de referencia: período para el cual se obtiene la información de las unidades de muestreo.

j. Cuestionario u hoja de encuesta. Con función de la información a obtener $\rightarrow$ definir presentación encuesta. Las preguntas deben ser claras y sin ambigüedades y al punto. Deben evitarse preguntas que orienten las respuestas. Orden de preguntas. Preguntas control. Prueba piloto. Definir los términos técnicos adecuadamente. Cuestionarios precodificados.

k. La capacitación de los entrevistadores y supervisión, instrucciones detalladas en los métodos que se emplearán las mediciones.

l. Inspección de la información entregada. Control de calidad de la información.

m. Personas que se rehúsan responder. Elaborarse procedimientos para tratar con quienes no respondan.

- **PRESENTACIÓN Y ANÁLISIS DE DATOS.** Dicho análisis se realiza según el plan de tabulaciones diseñado y las técnicas estadísticas propuestas para cumplir con los objetivos previstos en la investigación.

- **INFORME Y PUBLICACIÓN DE RESULTADOS.** En esta última etapa se redacta el informe contentivo de los resultados de la investigación por muestreo y se ejecuta el plan de publicación de los mismos.

CONCEPTOS GENERALES

Población: es una colección de objetos acerca de los cuales deseamos hacer alguna inferencia. Un conjunto finito o infinito de elementos.

Elemento o unidad elemental o unidad de observación: objeto sobre el cual se realizan las mediciones de la característica. Es un objeto en el cual se toman las mediciones.

Unidades de muestreo: son colecciones no traslapadas de elementos de la población que cubran la población completa. Otra definición es: colecciones o grupos no solapados de unidades elementales. También es la unidad donde realizamos la muestra.

Ejemplo:

- Encuesta de Viviendas $\rightarrow$ Unidad de muestreo: manzanas definidas de tal manera que cada vivienda no pueda ser muestreada más de una vez y que cada vivienda tenga una oportunidad de ser seleccionada en la muestra.
- Encuesta sobre Ingreso Familiar $\rightarrow$ Unidad de muestreo: vivienda.
- Proporción de votantes que favorecieron la emisión de bonos $\rightarrow$ Unidad de muestreo: hogares. Unidad elemental: votantes.

En el muestreo de elementos cada unidad de muestreo contiene un solo elemento, por tanto, la Unidad de muestreo = Unidad elemental.

Marco muestral: es una lista de todas las unidades de muestreo.

Muestra: es un subconjunto de la población. Es una colección de unidades seleccionadas de un marco o de varios marcos. En una población infinita, una muestra aleatoria es una sucesión de variables aleatorias independientes e idénticamente distribuidas**.**

POBLACIÓN OBJETIVO		POBLACIÓN MUESTREADA
Población que se pretende cubrir. Definida por los objetivos de la encuesta. Es la colección completa de observaciones que deseamos estudiar.		Es la población de donde se extrae la muestra, es más restringida. Los resultados que se obtengan serán aplicados a la población muestreada.

TIPOS DE MUESTREO

1. Muestreo Probabilístico: Cuando de antemano se conoce la probabilidad asociada a cada muestra posible.

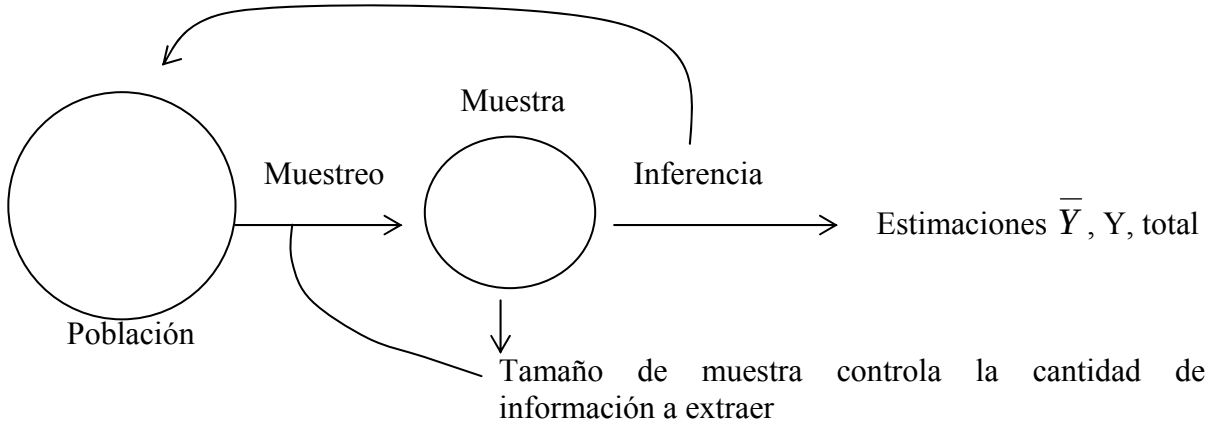
2. Muestreo No Probabilístico:
   - Intencional u opinático (representatividad subjetiva)
   - Sin norma circunstancial o errático
   - a capricho o comodidad.
   - por cuotas: a conveniencia adminis-trativa o económica.
   - Juicio: el investigador emplea su propio juicio para elegir la muestra.

Muestreo Aleatorio: Consiste en asignar a cada elemento poblacional una probabilidad no nula, de ser seleccionado. Con este muestreo podemos hacer estimaciones de las magnitudes de los errores de muestreo (valor estimado → Valor poblacional). Controlar la precisión de las estimaciones muestrales dentro de ciertos límites fijados con anticipación y con cierto grado de confianza.

TIPOS DE MUESTREO PROBABILÍSTICO

1) Muestreo Irrestricto Aleatorio.
2) Muestreo Estratificado.
3) Muestreo Sistemático.
4) Muestreo por Conglomerados o por Áreas.
5) Muestreo Bietápico.
6) Muestreo Polietápico.
7) Muestreo Mixto.

¿CÓMO SELECCIONAR LA MUESTRA?. El objetivo del muestreo es estimar parámetros de la población, tales como media, el total y proporción basándose en la información contenida en la muestra.



$\hat{\theta}$: Estimador. Medida estadística que describe cierta característica numérica de una muestra, siendo una magnitud variable de una muestra a otra.

θ : Parámetro. Medida estadística que describe cierta característica numérica de una población y que se considera constante y desconocida.

¿Cómo podemos determinar cual procedimiento usar y el número de observaciones a incluir en la muestra?. La respuesta depende de cuanta información se desee obtener. La

cantidad de información obtenida en la muestra depende del número de elementos muestreados y de la cantidad de variación en los datos. Debemos fijar un límite para el error de estimación menor que B.

Error de estimación = $\left|\hat{\theta} - \theta\right| < B$ = errores en el muestreo.

$B = e$ = límite de error de estimación.

$$\Pr\left(\left|\hat{\theta} - \theta\right| \leq e\right) = 1 - \alpha$$

El límite de error de estimación viene generalmente expresado en unidades de $\sigma\hat{\theta}$ $e = t_\alpha \sigma\hat{\theta}$ = error máximo admisible. t_α es dado a función 1- α ; 1- α = coeficiente confidencial.

El error de estimación se debe a que una muestra no proporciona información completa sobre una población. Esta clase de error se llama **error de muestreo**, el cual puede ser controlado por un diseño cuidadoso de la encuesta.

El margen de error dado en las encuestas es un expresión del **error de muestreo,** el cual resulta al considerar una muestra y no al examinar toda la población (Lohr, 2000, p. 15)

ERRORES AJENOS AL MUESTREO

Son aquellos que no se deben al muestreo, los cuales no se pueden atribuir a la variabilidad entre las muestras (Lohr, 2000) e influyen en la validez y confiabilidad de las estimaciones. Se pueden clasificar en:

a) Sesgos de estimación: debido al uso inadecuado de un estimador. Cuando se utilizan estimadores sesgados. Mal uso por parte del investigador.
b) Sesgos de selección: errores cometidos cuando el proceso de selección de la muestra no es totalmente aleatorio; pues incluye elementos opináticos y erráticos. Este ocurre cuando alguna parte de la población objetivo no está en la población muestreada.
c) Sesgo de medición: ocurre cuando el instrumento con el que se mide tiene una tendencia a diferir del valor verdadero en alguna dirección. Este debe ser minimizado en la etapa de diseño de la encuesta (Lohr, 2000).
d) Errores de observación o de medida: son el resultado de la interacción entre el observador, el instrumento y el individuo medido (sustituciones fortuitas pueden sesgar los resultados).
e) Errores por omisión: se refiere a la no respuesta, inaccesibilidad del elemento, o pérdida del dato.
f) Equivocaciones en el diseño de la encuesta.

Exactitud: se refiere a la magnitud de las desviaciones respecto a la media verdadera μ.

Precisión: se refiere a la magnitud de las desviaciones respecto a la media $\overline{Y}$ muestral.

TEMA 2. MUESTREO ALEATORIO SIMPLE

En una muestra aleatoria simple cada unidad o elemento de la población tiene una probabilidad de selección conocida; se emplea un método aleatorio para elegir las unidades a incluir en la muestra (Lohr, 2000). Los elementos o unidades podrán ser seleccionados de dos formas: con o sin reposición.

En el muestreo **aleatorio simple** con reemplazo o **con reposición** una unidad o elemento se puede incluir más de una vez en la muestra; mientras en el muestreo sin reemplazo o sin reposición, todas las unidades en la muestra son distintas.

Una muestra aleatoria con reposición, de tamaño *n* obtenida de una población de *N* unidades, se puede pensar como la extracción de *n* muestras independientes de tamaño 1. Cada unidad se extrae de la población al azar, por ser la primera unidad muestreada, con una probabilidad de 1/*N*, la cual se reemplaza en la población, y siguiente unidad se selecciona al azar con una probabilidad de 1/*N*. Este procedimiento se repite hasta que la muestra contenga las *n* unidades y puede tener duplicados.

El muestreo aleatorio sin reemplazo o sin reposición de poblaciones finitas se conoce con el nombre de **muestreo irrestricto aleatorio**, el cual consiste en la selección de n elementos sacados de una población con *N* unidades, de modo que todas las muestras posibles (distintas) $\binom{N}{n}$ de tamaño *n* tengan la misma probabilidad de ser seleccionada $P(S) = \frac{1}{\binom{N}{n}} = \frac{n!(N-n)!}{N!}$. $P(S)$ es la probabilidad de elegir cualquier muestra individual *S* de *n* unidades.

La probabilidad de la muestra también puede calcularse utilizando el cálulo de probabilidades:

1ª selección probabilidad $\frac{n}{N}$

2ª selección probabilidad $\frac{n-1}{N-1}$

Luego, la probabilidad de selección de una muestra $P(S)$ es:

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

Otra forma de calcularla es la que se presenta a continuación. Sea la muestra $S=\{u_1,u_2,...,u_n\}$, luego su probabilidad es una probabilidad condicional,

$$P(S)=P(u_1,u_2,...,u_n)=n!\,P(\{u_1,u_2,...,u_n\})=n!\,P(u_1)P(u_2/u_1)P(u_3/u_1u_2)...P(u_n/u_1u_2...u_{n-1})$$

$$P(S)=n!\frac{1}{N}\frac{1}{N-1}\frac{1}{N-2}\cdots\frac{1}{N-(n-1)}=n!\frac{1}{\frac{N!}{(N-n)!}}=\frac{n!(N-n)!}{N!}=\frac{1}{\frac{N!}{n!(N-n)!}}=\frac{1}{\binom{N}{n}}$$

En el cálculo anterior hemos supuesto que al no intervenir el orden en la colocación de los elementos, la muestra $S=\{u_1,u_2,...,u_n\}$ contiene las *n!* posibles ordenaciones de dicho conjunto.

PROBABILIDAD QUE TIENE UNA UNIDAD DE PERTENECER A LA MUESTRA

Se mencionó que los elementos que formarán la muestra pueden ser seleccionados de dos maneras:

1. **Con reposición**: en este procedimiento los elementos pueden ser seleccionados varias veces, y cada una de las *n* selecciones son independientes unas de otras, luego, la probabilidad de que un elemento forme parte de la muestra es 1/*N*. Por lo tanto, la probabilidad final de forme parte de la muestra de tamaño n es: $\frac{1}{N}+\frac{1}{N}+\cdots+\frac{1}{N}=\frac{n}{N}$. Este tipo de selección coincide con el muestreo de poblaciones infinitas.

2. **Sin reposición**: las unidades pueden ser seleccionadas una sola vez. Recibe el nombre muestreo irrestrictamente aleatorio, y la probabilidad que un elemento sea escogido en la i-ésima extracción estará condicionada a la probabilidad de que no haya sido escogido en los (i-1) sorteos anteriores, así cada selección y probabilidad es:

   1ª selección probabilidad $\frac{1}{N}$

   2ª selección probabilidad $\frac{1}{N-1}\cdot\frac{(N-1)}{N}=\frac{1}{N}$

3ª selección probabilidad $\frac{1}{N-2}\cdot\frac{(N-2)}{N-1}\cdot\frac{N-1}{N}=\frac{1}{N}$

nª selección probabilidad $\frac{1}{N-(n-1)}\cdot\frac{N-(n-1)}{N-(n-2)}\cdots\frac{N-1}{N}=\frac{1}{N}$

De allí que la probabilidad de que un elemento sea seleccionado en cualquiera de las n elecciones será igual a $\frac{1}{N}$ y la probabilidad final de que un elemento sea incluido en la muestra es $\pi_i=\frac{n}{N}$, aquí se aplica la sumatoria de las probabilidades de cada una de n selecciones en las que puede ser elegido el elemento i en la muestra.

También podemos decir que de las $\binom{N}{n}$ muestras posibles, de ellas $\binom{N-1}{n-1}$ contienen un elemento particular, por tanto, su probabilidad es:

$$\pi_i=\frac{\text{No. muestras favorables}}{\text{No. muestras posibles}}=\frac{\binom{N-1}{n-1}}{\binom{N}{n}}=\frac{n}{N}$$

Todo diseño muestral comprende las siguientes partes:

1. Método de selección de la muestra.
2. Estimadores a utilizar y propiedades.
3. Determinación del tamaño de muestra.
4. Modificaciones al diseño básico.

Forma de seleccionar una muestra irrestricta aleatoria

- La selección aleatoria garantiza:
  a. Inferencias estadísticas válidas.
  b. Mejoramientos acumulativos a través de la separación y evaluación objetivo de sus fuentes de error.

- Tablas de número aleatorios.
- Computadora

Este método de muestreo se usa en poblaciones suficientemente homogéneas, es decir, cuya varianza poblacional tienda a cero, exige disponer una lista enumerada de 1 a N y de allí mediante un experimento aleatorio seleccionar a cada uno de los n elementos de la muestra.

Dos factores afectan la cantidad de información contenida en la muestra y por tanto, la precisión (tamaño muestra y cantidad de variación que se controla por el tipo de muestreo).

ESTIMACIÓN DE LA MEDIA Y EL TOTAL

Simbología básica:

y_i = i - ésimo elemento de la muestra
N = total
u_i = elemento genérico de la población

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

Suponga que $y_1, y_{2, \dots}, y_n$ es una muestra irrestricta aleatoria (m.i.a) de una población de valores $u_1, u_{2, \dots}, u_N$,, (considere que y_i la muestra aleatoria es de tamaño uno).

$\mu = E(y_i) = \sum_{i=1}^{n} \frac{y_i}{N} = \mu =$ media poblacional $E(y_i) = \mu = \sum_{i=1}^{N} u_i \left(\frac{1}{N} \right)$

σ^2 varianza poblacional

$$V(Y_i) = E[Y_i - \mu]^2 = \sum (Y_i - \mu)^2 \left(\frac{1}{N} \right) = \frac{1}{N} \left(\sum_{i=1}^{N} Y_i^2 - N\mu^2 \right) = \frac{1}{N} \left[\sum Y_i^2 - \frac{\left(\sum Y_i \right)^2}{N} \right] = \sigma^2$$

La varianza muestral es:

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

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

La Covarianza poblacional en el m.i.a. es $\neq 0$

$$Cov(y_i, y_j) = E\left[(y_i - \mu)(y_j - \mu) \right] = E\left[y_i y_j - \mu y_i - \mu y_j + \mu^2 \right] = E\left[y_i y_j \right] - \mu^2$$

$$= \sum_{i \neq j}^{N} u_i u_j \left(\frac{1}{N(N-1)} \right) - \frac{1}{N^2} \left(\sum_{i=1}^{N} u_i \right)^2$$

$$= \frac{1}{N}\left[\frac{\sum_{i\neq j}^{N} u_i u_j}{N-1} - \frac{1}{N}\left(\sum_{i=1}^{N} u_i\right)^2\right]$$

Como $\left(\sum u_i\right)^2 = \sum_{i=1}^{N} u_i^{\,2} + \sum\sum_{i\neq j} u_i u_i$

Entonces $\sum_{i\neq j}^{N} u_i u_i = \left(\sum_{i=1}^{N} u_i\right)^2 - \sum_{i=1}^{N} u_i^{\,2}$

Sustituyendo en la covarianza se tiene

$$Cov(y_i, y_j) = \frac{1}{N}\left[\frac{\left(\sum_{i=1}^{N} u_i\right)^2 - \sum u_i^{\,2}}{N-1} - \frac{1}{N}\left(\sum u_i\right)^2\right]$$

$$= \frac{1}{N}\left[\left(\sum u_i\right)^2\left(\frac{1}{N-1} - \frac{1}{N}\right) - \frac{\sum u_i^{\,2}}{N-1}\right]$$

$$= \frac{1}{N}\left[\left(\sum u_i\right)^2\left(\frac{1}{N(N-1)}\right) - \frac{\sum u_i^{\,2}}{N-1}\right]$$

$$= -\frac{1}{N}\left[\frac{1}{N-1}\sum u_i^{\,2} - \frac{1}{N(N-1)}\left(\sum u_i\right)^2\right]$$

$$= -\frac{1}{N}\left[\frac{1}{N-1}\sum u_i^{\,2} - \frac{N\mu^2}{N-1}\right]$$

En definitiva, la covarianza queda igual,

$$Cov(y_i, y_j) = -\frac{1}{N(N-1)}\left(\sum u_i^{\,2} - N\mu^2\right) = -\frac{1}{N(N-1)}\sum\left(u_i - \mu\right)^2 = -\frac{1}{N-1}\sigma^2$$

ESTIMACIÓN DE LA MEDIA POBLACIONAL

El estimador de la media μ es $\hat{\mu} = \bar{y} = \frac{\sum y_i}{n}$

Por definición la varianza muestral es:

$S^2 = \frac{\sum y_i^2 - n\bar{y}^2}{n-1}$ entonces $S^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$

Consideremos que:

1) La media muestra es un estimador insesgado, es decir $E(\bar{y}) = \mu$

2) La varianza de la media es $V(\bar{y}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$ y su estimador es $\hat{V}(\bar{y}) = \frac{S^2}{n}\left(\frac{N-n}{N}\right)$ que también es insesgado.

Por definición, el límite de error de estimación es $e = B = t_\alpha \sqrt{\hat{V}(\bar{y})}$

Ahora vamos a demostrar los 2 puntos anteriores:

1. La media muestra es un estimador insesgado, es decir $E(\bar{y}) = \mu$

Considere que y_i es una muestra aleatoria es de tamaño uno.

Por definición la media muestral es $\bar{y} = \frac{\sum y_i}{n}$ al aplicar operador esperanza se tiene

$$E(\bar{y}) = E\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} E(y_i) = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{N} u_j \frac{1}{N}\right) = \frac{1}{n}(n\mu) = \mu$$

2. La varianza de la media es $V(\bar{y}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$ y su estimador es también insesgado, es decir que $E\left(\hat{V}(\bar{y})\right) = V(\bar{y})$.

A continuación vamos a realizar un ejercicio para estimar la media y determinar el error de estimación.

Ejercicio 4.7: Una muestra irrestricta aleatoria de n = 100 medidores de agua es controlada dentro de una comunidad para estimar el promedio de consumo de agua diario por casa, durante un periodo estacional seco. La media y la varianza muéstrales fueron $\bar{y} = 12.5$ y $S^2 = 1252$ Si suponemos que hay N = 10.000 casas dentro de la comunidad, estime μ, el promedio de consumo diario verdadero, y establezca un límite para el error de estimación. (Mendenhall, pag. 68.)

Datos

n = 100, $\bar{y} = 12.5$, $S^2 = 1252$, N = 10000

Se pide estimar μ y B

$$B = 2\sqrt{\hat{V}(\bar{y})} = 2\sqrt{\frac{\hat{S}^2}{n}\left(\frac{N-n}{N}\right)} = 2\sqrt{\frac{1252}{100}\left(\frac{10000-100}{10000}\right)} = 2 * 3.52$$

El error de estimación es $B = 7.04$

El intervalo de confianza para la Media Poblacional es:

$$\bar{y} \pm B \Rightarrow (\bar{y} - B; \bar{y} + B)$$

Sustituyendo los valores obtenidos se tiene que el intervalo es:

$$(12.5 - 7.04 \ ; \ 12.5 + 7.04) \Rightarrow (5.46 \ ; \ 19.54)$$

Se puede interpretar los resultados de la siguiente manera: "Se tiene por lo menos un 75 % de confianza que el verdadero valor del promedio diario de consumo de agua se encuentre entre 5.46 y 19.54".

Como el tamaño de la muestra es grande se puede emplear el teorema central del limite y asumir que la media se aproxima a una normal. En este ejemplo, el error de estimación es igual a: B = 1.96 * 3.52 =6.8992; y el intervalo es (5.40 ; 19.60) el cual indica que tenemos un 95% de confianza que el verdadero valor del consumo de agua promedio poblacional se encuentra entre 5.40 y 19.60.

A continuación vamos a estudiar la estimación del total poblacional, sus varianzas y la determinación del tamaño de muestra.

ESTIMACION DEL TOTAL POBLACIONAL τ

Ya sabemos que su estimador es $\hat{\tau} = N\bar{y} = \dfrac{N\sum_{i}^{n} y_i}{n}$

La varianza poblacional del total estimado $\hat{\tau}$ se obtiene al aplicar el operador varianza a la definición de dicho estimador, el cual queda igual a:

$V(\hat{\tau}) = N^2 \dfrac{S^2}{n}\left(\dfrac{N-n}{N}\right) = N^2 \dfrac{S^2}{n}(1-f)$, donde $f = n/N$ es la fracción de muestreo

La varianza estimada del total estimado $\hat{\tau}$ es:

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

El limite para error de estimación es $B = e = 2\sqrt{\hat{V}(\hat{\tau})}$ o $B = t_k \sqrt{\hat{V}(\hat{\tau})}$

Ejemplo 4.8: Usando los datos del ejercicio 4.7, estime el número total de galones de agua, τ, usado diariamente durante el periodo seco. Establezca un límite para el error de estimación. (Mendenhall, pag. 68.)

Solución:

n =100 medidores, N=10000, $\bar{y} = 12.5$, S^2=1252

$$\hat{T} = N\bar{y} = 10000 * 12.5 = 125000$$

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

$$\hat{V}\left(\hat{T}\right) = (10000)^2 \frac{(1252)}{100}\left(\frac{10000-100}{100}\right) = 1239.48 * (10000)^2 = 1.239.480.000$$

$$\sqrt{\hat{V}(\hat{T})} = \sqrt{123948 * 10^4} = 35206.25$$

$$B = t_k\sqrt{\hat{V}(\hat{T})} = 2\sqrt{\hat{V}(\hat{T})} = 2 * 35206.25 = 70412.5$$

Intervalo de confianza para el total de galones de agua usado durante el periodo seco.

$$\left(\hat{T} - B, \hat{T} + B\right) = (54587.5, 195412.5)$$

TAMAÑO DE LA MUESTRA PARA ESTIMAR LA MEDIA

Para determinar el tamaño de la muestra se despeja n de B con varianza poblacional (4) o varianza estimada (5)

$$e = B = t_\alpha \sqrt{\hat{V}(\bar{y})} = t_\alpha \sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)} \qquad (4)$$

$$e = B = t_\alpha \sqrt{\hat{V}(\bar{y})} = t_\alpha \sqrt{\frac{\hat{S}^2}{n}\left(\frac{N-n}{N-1}\right)} \qquad (5)$$

despejando *n* de (4) se tiene que:

$$e^2 = t_\alpha^2 \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \Rightarrow n\left(\frac{e^2}{t_\alpha^2}\right)(N-1) = N\sigma^2 - n\sigma^2$$

$$n\left(\frac{e^2}{t_\alpha^2}\right)(N-1) + n\sigma^2 = N\sigma^2$$

Finalmente, el tamaño de muestra queda igual a:

$$n = \frac{N\sigma^2}{\left(\frac{e^2}{t_\alpha^2}\right)(N-1) + \sigma^2} \Rightarrow n = \frac{N\sigma^2}{D(N-1) + \sigma^2}$$

TAMAÑO DE LA MUESTRA PARA ESTIMAR EL TOTAL

Por un procedimiento similar al de la media se determina el tamaño de muestra para estimar el total.

$$n = \frac{N\sigma^2}{(N-1)\frac{e^2}{t_\alpha^2 N^2} + \sigma^2} = \frac{N^3\sigma^2 t_\alpha^2}{(N-1)e^2 + \sigma^2 t_\alpha^2 N^2} = \frac{N\sigma^2}{(N-1)D + \sigma^2}$$

$$D = \frac{e^2}{t_\alpha^2 N^2} = \frac{B^2}{t_\alpha^2 N^2} \qquad t_\alpha^2 = 4$$

Tanto en el caso de muestras para estimar el total o la media se supone que el investigador debe conocer σ^2

FORMAS DE CALCULAR ESTIMACIONES DE σ^2

La estimación de la varianza poblacional σ^2 para calcular el tamaño de muestra se puede realizar a través de:

1. Estudios anteriores.
2. Muestra piloto.
3. Usando el rango de la variable (dos desviaciones de la media)
$$\sigma = \frac{rango\ \ de\, Y}{4}$$
4. Consideraciones prácticas acerca de la estructura poblacional.

ESTIMACIÓN DE LA PROPORCIÓN P

El investigador que realiza una encuesta por muestreo frecuentemente esta interesado en estimar la proporción de la población que posee una característica.

Ejemplo: proporción de personas que opinan que el servicio de BIECI es bueno.

Las propiedades de $\hat{P}$ son equivalentes a las de $\bar{y}$ en el muestreo irrestricto aleatorio.

Sea $y_i = 0$ si el i-ésimo elemento seleccionado no posee la característica especifica, y $y_i = 1$ si las posee.

$\hat{p} = \frac{\Sigma y_i}{n} = \bar{y}$ es el estimador de p

La varianza poblacional de la proporción es: $V(p) = \frac{PQ}{n}\left(\frac{N-n}{N-1}\right)$

Varianza estimada de $\hat{p}$ es: $\hat{V}(\hat{p}) = \frac{\hat{p}\hat{q}}{n-1}\left(\frac{N-n}{N}\right)$

A continuación vamos a demostrar: la varianza estimada de la proporción $\hat{V}(\hat{p})$:

Sabemos que $\bar{y} = \frac{\Sigma y_i}{n} = \hat{p} \Rightarrow \Sigma y_i = np$

Sea la cuasivarianza:

$$S^2 = \frac{\Sigma\left(y_i - \bar{y}\right)^2}{n-1} = \frac{\Sigma y_i^{\,2} - n\bar{y}^2}{n-1} = \frac{np - np^2}{n-1} = \frac{n}{n-1}\,p(1-p) \quad = \frac{n}{n-1}\,pq$$ y además

$\hat{V}(\bar{y}) = \frac{S^2}{n}\left(\frac{N-n}{N}\right)$, sustituyendo $S^2 = \frac{n}{n-1}\,pq$ se tiene

$$\hat{V}(\bar{y}) = \frac{n\frac{pq}{n-1}}{n}\left(\frac{N-n}{N}\right) = \frac{pq}{n-1}\left(\frac{N-n}{N}\right) = V(\hat{p})$$ esto es lo que queríamos demostrar.

$$\hat{V}(\bar{y}) = \frac{n\frac{\hat{p}\hat{q}}{n-1}}{n}\left(\frac{N-n}{N}\right) = \frac{\hat{p}\hat{q}}{n-1}\left(\frac{N-n}{N}\right) = \hat{V}(\hat{p})$$ es la varianza estimada de la proporción.

El error de estimación es:

$$e = B = t_\alpha\sqrt{V(\hat{p})}$$

Ejercicio 4.5: Las autoridades de un parque estatal están interesadas en la proporción de personas que acampan y que consideran que el espacio del área disponible para acampar en un terreno en particular es adecuado. Las autoridades decidieron tomar una muestra irrestricta aleatoria de n = 30 de los primeros N = 300 grupos acampados que visitan el campo. Sea $y_i = 0$ si jefe del i-ésimo grupo muestreado considera que el espacio del área disponible para acampar no es adecuado, y $y_i = 1$ si considera que es adecuado (i=1,2,...,30). Use los datos de la tabla adjunta para estimar p, la proporción de personas que acampan y que consideran que el espacio del área disponible para acampar es adecuado. Establezca un limite para el error de estimación (Mendenhall, pag. 67-68.)

Persona Muestreada	**Respuesta** y_i
1	1
2	0
3	1
. . .	. . .
29	1
30	1
	$\sum_{i=1}^{30} y_i = 25$

Solución:

$\sum y_i = 25 \qquad n = 30 \quad N{=}300$

$\hat{p} = \frac{\Sigma y_i}{30} = \frac{25}{30} = 0.8333$ y $\hat{q} = 1 - \hat{p} = 0.1667$

$$\hat{V}(\hat{p}) = \frac{\hat{p}\hat{q}}{n-1}\left(\frac{N-n}{N}\right) = \frac{(0.8333)(0.1667)}{30-1}\left(\frac{300-30}{300}\right) = 0.00431103$$

El error de estimación resulta igual a: *B*=2*0.065658=0.1313. Al calcular el intervalo de confianza queda igual a (0.702 ; 0.9646).

TAMAÑO DE LA MUESTRA PARA ESTIMAR p

Sabemos que el tamaño de muestra para estimar la media en el muestreo irrestricto aleatorio es:

$$n = \frac{N\sigma^2}{\left(\frac{e^2}{t^2{}_k}\right)(N-1)+\sigma^2} = \frac{N\sigma^2}{D(N-1)+\sigma^2}$$

Haciendo la varianza poblacional igual a $\sigma^2 = PQ$ y sustituyéndola se tiene:

$$n = \frac{NPQ}{\frac{e^2}{t^2{}_k}(N-1)+PQ}$$

$$e = t_k\sqrt{\hat{V}(p)}$$

$$\hat{V}(\hat{p}) = \frac{\hat{p}\hat{q}}{n-1}\left(\frac{N-n}{N}\right)$$

$$n = \frac{n_0}{1+\frac{n_0}{N}}; \qquad n_0 = \frac{t^2 pq}{e^2}$$

Ejemplo 4.6: Use los datos del Ejercicio 4.5 para determinar el tamaño de muestra requerido para estimar p con un límite para el error de estimación de magnitud B = 0.05. (Mendenhall, Pag 68.)

Solución:

$\sum y_i = 25 \qquad N = 300 \qquad n = ? \quad B = 0.05 = e \qquad t = 2$

$\hat{p} = \frac{\Sigma y_i}{30} = \frac{25}{30} = 0.8333$ y $\hat{q} = 1 - \hat{p} = 0.1667$

$$n = \frac{NPQ}{(N-1)\left(\frac{e^2}{t^2}\right) + PQ^2} = \frac{300(0.83333)(0.1667)}{(300-1)\left(\frac{0.05^2}{4}\right) + (0.83)(0.17)} = 127.90 \cong 128$$

MUESTREO CON PROBABILIDADES PROPORCIONALES AL TAMAÑO

Sea π_i = la probabilidad de que y_i aparezca en la muestra.

El Estimador del total T es: $\hat{T}_{pp} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i}{\pi_i}\right)$

Varianza estimada del $\hat{T}_{pp}$ es: $\hat{V}\left(\hat{T}_{pp}\right) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{\pi_i} - \hat{T}_{pp}\right)^2$

El limite para el error de estimación es: $t_k\sqrt{\hat{V}(\hat{T}_{pp})}$

Estimador de la media poblacional μ :

$$\hat{\mu}_{pp} = \frac{1}{Nn}\hat{T}_{pp} = \frac{1}{Nn}\sum_{i=1}^{n}\left(\frac{y_i}{\pi_i}\right)$$

Varianza estimada de $\hat{\mu}_{pp}$ es: $\hat{V}\left(\hat{\mu}_{pp}\right) = \frac{1}{N^2 n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{\pi_i} - T_{pp}\right)^2$

El limite para el error de estimación es $t_k\sqrt{\hat{V}\left(\hat{\mu}_{pp}\right)}$

VENTAJAS DEL MUESTREO ALEATORIO SIMPLE

1) Las ventajas derivadas de realizar muestreo.
2) Es relativamente simple determinar la precisión de las estimaciones que se hacen a partir de las observaciones muestrales.
3) Tiende a reflejar todas las características del universo, esto es, cuando el tamaño de la muestra crece, ésta se hace cada vez más representativa del universo o población.

DESVENTAJAS DEL MUESTREO ALEATORIO SIMPLE

1) Suponemos un listado completo.
2) Si la población es muy grande la numeración demanda tiempo y trabajos que pueden ser ahorrados si se emplea otro diseño muestral.
3) El tamaño de n estratificado es mejor que el tamaño de n aleatorio para el mismo nivel de confiabilidad.
4) Costos mayores con la dispersión espacial de las unidades muestreadas.

TEMA 3. MUESTREO ESTRATIFICADO

En el muestreo aleatorio simple la varianza del estimador depende del tamaño de la muestra y de la dispersión de la variable en estudio. Si la población es muy heterogénea y las consideraciones de costos limitan el tamaño de la muestra, podría ser imposible obtener una estimación lo suficientemente precisa tomando una muestra aleatoria simple. Es decir, el tamaño de la muestra aumenta para una precisión dada. Pero, si podemos clasificar los elementos de la población en grupos (estratos)de manera que se reduzca la variación de la variable Y dentro de cada estrato, por tanto, puede hacerse una mejor estimación.

Ejemplo: Cargos vacantes en las empresas.
Criterio de estratificación: tamaño de la empresa.

DEFINICIÓN: Una muestra aleatoria estratificada es la obtenida mediante la división de la población en subpoblaciones denominadas estratos, en la cual, dentro de cada estrato se selecciona en forma independiente una muestra irrestricta aleatoria. Calculándose para cada estrato sus estimadores y el estimador de la población se calcula como una ponderación adecuada de las estimaciones por estrato.

RAZONES PARA ESTRATIFICAR

1) Aumentar la precisión de las estimaciones al disminuir la variación dentro de los estratos. La estratificación puede producir un limite más pequeño para el error de estimación que el que se produciría con un muestreo aleatorio simple.

2) Disminuir los costos al estratificar y variar las fracciones de muestreo dentro de los estratos.
3) Permitir definir los estratos como dominios de estudio y obtener estimaciones con precisión conocida para los estratos.

¿CÓMO SELECCIONAR UNA MUESTRA ALEATORIA ESTRATIFICADA?

Dividir la población en estratos de acuerdo a las razones para estratificar, ubicar cada unidad muestral en su respectivo estrato, asignar el tamaño muestral de cada estrato n_i (de modo que si los L estratos y n es el tamaño de la muestra $\sum_{i=1}^{L} n_i = n$ y seleccionar muestras aleatorias simples en cada estrato de forma independiente.

La estratificación se realiza de acuerdo a la distribución de la variable en estudio o de acuerdo a una variable X altamente correlacionada con la variable en estudio o de acuerdo a un criterio de disminución de los costos.

En general, la precisión aumenta con el número de estratos si estos están bien elegidos, pero no es conveniente aumentar mucho el número de estratos si tal aumento no compensa las complicaciones de cálculo y la disminución del tamaño de la muestra dentro de los estratos.

NOTACIÓN

N = tamaño de la población.
L = número de estratos.
N_i = tamaño del i-ésimo estrato $i = 1,2,...,L$
N = tamaño de la muestra.

$$\sum_{i=}^{L} N_i = N \qquad \sum_{i=1}^{L} n_i = n$$

$W_i = N_i/N$ tamaño proporcional del estrato i $\sum W_i = 1$
$W_i = n_i/n$ proporción de la muestra en el estrato i $\sum w_i = 1$

ESTIMACIÓN DE LA MEDIA

Para estimar la media poblacional μ el estimador es: $\bar{y}_{st} = \frac{1}{N}\sum_{i=}^{L} N_i \bar{y}_i$

Sea $\bar{y}_i = \sum_{j=1}^{n_i} \frac{y_{ij}}{n_i}$ la media muestral del *i-ésimo* estrato

La Varianza poblacional de $\bar{y}_{st}$ es:

$$V(\bar{y}_{st}) = V\left[\frac{1}{N}\sum_{i=1}^{L} N_i y_i\right] = \frac{1}{N^2}\left[\sum_{i=1}^{L} N_i{}^2 V(\bar{y}_i)\right] = \frac{1}{N^2}\left[\sum_{i=1}^{L} N_i{}^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{S_i{}^2}{n_i}\right)\right]$$

La Varianza estimada de $\bar{y}_{st}$ es:

$$\hat{V}(\bar{y}_{st}) = V\left[\frac{1}{N}\sum_{i=1}^{L} N_i y_i\right] = \frac{1}{N^2}\left[\sum_{i=1}^{L} N_i{}^2 V(\bar{y}_i)\right] = \frac{1}{N^2}\left[\sum_{i=1}^{L} N_i{}^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{\hat{S}_i{}^2}{n_i}\right)\right]$$

Vamos a demostrar que la media es un estimador insesgado, es decir, $E(\bar{y}_{st}) = \mu$

El estimador $\bar{y}_{st}$ es un estimador insesgado puesto que los $\bar{y}_i$ son insesgados.

$$\mathrm{E}[\bar{y}_{st}] = \frac{1}{N}\sum_{i=1}^{L} N_i \mathrm{E}(\bar{y}_i) = \frac{1}{N}\sum N_i \bar{Y}_i = \frac{1}{N}\sum T_i = \frac{T}{N} = \bar{Y} = \mu$$

Tarea: Demostrar que la varianza estimada de $\bar{y}_{st}$ es un estimador insesgado de $V(\bar{y}_{st})$

Si las fracciones de muestreo n_i/N_i son despreciables $\to \infty$ en todos los estratos entonces

$$\hat{V}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{L} \frac{N_i{}^2 \hat{S}_i{}^2}{n_i}$$

Ejemplo 5.4: Se forma una comisión de Zonificación para estimar el valor promedio de avalúo en un suburbio residencial de una ciudad. El uso de ambos distritos de votantes en el suburbio como los estratos es conveniente porque se tienen disponibles listas separadas de las viviendas en cada distritos. De los datos presentados en la tabla acompañante, estime el valor promedio de avalúo para todas las casas en el suburbio, y establezca un límite para el error de estimación (nótese que se utilizó la asignación proporcional). (Mendenhall, pag. 113-114.)

ESTRATO I	ESTRATO II
$N_1 = 110$	$N_2 = 168$
$n_1 = 20$	$n_2 = 30$
$\sum_{i=1}^{n_1} y_i = 240.000$	$\sum_{i=1}^{n_2} y_i = 420.000$
$\sum_{i=1}^{n_1} y_i^2 = 2.980.000.000$	$\sum_{i=1}^{n_2} y_i^2 = 6.010.000.000$

Se pide determinar la media y el error de estimación, es decir, $\bar{y}_{st} = ?$ y $B = ?$

$$\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i \bar{y}_i$$

Sabemos que: $\bar{y}_i = \dfrac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$

$$\hat{S}_i^2 = \frac{\sum_{j=1}^{n_i} y_{ij^2} - n\bar{y}_i^2}{n_i - 1} = \frac{\sum y_{ij}^2 - \dfrac{\left(\sum y_{ij}\right)^2}{n_i}}{n_i - 1}$$

$$\bar{y}_1 = \frac{240.000}{20} = 12.000$$

$$S_1^2 = \frac{2.980.000.000 - 20 \times (12.000)^2}{19} = 5.263.157{,}895$$

$$\bar{y}_2 = \frac{420.000}{30} = 14.000$$

$$S_2^2 = \frac{6.010.000.000 - 30 \times (14.000)^2}{29} = 4.482.758{,}62$$

$\bar{y}_{st} = \dfrac{1}{278}\left[110 \times 12.000 + 168 \times 14000\right] = 13.208{,}63$ es el valor promedio de valúo para todas las casas del suburbio.

La varianza estimada es: $\hat{V}(\bar{y}_{st}) = \dfrac{1}{N^2}\left[\sum_{i=1}^{L} N_i{}^2 \left(\dfrac{N_i - n_i}{N_i}\right)\left(\dfrac{\hat{S}_i{}^2}{n_i}\right)\right]$ al sustituir los valores respectivos tenemos:

$$\hat{V}(\bar{y}) = \frac{1}{(278)^2}\left[(110)^2\left(\frac{110-20}{110}\right)\left(\frac{526157.895}{20}\right) + (168)^2\left(\frac{168-30}{168}\right)\left(\frac{4482758.62}{30}\right)\right]$$

$$\hat{V}(\bar{y}) = 7853.52$$

El error de estimación es: $B = t_k\sqrt{\hat{V}(\bar{y})} = 2\sqrt{7853.52} = 560.48$

Los límites de confianza son: $\bar{y} \pm B$ luego, en este ejemplo,

$\bar{y} \pm B$ (13208.63-560.48; 13208.63+560.48)

El intervalo resultante es: (12648.15; 13769.11). Es decir que se estima que con por lo menos un 75% de confianza el valor promedio de avalúo para todas las casas en el suburbio oscile entre 12648.15 y 13769.11$.

Como en este tipo de muestreo, las muestras en cada estrato son independientes, entonces se puede realizar estimaciones separadas, así:

Estrato 1		Estrato 2
$\bar{y}_1 \pm t_k \sqrt{\left(\frac{N_1 - n_1}{N_1}\right)\frac{\hat{s}_1^2}{n_1}}$		$\bar{y}_2 \pm t_k \sqrt{\left(\frac{N_2 - n_2}{N_2}\right)\frac{\hat{s}_2^2}{n_2}}$
12000 ± 928.03		14000 ± 700.69
(11071.97, 2928.03)		(13299.31, 14700.69)

ESTIMACIÓN DEL TOTAL

El estimador del total es: $\hat{T}_{st} = N\bar{y}_{st} = N\sum_i^L W_i \bar{y}_i = N\frac{1}{N}\sum_i^L N_i \bar{y}_i = \sum_i^L N_i \bar{y}_i$

La Varianza poblacional de $\hat{T}_{st}$:

$$V\left(\hat{T}_{st}\right) = \hat{V}\left(N\bar{y}_{st}\right) = N^2\hat{V}\left(\bar{y}_{st}\right) = \sum_{i=}^{L} N_i^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{S_i^2}{n_i}\right)$$

La Varianza estimada de $\hat{T}_{st}$:

$$\hat{V}\left(\hat{T}_{st}\right) = \hat{V}\left(N\bar{y}_{st}\right) = N^2\hat{V}\left(\bar{y}_{st}\right) = \sum_{i=}^{L} N_i^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{\hat{S}_i^2}{n_i}\right)$$

Ejemplo 5.3: Para el Ejercicio 5.2 estime el número total de horas-hombre perdidas durante el mes indicado y establezca un límite para el error de estimación. Use los datos de la tabla acompañante, obtenida en una muestra de 18 obreros, 10 técnicos y 2 administrativos. (Mendenhall, pag. 113.)

I (Obreros)			II (Técnicos)		III (Administrativos)
8	24	0	4	5	1
0	16	32	0	24	8
7	4	4	8	12	
9	5	8	3	2	
18	2	0	1	8	

Solución:

Se desea estimar el número total de horas - hombre perdidas.

I	II	III
Obreros	Técnicos	Administrativos
$n_1 = 18$	$n_2 = 10$	$n_3 = 2$
$\bar{y}_1 = 8{,}8333$	$\bar{y}_2 = 6{,}7$	$\bar{y}_3 = 4{,}5$
$S_1^2 = 81{,}5588$	$S_2^2 = 50{,}4556$	$S_3^2 = 24{,}5$
$N_1 = 132$	$N_1 = 92$	$N_3 = 27$

$$\hat{T}_{st} = \sum_{i=1}^{L} N_i \bar{y}_i = 132 \times 8{,}8333 + 92 \times 6{,}7 + 27 \times 4{,}5$$

$\hat{T}_{st} = 1903{,}8956 = 1903{,}9$ Número total de horas – hombres perdidas por accidente en un mes determinado.

$$\hat{V}\left(\hat{T}_{st}\right) = \sum_{i=1}^{L} N_i^{\,2} \left(\frac{N_i - n_i}{N_i}\right) \frac{S_i^2}{n_i}$$

$$= (132)^2 \left(\frac{132-18}{132}\right)\left(\frac{81{,}5588}{18}\right) + (92\ )^2 \left(\frac{92-10}{92}\right)\left(\frac{50{,}4556}{10}\right) + (27)^2 \left(\frac{27-2}{27}\right)\left(\frac{24{,}5}{2}\right)$$

$$\hat{V}\left(\hat{T}_{st}\right) = 114.515{,}61$$

$$B = e = t_k \sqrt{\hat{V}(T_{st})} \Rightarrow e = 2\sqrt{114.515{,}61} = 2 \times 338{,}402 = 676{,}803 \cong 676{,}8$$

Intervalo de confianza del Total $\hat{T}_{st}$

$\hat{T}_{st} \pm B$

$(1903{,}9 \pm 676{,}8)$

El verdadero número total de horas perdidas por enfermedad está en el intervalo $(1227{,}1\ ;\ 2580{,}7)$

La estimación separada del total para el estrato 1 es: $\hat{T}_1 \pm t_k \sqrt{N_1^{\,2}\left(\frac{S_1^2}{n_1}\right)\left(\frac{N_1 - n_1}{N_1}\right)}$

$(1165{,}996 \pm 2\sqrt{68.183{,}157})$

$(1165{,}996 \pm 552{,}24)$

$(643{,}76;1688{,}23)$

El límite de error 552,24 es muy grande porque S_1^2 es grande y por tanto se obtiene una estimación deficiente.

Si se desea una estimación para un estrato particular, la muestra del estrato debe ser lo suficientemente grande para proporcionar un límite razonable para el error de estimación.

SELECCIÓN DEL TAMAÑO DE LA MUESTRA PARA ESTIMAR μ

Prefijados el error máximo admisible (precisión mínima del estimador) indicado por, $e = t_k \sqrt{\hat{V}(\bar{y})}$. El coeficiente de confianza 1-α determina el valor de t_α (acorde a la forma de distribución del estimador) y la variabilidad de la población (paradoja de Friedman).

Si $1\text{-}\alpha \uparrow \Rightarrow t_\alpha \uparrow$

$\sigma_{\bar{x}} \downarrow si \quad n \uparrow$

En este diseño supones conocidos: $N, N_1, N_2, \ldots, N_l, n$, $w_i = n_i/n$

$$e = t_\alpha \sqrt{\hat{V}(\bar{y}_{st})} = t_\alpha \frac{1}{N^2}\left(\sum_{i=1}^{L} N_i^{\,2}\left(\frac{N_i - n_i}{N_i}\right)\frac{S_i^2}{n_i}\right)^{\frac{1}{2}}$$

Para determinar el tamaño de muestra se fija el nivel del error de estimación que se está dispuesto a cometer. También, se supone que $w_i = n_i/n$ para poder despejar $n \Rightarrow$ haciendo $n_i = w_i n$ y se sustituye:

$$e^2 = t_\alpha^{\,2} \frac{1}{N^2} \sum_{i=}^{L} N_i^2 \left(\frac{N_i - w_i n}{N_i}\right)\frac{S_i^2}{w_i n} \Rightarrow \frac{N^2 e^2}{t_\alpha^2} = \sum_{i=1}^{L} \frac{N_i^2 S_i^2}{w_i n} - \sum_{i=1}^{L} N_i S_i^2$$

$$\frac{N^2e^2}{t_\alpha^2}+\sum_{i=}^{L}N_iS_i^2=\frac{1}{n}\sum_{i=}^{L}\frac{N_i^2S_i}{w_i}^2$$

El tamaño de la muestra aproximado para estimar μ es:

$$n=\frac{\sum_{i=1}^{L}\frac{N_i^2S_i^2}{w_i}}{\frac{N^2e^2}{t_\alpha^2}+\sum_{i=1}^{L}N_iS_i^2}$$

Como $W_i = N_i/N$ y si dividimos ambos miembros por N^2 tenemos:

$$n=\frac{\frac{1}{N^2}\sum_{i=1}^{L}\frac{N_i^2S_i^2}{w_i}}{\frac{N^2e^2}{N^2t_\alpha^2}+\frac{1}{N^2}\sum_{i=1}^{L}N_iS_i^2}=\frac{\sum_{i=1}^{L}W_i^2\frac{S_i^2}{w_i}}{\frac{e^2}{t_\alpha^2}+\frac{1}{N^2}\sum_{i=1}^{L}N_iS_i^2}=\frac{\sum_{i=1}^{L}W_i^2\frac{S_i^2}{w_i}}{\frac{e^2}{t_\alpha^2}+\frac{1}{N}\sum W_i\ S_i^2}$$

es el tamaño de muestra aproximado para estimar la media.

El tamaño de muestra para una población que tiende a infinito es:

$$n_0=\left(\sum_{i=}^{L}W_i^2\frac{S_i^2}{w_i}\right)\frac{t_\alpha^2}{e^2}$$

$V = (e/t)^2$ es una varianza especificada en función del margen de error, también se denomina varianza anticipada.

$$n=\frac{\sum_{i=1}^{L}w_i^2\frac{S_i^2}{w_i}}{V+\frac{1}{N}\sum_{i=1}^{L}N_iS_i^2}$$

$$n_0=\left(\sum_{i=}^{L}W_i^2\frac{S_i^2}{w_i}\right)\frac{t_\alpha^2}{e^2}=\frac{1}{V}\left(\sum_{i=1}^{L}W_i^2\frac{S_i^2}{w_i}\right)$$

luego, el tamaño de muestra es:

$$n=\frac{n_0}{1+\frac{1}{NV}\sum_{i=1}^{L}W_iS_i^2}$$

TAMAÑO DE LA MUESTRA PARA EL TOTAL
(TAMAÑO APROXIMADO)

Este tamaño de muestra se obtiene de igual forma, partiendo del error de estimación para estimar el total:

$$e^2 = T_\alpha^2 \hat{V}\left(\hat{T}\right) = t_\alpha^2 \sum_{i=1}^{L} N_i^2 \left(\frac{N_i - n_i}{N_i}\right)\frac{S_i^2}{n_i} \qquad \text{si } n_i = nw_i$$

$$\frac{e^2}{t_\alpha^2} = \sum_{i=1}^{L} N_i \left(N_i - nw_i\right)\frac{S_i^2}{nw_i}$$

$$\frac{e^2}{t_\alpha^2} = \sum_{i=1}^{L} \frac{N_i^2 S_i^2 - N_i S_i^2 nw_i}{nw_i} = \sum_{i=1}^{L} \frac{N_i^2 S_i^2}{nw_i} - \sum_{i=}^{L} N_i S_i{}^2$$

$$\frac{e^2}{t_\alpha^2} + \sum_{i=1}^{L} N_i S_i^2 = \frac{1}{n}\sum_{i=1}^{L} \frac{N_i^2 S_i^2}{w_i}$$

Despejando *n* se obtiene el tamaño de muestra aproximado para estimar el total:

$$n = \frac{\sum_{i=1}^{L} \frac{N_i^2 S_i^2}{w_i}}{\frac{e^2}{t_\alpha^2} + \sum_{i=1}^{L} N_i S_i^2} = \frac{\sum_{i=1}^{L} \frac{N_i^2 S_i^2}{w_i}}{V + \sum_{i=1}^{L} N_i S_i^2}$$

Ejemplo: A continuación se realiza un ejemplo del cálculo del tamaño de muestra necesario para determinar la Calidad de la Leche (variable: acidez). Suponga que se realizó una muestra piloto y se obtuvo los siguientes datos:

n_i	*Estratos*	N_i = *fincas*	$W_i = N_i/N$	S_i^2	$W_i^2 S_i^2 / w_i$	$w_i = n_i/n$
10	Urdaneta	13	0,194	1,676	0,133	0,476
3	Valera	7	0,1045	2,333	0,178	0,143
1	Escuque	3	0,0445	0	0	0,048
6	Boconó	32	0,4776	1,14	0,909	0,286
1	Trujillo	12	0,1791	0	0	0,048
		N = 67			$\sum W_i^2 S_i^2 / w_i = 1{,}22$	

$$V = \frac{1}{n}\sum_{i=1}^{L}\frac{W_i^2 S_i^2}{w_i} - \frac{1}{n}\sum_{i=1}^{L} W_i S_i^2 = \frac{1}{21}(1{,}22) - \frac{1}{67}(1{,}113) = 0{,}041$$

$$n = \frac{\sum_{i=1}^{L}\frac{W_i^2 S_i^2}{w_i}}{V + \frac{1}{N}\sum_{i=1}^{L} W_i S_i^2} = \frac{1{,}22}{0{,}041 + \frac{1}{67}(1{,}113)} \cong 21 \text{ Fincas}$$

ASIGNACIÓN DE LA MUESTRA

Se denomina asignación o afijación al reparto o distribución del tamaño de la muestra n entre los diferentes estratos, es decir, la determinación de los L valores n_i de modo que $n_i + n_2 + ... + n_l = n$. Cada asignación puede originar una varianza diferente al estimador, nuestro objetivo es determinar un esquema de asignación que aumente la precisión y minimice los costos .

1) Los factores que influyen en la asignación son:
2) El número total de elementos en cada estrato.
3) La dispersión en cada estrato y.
4) El costo de observación en cada estrato.

TIPOS DE ASIGNACIÓN.

1. Igual $n_i = \frac{n}{L}$
2. Optima.
3. Proporcional.

ASIGNACIÓN OPTIMA: en el muestreo estratificado los valores de los tamaños de la muestra por estrato puede ser asignados con la finalidad de minimizar la variabilidad del estimador para un costo fijo o para minimizar el costo para un valor específico de la varianza de la media $\hat{V}(\bar{y}_{st})$.

La función de costo fijo más sencilla es $C = c_0 + \sum_{i=1}^{L} c_i n_i$. Dentro de cualquier estrato el costo es proporcional al tamaño de la muestra, pero el costo por cada unidad c_i puede variar entre los estratos.

Por tanto, C_0 representa un costo general y c_i el costo por unidad encuestada en el estrato *i.*

Sabemos que la varianza estimada de la media es:

$$\hat{V}(\bar{y}_{st}) = \left(\frac{1}{N^2}\right)\sum_{i=1}^{L} N_i^2 \frac{(N_i - n_i)}{N_i}\frac{S_i^2}{n_i} = \sum_{i=1}^{l}\frac{N_i^2 S_i^2}{N^2 n_i} - \sum_{i=1}^{L}\frac{N_i S_i^2}{N^2}$$

haciendo $N_i / N = W_i$ obtenemos:

$$\hat{V}(\bar{y}_{st}) = \sum\frac{W_i^2 S_i^2}{n_i} - \frac{1}{N}\sum W_i S_i^2$$

Ahora, vamos a minimizar la varianza $\hat{V}(\bar{y}_{st})$ sujeto a la restricción $c_1 n_1 + c_2 n_2 + ... + c_L n_L = C - C_0$.

$$C_0 + \sum_{i=1}^{L} c_1 n_i - C = 0$$

Usando el método de los multiplicadores de Lagrange debemos minimizar, la función:

$$\phi(n_i) = \sum_{i=1}^{L}\frac{W_i^2 S_i^2}{n_i} - \sum_{i=1}^{L}\frac{W_i^2 S_i^2}{N_i} + \lambda\left(\sum n_i c_i - c + c_0\right)$$

Diferenciando con respecto a n_i, en cada uno de L estratos, $i = 1,2,....,L$ e igualando a cero, las L ecuaciones obtenidas son:

$$-\frac{W_i^2 S_i^2}{n_i^2} + \lambda c_i = 0 \qquad \Rightarrow \quad \frac{W_i^2 S_i^2}{n_i^2} = \lambda c_i$$

Extrayendo la raíz cuadrada,

$$\frac{W_i S_i}{\sqrt{c_i}} = n_i\sqrt{\lambda} \quad (1)$$

Sumando (¡) sobre i se obtiene:

$$\sum_{i=1}^{L} n_i\sqrt{\lambda} = \sum\frac{W_i S_i}{\sqrt{c_i}} \qquad \Rightarrow \qquad n\sqrt{\lambda} = \sum\frac{W_i S_i}{\sqrt{c_i}} \quad (2)$$

haciendo el cociente de (1) y (2) para eliminar λ

$$\frac{n_i\sqrt{\lambda}}{n\sqrt{\lambda}} = \sum\frac{W_i S_i/\sqrt{c_i}}{\sum_{i=1}^{L}\frac{W_i S_i}{\sqrt{c_i}}} \qquad \Rightarrow \quad \frac{n_i}{n} = \frac{W_i S_i/\sqrt{c_i}}{\sum_{i=1}^{L}\frac{W_i S_i}{\sqrt{c_i}}}$$

i es un valor específico del estrato.

Como $w_i = n_i / n \Rightarrow \dfrac{n_i}{n} = \dfrac{\frac{N_i}{N} S_i / \sqrt{c_i}}{\frac{1}{N} \sum_{i=1}^{L} \frac{N_i S_i}{\sqrt{c_i}}}$

Entonces $\dfrac{n_i}{n} = \dfrac{N_i S_i / \sqrt{c_i}}{\sum_{i=1}^{L} \frac{N_i S_i}{\sqrt{c_i}}}$ luego $n_i = n \dfrac{N_i S_i / \sqrt{c_i}}{\sum_{i=1}^{L} \frac{N_i S_i}{\sqrt{c_i}}}$

Este resultado nos indica que en un estrato dado se debe tomar una muestra grande si:

- El estrato es grande ($N_i \uparrow$).
- El estrato es más variable internamente.
- El muestreo es mas barato en el estrato.

1.1 TAMAÑO DE LA MUESTRA PARA LA ASIGNACIÓN OPTIMA.

El tamaño de muestra según asignación o afijación optima, a su vez depende de:

a) Si la muestra es escogida para satisfacer un costo total *C* especifico, o
b) Para dar una varianza de $\left(\bar{y}_{st}\right)$ especifica.

a) En el primer caso, Si el costo es fijo, entonces, se minimiza $V\left(\bar{y}_{st}\right)$. Quiere decir que en la función de costos sustituimos el valor de n_i .

$$C = c_0 + \sum_{i=}^{L} c_i n_i \quad \Rightarrow \quad C - c_0 = \sum_{i=1}^{L} c_i n \left(\frac{N_i S_i / \sqrt{c_i}}{\sum_{i=1}^{l} N_i S_i / \sqrt{c_i}} \right)$$

$$(C - c_0) \sum_{i=1}^{L} N_i S_i / \sqrt{c_i} = \sum_{i=1}^{L} c_i n \left(N_i S_i / \sqrt{c_i} \right)$$

$$(C - c_0) \sum_{i=1}^{L} N_i S_i / \sqrt{c_i} = n \sum_{i=1}^{L} N_i S_i \sqrt{c_i}$$

Despejando n se tiene: $n = \dfrac{(C - c_0)\sum_{i=1}^{L} N_i S_i / \sqrt{c_i}}{\sum_{i=1}^{L} \left(N_i S_i \sqrt{c_i}\right)}$

b) Si la varianza se fija con anticipación, al sustituir $w_i = n_i / n$ en la fórmula de la varianza de la media $V(\bar{y}_{st}) = V = \dfrac{1}{N}\sum \dfrac{w_i^2 s_i^2}{w_i} - \sum \dfrac{w_i^2 s_i^2}{N_i}$ o en la fórmula de tamaño aproximado de la muestra, tenemos:

$$n = \frac{\sum_{i=1}^{L} W_i^2 S_i / w_i}{V + \frac{1}{N}\sum_{i=1}^{L} W_i S}$$ donde: $V = e^2/t^2$ es la varianza anticipada

Sustituyendo

$$w_i = n_i / n = \frac{W_i S_i / \sqrt{c_i}}{\sum_{i=1}^{L} \left(W_i S_i / \sqrt{c_i}\right)} = \frac{N_i S_i / \sqrt{c_i}}{\sum_{i=1}^{L} \left(N_i S_i / \sqrt{c_i}\right)}$$ nos queda

$$n = \frac{\left(\sum W_i S_i \sqrt{c_i}\right)\left(\sum W_i S_i \sqrt{c_i}\right)}{V + \frac{1}{N}\sum_{i=1}^{L} W_i S_i^2}$$

para expresarla en términos de N_i, se sustituye $W_i = N_i / N$

$$n = \frac{\frac{1}{N^2}\left(\sum N_i S_i \sqrt{c_i}\right)\left(\sum N_i S_i / \sqrt{c_i}\right)}{V + \frac{1}{N^2}\sum_{i=1}^{L} N_i S_i^2} = \frac{\left(\sum N_i S_i \sqrt{c_i}\right)\left(\sum N_i S_i / \sqrt{c_i}\right)}{N^2 V + \sum_{i=1}^{L} N_i S_i^2}$$

1.2 TAMAÑO DE LA MUESTRA, CASO DE COSTOS IGUALES POR ESTRATO (ASIGNACIÓN DE NEYMANN)

En algunos problemas el costo para obtener información en cada uno de loa estratos es el mismo, así $C_1 = C_2 = ...= C_L = C$. Si los costos son conocidos se puede suponer que los costos son iguales.

$n_i = \frac{nN_iS_i}{\sum N_iS_i} \quad \Rightarrow \qquad n = \frac{W_iS_i}{\sum W_iS_i}$ este tipo de asignación se conoce como asignación de Neyman (asignación óptima supuesta).

En este caso $n = \frac{\left(\sum N_iS_i\right)^2}{N^2V + \sum N_iS_i^2}$ $\qquad V = \frac{e^2}{t^2}$

$$n = \frac{\left(\sum W_iS_i\right)^2}{V + \frac{1}{N}\sum_{i=1}^{L} W_i\ S_i^2}$$

1.3 TAMAÑO DE MUESTRA PARA COSTOS IGUALES, VARIANZAS IGUALES, ASIGNACIÓN PROPORCIONAL.

Este método de asignación de la muestra se denomina asignación proporcional porque los tamaños de la muestra n_1 ,n_2,.....,n_L, se distribuyen de acuerdo al peso del estrato en la población, por tanto:

$$\bar{Y}_{st} = \frac{1}{N}\sum N_i\bar{y}_i = \sum W_i\bar{y}_i$$

$$W_i = \frac{N_i}{N} = \frac{n_i}{n}$$

Sea la definición de varianza:

$$V(\bar{y}_{st}) = \sum W_i^2 V(\bar{y}_i) = \sum W_i^2\left(\frac{N_i - n_i}{N_i}\right)\frac{S_i^2}{n_i}$$

Al sustituir en términos de los pesos de los estratos poblacionales,

$$= \sum W_i^2\ \frac{(NW_i - nW_i)S_i^2}{NW_in_i} = \sum W_i^2\left(\frac{N-n}{N}\right)\frac{S_i^2}{nW_i}$$

$$V(\bar{y}_{st}) = \frac{(N-n)}{Nn}\sum W_i^2\ \frac{S_i^2}{W_i} = \frac{(N-n)}{Nn}\sum W_i\ S_i^2$$

Para determinar el tamaño de muestra, hacemos $\frac{n_i}{n} = \frac{N_i}{N} = w_i = W_i$ y luego al sustituir en

$$n_i = n\left(\frac{N_i S_i / \sqrt{c_i}}{\sum_{i=1}^{L} N_i S_i / \sqrt{c_i}}\right) \qquad S_1 = S_2 = \ldots\ldots = S_L \qquad \text{y} \quad c_1 = c_2 = \ldots = c_L$$

Se tiene $n_i = n\dfrac{N_i}{\sum N_i}$ y el valor de $\sum N_i = N$

Empleando la formula de tamaño aproximado para estimar la media:

$$n = \frac{\sum N_i^2 S_i / w_i}{N^2 V + \sum_{i=1}^{L} N_i S_i^2} = \frac{\sum N_i^2 S_i / W_i}{N^2 V + \sum_{i=1}^{L} N_i S_i^2} = \frac{\sum N_i S_i^2}{NV + \frac{1}{N}\sum_{i=1}^{L} N_i S_i^2}$$

si dividimos numerador y denominador por *NV*.

$$n = \frac{\sum N_i S_i^2}{NV + \frac{1}{N}\sum_{i=1}^{L} N_i S_i^2} = \frac{\frac{\sum N_i S_i^2}{NV}}{1 + \frac{1}{N}\frac{\sum_{i=1}^{L} N_i S_i^2}{NV}} = \frac{\frac{1}{V}\sum W_i S_i^2}{1 + \frac{1}{NV}\sum_{i=1}^{L} W_i S_i^2}$$

haciendo $n_0 = \dfrac{1}{V}\sum W_i S_i^2$ queda $n = \dfrac{n_0}{1 + \dfrac{n_0}{N}}$

Esta asignación puede utilizarse también cuando los costos y las varianzas no son iguales (pero no son tomados en cuenta al momento de fijar los tamaños de la muestra), una ventaja al usar esta descomposición es que $\overline{y}_S = \overline{y}$.

Comparación de la precisión del muestreo aleatorio estratificado con relación al muestreo irrestricto aleatorio.

Si se usa inteligentemente la estratificación, es decir, si es el modelo de muestreo adecuado, entonces, da como resultado una varianza más pequeña para el estimador que la obtenida mediante muestreo aleatorio simple. Sin embargo, no es verdad que el muestreo estratificado dé siempre una varianza menor que en el muestreo aleatorio simple.

Para obtener la varianza S^2 en el muestreo irrestricto aleatorio (m.i.a.), asumamos que tenemos una población estratificada y por tanto, la variación total se divide en dos fuentes de variación: entre y dentro de los estratos.

$$\sum_{i=1}^{L}\sum_{i=1}^{L}\left(Y_{ij}-\overline{Y}\right)^2=\sum_{i=1}^{L}\sum_{j=1}^{N_i}\left(Y_{ij}-\overline{Y}_i\right)^2+\sum_{i=1}^{L}N_i\left(\overline{Y}_i-\overline{Y}\right)^2 \qquad \text{asumamos que } \overline{Y}_S=\overline{Y}$$

$$(N-1)S^2=\sum_{i=1}^{L}(N_i-1)S_i^{\,2}+\sum_{i=1}^{L}N_i\left(\overline{Y}_i-\overline{Y}\right)^2$$

$$S^2=\frac{\sum_{i=1}^{L}(N_i-1)S_i^{\,2}+\sum_{i=1}^{L}N_i\left(\overline{Y}_i-\overline{Y}\right)^2}{(N-1)}$$

Sabemos que en el muestreo aleatorio simple sin reposición la varianza de la media es:

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

Ahora, en el muestreo estratificado la varianza de la media es:

$$\hat{V}(\overline{y}_{st})=\frac{1}{N^2}\left[\sum_{i=1}^{L}N_i^{\,2}\left(\frac{N_i-n_i}{N_i}\right)\left(\frac{\hat{S}_i^{\,2}}{n_i}\right)\right]$$

Si en esta definición realizamos las sustituciones de acuerdo al tipo de afijación de la muestra, obtendremos la varianza de la media según ese tipo de asignación o afijación. Es el caso de la afijación proporcional, al sustituir los pesos de los estratos, se obtiene la varianza proporcional, así:

$$W_i=\frac{N_i}{N}=\frac{n_i}{n}=w_i$$

$$V_{prop}=V(\overline{Y}_{st})=\frac{1}{n}\left(\sum W_iS_i^2\right)=\frac{(1-f)}{Nn}\sum N_iS_i^2$$

La varianza de la media según la asignación óptima (Neymann) es:

$$V_{opt}=V(\overline{Y}_{st})=\frac{1}{n}\left(\sum W_iS_i\right)^2-\frac{1}{n}\left(\sum W_iS_i^2\right)=\frac{1}{nN^2}\left(\sum W_iS_i\right)^2-\frac{1}{N^2}\sum W_iS_i^2$$

Expresada en términos de los tamaño de los estratos, es:

$$V_{opt} = V(\bar{Y}_{st}) = \frac{1}{nN^2}\left(\sum N_i S_i\right)^2 - \frac{1}{N^2}\sum N_i S_i^2$$

Teorema (Cochran): $V_{opt} \le V_{prop} \le V_{ran}$ (tarea demostrar)

Se puede medir la eficiencia del diseño de muestreo estratificado utilizando el siguiente cociente: V_{prop} / V_{ran} = mide el efecto del diseño. También se puede calcular con relación a la varianza óptima, es decir, V_{op} / V_{ran}. Si el resultado es menor que 1 indica que es eficiente, si es igual a 1 es preferible usar el muestreo aleatorio simple y si es mayor que 1 no es eficiente.

ESTIMACIÓN DE LA PROPORCIÓN p

Si queremos estimar la proporción de unidades de la población que posee una característica, la estratificación ideal es dividir la población en dos estratos, uno el de todas las unidades que poseen la característica y el otro, las que no lo poseen. Esto es en general imposible por ello trataremos de construir estratos que la proporción varíe tanto como sea posible de estrato a estrato. Sabemos que la proporción $\hat{p}$ es un caso particular de $\bar{y}$, para una variable dicotómica.

$$\hat{p}_S = \frac{1}{N}\left(N_1\hat{p}_i + N_2\hat{p}_2 + \ldots + N_L\hat{p}_L\right) = \frac{1}{N}\sum_{i=1}^{L} N_i\hat{p}_i$$

$$\hat{V}(\hat{p}_S) = \frac{1}{N^2}\sum_{i=1}^{L} N_i{}^2\hat{V}\left(\hat{p}_i\right) = \frac{1}{N^2}\sum_{i=1}^{L} N_i{}^2\left(\frac{N_i - n_i}{N_i}\right)\frac{\hat{p}_i\hat{q}_i}{(n_i - 1)}$$

TAMAÑO DE LA MUESTRA PARA LA PROPORCIÓN

Las fórmulas de calculo del tamaño de la muestra para la proporción son iguales a la de la media excepto en que $S_i^2 = \hat{p}_i\hat{q}_i$

$n = \dfrac{\sum N_i{}^2\hat{p}_i\hat{q}_i / w_i}{N^2\left(\dfrac{e^2}{t^2}\right) + \sum_{i=1}^{L} N_i\,\hat{p}_i\hat{q}_i}$ es el tamaño de muestra aproximado para estimar la proporción.

Asignación optima: que minimiza el costo para $\hat{V}(\hat{p}_S)$ dada o minimiza $\hat{V}(\hat{p}_S)$ para el costo dado.

$$n_i = \frac{nN_i \sqrt{\hat{p}_i\hat{q}_i / c_i}}{\sum_{i=1}^{L} N_i \sqrt{\hat{p}_i\hat{q}_i / c_i}}$$

El tamaño de la muestra para satisfacer un costo total C es:

$$n = \frac{(c - c_0)\sum N_i \sqrt{\hat{p}_i\hat{q}_i / c_i}}{\sum_{i=1}^{L} N_i \sqrt{\hat{p}_i\hat{q}_i / c_i}}$$

El tamaño de muestra óptimo que minimiza la varianza es:

$$n = \frac{\frac{1}{N^2}\left(\sum N_i \sqrt{\hat{p}_i\hat{q}_i c_i}\right)\left(\sum N_i \sqrt{\hat{p}_i\hat{q}_i / c_i}\right)}{V + \frac{1}{N^2}\sum_{i=1}^{L} N_i \hat{p}_i\hat{q}_i} = \frac{\left(\sum N_i \sqrt{\hat{p}_i\hat{q}_i c_i}\right)\left(\sum N_i \sqrt{\hat{p}_i\hat{q}_i / c_i}\right)}{N^2 V + \sum_{i=1}^{L} N_i \hat{p}_i\hat{q}_i}$$

donde $V = \frac{e^2}{t^2}$ es la varianza anticipada.

Si los costos son iguales, el tamaño de muestra n óptimo (Neymann) es:

$$n_i = \frac{nN_i \sqrt{\hat{p}_i\hat{q}_i}}{\sum_{i=1}^{L} N_i \sqrt{\hat{p}_i\hat{q}_i}} \text{ en este caso } \qquad n = \frac{\left(\sum_{i=1}^{L} N_i \sqrt{\hat{p}_i\hat{q}_i}\right)^2}{N^2\left(\frac{e^2}{t^2}\right) + \sum_{i=1}^{L} N_i \hat{p}_i\hat{q}_i}$$

Asignación proporcional: se presenta cuando existen costos iguales y también las varianzas de los estratos son iguales.

$$n_i = n\left(\frac{N_i}{N}\right) \qquad \text{y} \qquad n = \frac{\sum_{i}^{L} N_i \hat{p}_i\hat{q}_i}{N^2\left(\frac{e^2}{t^2}\right) + \frac{1}{N}\sum_{i=1}^{L} N_i \hat{p}_i\hat{q}_i}$$

TEMA 4. MUESTREO POR CONGLOMERADOS

Se caracteriza porque las unidades de muestreo contienen a dos o más unidades primarias (últimas). La población se subdivide en subpoblaciones y algunas de ellas, que denominaremos conglomerados, pero no todas serán incluidas en la muestra.

El muestreo por conglomerados es similar al muestreo aleatorio simple, pero se diferencian en que la unidad de muestreo es un conjunto de unidades primarias o elementales.

A diferencia con el muestreo estratificado, donde la población también se subdivide en subpoblaciones, pero siempre todos los estratos están representados en la muestra. Mientras que el muestreo estratificado es diseñado y utilizado fundamentalmente con el objeto de reducir la varianza de los estimadores, el muestreo por conglomerados es utilizado debido a que muestrear directamente sobre las unidades primarias, el costo es exageradamente alto.

Este muestreo es, en muchos casos, un muestreo efectivo para obtener la información deseada a un menor costo, aunque el uso de los conglomerados conlleve en algunos casos a una varianza mayor de los estimadores.

Los casos en los cuales se justifica la aplicación de este diseño muestral son:

1) Donde existe un alto costo por la movilización o traslado entre las unidades primarias; el muestreo por conglomerado permite disminuir las distancias; pues por lo general, los conglomerados son áreas físicas o geográficas, donde las unidades primarias están contiguas.
2) Cuando no existe lista de las unidades primarias (o últimas) sobre los cuales hay que tomar las observaciones, y el costo de levantar un marco muestral de estas unidades es alto, en comparación con el costo de muestrear sobre conglomerados, los cuales si pueden disponer de un marco o directorio.
3) Para pequeñas unidades donde puede ser difícil fijar con precisión sus limites, sin embargo, puede ser posible y fácil, dividir con población en unidades mayores y luego muestrear y medir aquellas unidades mayores seleccionadas. Ejemplo: animales.
4) También, pueden existir consideraciones administrativos que jueguen papel importante en la colección del diseño a utilizar.

La diferencia de objetivos entre estratificación y conglomerados conduce a diferentes criterios para establecer los conglomerados o los estratos. En contraste, con el estratificado, la varianza del estimador se hace pequeña al hacer el conglomerado, tanto como sea posible, representativo de la diversidad de toda población, y todas los conglomerados deben ser en lo posible construidos de modo que sean lo más semejante entre sí. A diferencia del muestreo estratificado, donde los estratos deben ser homogéneos dentro de sí y heterogéneos entre sí.

¿CÓMO SELECCIONAR UNA MUESTRA POR CONGLOMERADOS?

- Definir el conglomerado tipo (tamaño del conglomerado). El número de elementos que integran un conglomerado se denomina tamaño. En la mayoría de los métodos por conglomerados, los conglomerados son de tamaños diferentes unas de otras, los conglomerados de igual tamaño, rara vez se logran en la practica, pero se constituyen una introducción sencilla al estudio del método por muestreo, y pueden resultar en situaciones practicas donde las condiciones fueran las indicadas, tales como: procesos de producción (control de calidad).

El problema de elegir un tamaño de conglomerado (m_i) apropiado puede ser un proceso un tanto complicado. El tamaño óptimo de los conglomerados no es una característica que depende exclusivamente de la población, sino también de la estructura de costos de la investigación. El tamaño del conglomerado óptimo es aquel para el cual la varianza del estimador es mínimo donde el costo de la investigación o el costo de la encuesta es mínimo dada la varianza. Así, por ejemplo, el tamaño del conglomerado se hace más pequeño cuando aumenta la duración de la entrevista, cuando el traslado entre las unidades primarias es barato, cuando la densidad del conglomerado es mayor y cuando el presupuesto del gasto aumenta.

- Formar el marco muestral, listando los conglomerados en los cuales se ha particionando la población. Resolviendo las imperfecciones que el marco pueda tener y garantizando que todos las unidades primarias que están en los conglomerados esta en uno y solo uno de los conglomerados.
- Seleccionar los conglomerados que van en la muestra utilizando un muestreo irrestricto aleatorio.

Si hacemos un muestreo o encuestamos todas las unidades de los conglomerados seleccionados, el muestreo se denomina muestreo por conglomerados **Monoetápico.**

Si en vez de entrevistar u observar a todos los individuos o unidades primarias del conglomerado observado en la muestra a su vez tomamos muestras de estas unidades primarias de los conglomerados seleccionados, el muestreo se denomina **Bietápico,** pues la muestra se selecciona en dos etapas.

Este proceso se puede generalizar a más de dos etapas y el muestreo se denomina **Polietápico.**

Notación:

N= números de conglomerados en la población
n= números de conglomerados en la muestra
m_i = números de unidades elementales (primarias) en el i-ésimo conglomerado

$M=\sum_{i=1}^{N} m_i$ Total de elementos en la población.

$\overline{M} = \frac{M}{N}$ Tamaño promedio del conglomerado en la población

y_i = total del conglomerado i-ésimo

$\overline{m} = \frac{\sum_{i}^{n} m_i}{n}$ Tamaño promedio del conglomerado en la muestra.

ESTIMACIÓN DE LA MEDIA POBLACIONAL

Por definición la media poblacional es: $\mu = \frac{\sum_{i=1}^{N} y_i}{\sum_{i}^{N} m_i}$

Luego, la estimación de la Media Poblacional es: $\hat{\mu} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}$

Este estimador de la media tiene la forma de un estimador de razón, por lo tanto, la varianza de la media tiene la forma de la varianza del estimador de razón, así:

$$\hat{V}(\bar{y}) = \left(\frac{N-n}{Nn\overline{M}}\right)\frac{\sum_{i}^{n}(y_i - \bar{y}m_i)^2}{n-1}$$

Si se desconoce el total de elementos en la población M, entonces, $\overline{M}$ puede ser estimado con $\overline{m} = \frac{\sum_{i}^{n} m_i}{n}$

El límite para el error de estimación es:

$$e = B = t_k\sqrt{\hat{V}(\bar{y})} = t_k\sqrt{\left(\frac{N-n}{Nn\overline{M}^2}\right)\frac{\sum(y_i - \bar{y}m_i)^2}{n-1}}$$

Los límites de confianza son: $\bar{y} \pm e$

En el muestreo por conglomerados Monoetápico distinguiremos dos casos:

1. Todos los conglomerados son de igual tamaño.
2. Todos los conglomerados son de tamaño diferentes

ESTIMACIÓN DEL TOTAL POBLACIONAL

El total poblacional τ puede ser determinado por $M\mu$ porque M denota el total de elementos en la población. Por lo tanto, así como en el muestreo aleatorio simple, el total puede ser estimado por:

$$\hat{\tau} = M\bar{y} = M \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}$$

La varianza estimada de $\hat{\tau} = M\bar{y}$:

$$\hat{V}(M\bar{y}) = M^2 \hat{V}(\bar{y}) = N^2 \left(\frac{N-n}{Nn} \right) \frac{\sum_{i=1}^{n} (y_i - \bar{y}m_i)^2}{n-1}$$

El límite para el error de estimación es:

$$e = B = t_k \sqrt{\hat{V}(\bar{y})} = t_k \sqrt{N^2 \left(\frac{N-n}{Nn} \right) \frac{\sum_{i}^{n} (y_i - \bar{y}m_i)^2}{n-1}}$$

Observe que este estimador $\hat{\tau} = M\bar{y}$ es útil solo cuando se conoce M el total de elementos de la población.

Sin embargo, a menudo ese número de elementos de la población no se conoce, por tanto se debe utilizar <u>otro tipo de estimador</u>, el cual no depende de M :

$$\hat{\tau} = N\bar{y}_t = \frac{N}{n} \sum_{i=1}^{n} y_i$$

donde:

N/n =factor de expansión

$\bar{y}_t = \frac{1}{n}\sum_{i=1}^{n} y_i$ es el promedio de totales de conglomerado para la muestra seleccionada.

La varianza estimada de $\hat{\tau} = N\bar{y}_t$:

$$\hat{V}(N\bar{y}_t) = N^2\hat{V}(\bar{y}_t) = N^2\left(\frac{N-n}{Nn}\right)\frac{\sum_{i=1}^{n}(y_i - \bar{y}_t)^2}{n-1}$$

El límite para el error de estimación es:

$$e = B = t_k\sqrt{\hat{V}(N\bar{y}_t)} = t_k\sqrt{N^2\left(\frac{N-n}{Nn}\right)\frac{\sum_{i}^{n}(y_i - \bar{y}_t)^2}{n-1}}$$

Este estimador $\hat{\tau}$ tiene a menudo el inconveniente de ser poco preciso, pues por lo general, las medias de los conglomerados varían poco y los m_i varían mucho. En este caso el total del conglomerado $y_i = m_i\bar{y}_i$, también varia mucho de unidad a unidad y entonces $V(\hat{\tau})$ es muy grande, sin embargo, este estimador es a veces utilizado, pues tiene la ventaja de que no es necesario conocer el tamaño de la población. $M = \sum_{i=1}^{N} m_i$

Los estimadores de μ y τ poseen propiedades especiales cuando todos los tamaños de los conglomerados son de igual tamaño, es decir, $m_1 = m_2 = ... = m_N = m$:

1. El estimador $\bar{y}$ es un estimador insesgado de μ.

2. La varianza estimada $\hat{V}(\bar{y}) = \left(\frac{N-n}{Nn\bar{M}^2}\right)\frac{\sum_{i}^{n}(y_i - \bar{y}m_i)^2}{n-1}$ es un estimador insesgado de la varianza poblacional $V(\bar{y}) = \left(\frac{N-n}{Nn\bar{M}^2}\right)\frac{\sum_{i}^{N}(y_i - \bar{y}m_i)^2}{n-1}$

3. Los estimadores del Total Poblacional $\hat{\tau} = M\bar{y}$ y $\hat{\tau} = N\bar{y}_t$ son equivalentes.

SELECCIÓN DEL TAMAÑO DE MUESTRA

1. Para estimar la Media Poblacional:

Por definición el error de estimación es:

$$e = B = t_k\sqrt{V(\bar{y})} = t_k\sqrt{\left(\frac{N-n}{Nn}\right)\sigma_c{}^2} = t_k\sqrt{V(\bar{y})}, \text{ donde}$$

$V(\bar{y}) = \left(\frac{N-n}{Nn\overline{M}^2}\right)\sigma_c^2$ es la varianza poblacional y $\hat{V}(\bar{y}) = \left(\frac{N-n}{Nn\overline{M}^2}\right)S_c^2$ es la varianza estimada.

Al despejar de la formula del error de estimación el valor de *n*, se tiene que el tamaño de muestra es:

$$n = \frac{N\sigma^2 c}{ND + \sigma^2 c}, \text{ donde}$$

$D = \left(\frac{e^2}{t^2\alpha}\right)\overline{M}^2$ es la varianza anticipada

2. Para estimar el Total Poblacional. En este caso tenemos dos tipos de estimadores:

a. $\hat{\tau} = M\bar{y}$

$$n = \frac{N\sigma^2 c}{ND + \sigma^2 c}, \text{ donde}$$

$$D = \left(\frac{e^2}{t^2{}_k N^2}\right)$$

b. $\hat{\tau} = N\bar{y}_t$

$$n = \frac{N\sigma_t{}^2}{ND + \sigma_t{}^2}, \text{ donde}$$

$$D = \left(\frac{e^2}{t^2{}_k N^2}\right)$$

σ_t^2, esta varianza es estimada por $S_t^2 = \dfrac{\sum_{i=1}^{n}\left(y_i - \bar{y}_t\right)^2}{n-1}$ que es la cuasivarianza de totales de conglomerados en la muestra.

ESTIMADOR DE LA PROPORCIÓN

La proporción es un parámetro muy frecuentemente estimado en las investigaciones por muestreo.

La proporción no es mas que la media de una población dicotómica. El estimador usual en la proporción es el número de éxitos en la muestra entre el total de la muestra.

Como estimador se utiliza $\hat{p} = \dfrac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} m_i}$ donde a_i es el total de "éxitos" en el i-ésimo conglomerado. Este estimador lo podemos considerar como un estimador de razón como el de la media, visto anteriormente.

Y así, su varianza es $V(\hat{p}) = \dfrac{N-n}{Nn\bar{M}^2}\sigma_c^2$, donde: $\sigma_c^2 = \dfrac{\sum_{i}^{N}\left(ai - \hat{p}m_i\right)^2}{N-1}$

Y el estimador de la varianza de la proporción es $\hat{V}(\hat{p}) = \dfrac{N-n}{Nn\bar{M}^2}S_c^2$

Donde:

$S_c^2 = \dfrac{\sum_{i}^{n}\left(a_i - \hat{p}m_i\right)^2}{n-1}$

SELECCIÓN DEL TAMAÑO DE MUESTRA PARA ESTIMAR LA PROPORCIÓN

Para obtener el tamaño de muestra para estimar la proporción se fija el error máximo admisible $e=B$ y el multiplicador de confianza t_k.

Por definición este error es:

$B = e = t_k\sqrt{\hat{V}(\hat{p})}$

Al elevarlo al cuadrado se tiene:

$$e^2 = t_k{}^2 \hat{V}(\hat{p}) \Rightarrow \left(\frac{e^2}{t_k^2}\right) = \hat{V}(\hat{p}) \Rightarrow \left(\frac{e^2}{t_k^2}\right) = \frac{N-n}{Nn\overline{M}}\sigma_c^2$$

Al despejar se encuentra el tamaño de muestra:

$$n = \frac{N\sigma_c^2}{\frac{e^2}{t^2}N\overline{M}^2 + \sigma_c^2} = \frac{N\sigma_c^2}{ND + \sigma_c^2}$$

D= es la varianza anticipada

La varianza σ_c^2 puede ser estimada por S_c^2 o proviene de:

(a) Muestras pilotos.
(b) Censos anteriores, y
(c) De otras estimaciones.

BIBLIOGRAFÍA

Cochran, W. (1980) **Muestreo**. Trillas.

Lohr, Sharon. (2000) **Muestreo: Diseño y Análisis**. International Thomson Editores, México.

Pérez, César. (2000) **Muestreo con aplicaciones informáticas**. Madrid.

Scheaffe, R., Mendenhall, W., y Ott, L. (1991) **Elementos de Muestreo**. Duxbury Press, Boston.