UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE ESTUDIOS SUPERIORES ACATLÁN

"Introducción al muestreo probabilístico sin reemplazo"

Tesina

QUE PARA OBTENER EL TITULO DE

Actuario

PRESENTA

AYALA PÉREZ CÉSAR ANTONIO

Asesor:Mtro. Suárez Madariaga Jorge Luis

OCTUBRE 2012

Agradecimientos

La elaboración de este trabajo fue un gran reto para mí, la dedicación y el esfuerzo que le dediqué significó dar un gran paso personal, el ser constante y saber que nadie estará ahí para presionar o exigir más que yo mismo me hizo crecer. Tener este trabajo concluido me hace anhelar el siguiente paso en mi vida profesional, seguir creciendo y aprendiendo como actuario es el mejor sabor de boca que me deja cerrar este ciclo.

Agradezco a mi familia y a mi madre en especial por creer en mí y apoyarme siempre, sin sus consejos y educación no hubiera podido llegar hasta donde estoy ni aspirar hasta donde quiero llegar.

Con este trabajo cierro el ciclo de la universidad donde pude aprender infinidad de cosas que me hicieron dar cuenta lo poco que aún sé. Una época donde llegué a conocer amistades únicas como Nash, Rocío, Lalo y muchas más que aún siguen siendo parte de mi vida y que fueron un gran apoyo en la realización de este trabajo y les agradezco por lo mismo. También conocí profesores únicos y que también agradezco por compartir sus conocimientos, en particular al maestro Jorge Luis Madariaga que además de ser mi asesor me tuvo la paciencia en todo este proceso.

A todos los profesores y personas que llegué a conocer a lo largo de los años en la universidad y me dieron consejos, aprendizaje, risas, crecimiento para la culminación de esta tesina les doy las gracias.

Contenido

Introducción

El muestreo es una parte de la estadística donde su objetivo principal es tomar sólo algunos individuos (muestra) de la población total para poder dar conclusiones del total, media, proporciones de la población entre otros con un error cuantificable y que se puede controlar. Si se habla de muestreo probabilístico entonces se habla de una mayor precisión en las conclusiones [1] pues está basado en inferencia estadística principalmente.

Sin saberlo muchas veces hacemos muestreo para dar conclusiones sobre poblaciones totales. Un ejemplo sencillo de lo anterior es en la comida, supongamos que en una cena de navidad donde el plato principal es una pavo relleno, entonces por simple vista y olor se puede dar una idea de que todo el pavo está sabroso (o no), para confirmarlo se puede agarrar un pedazo de éste (muestra) y al sentir que tiene una textura y sabor excelentes, entonces se concluye que efectivamente el pavo está sabroso sin tener que probar el resto de éste.

La más grande ventaja del muestreo es que a comparación de los censos es mucho más económico, pues no se tiene que ir individuo por individuo como en cualquier censo, sino que sólo se tiene que ir con cada individuo pero de la muestra seleccionada, sin embargo, el tamaño de la muestra puede ser un inconveniente y depende mucho del tipo de muestreo que se esté realizando y del error que se está dispuesto a soportar para que la muestra represente de manera amplia a la población, es decir, la forma de obtener la muestra es un aspecto de gran importancia para la representatividad de la población total. Además de lo económico, al tomar muestras se tiene la ventaja de que se puede reunir más rápido información que con censos.

Algunas de las posibles aplicaciones del muestreo son en:

- Ciencias biológicas
    - *Proporción* de semillas “malas” en un total de semillas
    - *Cantidad total* de impurezas en un vagón de ferrocarril cargado de trigo
- Industria
    - Control de calidad de muestreo de lote de producción para determinar si se cumple con las especificaciones requeridas en el proceso.

Existen varios tipos de muestreo probabilístico y no probabilístico algunos ejemplos son:

[1] Siempre y cuando el muestreo esté realizado de forma correcta, tomando en cuenta error y diseño muestral

- Probabilísticos

    - Muestreo aleatorio simple
    - Muestreo aleatorio sistemático
    - Muestreo aleatorio estratificado
    - Muestreo aleatorio por conglomerados

- No probabilísticos

    - Muestreo por cuotas
    - Muestreo de conveniencia
    - Bola de nieve
    - Muestreo discrecional

Podemos definir en general las ventajas y desventajas de los tipos de muestreo:

Muestreo	Ventajas	Desventajas
Probabilístico	▪ Selección correcta de la muestra. ▪ Representatividad en la muestra seleccionada. ▪ Conclusiones con mayor veracidad. ▪ Estimaciones del universo basada en teoría estadística demostrada. ▪ Capacidad de obtener el error de muestreo deseado.	▪ Requiere recursos altos en comparación con el no probabilístico. ▪ Mayor tiempo de generación. ▪ Requiere datos específicos del universo (listado de la población, distribución de la población, etc.).
No probabilístico	▪ El costo económico es mínimo. ▪ Exploración de una variable de interés. ▪ Costos humanos bajos. ▪ El tiempo es menor que el muestreo probabilístico.	▪ No se puede inferir totales de la población. ▪ No se tiene control en el error de la muestra. ▪ Muestra no representativa.

Por lo tanto se recomienda usar el muestreo probabilístico cuando se pretende hacer un estudio serio de alguna(s) variable(s) de interés. Si la intención sólo es de exploración, el muestreo no probabilístico podría ser una opción.

Capítulo I. Antecedentes del Muestreo

Neyman es uno de los padres del muestreo probabilístico como lo conocemos hoy, pues su artículo publicado en 1934 es uno de los pilares del muestreo.
Al respecto Leslie Kish, asegura que Neyman hizo siete grandes contribuciones al muestreo:

1.- Propuso la asignación de Neyman para el tamaño de muestra con diseños estratificados.

2.- Descubrió que el muestreo por conglomerados puede hacerse basado en un esquema probabilístico tal que la varianza de los estimadores pudieran ser calculados o estimados.

3.- Para que lo anterior se tuviera, se necesita una muestra grande de unidades.

4.- Para seleccionar una muestra grande es de suma importancia definir un marco de selección de número aleatorios.

5.- Para saber cómo formar estratos, el conocimiento a priori del comportamiento de la población puede facilitarlo.

6.- La importancia de selección probabilística en vez de selección a conveniencia.

7.- Para convencer a los escépticos acerca de la validez de sus afirmaciones, hizo ejemplos prácticos en la vida con encuestas verdaderas a gran escala.

Sin embargo la primera persona que se interesó por métodos representativos (muestreo) fue el director de estadística noruego A.N. Kaiser (1897) pues demostró empíricamente que haciendo estratos se obtenía unos estimadores mejores de las medias y los totales, sin embargo sus conclusiones no fueron bien recibidas por todos, pues los escépticos al muestreo comentaban que era peligroso y que las muestras no daban la misma información que un censo.

No fue sino hasta las décadas de 1930-1940 cuando las muestras probabilísticas fueron bien recibidas, esto se debió al trabajo de Neyman arriba mencionado, al de Bowley (1906) que propuso la fórmula para la varianza de diseño del muestreo estratificado. La publicación de las tablas de números aleatorios de Tippett (1927) también facilitó la selección de muestras probabilísticas, entre otros.

A partir de que el muestreo tuvo una mejor aceptación en la estadística se han ido realizando más avances sobre el mismo, como el caso de Cochran que introduce el estimador de razón y también desarrolla la teoría de totales y medias por regresión. En el 44, Madow y Madow muestran el muestreo sistemático entre otros.

Capítulo II. Definiciones básicas.

Sin embargo para empezar a hablar formalmente de muestreo son necesarias definiciones de términos importantes que se estarán usando a lo largo del trabajo, términos que aclararemos a continuación.
El término "muestreo probabilístico" es una investigación estadística con las siguientes características y definiciones:

- Una muestra es una extracción de una población finita, es decir una población con N elementos, y cada elemento tiene un etiqueta que lo identifica, generalmente le pondré números o letras para identificarlo, por ejemplo la k-ésima persona de la población será identificada con la letra k. El objetivo de una muestra es tener todas las características de la población total de donde fue extraída, sin embargo tener todas éstas características en una muestra es difícil de lograr, por lo que se considera "buena" la muestra si tiene las características de interés del estudio que estemos realizando.
- Cada elemento de la población debe contener información de una o varias variables de interés. "Lo que se quiere al hacer muestreo es obtener características de la población que se desconocen a base de la muestra, medidas cuantitativas de interés para el investigador" (Carl-Erik Särndal, 2003), por ejemplo, cantidad de desempleados, peso total de la población, etc.
- **Marco muestral** es una lista que nos da el acceso y observaciones de cada elemento de la población, es un dispositivo que nos permite asociar los elementos de la población con las unidades de muestreo, es decir con las unidades donde se hace la muestra.
- Los elementos de la población de los cuales las variables de estudio son medidas y anotadas son llamados **unidades de observación**.
- El siguiente término es una parte fundamental para las estadísticas resultantes y en general para los resultados, el conjunto de observaciones que se quiere estudiar es llamado **población objetivo**. Supongamos una encuesta de automóviles, ¿la población objetivo serían todas las personas que pueden y saben manejar?, ¿las personas que tienen un automóvil en ese momento?, ¿las personas que han tenido un automóvil? ¿las personas que han manejado un vehículo?
- Para el conjunto donde se obtiene la muestra es llamado **población muestreada**.
- Al momento de hacer muestreo cada unidad de la población tiene una **variable respuesta**, que será la respuesta de la característica que se está estudiando.
- Llamamos **cantidades poblacionales** (estadísticos de la población) al promedio de la población, el total o la proporción de la misma.

Se pondrá de una forma más clara lo anterior. Supongamos que se quiere sacar una encuesta para cierta variable de interés en niños entre cierto rango de edad, donde el único marco muestral que tenemos es uno que considera a todas las personas de la población de hace 1 mes.

El inconveniente del marco muestral en este caso es que no está tomando en cuenta a niños extranjeros que están de vacaciones o se cambiaron de hogar al lugar donde se desea realizar la encuesta.
Como en el marco tenemos a todas las personas y no sólo los niños deseados para la encuesta, entonces nuestra población muestreada se reduce, de igual forma es reducida por niños que se observan en el marco muestral pero que actualmente se cambiaron de casa, fallecieron, están fuera de la ubicación, o simplemente no quieren contestar y no es posible obtener información de éstos.
En lo particular para esta encuesta se tendrán sesgos de selección pues el marco muestral no incluye a todos los niños que se desean estudiar es decir no incluye a toda la población objetivo, lo cual hará poco confiables las estimaciones que se realicen. En general los sesgos de medición se presentan mucho en muestreo no probabilístico, pues por ejemplo se suele encuestar a unidades elegidas por los investigadores sin alguna probabilidad y de fácil acceso.

Todos estos conceptos son de suma importancia, pues es indispensable saber bien quiénes son las unidades, población objetivo etc. para hacer las conclusiones de forma correcta. A continuación, algunos problemas resueltos del tema.

Problemas resueltos:

Para cada una de las siguientes encuestas, describa la población objetivo, la población muestreada, el marco muestral, la unidad de muestreo, la unidad de observación, la variable respuesta y las cantidades poblacionales.

1. Una persona selecciona dos páginas de cada uno de los primeros n capítulos de un libro que consta de N capítulos (n<N) y cuidadosamente cuenta el número de errores en cada página seleccionada. La persona entonces reporta que el número total de errores en el libro se encuentra entre 85 y 95.

Población objetivo: todas las páginas contenidas en los N capítulos del libro.
Población muestreada: las páginas de los n capítulos seleccionados.
Marco muestral: Una lista que contenga los números de página de los n capítulos, el índice del libro puede ser esa lista.
Unidad de muestreo: Las páginas de los n capítulos.
Unidad de observación: Las dos páginas seleccionadas en la muestra de cada uno de los n primeros capítulos.
Variable respuesta: los números de errores arrojados en cada hoja seleccionada.
Cantidades poblacionales: total de errores en los primeros n capítulos.

2. El año pasado la FES Acatlán realizó una encuesta que arrojó como resultados que entre el 10% y 15% de los graduados de Actuaría trabajan en puestos en alguna dependencia del gobierno. Los datos fueron recolectados por medio de cuestionarios enviados por correo postal a graduados de Actuaría de los últimos 5 años.

Población objetivo: Graduados de Actuaría.
Población muestreada: Los graduados de actuaría de los últimos 5 años de los cuales se tenía la dirección para posiblemente enviar los cuestionarios.
Marco muestral: Listado con nombres y direcciones de los graduados de los últimos 5 años.
Unidad de muestreo: Cada uno de los nombres y direcciones de los graduados de los últimos 5 años.
Unidad de observación: Cada uno de los nombres y direcciones de los graduados de los últimos 5 años que fueron seleccionados para enviarles el cuestionario (fueron parte de la muestra).
Variable respuesta: variable dicotómica donde:
1 si trabaja en el gobierno
0 si no trabaja en el gobierno

Cantidades poblacionales: Proporción de graduados de Actuaría de los últimos 5 años que trabajan en gobierno.

3. Una encuesta es llevada a cabo para investigar el porcentaje de consumidores de cigarrillos en el estado de Guanajuato que fuman regularmente los cigarros de marca X. la información es obtenida mediante llamadas telefónicas y los números telefónicos son seleccionados a partir del directorio telefónico de la ciudad.

Población objetivo: Fumadores en Guanajuato.
Población muestreada: Fumadores de Guanajuato que tengan teléfono y su número telefónico esté en el directorio.
Marco muestral: Directorio telefónico de Guanajuato.
Unidad de muestreo: Los fumadores que vivan en las casas de los números telefónicos que aparezcan en el marco muestral.
Unidad de observación: Los fumadores que vivan en las casas de los números telefónicos que fueron seleccionados para encuestar.
Variable respuesta: variable dicotómica donde:
1 si fuman la marca X
0 si no fuman la marca X

Cantidades poblacionales: Porcentaje de fumadores que están en el listado y fuman la marca X.

II.I. Estimadores.

La importancia de los estimadores se ilustrará comenzando con un ejemplo sencillo.
Sea una población (N) con 4 elementos:

U = { 1, 2, 3, 4} N = 4

Supongamos que se necesita seleccionar una muestra de tamaño n = 3.

Sea y = { 4, 2, 5, 9} la variable respuesta de interés de cada uno de los elementos de la población, es decir la unidad 1 de la población tiene como variable respuesta 4, la unidad 2 tiene 2 y así para cada una unidad.
Por lo tanto el total de y = 4+2+5+9 = 20
Es claro que conocemos todos los elementos de y de la población pero en la vida real esto nos es posible saber siempre y es por eso que se recurre a los estimadores para por ejemplo estimar el total de la variable y.
Supongamos entonces que no se sabe el valor total de y, por lo tanto deseamos estimarlo. Dado que la muestra es de 3 (n) entonces las posibles muestras son:

$$S_1 = \{1, 2, 3\} \qquad S_2 = \{1, 2, 4\}$$

$$S_3 = \{1, 3, 4\} \qquad S_4 = \{2, 3, 4\}$$

Y la probabilidad de seleccionar una de las 4 muestras es de ¼.[2]
La probabilidad de seleccionar a una unidad de la población en la muestra la llamaremos $\pi_\alpha \; con \; \alpha \in U$[3]
Dadas el número posible de muestras con n elementos y suponiendo que es sin remplazo la extracción de unidades se puede deducir que:

$$\pi_1 = \pi_2 = \pi_3 = \pi_4 = \frac{3}{4}$$

Las respectivas respuestas de cada muestra son las siguientes:

$$y_{S_1} = \{4, 2, 5\} \qquad y_{S_2} = \{4, 2, 9\}$$

$$y_{S_3} = \{4, 5, 9\} \qquad y_{S_4} = \{2, 5, 9\}$$

Ahora se proponen los siguientes estimadores del total de la población de y (t_u)

$$\hat{t}_1 = \sum_{k \in S} y_k$$

$$\hat{t}_2 = \; N\frac{\sum_{k \in S} y_k}{n} = N\bar{y} = 4 \; \cdot \frac{\sum_{k \in S} y_k}{3}$$

Estos dos estimadores son estadísticos (función de valores de una muestra).
La mayor parte de los resultados en muestreo se buscan en la distribución de muestreo de un estadístico, la usual es la distribución de los distintos valores de la estadística obtenida al considerar todas las muestras posibles de la población.
La distribución de muestreo es un ejemplo de una distribución de probabilidad discreta.
Ahora supongamos que tomamos el primer estimador, por lo tanto para $\hat{t}_1$:

$$\hat{t}_{1S_1} = 4 + 2 + 5 = 11 \qquad \hat{t}_{1S_2} = 4 + 2 + 9 = 15$$

$$\hat{t}_{1S_3} = 4 + 5 + 9 = 18 \qquad \hat{t}_{1S_4} = 2 + 5 + 9 = 16$$

Distribución de muestreo de $\hat{t}_1$

$$\frac{x = \hat{t}_1}{P(\hat{t}_1 = x)} = \frac{11}{\frac{1}{4}} \quad \frac{15}{\frac{1}{4}} \quad \frac{18}{\frac{1}{4}} \quad \frac{16}{\frac{1}{4}}$$

[2] Para fines de este trabajo se maneja con muestras equiprobables.
[3] También llamada probabilidad de inclusión de primer orden.

$$=> E(\hat{t}_1) = \frac{1}{4} \cdot 11 + \frac{1}{4} \cdot 15 + \frac{1}{4} \cdot 18 + \frac{1}{4} \cdot 16 = 15$$

$E(\hat{t}_1) \neq t_u \quad \therefore \ \hat{t}_1 \ es \ sesgado$

Sesgo de $\hat{t}_1 = 15 - 20 = -5$

$$Var(\hat{t}_1) = E[(\hat{t}_1 - E(\hat{t}_1)]^2 = \sum_{todas\ las\ muestras\ posibles} P(S)[(\hat{t}_s - E(\hat{t}_1)]^2$$

$$= \frac{1}{4}[11 - 15]^2 + \frac{1}{4}[15 - 15]^2 + \frac{1}{4}[18 - 15]^2 + \frac{1}{4}[16 - 15]^2 \approx 6.5$$

ECM ($\hat{t}_1$)= V ($\hat{t}_1$) + (Sesgo ($\hat{t}_1$))^2 = 6.5 + (-5)^2 = 6.5 + 25 = 31.5

Ahora veamos para $\hat{t}_2$:

$\hat{t}_{2S_1} = 4 \cdot \frac{(4+2+5)}{3} = 14.66$ $\qquad$ $\hat{t}_{2S_2} = 4 \cdot \frac{(4+2+9)}{3} = 20$

$\hat{t}_{2S_3} = 4 \cdot \frac{(4+5+9)}{3} = 24$ $\qquad$ $\hat{t}_{2S_4} = 4 \cdot \frac{(2+5+9)}{3} = 21.33$

La distribución de muestreo para $\hat{t}_2$

$$\frac{x = \hat{t}_2}{P(\hat{t}_2 = x)} = \frac{14.66}{\frac{1}{4}} \quad \frac{20}{\frac{1}{4}} \quad \frac{24}{\frac{1}{4}} \quad \frac{21.33}{\frac{1}{4}}$$

$$E(\hat{t}_2) = \frac{1}{4} \cdot 14.66 + \frac{1}{4} \cdot 20 + \frac{1}{4} \cdot 24 + \frac{1}{4} \cdot 21.33 = 20$$

$$\therefore \ \hat{t}_2 \ es \ insesgado$$

Ahora

$$Var(\hat{t}_2) = \frac{1}{4} \cdot \left[\frac{44}{3} - 20\right]^2 + \frac{1}{4}[20 - 20]^2 + \frac{1}{4} \cdot [24 - 20]^2 + \frac{1}{4} \cdot [21.33 - 20]^2 \approx 11.55$$

ECM ($\hat{t}_2$)=11.55 < ECM ($\hat{t}_1$)

Un estimador puede ser:

- Insesgado si E($\hat{t}$)= t
- Preciso si V ($\hat{t}$) =E [($\hat{t}$-E ($\hat{t}$))^2] es pequeña.
- Exacto si ECM ($\hat{t}$)= Var ($\hat{t}$)+ [sesgo ($\hat{t}$)]^2 es pequeño.

Para analizar lo preciso que es un estimador se usan los conceptos de varianza, ECM (error cuadrático medio) y sesgo. Se suele llamar preciso a un estimador a lo acertado que éste sea. Sin embargo el sesgo y varianza son valores importantes que pesan en la precisión del estimador, y pueden relacionarse a partir del ECM

$$\begin{aligned} ECM(\hat{\theta}) &= E\left[(\hat{\theta}-\theta)^2\right] \\ &= E\left[(\hat{\theta}-E(\hat{\theta})+E(\hat{\theta})-\hat{\theta})^2\right] \\ &= E[(\hat{\theta}-E(\hat{\theta})]^2+\left(E[\hat{\theta}]-\theta\right)^2+2E\left[(\hat{\theta}-E[\hat{\theta}])(E(\hat{\theta})-\theta)\right] \\ &= V(\hat{\theta})+sesgo(\hat{\theta})^2 \end{aligned}$$

En la práctica, se considera que el sesgo de $\hat{\theta}$ no es influyente cuando

$$\left|\frac{Sesgo(\hat{\theta})}{\sigma(\hat{\theta})}\right| < \frac{1}{10}$$

A partir del ECM es fácil observar que para un estimador Insesgado $\hat{\theta}_1$ de θ_1 su error cuadrático medio es entonces

$$ECM(\hat{\theta}_1) = V(\hat{\theta}_1)$$

Y como bien se sabe, puede haber más de un estimador insesgado para un parámetro, por tanto, para ver qué estimador insesgado es más preciso uno de otro basta con considerar el error de muestreo también llamado **error estándar** o bien la desviación estándar

$$\sqrt{V(\hat{\theta}_1)} = \sigma(\hat{\theta}_1)$$

Se considera como un mejor estimador al que tenga menor error estándar.

Mientras que para estimadores sesgados, la forma general para analizar su exactitud es el ECM. Y por lo mismo, para comparar varios estimadores sesgados en cuanto a precisión, el que tenga

menor ECM será el que sea más preciso. Sin embargo en la práctica el ECM puede ser complicado calcular y por lo tanto se calcula a cada estimador insesgado la cantidad de

$$\left|\frac{Sesgo(\hat{\theta})}{\sigma(\hat{\theta})}\right|$$

Siendo mejor (más preciso) estimador que tenga menor relación del sesgo con su error estándar.

Por lo tanto y estrictamente hablando para evaluar estimadores insesgados se usa el error estándar mientras que para sesgados es más apropiado usar el error cuadrático medio. Sin embargo si el sesgo de los estimadores es no influyente o despreciable se puede usar también el error estándar como medida de comparación.

II.II. Errores de muestreo y no de muestreo.

Todas las encuestas llevadas a cabo tienen un error que es conocido como error de muestreo, contrario a las otras fuentes de error, el error muestral puede ser medido, controlado e incluso reducido al aumentar la cantidad de muestra (n). Es decir que este error ocurre por no seleccionar a toda la población, y por lo tanto varía dependiendo de la muestra.

Sin embargo cuando se trata de sesgo de selección, imprecisión de respuestas, malos encuestadores, preguntas de la encuesta hechas de forma no clara, entre otros ejemplos estamos hablando de errores no muestrales, que no dependen de la muestra. Aún así este tipo de error es posible reducirlo al momento de elaborar encuestas, capacitar encuestadores, etc. Por este tipo de errores es por el cual a veces la gente no cree en la teoría del muestreo, pues muchas encuestas afirman que tanto por ciento de la población prefiere tal o cual cosa, sin embargo no consideran este tipo de error y por lo mismo los resultados que arroja son probablemente falsos.
Estos errores que no son propios de la teoría del muestreo se pueden controlar dándole la importancia necesaria a la construcción del cuestionario a realizar y a la buena supervisión al momento de levantar las encuestas.

El error de muestreo entonces está definido como

$$|\theta - \hat{\theta}|$$

Pero se puede controlar con el diseño del muestreo.

II.III. Probabilidad de inclusión

En muestreo probabilístico cada elemento del universo donde se pretende sacar una muestra tiene una probabilidad de que pertenezca a ella. El orden de inclusión depende del número de elementos que se desea saber su probabilidad de inclusión en la muestra.

II.III.I. De primer orden

La probabilidad de inclusión π_k (o probabilidad de inclusión de primer orden) es la probabilidad de que el elemento k pertenezca a la muestra S, esto es:

$$\pi_k = P(k \epsilon S) \qquad \forall\, k = 1, 2, \dots, N$$

Generalmente se asume que $\pi_k > 0$.
El inverso de la probabilidad de inclusión, es el llamado "*peso de muestreo*" de la unidad k.

II.III.II. De segundo orden

La probabilidad de segundo orden π_{kl} es la probabilidad de que las unidades k y l estén en la muestra S, es decir

$$\pi_{kl} = P(k, l \epsilon S) \qquad \forall\, k \neq l$$

La variable indicadora I_k (k= 1, … , N) es una variable aleatoria y está definida por:

$$I_k = \begin{cases} 1\ si & k \epsilon S \\ 0\ otro\ caso & \end{cases}$$

Por lo tanto es fácil deducir que

$$\sum_{k=1}^{N} I_k = n$$

Nótese que las únicas cantidades aleatorias en muestreo son las I_k pues las variable respuesta de muestreo para la k-ésima unidad del universo (y_k) son valores fijos y solamente se conoce aquellos que pertenecen a S.
Podría pensarse que $\pi_{kl} = \pi_k \cdot \pi_l$ sin embargo esto es falso, una forma de demostrarlo es la siguiente:

$$\begin{aligned} Cov(I_k, I_l) &= E(I_k \cdot I_l) - E(I_k)E(I_l) \\ &= \big(1 \cdot P(I_k, I_l = 1) + 0 \cdot P(I_k, I_l = 0)\big) - \big(1 \cdot P(I_k = 1) + 0 \cdot P(I_k = 0)\big) \\ &\quad \cdot \big(1 \cdot P(I_l = 1) + 0 \cdot P(I_l = 0)\big) \end{aligned}$$

Ahora si observamos $P(I_k = 1)$, es la probabilidad de que k pertenezca a S, es decir π_k. Por lo tanto

$$Cov(I_k, I_l) = P(k, l\epsilon S) - P(k\epsilon S) \cdot P(l\epsilon S)$$
$$= \pi_{kl} - \pi_k \pi_l$$

por lo tanto $\pi_{kl} = \pi_k \pi_l + Cov(I_k, I_l)$

II.IV. Diseño muestral

El diseño muestral juega un papel muy importante en la teoría del muestreo, pues determina las propiedades esenciales, de cantidades aleatorias como la media o la varianza muestral.

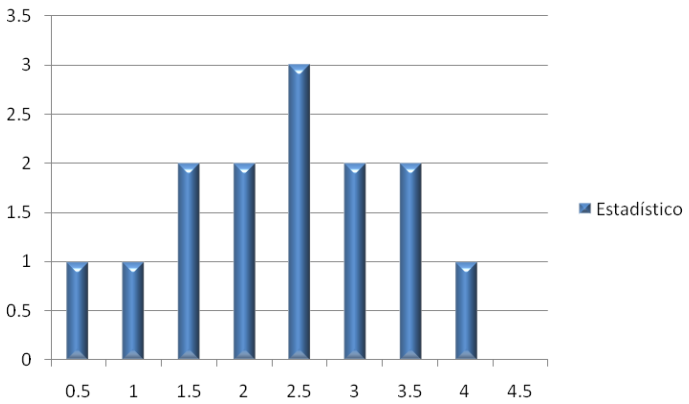
Es el tipo de muestreo que se realizará para la encuesta en cuestión, es decir, es la probabilidad de selección de la muestra hecha bajo cierto esquema. Cada uno de éstos tiene propiedades únicas que los diferencían entre los otros, y por lo mismo cambian las fórmulas de sus estadísticos e incluso el tamaño de la muestra.
Existen diseños de muestreo con reemplazo, sin reemplazo y de tamaño fijo.

II.V. Distribución muestral

La distribución muestral es la distribución de algún estadístico de la muestra, tal como media, total, etc.

Dado un universo y un tamaño de muestra, se pueden sacar muchos promedios muestrales dependiendo de la muestra elegida, la distribución muestral entonces es el histograma de los posibles valores de la media muestral en cada muestra.

valor del estadístico	muestras con este valor (frecuencia)
0.5	1
1	1
1.5	2
2	2
2.5	3
3	2
3.5	2
4	1
4.5	0



Capítulo III. Muestreo Aleatorio Simple (MAS) sin remplazo

"El muestreo aleatorio simple es la forma más sencilla de muestreo de probabilidad y proporciona la base teórica de las formas más complejas" (Sharon L. Lohr, 2005)

En esta tesina no se hablará de muestreo con reemplazo pues la intención de la misma es para casos reales y con poblaciones finitas, por lo que un muestreo con un objeto de interés que se repite 2 o más veces no proporciona mejor información que si aparece una vez.

Se hablará de las ventajas y desventajas de este tipo de muestreo, para totales, proporciones, y promedios, la forma de sacar un tamaño adecuado de muestra para que los resultados del estudio sean confiables.

Ya que se han definido los conceptos más importantes del muestreo, podemos entonces pensar en cómo sacar una muestra. Bien podríamos simplemente usar un sistema que genere números aleatorios y empezar a sacar unidades de nuestra población para una muestra. Sin embargo eso no sería un muestreo probabilístico.

III.I. Notación y definición

N = Número de unidades en la población.
n = Número de unidades en la muestra.
y_i = Valor de la variable estudiada del i-ésimo unidad de muestreo o población
U= Población de donde se seleccionará una muestra
S= Muestra seleccionada a partir de U.

La característica representativa del muestreo aleatorio simple (MAS) es que toda muestra tiene la misma probabilidad de ser seleccionada, y cada individuo podrá ser seleccionado sólo una vez (sin remplazo). Por lo anterior y dado que el orden de la muestra S no importa se puede deducir que el número posibles de muestras de tamaño n que podemos obtener es:

$$\binom{N}{n}$$

Supongamos que se tienen 5 unidades en nuestra población y se quiere una muestra de 4. Entonces.

U:= {1, 2, 3, 4, 5}

Las posibles muestras son 5:

$$S_1 = \{1, 2, 3, 4\}$$
$$S_2 = \{1, 2, 3, 5\}$$
$$S_3 = \{2, 3, 4, 5\}$$
$$S_4 = \{1, 3, 4, 5\}$$
$$S_5 = \{1, 2, 4, 5\}$$

Todas las combinaciones.

III.II. Probabilidad de muestra

Dado que todas las muestras tienen la misma probabilidad de ser seleccionadas bajo MAS, la función de probabilidad del diseño muestral para una muestra S es la siguiente:

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

Una forma de probar esto es la siguiente:

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

$$= \frac{1}{N \cdot N-1 \cdot N-2 \cdot \ldots \cdot N-n+2 \cdot N-n+1} \cdot n!$$

Ahora, nótese que:

$$\frac{1}{N \cdot N-1 \cdot N-2 \cdot \ldots \cdot N-n+2 \cdot N-n+1}$$

Es la probabilidad de extraer una muestra cualquiera de n elementos

S = {1, 2,…, n-1, n}

Y dado que no importa el orden de los elementos de la muestra, es decir la muestra anterior es la misma que la siguiente:

S = {n, 2,…, 1, n-1}

Y como ésta hay n! posibles formas de ordenar la primera, entonces la probabilidad de extraer esta muestra en particular es:

$$\frac{1}{N \cdot N-1 \cdot N-2 \cdot \ldots \cdot N-n+2 \cdot N-n+1} \cdot n! = \frac{1}{\binom{N}{n}}$$

Aunque el diseño muestral juega un papel muy importante, en la mayoría de los casos no es una herramienta práctica para seleccionar una muestra pues para poblaciones más grandes es poco práctico poner todas las posibles muestras, si por ejemplo nuestro universo cuenta con 50 unidades y nuestra muestra necesita 5 unidades entonces las posibles muestras son 2,118,760.

En MAS todas las unidades tienen la misma probabilidad de pertenecer a la muestra. Pues para una primera extracción, la probabilidad de que una unidad sea seleccionada es:

$$P(i \in S \mid \#S = 0) = \frac{1}{N}$$
$$\forall i \in \Omega$$

La probabilidad de que una unidad sea seleccionada en la segunda extracción está dada por.

$$P(i \in S \mid \#S = 1) = \frac{1}{N-1} \cdot P(i \notin S \mid \#S = 0)$$

$$P(i \notin S \mid \#S = 0) = 1 - P(i \in S \mid \#S = 0) = 1 - \frac{1}{N} = \frac{N-1}{N}$$

$$\therefore \quad P(i \in S \mid \#S = 1) = \frac{1}{N-1} \cdot \frac{N-1}{N} = \frac{1}{N}$$

Es decir, es la probabilidad de ser seleccionada en la muestra dado que ésta ya contiene un elemento[4] y por lo tanto ya se eligió una unidad del universo por la probabilidad de que la unidad no haya sido seleccionada en la primera extracción.

Usando la misma lógica, para la n-ésima extracción

[4] El símbolo número # representa la cardinalidad del conjunto, es decir cuántos elementos contiene el conjunto.

$$P(i \in S \mid \#S = n-1) = \frac{N-1}{N} \cdot \frac{N-2}{N-1} \cdot \ldots \cdot \frac{1}{N-(n-1)} = \frac{1}{N}$$

Por lo tanto la probabilidad de que una unidad esté en la muestra, es la suma de probabilidades de que esté en la 1ª, 2ª, 3ª,…, nª extracción

$$\overbrace{\frac{1}{N} + \frac{1}{N} + \cdots + \frac{1}{N}}^{n} = \frac{n}{N}$$

Es decir que en MAS $\pi_k = \frac{n}{N}$.

Ahora se darán unas importantes definiciones de la población y la muestra.

Para la población total definiremos el término a continuación presentado.

$$t_u = \sum_{k \epsilon U} y_k$$

La media poblacional μ_y está definida como

$$\mu_y = \frac{1}{N} \sum_{k \epsilon U} y_k$$

También denotada por $\bar{y}_u$.

La varianza de y en la población es

$${S_u}^2 = \frac{1}{N-1} \sum_{k \epsilon U} (y_k - \bar{y}_u)^2$$

Por lo tanto la desviación estándar de y en la población (U)

$$S_u = \sqrt[2]{{S_u}^2}$$

Observando las definiciones anteriores para la población total, también se puede definir para la muestra de la misma forma, es decir:

$\bar{y}_s = \frac{1}{n}\sum_{k\epsilon S} y_k$ Es la definición de la media muestra.

${S_S}^2 = \frac{1}{n-1}\sum_{k\epsilon S}(y_k - \bar{y}_S)^2$ Así es definida la varianza muestral.

Nótese que la notación usada asume que el tamaño de muestra n es fijo, ie, el número de unidades que será seleccionado es conocido antes de hacer la selección de la muestra.

Sin embargo en la práctica casi nunca se saben estas cantidades poblacionales y mucho menos todas, pues si se supieran no tendría mucho sentido realizar muestreo. Por lo mismo, a partir de una muestra se puede obtener estimadores para la población total.

Un buen estimador para la media poblacional es la media muestral, pues es un estimador insesgado para la media poblacional es decir $E(\bar{y}_S) = \bar{y}_U$. Ya que por definición de esperanza se tiene

$$E(\bar{y}_S) = E\left(\frac{1}{n}\sum_{k\epsilon S} y_k\right)$$

Por otro lado

$$\sum_{k\epsilon S} y_k = \sum_{k\epsilon S} 1 \cdot y_k + \sum_{k\epsilon S^C} 0 \cdot y_k = \sum_{k=1}^{N} I_k y_k$$

Y además

$$E(I_k) = 1 \cdot p(k \in S) + 0 \cdot p(k \in S^C) = 1 \cdot \pi_k = \pi_k = \frac{n}{N}$$
$$E\left({I_k}^2\right) = 1^2 \cdot p(k \in S) + 0^2 \cdot p(k \in S^C) = \frac{n}{N}$$

Ahora se observa que bajo MAS también ocurre lo siguiente

$$E(I_k I_l) = P(k \in S, l \in S) = \frac{n-1}{N-1}\left(\frac{n}{N}\right)$$

Por lo tanto

$$Cov(I_k, I_l) = E(I_k I_l) - E(I_k)E(I_l) = \frac{n-1}{N-1}\left(\frac{n}{N}\right) - \left(\frac{n}{N}\right)^2$$

$$E\left(\frac{1}{n}\sum_{k\epsilon S} y_k\right) = E\left(\frac{1}{n}\sum_{k=1}^{N} I_k y_k\right) = \frac{1}{n}\sum_{k=1}^{N} E(I_k) y_k = \frac{1}{n}\sum_{k=1}^{N} \frac{n}{N} y_k = \sum_{k=1}^{N} \frac{1}{N} y_k = \bar{y}_U$$

■

III.III. Estimador Horvitz-Thompson

El estimador de Horvitz-Thompson o también conocido como estimador H-T es un estimador insesgado del total de la población (t_U) y se define:

$$\hat{t}_\pi = \sum_{k\epsilon S} \frac{y_k}{\pi_k} = \sum_{k\epsilon S} w_k \cdot y_k$$

Siendo $w_k = \frac{1}{\pi_k}$. Es decir, es el peso de la k-ésima unidad.

Ahora se demostrará que efectivamente es un estimador insesgado de t_U, por definición de estimador insesgado tendríamos que llegar a:

$$E(\hat{t}_\pi) = t_U = \sum_{k\epsilon U} y_k$$

Para esta demostración se usará nuevamente la variable indicadora I_k .

$$\sum_{k\epsilon S} w_k \cdot y_k = \sum_{k\epsilon S} w_k \cdot y_k + \sum_{k\epsilon\{U-S\}} 0 = \sum_{k\epsilon U} w_k \cdot y_k \cdot I_k$$

$$E(\hat{t}_\pi) = E\left(\sum_{k\epsilon S} w_k \cdot y_k\right) = E\left(\sum_{k\epsilon U} w_k \cdot y_k \cdot I_k\right) = \sum_{k\epsilon U} w_k \cdot y_k \, E(I_k) = \sum_{k\epsilon U} w_k \cdot y_k \cdot \pi_k$$

Por definición de w_k se sigue que

$$E(\hat{t}_\pi) = \sum_{k\epsilon U} y_k = t_U \qquad ■$$

La varianza por otro lado está dada por:

$$V(\hat{t}_\pi) = \sum_{k\epsilon U}\sum_{l\epsilon U}(\pi_{k,l} - \pi_k\pi_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

Para la demostración de la varianza, se usa el mismo método que se usó en la demostración anterior.

$$V(\hat{t}_\pi) = V\left(\sum_{k\epsilon S}\frac{y_k}{\pi_k}\right) = V\left(\sum_{k\epsilon U} w_k \cdot y_k \cdot I_k\right) \quad \text{si no son independientes entonces}$$

$$= \sum_{k\epsilon U} V(w_k \cdot y_k \cdot I_k) + \sum_{k\epsilon U}\sum_{l\epsilon U} Cov\left(\frac{y_k}{\pi_k}I_k, \frac{y_l}{\pi_l}I_l\right) \qquad k \neq l$$

$$= \sum \left(\frac{y_k}{\pi_k}\right)^2 V(I_k) + \sum_{k\epsilon U}\sum_{l\epsilon U}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}Cov(I_k, I_l)$$

Por otro lado $V(I_k) = E\left({I_k}^2\right) - E(I_k)^2 = E(I_k) - E(I_k)^2$ dado que ${I_k}^2 = I_k$ por lo tanto $V(I_k) = \pi_k - {\pi_k}^2 = \pi_k(1 - \pi_k)$.

Regresando entonces…

$$= \sum \left(\frac{y_k}{\pi_k}\right)^2 \pi_k(1 - \pi_k) + \sum_{k\epsilon U}\sum_{l\epsilon U}\frac{y_k}{\pi_k}\frac{y_l}{{\pi_l}_k}\left(\pi_{k,l} - \pi_k\pi_l\right) \qquad k \neq l$$

$$= \sum_{k\epsilon U}\frac{y_k}{\pi_k}\cdot\frac{y_k}{\pi_k}(\pi_{k,k} - \pi_k\pi_k) + \sum_{k\epsilon U}\sum_{l\epsilon U}\frac{y_k}{\pi_k}\frac{y_l}{{\pi_l}_k}\left(\pi_{k,l} - \pi_k\pi_l\right)$$

$$= \sum_{k\epsilon U}\sum_{l\epsilon U}(\pi_{k,l} - \pi_k\pi_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} = V(\hat{t}_\pi)$$

Al igual que los estimadores anteriores, por lo general nunca se tendrá conocimiento total de U, por lo tanto un buen estimador de la varianza depende de la muestra y no de la población, para fines prácticos se puede usar la fórmula sobre la muestra en vez de la población total.

$$\hat{V}(\hat{t}_\pi) = \sum_{k\epsilon S}\sum_{l\epsilon S}\left(\frac{\pi_{k,l} - \pi_k\pi_l}{\pi_{k,l}}\right)\cdot\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

Se mostrará ahora que dicho estimador es insesgado.

$$E\left(\sum_{k\epsilon S}\sum_{l\epsilon S}\left(\frac{\pi_{k,l} - \pi_k\pi_l}{\pi_{k,l}}\right)\cdot\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\right) = E\left(\sum_{k\epsilon U}\sum_{l\epsilon U}\left(\frac{\pi_{k,l} - \pi_k\pi_l}{\pi_{k,l}}\right)\cdot\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}I_kI_l\right)$$

$$= \sum_{k\epsilon U}\sum_{l\epsilon U}\left(\frac{\pi_{k,l} - \pi_k\pi_l}{\pi_{k,l}}\right)\cdot\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\mathrm{E}(I_kI_l) = \left(\sum_{k\epsilon U}\sum_{l\epsilon U}\left(\frac{\pi_{k,l} - \pi_k\pi_l}{\cancel{\pi_{k,l}}}\right)\cdot\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\right)\cdot\cancel{\pi_{k,l}} = V(\hat{t}_\pi)$$

Por lo tanto es un estimador insesgado.

Otra forma de plantear la varianza del estimador H-T es la siguiente.

$$V(\hat{t}_\pi) = -\frac{1}{2}\sum_{k\epsilon U}\sum_{l\epsilon U}(\pi_{k,l} - \pi_k\pi_l)\cdot\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2$$

Hasta ahora se han dado y demostrado algunas fórmulas del estimador H-T y sus estadísticas básicas, sin embargo cuando se mezcla MAS con H-T se reducen muchas fórmulas simplificadas por las hipótesis de MAS a continuación se mostrará qué pasa con el estimador de Horvitz Thompson en muestreo aleatorio simple.

Se sabe que $\hat{t}_\pi = \sum_{k=1}^{n}\frac{y_k}{\pi_k}$ y que en MAS $\pi_k = \frac{\binom{N-1}{n-1}}{\binom{N}{n}}$

Por lo tanto $\pi_k = \frac{n}{N}$ provocando que el estimador H-T se convierta a

$$\hat{t}_\pi = \sum_{k=1}^{n}\frac{N}{n}y_k$$

Recordando que $\bar{y}_S = \frac{1}{n}\sum_{k\epsilon S}y_k$

$$\boxed{\hat{t}_\pi = N\bar{y}_S}$$

Es importante observar que a partir de esta fórmula se puede obtener la siguiente expresión para el promedio muestral $\bar{y}_S$.

$$\bar{y}_S = \frac{\hat{t}_\pi}{N}$$

Regresando con el estimador H-T, se observa que

$$\pi_{k,l} = P(k, l \epsilon S) = P(I_k = 1, I_l = 1) = P(I_k = 1 | I_l = 1) P(I_l = 1)$$

$$P(I_k = 1 | I_l = 1) = \frac{\binom{N-1}{n-2}}{\binom{N-1}{n-1}} = \frac{n-1}{N-1}$$

$$\pi_{k,l} = \frac{n}{N} \cdot \frac{n-1}{N-1}$$

Obtenidas estas dos expresiones, podemos sustituirlas en los estadísticos demostrados anteriormente:

A partir de

$$V(\hat{t}_\pi) = -\frac{1}{2} \sum_{k \epsilon U} \sum_{l \epsilon U} (\pi_{k,l} - \pi_k \pi_l) \cdot \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2$$

Sustituyendo las probabilidades de inclusión en MAS se pueden obtener 3 resultados equivalentes para la varianza del estimador llegando a

$$V(\hat{t}_\pi) = \begin{cases} \frac{N-n}{n(N-1)} \cdot N(N-1) {S_u}^2 \\ (N-n) \cdot N \frac{{S_u}^2}{n} \cdot \frac{N}{n} \\ N^2 \left(\frac{N-n}{N} \right) \frac{{S_u}^2}{n} \end{cases}$$

De la misma manera se puede obtener la siguiente expresión para

$$\hat{V}(\hat{t}_\pi) = N^2\left(\frac{1}{n} - \frac{1}{N}\right){S_s}^2$$

En donde ${S_S}^2 = \frac{1}{n-1}\sum_{k\epsilon S}(y_k - \bar{y}_S)^2$ recordando también que $\bar{y}_S$ es la media muestral, donde ésta también puede tener nuevas representaciones para sus estadísticos bajo lo demostrado anteriormente en MAS

$$\hat{V}_{M.A.S.}(\hat{t}_\pi) = N^2\left(\tfrac{1}{n} - \tfrac{1}{N}\right){S_s}^2 \quad \text{de aquí que}$$

$$\bar{y}_S = \frac{1}{N}\hat{t}_\pi \qquad y \qquad \hat{t}_\pi = N\bar{y}_S$$

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

Como se había mencionado y demostrado anteriormente $\bar{y}_S$ es un estimador insesgado para la media poblacional, sin embargo hay una forma más rápida demostrarlo con el estimador H-T.

$$E(\bar{y}_S) = E\left(\frac{\hat{t}_\pi}{N}\right) = \frac{1}{N}E(\hat{t}_\pi) = \frac{1}{N}t_u = \bar{y}_U$$

III.IV. Intervalos de Confianza

Generalmente para explicar un intervalo de confianza al 95% de confianza se dice que si se extraen muestras de la población y construimos un intervalo de confianza para cada una de éstas, se espera que al menos el 95% de los intervalos de confianza contengan el valor del parámetro, es decir, si se hacen 100 muestras, cada una con su intervalo de confianza entonces se espera que 95 de éstos contengan el valor real del estimador.

*"*Obtener una estimación por intervalos para un parámetro poblacional Θ al nivel de confianza α consiste en hallar un intervalo real para el que se tiene una probabilidad de 1-α de que el verdadero valor del parámetro Θ caiga dentro del citado intervalo. El valor (1-α) suele denominarse coeficiente de confianza*"* (César Pérez López, 2005, pág. 8)

Un intervalo de confianza depende mucho de la distribución del estimador, que por lo general en muestreo y en particular MAS se supone una distribución normal, sin embargo no siempre es de esta manera pues influye mucho el conocimiento previo que se tenga de la variable en estudio pues pueden haber variables con comportamiento asimétrico como: tamaños de las ciudades, empresas, tiendas el ingreso de la población, etc. y son estos los que necesitan un mayor número de muestra para considerar normal su distribución a comparación de las variables con comportamiento simétrico.

La probabilidad de que el estimador de una variable Θ esté alejado del valor real sea cuando más δ es de 1-α. δ es conocido como error absoluto y 1-α es conocido como confianza.

Esto es:

$$P\left(\theta - \delta \leq \hat{\theta} \leq \theta + \delta\right) = 1 - \alpha \quad ó \quad P\left(\left|\hat{\theta} - \theta\right| < \delta\right) = 1 - \alpha$$

Es decir; la probabilidad de que la diferencia entre el estimador y el valor real sea a lo más δ es $1 - \alpha$.

Suponiendo que los elementos de la muestra (n) sean lo suficientemente grandes para que se distribuyan como una normal[5].

De no cumplirse la normalidad se puede usar el percentil de una t-student con N-1 grados de libertad.

[5] En general si n>30 se supone normalidad, sin embargo si la variable se distribuye de forma normal entonces con n=1 bastará.

III.V. Teorema del límite central

En forma general el teorema del límite central (TLC) dice que los promedios de muchas muestras probabilísticas de una población tienden, al aumentar el tamaño de la muestra (n) a tomar una distribución normal, a pesar de que la variable que se mida no tenga una distribución normal en la población.

Para saber con qué rapidez se acerca la variable estudiada a una distribución normal depende mucho de la forma que tiene dicha variable (normal, asimétrica, uniforme, etc.).

Teniendo en cuenta el teorema, podemos aplicarlo a los intervalos de confianza que se venían manejando.

Ahora si la media muestral se distribuye de la siguiente forma

$$\bar{y}_S \sim N\left(\bar{y}_U, \left(1-\frac{n}{N}\right)\frac{{S_U}^2}{n}\right)$$

Y teniendo a Z como una variable normal (0,1)

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

Recordando que

$$\frac{\bar{y}_S - \bar{y}_U}{\sqrt{\left(1-\frac{n}{N}\right)\frac{{S_U}^2}{n}}} = Z$$

Entonces

$$P\left(-Z_{\left(1-\frac{\alpha}{2}\right)} < \frac{\bar{y}_S - \bar{y}_U}{\sqrt{\left(1-\frac{n}{N}\right)\frac{{S_U}^2}{n}}} < Z_{\left(1-\frac{\alpha}{2}\right)}\right) = 1-\alpha$$

$$P\left(\bar{y}_S - Z_{\left(1-\frac{\alpha}{2}\right)}\sqrt{\left(1-\frac{n}{N}\right)\frac{{S_U}^2}{n}} < \bar{y}_U < \bar{y}_S + Z_{\left(1-\frac{\alpha}{2}\right)}\sqrt{\left(1-\frac{n}{N}\right)\frac{{S_U}^2}{n}}\right) = 1-\alpha$$

Esto un intervalo del 100(1-α) %.

$$\bar{y}_U \epsilon \left(\bar{y}_S - Z_{\left(1-\frac{\alpha}{2}\right)} \sqrt{\left(1-\frac{n}{N}\right)\frac{{S_U}^2}{n}}, \bar{y}_S + Z_{\left(1-\frac{\alpha}{2}\right)} \sqrt{\left(1-\frac{n}{N}\right)\frac{{S_U}^2}{n}} \right)$$

También es común escribir los I.C. (intervalo de confianza) de la siguiente manera

$$\bar{y}_S \mp Z_{\left(1-\frac{\alpha}{2}\right)} \sqrt{\left(1-\frac{n}{N}\right)\frac{{S_U}^2}{n}}$$

Recordando que en MAS

$$\hat{t}_\pi = N\bar{y}_S$$

Se puede obtener el IC del total

$$\hat{t}_\pi \mp NZ_{\left(1-\frac{\alpha}{2}\right)} \sqrt{\left(1-\frac{n}{N}\right)\frac{{S_U}^2}{n}}$$

A menudo no se conoce S_U por lo que se puede sustituir por S_S.

De forma muy general tenemos que un I.C. para un estimador $\hat{\theta}$ es el siguiente (asumiendo normalidad).

$$\hat{\theta} \pm Z_{\left(1-\frac{\alpha}{2}\right)} \sqrt{\hat{V}(\hat{\theta})} = \hat{\theta} \pm Z_{\left(1-\frac{\alpha}{2}\right)} EE(\hat{\theta})$$

En caso de no cumplirse la normalidad:

$$\hat{\theta} \pm t_{\left(1-\frac{\alpha}{2}, N-1\right)} \sqrt{\hat{V}(\hat{\theta})}$$

Por lo tanto para un intervalo de confianza lo que se necesita es:

- El estimador de nuestro parámetro. $\hat{\theta}$
- El estimador de la varianza del estimador del parámetro. $\hat{V}(\hat{\theta})$

III.VI. Estimación del tamaño de la muestra

Una de las preguntas que surgen a la mente al momento de realizar una encuesta es ¿Cuántas personas necesito para que mi muestra sea buena? (en caso de que sean personas a las que se hará la encuesta), es decir ¿De qué tamaño será n?, esta pregunta va muy de la mano con la siguiente pregunta: ¿Qué tan exacto se desea el resultado? Un error que se suele hacer es intentar acomodar el presupuesto al tamaño de muestra, sin embargo lo correcto es ajustar el tamaño de muestra con los recursos que se tengan y con la precisión que se desea obtener. Si al momento de intentar ajustar el tamaño de muestra con los recursos y la precisión deseada se observa que necesitaría más recursos para una mejor muestra, en ese caso se deberá pensar en si se realiza la encuesta o no.

Un tamaño de muestra pequeño implica:

- Insuficientes elementos para hacer una buena inferencia del total.
- Bajo costo en el estudio.
- Poca precisión de los estimadores.
- Los I.C. suelen ser muy grandes con muestras pequeñas (ineficientes).

Por otro lado una muestra grande implica:

- Costos elevados para el estudio
- Un posible aumento en los errores que no son de muestreo.
- La precisión mejora.

Generalmente la precisión deseada es expresada en términos absolutos.

$$P(|\hat{y}_S - y_U| \leq \delta) = 1 - \alpha$$

A δ se le suele llamar *margen de error*.

Para encontrar una ecuación que relacione esta probabilidad con el tamaño de muestra es la ecuación de los intervalos de confianza. Para determinar una precisión absoluta, se necesita determinar un valor de n que cumpla con lo siguiente. Suponiendo para un promedio.

$$\delta^2 = Z_{\left(1-\frac{\alpha}{2}\right)}\sqrt{1-\frac{n}{N}}\cdot\frac{S_U}{\sqrt{n}}$$

Al despejar n

$$\delta^2 = Z^2_{\left(1-\frac{\alpha}{2}\right)}\left(1-\frac{n}{N}\right)\frac{{S_U}^2}{n}$$

$$\frac{1}{n} = \frac{\delta^2}{Z^2_{\left(1-\frac{\alpha}{2}\right)}{S_U}^2} + \frac{1}{N}$$

$$\frac{1}{n} = \frac{N\delta^2 + Z^2_{\left(1-\frac{\alpha}{2}\right)}{S_U}^2}{Z^2_{\left(1-\frac{\alpha}{2}\right)}{S_U}^2 N}$$

$$n = \frac{Z^2_{\left(1-\frac{\alpha}{2}\right)}{S_U}^2 N}{N\delta^2 + Z^2_{\left(1-\frac{\alpha}{2}\right)}{S_U}^2}$$

Ahora si se divide sobre el tamaño de población…

$$n = \frac{Z^2_{\left(1-\frac{\alpha}{2}\right)}{S_U}^2}{\delta^2 + \frac{Z^2_{\left(1-\frac{\alpha}{2}\right)}{S_U}^2}{N}}$$

Sea

$$n_0 = \frac{Z^2_{\left(1-\frac{\alpha}{2}\right)}{S_U}^2}{\delta^2}$$

Por lo tanto la fórmula final para el tamaño de muestra es la siguiente.

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

Para el caso de una proporción la variable de estudio tiene distribución de una bernoulli por lo tanto ${S_U}^2 = p(1-p)$ para poblaciones grandes. Ésta ecuación toma su máximo valor cuando p=.5 y

$$n_o = \frac{Z^2\left(1-\frac{\alpha}{2}\right)\frac{N}{N-1}{S_U}^2}{\delta^2}$$

Para la estimación de un total se obtiene con un procedimiento similar al anterior la siguiente fórmula:

$$n_o = \frac{Z^2\left(1-\frac{\alpha}{2}\right)N^2{S_U}^2}{\delta^2}$$

Ejemplo:

Una ciudad tiene 15,000 habitantes. Una encuesta es llevada a cabo para estimar el número total de gente en la ciudad quien desea que se construya un estadio ahí. ¿Qué tamaño de muestra es necesario para estimar el número total con un margen de 300 personas con 95% de confianza? Se sabe que la construcción de un estadio es una idea innovadora y se cree que al menos 60% de los residentes del pueblo les gustaría verlo construido.

Respuesta.

Es importante observar que la variable a estudiar tiene la siguiente estructura:

$$y_k = \begin{cases} 1 \; si \; la \; k - ésima \; persona \; quiere \; el \; estadio. \\ 0 \;\; otro \; caso. \end{cases}$$

En este caso y con la información que se da de que el 60% de personas quieren el estadio obtenemos que p (1-p) produce la varianza máxima con el valor de .6

Obtenida la varianza podemos simplemente sustituir los valores.

$$n_o = \frac{Z^2 \left(1-\frac{\alpha}{2}\right) \frac{N}{N-1} {S_U}^2}{\delta^2}$$

$$n_o = \frac{1.96^2 \left(\frac{15000}{14999}\right)[(0.6)(0.4)]}{300^2} = 2305.11$$

$$n = \frac{2305.11}{1 + 2305.11/15000} = 1998.060$$

Como se pudo observar en el ejemplo anterior al calcular el tamaño de la muestra es necesario conocer el valor de la varianza y especificar el error absoluto. Para lograr esto se puede usar uno de los siguientes métodos.

- Muestra Piloto
- Consultar estudios anteriores o datos disponibles de referencia
- Asumir distribución hipotética de la variable Y.

Un ejemplo de esto es un caso muy común en la vida rutinaria de cualquier persona, cuando se está organizando una fiesta muchas veces no sabemos la cantidad necesaria de bebidas, por ejemplo cervezas, se puede intuir este número a base de conocer a las personas que irán a la fiesta, es decir si la persona que está organizando la fiesta sabe que irá su amigo más borracho entonces tal vez debería comprar más cervezas, o si sabe que irá la persona que siempre lleva "gorrones" consigo entonces también se toma en cuenta al momento de decidir la cantidad de cervezas a comprar.

III.VII. Algoritmo MAS.

En caso de contar con una computadora, se puede usar el siguiente algoritmo para seleccionar una muestra MAS

1. Generar U_1, … , U_N variables aleatorias idénticamente distribuidas Unif (0,1)
2. La unidad k=1 es seleccionada si $U_1 < \frac{n}{N}$
3. Las unidades k=2,3,…, N son seleccionadas si $U_k < \frac{n-a_k}{n-k+1}$ donde a_k es el número de unidades seleccionadas entre los primeros k-1 elementos de la población.
4. Repetir paso 3 hasta que se obtenga la cantidad total de elementos de la muestra. Es decir cuando a_k=n

Ejemplo:

Supongamos que N=4 y n=3

Paso 1

U	U_k	$k \epsilon S$	
1	0.3	1	→
2	0.9	0	
3	0.1	1	→ Pertenecen a S
4	0.2	1	→

Paso 2

El primer elemento es seleccionado si $U_1 < \frac{n}{N} = \frac{3}{4} = .75$

$U_2 < \frac{n-a_k}{N-k+1} = \frac{3-1}{4-2+1} = \frac{2}{3}$

$U_3 < \frac{3-1}{4-3+1} = \frac{2}{2} = 1$

$U_4 < \frac{3-2}{4-4+1} = \frac{1}{1} = 1$

Capítulo IV. Muestreo aleatorio estratificado

El muestreo aleatorio estratificado es cuando a la población se divide en grupos de tal forma que sean homogéneos (respecto a la variable de interés) en su interior y heterogéneos entre sí, además de ser mutuamente excluyentes y exhaustivos[6]. Los estratos elegidos son con un conocimiento a priori de la variable de interés para que pueda resultar efectiva la división.

Se usa un muestreo estratificado cuando…

- Queremos reducir la probabilidad de obtener una mala muestra

  Suponga que se tiene una población de N habitantes, con la mitad hombres y la otra mitad mujeres. Si usáramos MAS se podría obtener una muestra con X% de hombres y Y% de mujeres, donde es probable que no sea una muestra representativa (dependiendo del estudio a realizar). Sin embargo si estratificamos nuestra población en 2 grupos (estratos) por género tendríamos una mejor muestra, garantizando representatividad de la muestra por género.

- Una muestra estratificada es posible de administrar de manera más conveniente a un menor costo

  En un estudio de Autoservicios podría usarse una entrevista personal para tiendas grandes y tal vez una entrevista telefónica para tiendas pequeñas, esto por la participación de las tiendas en el mercado.

- Si se realiza correctamente el muestreo estratificado, los resultados darán estimaciones con menos varianza, es decir, más precisas. Esto dado que los estratos son homogéneos por dentro y heterogéneos entre sí.

  Si se tiene un estudio de videojuegos, es probable que convenga estratificar por edades ya que se sabe que a la edad joven le gusta más este tipo de entretenimiento.

Este tipo de muestreo resulta particularmente útil cuando la variable en estudio toma valores promedio distintos en los diferentes grupos. Se recomienda por lo tanto usar cuando se conozca que la división por alguna variable (género, ubicación geográfica, nivel socioeconómico etc.) pueda ayudar al estudio.

[6] Que la unión de todos los estratos de como resultado el universo y que cada elemento pertenezca a sólo un estrato.

IV.I. Notación

La población se representará con U, N seguirá siendo el número de unidades dentro de U
La letra H es el número de estratos de U, siendo los estratos (U_h) mutuamente excluyentes, donde cada U_h tiene N_n unidades tal que

$$U = \bigcup_{n=1} U_h \qquad y \qquad \sum_{n=1}^{H} N_h = N$$

Dentro de cada estrato, una muestra aleatoria S_h c U_h de tamaño n_h es seleccionada de acuerdo al diseño muestral. Teniendo una probabilidad de ser seleccionada de P_h (S_h).
La selección en un estrato es independiente de la selección en otro estrato, resultando en el siguiente diseño muestral:

$$P(S) = P_1(S_1) \cdots P_h(S_h)$$

Siendo entonces la varianza

$$Var(\hat{t}) = \sum_{h=1}^{H} Var(\hat{t}_h)$$

Un diferente diseño muestral puede ser usado en cada estrato, en la práctica sin embargo, es un caso raro. Los tamaños de muestra en cada estrato pueden ser distintos entre sí.

IV.I.I. Propiedades de muestreo estratificado

Para el estrato U_h el total de un estrato es:

$$t_{U_h} = \sum_{k \in U_h} y_k$$

La media para el mismo estrato:

$$\bar{y}_{U_h} = \frac{1}{N_h} t_{U_h} = \frac{1}{N_h} \sum_{k \in U_h} y_k$$

La variabilidad entre cada estrato, es decir la varianza del estrato en la población está definida por:

$$S_{U_h}{}^2 = \frac{1}{N_h - 1} \sum_{k \in U_h} (y_k - \bar{y}_{U_h})^2$$

El tamaño relativo del estrato h está denotado por W_h y se define como:

$$W_h = {}^{N_h}/_{N} \qquad ; \qquad \sum_{h=1}^{H} W_h = 1$$

De lo anterior podemos decir que para el total de una población puede expresarse como:

$$t_u = \sum_{h=1}^{H} t_{U_h} = \sum_{h=1}^{H} \sum_{k \in U_h} y_k = \sum_{k \in U} y_k$$

$$\bar{y}_U = M_y = \frac{1}{N} \sum_{h=1}^{H} (N_h \bar{y}_{U_h}) = \sum_{h=1}^{H} W_h \bar{y}_{U_h}$$

Para la muestra S_h, se tiene la media muestral

$$\bar{y}_{S_h} = \frac{1}{n_h} \sum_{k \in S_h} y_k$$

Además la varianza muestral

$$S^2_{S_h} = \frac{1}{n_h - 1} \sum_{k \in S_h} (y_k - \bar{y}_{S_h})^2$$

IV.I.II. Horvitz – Thompson en muestreo estratificado

Por la definición del muestreo estratificado podemos decir lo siguiente:

$$\pi_k = P(k \in S) = P(k \in S_h)$$

Por lo mismo el estimador H-T del total en la población es la suma de los estimadores H-T en cada estrato donde

$$\hat{t}_{\pi_h} = \sum_{k \in S_h} \frac{y_k}{\pi_k}$$

Dado que los estratos son independientes la varianza del estimador H-T está representada de la siguiente forma:

$$V(\hat{t}_\pi) = \sum_{h=1}^{H} V(\hat{t}_{\pi_h}) = \sum_{h=1}^{H} \left(\sum_{k \in S_h} \sum_{l \in S_h} (\pi_{k,l} - \pi_k \pi_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \right)$$

Un estimador para la varianza en muestreo estratificado:

$$\hat{V}(\hat{t}_\pi) = \sum_{h=1}^{H} \hat{V}(\hat{t}_{\pi_h}) = \sum_{h=1}^{H} \left(\sum_{k \in S_h} \sum_{l \in S_h} \left(\frac{\pi_{k,l} - \pi_k \pi_l}{\pi_{k,l}} \right) \cdot \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \right)$$

Una observación importante en muestreo estratificado es que

$$\pi_{k,l} = \pi_k \pi_l > 0$$

Para $k \in U_h$ y $l \in U_{\grave{h}} \quad (h \neq \grave{h})$

IV.II. MAS Estratificado

Bajo el diseño muestral MAS estratificado, primero dividimos a la población en H estratos después se selecciona de manera independiente una muestra y en este caso nuestras probabilidades de inclusión son:

$$\pi_k = \frac{n_h}{N_h} \qquad para\ k \in U_h$$

$$\pi_{k,l} = \begin{cases} \frac{n_h}{N_h} \cdot \frac{n_{\grave{h}}}{N_{\grave{h}}} & si\ l \in U_{\grave{h}}\ y\ \ k \in U_h\ \ \grave{h} \neq\ h \\ \frac{n_h}{N_h} \cdot \frac{n_h - 1}{N_h - 1} & si\ \ k, l \in U_h \end{cases}$$

Sustituyendo éstos valores para el estimador H-T se llega a

$$\hat{t}_\pi = \sum_{h=1}^{H} N_h \bar{y}_{S_h}$$

$$\bar{y}_\pi = \frac{1}{N} \sum_{h=1}^{H} N_h \bar{y}_{S_h} = \sum_{h=1}^{H} W_h \bar{y}_{S_h} \quad W_h = \frac{N_h}{N}$$

Además $\bar{y}_\pi \neq \bar{y}_S = \frac{1}{n} \sum_{h \in H} n_h \bar{y}_{S_h}$ $\qquad S = \cup_{h=1}^{H} S_h$

Es decir en general $\bar{y}_\pi$no coincide con la media muestral, esto sólo para cuando se cumple que la fracción de muestreo para el estrato h es

$$\frac{n}{N} = \frac{n_h}{N_h}$$

Con la propiedad de independencia entre estratos, la varianza de H-T bajo MAS estratificado es:

$$V_{ST}(\hat{t}_\pi) = \sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_{U_h}^2$$

De igual forma si se divide por N^2 se llega a

$$V_{ST}(\bar{y}_\pi) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_{U_h}^2 \qquad llegando\ a\ V_{ST}(\hat{t}_\pi) = V_{ST}(\bar{y}_\pi N)$$

Estimadores insesgados de esta varianza para usos prácticos son los siguientes:

$$\hat{V}_{ST}(\hat{t}_\pi) = \sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_{S_h}^2$$

$$\hat{V}_{ST}(\bar{y}_\pi) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_{S_h}^2$$

Son insesgados a causa de que el estimador de H-T y su varianza son insesgados. Por lo tanto el intervalo de confianza para el estimador H-T queda de la siguiente forma:

$$\hat{t}_\pi \pm Z_{\left(1-\frac{\alpha}{2}\right)} \sqrt{\sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_{S_h}^2}$$

En este caso la aproximación normal se suele asumir cuando los elementos de la muestra de cada estrato (n_h) son grandes o cuando hay muchos estratos, es decir H es grande. También si la distribución de la variable de interés es aproximadamente normal.

IV.III. Asignación del tamaño de muestra en muestreo estratificado.

Como en MAS se comentó, la asignación del tamaño de muestra es crucial para los resultados y gastos de la encuesta. Sin embargo se tiene que considerar lo siguiente para MAS estratificado

$$\sum_{h=1}^{H} n_h = n$$

A continuación se enuncian algunos métodos de asignación.

IV.III.I. Asignación Igual

Como el nombre lo enuncia, éste método es el más cómodo y sencillo de usar pues da la misma cantidad de selección a cada estrato, es decir saca un promedio de muestra para cada estrato.

$$n_h = \frac{n}{H}$$

Sin embargo como es de esperarse este método es el que presenta las mayores varianzas para los estimadores. Además ¿a qué recurriría si se llegara a $n_h > N_h$?

IV.III.II. Asignación Proporcional

$$n_h = W_h n = \frac{N_h}{N} n$$

Como se puede observar, este método toma en cuenta el peso de cada estrato y lo respeta para la muestra del estrato, es decir que si un estrato es el 70% del total, así también será n_h respecto a n.
Dado que los tamaños de los estratos son generalmente conocidos, esta asignación es sencilla de usar y siempre puede ser usado.

Bajo esta asignación la varianza de la media…

$$V(\bar{y}_\pi) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_{U_h}^2$$

al sustituir el tamaño de muestra del estrato h

$$= \sum_{h=1}^{H} W_h^2 \left(\frac{N_h}{N_h\left(\frac{n}{N}\right)} - 1\right) S_{U_h}^2$$

$$= \sum_{h=1}^{H} \frac{N_h}{N} \left(\frac{1}{n} - \frac{1}{N}\right) S_{U_h}^2$$

$$V(\bar{y}_\pi) = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{h=1}^{H} W_h S_{U_h}^2$$

A partir de esto podemos obtener el valor para nuestra muestra general (n)

$$\underbrace{P\left(\left|\bar{y}_\pi - \mu_y\right|}_{} \leq \delta\right) = 1 - \alpha$$

$$Z_{\left(1-\frac{\alpha}{2}\right)} \sqrt{V(\bar{y}_\pi)} \leq \delta$$

$$Z_{\left(1-\frac{\alpha}{2}\right)}^2 V(\bar{y}_\pi) \leq \delta^2$$

$$Z_{\left(1-\frac{\alpha}{2}\right)}^2 \left[\left(\frac{1}{n} - \frac{1}{N}\right) \sum_{h=1}^{H} W_h S_{U_h}^2\right] \leq \delta^2$$

$$\frac{1}{n} \leq \frac{\delta^2 + Z_{\left(1-\frac{\alpha}{2}\right)}^2 \left(\frac{1}{N}\right) \sum_{h=1}^{H} W_h S_{U_h}^2}{Z_{\left(1-\frac{\alpha}{2}\right)}^2 \sum_{h=1}^{H} W_h S_{U_h}^2}$$

$$si\ n_0 = \frac{Z_{\left(1-\frac{\alpha}{2}\right)}^2 \sum_{h=1}^{H} W_h S_{U_h}^2}{\delta^2}$$

Si se divide todo entre δ^2

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

IV.III.III. Asignación Neyman

Esta asignación es óptima en el sentido en que proporciona la menor varianza. Sin embargo, requiere que las $S_{U_h}^2$ sean conocidas, en la práctica esto es poco común.

$$n_h = n\left[\frac{W_h S_{U_h}}{\sum_{h=1}^{H} W_h S_{U_h}}\right]$$

Se observa entonces que el tamaño de la muestra es directamente proporcional a S_{U_h}.

De la misma forma nuestra varianza queda de la siguiente manera.

$$V(\bar{y}_\pi) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_{U_h}^2$$

$$= \frac{1}{n}\left[\sum_{h=1}^{H} W_h S_{U_h}\right]^2 - \frac{1}{N}\sum_{h=1}^{H} W_h S_{U_h}^2$$

Para calcular n con esta varianza, queda de la siguiente manera

$$P\left(\left|\bar{y}_\pi - \mu_y\right| \leq \delta\right) = 1 - \alpha$$

$$Z_{\left(1-\frac{\alpha}{2}\right)}^2 V(\bar{y}_\pi) \leq \delta^2$$

$$Z_{\left(1-\frac{\alpha}{2}\right)}^2 \frac{1}{n}\left[\sum_{h=1}^{H} W_h S_{U_h}\right]^2 - Z_{\left(1-\frac{\alpha}{2}\right)}^2 \frac{1}{N}\sum_{h=1}^{H} W_h S_{U_h}^2 \leq \delta^2$$

$$\frac{1}{n} \leq \frac{\delta^2 + Z_{\left(1-\frac{\alpha}{2}\right)}^{\ 2} \frac{1}{N} \sum_{h=1}^{H} W_h S_{U_h}^2}{Z_{\left(1-\frac{\alpha}{2}\right)}^{\ 2} \left[\sum_{h=1}^{H} W_h S_{U_h}\right]^2}$$

$$n \geq \frac{\left[\sum_{h=1}^{H} W_h S_{U_h}\right]^2}{\frac{\delta^2}{Z_{\left(1-\frac{\alpha}{2}\right)}^{\ 2}} + \frac{\sum_{h=1}^{H} W_h S_{U_h}^2}{N}}$$

Una situación importante de mencionar es que usando este método puede resultar que la muestra en un estrato llegue a ser mayor que el total de unidades del estrato. Si esto ocurre una posible solución es que se haga en censo en todo el estrato y después se reajuste el cálculo para los demás estratos.

IV.III.IV. Situaciones donde la asignación óptima produce mejores ganancias de precisión.

- La población consta de elementos que varían mucho en tamaño.
- Las principales variables a medir están íntimamente relacionadas con los tamaños de los elementos de la población
- Se cuenta con una buena medición de tamaño de los elementos para establecer los estratos.

Ejemplo

Suponga un minero de datos de una empresa de telefonía móvil, quiere estimar el promedio de ventas mensuales de un nuevo producto. Suponga que dicha empresa sólo consta de 4 ciudades con distribuidores que venden el producto, entonces el minero de datos decide usar MAS estratificado, siendo cada ciudad un estrato distinto con los siguientes distribuidores

N_1=24, N_2=36, N_3=N_4=30; N=120

Se decide tomar una muestra de tamaño 20 de distribuidores. Dado que no existía información previa sobre las varianzas en los estratos, se decidió usar asignación proporcional.
Sustituyendo los valores…

$$n_1 = n\frac{N_1}{N} = 20 \cdot \frac{24}{120} = 4$$
$$n_2 = nW_2 = 6$$
$$n_3 = 5 = n_4$$

Los 20 distribuidores son seleccionando usando MAS dentro de cada ciudad (estrato). Las ventas del último mes para cada distribuidor fueron las siguientes

H=1	H=2	H=3	H=4
Y_k= 94	91	108	92
90	99	96	110
102	93	100	94
110	105	93	91
	111	93	113
	101		

Sacando las medias para cada estrato se obtiene lo siguiente:

H	$\bar{y}_{S_h}$
1	$\frac{(94+90+102+110)}{4} = 99$
2	100
3	98
4	100

La varianza muestral de cada estrato queda de la siguiente manera:

H	$S^2_{S_h}$
1	$\frac{1}{(4-1)}\sum_{k=1}^{4}(y_k - 99)^2 = 78.67$
2	55.6
3	39.5
4	112.5

$$\bar{y}_\pi = \frac{1}{N}\sum_{h=1}^{4} N_h \bar{y}_{S_h}$$

$$= \frac{1}{120}[24(99) + 36(100) + 30(98) + 30(100)] = 99.3$$

$$\hat{V}(\bar{y}_\pi) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{h=1}^{H} W_h S^2_{S_h}$$

$$= \frac{1}{120}\left(\frac{1}{20} - \frac{1}{120}\right)(24 \cdot 78.67 + 36 \cdot 55.6 + 30 \cdot 39.5 + 30 \cdot 117.5) = \mathbf{2.93}$$

Por lo tanto el intervalo de confianza al 95% de $\bar{y}_\pi$ queda de la siguiente forma

$$99.3 \mp 1.96\sqrt{2.93}$$

$$99.3 \mp 3.35$$

Si sólo hubiéramos ocupado MAS

$$\bar{y}_\pi = \bar{y}_S = 99.3$$

$$\hat{V}_{MAS}(\bar{y}_S) = \left(\frac{1}{n} - \frac{1}{N}\right) S_S^2 = 2.49$$

Entonces ¿porqué resultó menor la varianza para MAS que para estratificado con asignación proporcional? La respuesta es que para estos casos, las varianzas $S_{S_h}^2$ son usualmente mayores para instituciones grandes que para las pequeñas, lo cual hace ineficiente la asignación proporcional. De hecho en este caso donde la varianza sea grande dentro de algún estrato, es decir, que el estrato sea muy heterogéneo respecto a la variable de interés se recomienda usar la asignación Óptima o cambiar de muestreo.

IV.IV. Post-Estratificación

Cuando se sabe de una buena variable para poder estratificar la muestra pero no se tiene en el marco de muestreo se puede realizar primero el muestreo MAS y a la par de hacerlo se obtiene la variable de interés, por lo que al final se puede hacer una estimación de forma pseudo-estratificada como se tenía pensado desde el principio, lo anterior es el concepto del muestreo post-estratificado[7].

Usamos este tipo de estratificación cuando:

1. Es imposible estratificar la población U antes de seleccionar la muestra S, es decir el estrato al cual pertenece una unidad de muestreo no se conoce, sino hasta que la muestra ah sido hecha.
2. A pesar de eso, los N_h´s son conocidos (ó al menos los W_h´s) antes de seleccionar la muestra.
3. Estas características pasan generalmente cuando la variable usada para estratificar la población se refiere a características personas tales como género, edad, educación etc. por ejemplo no se conoce el género de un entrevistado sino hasta ser entrevistado, pero sabemos que W_1=W_2 =0.5

Después de tomar las observaciones, se puede asumir asignación proporcional[8]:

[7] Se debe de tener cuidado al hacer este tipo de muestreo ya que se pueden obtener varianzas muy pequeñas si se escogen de forma directa los grupos (estratos) una vez obtenida la muestra.
[8] Sólo si n es grande y n_h >29

$$\bar{y}_{POST} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_{S_h} = \sum_{h=1}^{H} W_h \, \bar{y}_{S_h}$$

$$\hat{V}_{POST}(\bar{y}_{POST}) = \left(1 - \frac{n}{N}\right) \sum_{h=1}^{H} W_h \frac{S_{U_h}^2}{h}$$

Es importante mencionar que esto es válido si

1. Se conoce $\frac{N_h}{N}$
2. n_h es razonablemente grande (n>30)

Ejemplo

Supóngase que un marco muestral enumera a todas las familias de una ciudad y que se necesita estimar la cantidad gastada en comida en un mes.

- Podemos usar el tamaño de las familias como una variable de estratificación, por lo tanto si la familia es grande implica que gastará más que una pequeña. Sin embargo no sabemos la cantidad de integrantes de las familias solo tenemos la numeración de éstas.
- Las fuentes oficiales reportan la distribución del tamaño de las familias en la ciudad.

# personas en la familia	**% de familias**
1	27.75
2	31.17
3	17.5
4	15.58
5+	10

- Por lo tanto no se puede estratificar por el tamaño de la familia antes de formar la muestra
- Si se puede tomar una muestra MAS de familias y registrar la cantidad gastada en comida y el tamaño de familias.

IV.V. Estratos en variables continuas

Ejemplo:

Suponga que los siguientes datos muestran la distribución de frecuencia del porcentaje de préstamos bancarios dedicados a préstamos de cualquier índole en una población con 13,435 bancos de algún país. La distribución es oblicua, con su moda en el extremo inferior.

(Préstamos de interés)% RESPECTO AL TOTAL	**f(y)**	**Acumulativo** $\sqrt{f(y)}$
0-5	3464	58.9
5-10	2516	109.1
10-15	2157	155.5
15-20	1581	195.3
20-25	1142	229.1
25-30	746	256.4
30-35	512	279
35-40	376	298.4
40-45	265	329.1
45-50	207	329.1
50-55	126	340.3
55-60	107	350.6
60-65	82	359.7
65-70	50	366.8
70-75	39	373
75-80	25	378
80-85	16	382
85-90	19	386.4
90-95	2	387.8
95-100	3	389.5

Donde la columna de **acumulativo** $\sqrt{f(y)}$ está formada de la siguiente manera

$58.9 = \sqrt{3464}$ $109.1 = \sqrt{3464} + \sqrt{2516}$ y así sucesivamente

Supóngase que se quieren 5 estratos.
Entonces cuando queremos formar estratos con este tipo de variables continuas, dada f (y), la regla es formar la acumulativa de $\sqrt{f(y)}$ y escoger las y_h de tal manera que éstas formen intervalos iguales en la escala de acumulativa $\sqrt{f(y)}$.

$$\frac{total\ de\ cum\ \sqrt{f(y)}}{5} = \frac{389.5}{5} = 77.9$$

Ahora sacamos los límites de nuestros estratos

Límite 1 = 77.9
Límite 2=2(77.9)=155.8
Límite 3=3(77.9)=233.7
Límite 4=4(77.9)=311.6

Si a nuestra acumulativa$\sqrt{\boldsymbol{f(y)}}$ la dividimos entre estos 4 límites nos darán 5 grupos que serán nuestros estratos. Buscamos el valor en la tabla más cercano al límite 1.
Se tiene entonces que el primer estrato es el intervalo de 0-5% de préstamos bancarios de interés respecto al total de sus préstamos pues la raíz de la cantidad de bancos que cumplen con esto (acumulativo $\sqrt{\boldsymbol{f(y)}}$) es 58.9 es decir que $\sqrt{3464} = 58.9$ siendo la cifra más cercana a 77.9 donde dicho estrato contiene 3464 bancos (N_1).
Teniendo entonces los siguientes estratos

ESTRATOS	**%PRÉSTAMOS**	**N_h**
1	0-5	3464
2	5-15	2516+2157=4673
3	15-25	2723
4	25-45	1899
5	45-100	676

Capítulo V. Muestreo por métodos indirectos.

En general en muestreo se tiene una dificultad universal, el error de muestreo. Hasta ahora se han mostrado procedimientos para estimadores lineales. Sin embargo existen otros estimadores donde se usa un desarrollo de Taylor para poder obtener una aproximación lineal del estimador del parámetro.

Estimación de varianza para estimadores no lineales.

Sea $\hat{\theta}$ un estimador no lineal de un parámetro de interés θ obtenido de una muestra de tamaño n. El plan es expresar dicho estimador como función de una serie de estimadores, ie, $\hat{\theta} = f(y_1, y_2, y_3, \dots, y_n) = \varphi(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \dots, \hat{\theta}_k)$. Esto para poder calcular las varianzas del estimador.

Usando el desarrollo de Taylor para este estimador en el punto $(\theta_1, \dots, \theta_k)$:

$$\varphi(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \dots, \hat{\theta}_k) = \varphi(\theta_1, \theta_2, \theta_3, \dots, \theta_k) + d\varphi(\hat{\theta}_1, \dots, \hat{\theta}_k)|_{(\theta_1,\dots,\theta_k)} + T_n$$

Donde dependiendo del entorno T_n puede ser o no ser eliminado y así obtener una aproximación válida. Si esto se cumple por lo tanto

$$\hat{\theta} - \theta \approx d\varphi(\hat{\theta}_1, \dots, \hat{\theta}_k)|_{(\theta_1,\dots,\theta_k)} = \sum_{r=1}^{k} \left(\frac{\partial \varphi(\hat{\theta}_1, \dots, \hat{\theta}_k)}{\partial \hat{\theta}_r} \right)_{(\theta_1,\dots,\theta_k)} (\hat{\theta}_r - \theta_r)$$

Finalmente elevando al cuadrado y sacando la esperanza se llega a una aproximación de la varianza.

$$Var(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2 \right] \approx$$

$$E\left[\left(\sum_{r=1}^{k} \left(\frac{\partial \varphi(\hat{\theta}_1, \dots, \hat{\theta}_k)}{\partial \hat{\theta}_r} \right)_{(\theta_1,\dots,\theta_k)} (\hat{\theta}_r - \theta_r) \right)^2 \right]$$

$$= \sum_{r=1}^{k} \sum_{l=1}^{k} \left(\frac{\partial \varphi(\hat{\theta}_1, \dots, \hat{\theta}_k)}{\partial \hat{\theta}_r} \right)_{(\theta_1,\dots,\theta_k)} \left(\frac{\partial \varphi(\hat{\theta}_1, \dots, \hat{\theta}_k)}{\partial \hat{\theta}_l} \right)_{(\theta_1,\dots,\theta_k)} Cov(\hat{\theta}_r, \hat{\theta}_l)$$

Para fines de este proyecto usaremos $R = \frac{\alpha}{\beta}$ como parámetro de interés y por lo tanto $\hat{R} = \frac{\hat{\alpha}}{\hat{\beta}}$.
Es decir $\hat{R} = (\hat{\alpha}, \hat{\beta})$, podemos entonces usar la varianza aproximada obtenida anteriormente llegando a

$$Var(\hat{R}) \approx \frac{1}{\beta^2}\left[Var(\hat{\alpha}) + R^2 Var(\hat{\beta}) - 2RCov(\hat{\alpha}, \hat{\beta})\right]$$

V.I. Estimadores de razón

En este tipo de estimadores se considera que a cada unidad k de la población se le asocian 2 variables numéricas Y_k y X_k. Si se tiene conocimiento "a priori" en toda la población, se puede usar este conocimiento para construir mejores estimadores.

Sea entonces Y el parámetro de interés, supongamos también que se conoce la variable auxiliar X en toda la población.

La fórmula general de los estimadores indirectos es:

$$f(\hat{Y}_G) = f(\hat{Y}) + b_0\left(f(X) - f(\hat{X})\right)$$

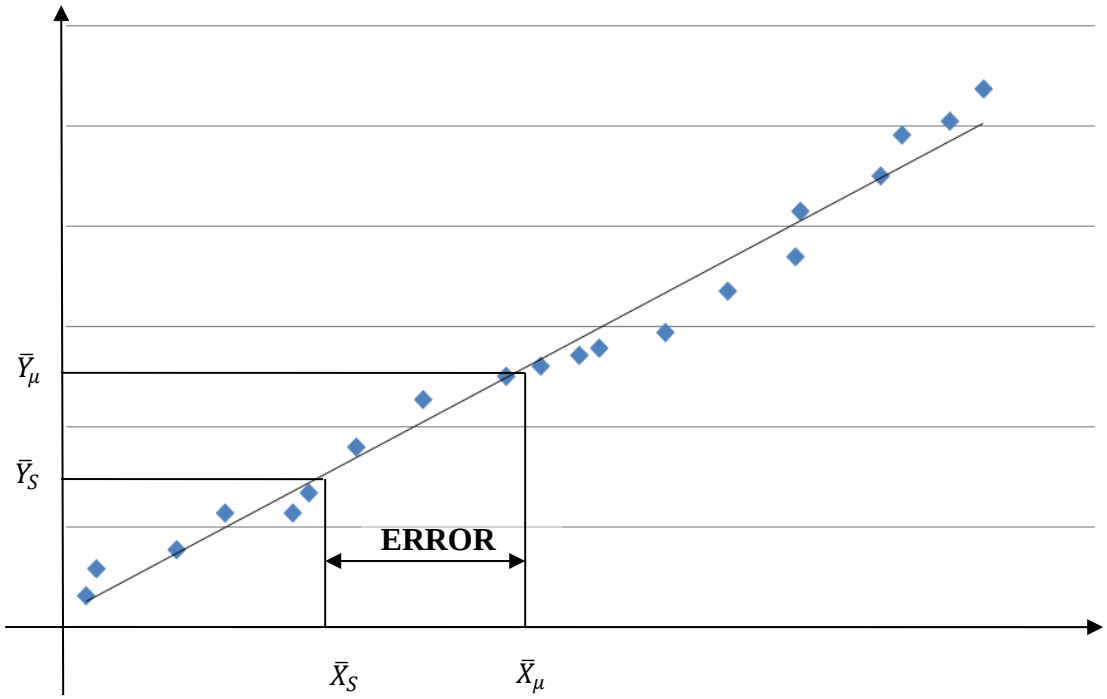
Con $\hat{Y}_G$ como el estimador indirecto de Y, $\hat{Y}$ y $\hat{X}$ los estimadores directos de Y y X. b_0 es un coeficiente de corrección. A partir de éste último, dependiendo de su valor es el tipo de estimador indirecto; Algunos de los casos más frecuentes son:

- $b_0 = 0$ Se tiene entonces que el estimador indirecto es igual al directo.
- $b_0 = 1$ se tiene $\hat{Y}_G = \hat{Y} + (X - \hat{X})$, es conocido como estimador de la diferencia.
- $b_0 = \frac{\hat{Y}}{\hat{X}}$ Este en particular es el estimador de razón
$$\hat{Y}_G = \hat{Y} + \frac{\hat{Y}}{\hat{X}}\left(X - \hat{X}\right) = \hat{R}X = \hat{Y}$$
- $b_0 = b$ Es el estimador de regresión.

Entonces si hablamos de estimadores de razón se tiene que en general no son insesgados, pero tienen varianza pequeña. Pero para esto es necesario conocer X ó $\bar{X}$ para poder calcularlo.

A pesar de que $\hat{R}$ es consistente se tiene que es sesgado, aunque bajo ciertas condiciones se cumple que sea insesgado.

1. $\hat{R}\ y\ X$ no se correlacionan en el muestreo.
2. La recta de regresión entre la variable de interés y la auxiliar pasa por el origen



Entonces

$$\frac{\bar{y}_U}{\bar{y}_S} = \frac{\bar{x}_U}{\bar{x}_S}$$

$$\hat{\bar{y}}_U = \bar{y}_S \frac{\bar{x}_U}{\bar{x}_S} = \hat{R}\bar{x}_U$$

Existen 2 aplicaciones de los estimadores de razón:

1. Se necesita estimar t_{ym} ó $\bar{y}_U$ bajo el supuesto de que y_k α x_k y además bajo el supuesto de que se conoce t_{xm} (ó $\bar{x}_U$) es decir se conoce el total y/o la media poblacional. No queremos realmente estimar R, pero lo estimamos para estimar el total de la variable Y (t_{ym}). En este tipo de estimaciones se aprovecha la existente correlación entre las variables X y Y en la población; mientras mejor es la correlación mejor la estimación.

   Por lo que

$$R = \frac{t_{y_\mu}}{t_{x_\mu}} = \frac{\frac{1}{N}\sum_{k\in U} y_k}{\frac{1}{N}\sum_{k\in U} x_k} = \frac{\bar{y}_U}{\bar{x}_U}$$

De esta manera, los estimadores del total y del promedio de y quedan definidos de la siguiente forma:

$$\hat{t}_{y_\mu} = \hat{R} t_{x_U}$$

$$\hat{\bar{y}}_\mu = \hat{R} \bar{x}_U$$

$$\hat{R} = \frac{\sum_{k \in S} \frac{y_k}{\pi_k}}{\sum_{k \in S} \frac{x_k}{\pi_k}}$$

2. No se conocen las x_i. Además, resulta de interés estimar a R. Además también nos sirve para estimaciones de proporciones si y_k toma valores 0 y 1, al igual que x_k

Podemos poner un ejemplo sencillo, hay que suponer que se necesita estimar el total de una población pero que no conocemos N, por lo que aún sabiendo la media no podemos usar la clásica estimación $t_y = N\bar{y}_U$. Pero con ayuda de otra variable podemos saber $N = t_x/\bar{x}_U$

En una fábrica que produce televisores se quiere estimar la cantidad total de televisores que fueron defectuosos en un mes (y), se tiene sólo la muestra de televisores (n) pero no se conoce el total de televisores que fueron fabricados (N); por lo mismo no se puede multiplicar por N a la proporción de televisores que resultaron defectuosos de la muestra. Sin embargo se sabe que hay una cantidad fija de trabajadores quienes fabrican los televisores, es decir, la cantidad de horas usadas para fabricar un televisor (x), y se sabe la cantidad de horas totales que trabajan, se puede entonces tomar esto como una variable auxiliar a la de interés. Si se multiplica la cantidad de horas por día trabajadas por los empleados por el total de días trabajados en el mes, se tiene el total de horas que usaron para fabricar a todos los televisores de la población total (t_x) y como se tiene el registro por televisor solo de la muestra de cuántas horas se usaron para fabricar cada televisor, se puede obtener el promedio de horas usadas para fabricar cada televisor de la muestra ($\bar{x}_S$). Por lo tanto se puede sacar el promedio de televisores defectuosos de la muestra ($\bar{y}_S$) y multiplicar por la razón de la variable auxiliar esto es:

$$\hat{t}_{y_U} = \bar{y}_S \frac{t_x}{\bar{x}_S}$$

La peculiaridad de este tipo de estimaciones es que hay veces que lo usamos y no nos damos cuenta pues el denominador de las estimaciones parece ser una constante, para saber cuándo usar este tipo de estimaciones tenemos que pensar si cambia el denominador al momento de cambiar la muestra, de ser esto correcto entonces se habla de estimaciones de razón.

Para calcular la varianza del estimador de razón se usa la definición de varianza de un estimador, es decir, como la esperanza del estimador menos su valor real al cuadrado.

$$Var(\hat{R}) = E\left[(\hat{R} - R)^2\right]$$

$$= E\left[\left(\frac{\bar{y}_s}{\bar{x}_s} - R\left(\frac{\bar{x}_s}{\bar{x}_s}\right)\right)^2\right]$$

$$= E\left[\left(\frac{\bar{y}_s - R\,\bar{x}_s}{\bar{x}_s}\right)^2\right]$$

Sea δ un número muy pequeño tal que

$$\delta\bar{x}_s = \bar{x}_s - \bar{x}_U$$

$$Var(\hat{R}) = E\left[\left(\frac{\bar{y}_s - R\,\bar{x}_s}{\bar{x}_U + \delta\bar{x}_s}\right)^2\right]$$

A continuación definimos f(θ) como:

$$f(\theta) = E\left[\left(\frac{\bar{y}_s - R\,\bar{x}_s}{\bar{x}_U + \theta\delta\bar{x}_s}\right)^2\right] = Var(\hat{R}) \iff \theta = 1$$

Si a esta expresión se le expresa como la expansión de Taylor en torno a cero se tiene que

$$f(\theta) = f(0) + \theta f'(0) + \frac{\theta^2 f''(0)}{2!} + \cdots$$

$$f(0) = E\left[\left(\frac{\bar{y}_s - R\,\bar{x}_s}{\bar{x}_U}\right)^2\right]$$

$$f'(0) = -2E\left[\frac{(\bar{y}_s - R\,\bar{x}_s)^2\delta\bar{x}_s}{{\bar{x}_U}^3}\right]$$

Cuando $\theta = 1$ se llega a que

$$Var(\hat{R}) = f(1) = E\left[\left(\frac{\bar{y}_s - R\,\bar{x}_s}{\bar{x}_U}\right)^2\right] - 2E\left[\frac{(\bar{y}_s - R\,\bar{x}_s)^2\delta\bar{x}_s}{{\bar{x}_U}^3}\right] + \cdots$$

La expansión entonces procede mediante potencias de $\delta\bar{x}_s$ siendo los términos que incluyen a $\delta\bar{x}_s$ cada vez más pequeños, por lo tanto si sólo se usa el primer término se tiene una buena aproximación.

$$Var(\hat{R}) \approx E\left[\left(\frac{\bar{y}_s - R\,\bar{x}_s}{\bar{x}_U}\right)^2\right] = \frac{1}{\bar{x}_U{}^2} E[(\bar{y}_s - R\,\bar{x}_s)^2]$$

$$\text{sea} \quad d_k = y_k - Rx_k$$

$$\bar{d}_s = \bar{y}_s - R\bar{x}_s$$

Así

$$E[(\bar{y}_s - R\,\bar{x}_s)^2] = E[\bar{d}_s{}^2]$$

Además con

$$Var(\bar{d}_s) = E\left(\bar{d}_s{}^2\right) - E^2(\bar{d}_s)$$

$$E(\bar{d}_s) = E(\bar{y}_s - R\,\bar{x}_s) = E\left(\bar{y}_s - \frac{\bar{y}_U}{\bar{x}_U}\,\bar{x}_s\right)$$

$$= E(\bar{y}_s) - \frac{\bar{y}_U}{\bar{x}_U} E(\bar{x}_s) = \bar{y}_U - \frac{\bar{y}_U}{\bar{x}_U}\bar{x}_U = 0$$

Por lo tanto la varianza quedaría

$$Var(\bar{d}_s) = E\left(\bar{d}_s{}^2\right) = E[(\bar{y}_s - R\,\bar{x}_s)^2]$$

Teniendo en cuenta que estamos usando MAS sabemos que

$$Var(\bar{y}_s) = \left(1 - \frac{n}{N}\right)\frac{S_U^2}{n} = \left(1 - \frac{n}{N}\right)\frac{1}{n}\frac{1}{N-1}\sum_{k\in U}(y_k - \bar{y}_U)^2$$

Por lo tanto aplicando lo anterior a $\bar{d}_s$

$$Var(\bar{d}_s) = \left(1 - \frac{n}{N}\right)\frac{1}{n}\frac{1}{N-1}\sum_{k\in U}(d_k - \bar{d}_U)^2$$

$$\bar{d}_U = \frac{1}{N}\sum d_k$$

$$= \frac{1}{N}\sum[\bar{y}_k - R\bar{x}_k]$$

$$= \frac{1}{N}\left[\sum \bar{y}_k - \frac{\sum \bar{y}_k}{\sum \bar{x}_k}\sum \bar{x}_k\right] = 0$$

Llegando entonces a

$$Var(\bar{d}_s) = \left(1 - \frac{n}{N}\right)\frac{1}{n}\frac{1}{N-1}\sum_{k\in U}(d_k)^2$$

Regresando con R

$$Var(\hat{R}) \approx \frac{1}{\bar{x}_U^2}Var(\bar{d}_s)$$

$$= \frac{1}{\bar{x}_U^2}\left(1 - \frac{n}{N}\right)\frac{1}{n}\frac{1}{N-1}\sum_{k\in U}(y_k - Rx_k)^2$$

Sin embargo como no sabemos el valor de R, para poder sacar una varianza práctica se usa

$$c_k = y_k - \hat{R}x_k$$

$$\hat{V}ar(\hat{R}) = \frac{1}{\bar{x}_S^2}\left(1 - \frac{n}{N}\right)\frac{1}{n}\frac{\sum_{k\in S}(y_k - \hat{R}x_k)^2}{n-1}$$

Sea ahora

$$S_e^2 = \sum_{k\in S}\frac{(y_k - \hat{R}x_k)^2}{n-1} \rightarrow \text{ Error cuadrado medio.}$$

Para el caso de $\hat{t}_y$

$$\hat{V}(\hat{t}_{y_U}) = \hat{V}(\hat{R}t_{x_U}) = N^2\left(1 - \frac{n}{N}\right)\frac{S_e^2}{n}$$

Prácticamente usamos una variable aleatoria fuertemente correlacionada con la variable de estudio para estimar valores deseados de nuestra variable que nos interesa.

Es importante observar que la estimación por razones funciona mejor si la línea que ajustan a los datos de las variables x, y pasan por el origen. Si éstos datos se ajustan mejor a una línea recta que no pasa por el origen convendría usar estimación por regresión (y=a+bx)

V.I.I. Estimadores de razón en muestreo estratificado.

Usamos estimadores de razón en muestreo estratificado cuando se tienen pocos estratos y/o los tamaños de muestra en cada estrato son grandes. Supone que las razones en cada estrato son similares, para este caso usamos el *estimador de razón por separado*.
El cual tiene los siguientes estimadores:

$$\hat{R}_S = \sum_{h=1}^{H} \frac{N_h}{N} \hat{R}_h$$

$$\hat{V}(\hat{R}_S) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \hat{V}(\hat{R}_h)$$

$$= \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{1}{\bar{x}_h^2}\left(1 - \frac{n_h}{N_h}\right)\frac{1}{n_h}\frac{\sum_{k\in S_h} (y_{k_h} - \hat{R}_h x_{k_h})^2}{n_h - 1}$$

Los sesgos de los estimadores de razón en cada estrato se suman, por lo que este estimador puede tener un sesgo muy grande, es por eso que conviene usarlo cuando los tamaños de muestra son grandes.

Si por lo contrario se tienen muchos estratos y/o los tamaños de muestra por cada estrato son pequeños, se usa el estimador de razón combinado.

Capítulo VI. Muestreo por conglomerados

El muestreo por conglomerados es una técnica similar en primera estancia al muestreo estratificado, sin embargo tiene aplicaciones y consecuencias distintas ya que frecuentemente este muestreo disminuye la precisión respecto a MAS considerando la precisión del muestreo estratificado.

Usamos muestreo por conglomerados entre otras cosas por los siguientes puntos

- La construcción de un listado que contenga a todos los elementos de la población puede ser difícil, o hasta imposible.

  por ejemplo, en una empresa de telefonía celular, es muy complicado y caro tener un listado de todos los clientes que pueden comprar un celular, sin embargo existe un listado de los canales de venta de la empresa.

- La dispersión geográfica de la población, o aparecer en agrupaciones como familias o escuelas. Por ejemplo si queremos una muestra de estudiantes podríamos usar como población objetivo escuelas y entrevistar a todos los estudiantes de las escuelas seleccionadas por el muestreo a usar un MAS sobre los estudiantes directamente, ya que esto puede provocar que tengamos que ir a una escuela a entrevistar sólo a un estudiante, lo cual lo hace lento, poco práctico y caro.

Por lo tanto se recomienda usar prácticamente por ser más práctico[9] en la vida real que otro muestreo. Se usa también cuando los conglomerados tienen promedios muy similares entre sí (que sean homogéneos).

En el muestreo por conglomerados a diferencia de otros muestreos tenemos que:

- Los individuos de la población solamente participan en la muestra si pertenecen a un conglomerado elegido en la muestra.
- La unidad primaria de muestreo no es necesariamente la unidad de observación, sino es la unidad secundaria del muestreo.
- Dado que no se tiene un marco muestral definido para la población objetivo las estimaciones son menos precisas, contrario al muestreo por estratos.

[9] Ya que logísticamente ayuda a que el encuestador no se tenga que mover mucho, provocando además que no se necesiten a muchas personas para levantar las encuestas.

VI.I. Notación

Dado que en conglomerados nuestros elementos observados no son nuestras unidades de interés sino que vienen siendo las unidades primarias y las unidades secundarias (dentro de las unidades primarias) nuestras unidades de observación.

Nivel primario.

y_{ij} Es la respuesta del elemento j dentro de la unidad primaria i.

N Número de unidades primarias en la población.

M_i Número de unidades secundarias en unidad primaria i

$k=\sum_{i=1}^{N} M_i$ Cantidad total de unidades secundarias en la población.

$t_i = \sum_{i=1}^{M_i} y_{ij}$ Total de respuestas dentro de la undad primaria i

$$t_U = \sum_{i=1}^{N} t_i$$

$$\bar{y}_U = \sum_{j=1}^{N} \frac{\sum_{i=1}^{M_i} y_{ij}}{k}$$

$S_t^2 = \frac{\sum_{i=1}^{N} \left(t_i - \frac{t_U}{N}\right)^2}{N-1}$ Varianza poblacional de las unidades primarias.

Nivel secundario:

$\bar{y}_{iU} = \sum_{j=1}^{M_i} \frac{y_{ij}}{M_i} = \frac{t_i}{M_i}$ Media poblacional de la unidad primaria i.

$S^2 = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{\left(y_{ij} - \bar{y}_U\right)^2}{k-1}$ Varianza de la variable respuesta y en toda la población.

$S_i^2 = \sum_{j=1}^{M_i} \frac{\left(y_{ij} - \bar{y}_{iU}\right)^2}{M_i - 1}$ Varianza poblacional de y dentro de la unidad primaria i.

m_i número de unidades secundarias en muestra de la unidad primaria i

$\bar{y}_i = \sum_{j \in S_i} \frac{y_{ij}}{m_i}$ media muestral para la unidad primaria i

$$\hat{t}_i = \sum_{j \in S_i} \frac{M_i\, y_{ij}}{m_i}$$ total estimado para la unidad primaria i.

$$\hat{t}_\pi = \sum_{j \in S} \frac{N}{n} \hat{t}_i$$ Estimador H-T del total de la población

$$S^2_{t_s} = \frac{1}{n-1} \sum_{i \in S} \left(\hat{t}_i - \frac{\hat{t}_\pi}{N} \right)^2$$ Varianza del estimador H-T

VI.II. Muestreo por conglomerados en 1 etapa

Este tipo de muestreo es donde si un conglomerado (unidad primaria) está en la muestra entonces todas sus unidades secundarias son seleccionadas. Bajo este esquema:

- En la población de N unidades primarias, la unidad i contiene M_i unidades secundarias.
- Extraemos una muestra MAS de las N unidades primarias y medimos la variable de interés para cada elemento ó unidad secundaria de la unidad primaria.
- Para este caso $M_i = m_i$

Se tiene una muestra MAS de n observaciones y para todos los elementos de la unidad primaria i. Entonces $\hat{t}_s = \frac{1}{n} \sum_{i \in S} t_i$ es el estimador del promedio de cada total por conglomerado, por ejemplo si se desea estimar el ingreso total de las familias de una comunidad, las observaciones individuales y_{ij} son los ingresos de cada persona j dentro de la familia i, entonces $t_i = \sum_{j=1}^{M_i} y_{ij}$ es el ingreso total.

$$\hat{t}_\pi = \sum_{i \in S} \frac{N}{n} t_i = \frac{N}{n} \sum_{i \in S} \sum_{j=1}^{M_i} y_{ij}$$

Es el estimador del ingreso total, por lo tanto se pueden aplicar las fórmulas de MAS a $\hat{t}_\pi$ pues se tiene una muestra MAS de n unidades extraídas de una población de N unidades; entonces:

$$V(\hat{t}_\pi) = N^2 \left(1 - \frac{n}{N}\right) \frac{S_t^2}{n}$$

$$S_t^2 = \frac{\sum_{i=1}^{N} \left(t_i - \frac{t_U}{N} \right)^2}{N-1}$$

$$\hat{V}(\hat{t}_\pi) = N^2 \left(1 - \frac{n}{N}\right) \frac{S_{t_s}^2}{n}$$

$$S_{t_s}^2 = \frac{1}{n-1} \sum_{i \in S} \left(t_i - \frac{\hat{t}_\pi}{n} \right)^2$$

Para estimar el ingreso promedio por persona, dividimos el total estimado entre el número de personas

$$\bar{y}_s = \frac{\hat{t}_\pi}{k}$$

Donde k es el número total de unidades secundarias o unidades de observación.

$$k = \sum_{i=1}^{N} M_i \qquad \hat{V}(\bar{y}_s) = \frac{1}{k^2}\hat{V}(\hat{t}_\pi) = \frac{1}{k^2}N^2\left(1 - \frac{n}{N}\right)\frac{S_{t_s}^2}{n}$$

VI.III. Muestreo por conglomerados en 2 etapas

Se suele usar este tipo de muestreo cuando:

1. Los elementos de un conglomerado son muy similares, por lo tanto analizar a todas las subunidades podría ser un desperdicio de recursos y tiempo.
2. Puede ser muy caro medir una unidad secundaria respecto a medir una unidad primaria, podría ser más barato tomar una submuestra dentro de cada unidad primaria.

Al decir 2 etapas nos referimos entonces a

1. Se elige una muestra aleatoria simple S de n unidades primarias de la población de N unidades primarias.
2. Se selecciona una muestra aleatoria simple de unidades secundarias dentro de cada unidad primaria ya seleccionada. Es una muestra de m_i elementos del conglomerado i, se denota como S_i.

En este tipo de muestreo, dado que no observamos todas las unidades secundarias en cada unidad primaria estimamos los totales individuales de la unidad primaria con lo siguiente

$$\hat{t}_i = \sum_{j \in S_i} \frac{M_i}{m_i} y_{ij} = M_i \sum_{j \in S_i} \frac{y_{ij}}{m_i} = M_i \bar{y}_i$$

Por lo tanto un estimador insesgado del total de la población es el estimador H-T que en este caso

$$\hat{t}_\pi = \sum_{i \in S} \frac{N}{n}\hat{t}_i = \sum_{i \in S} \frac{N}{n} \sum_{j=1}^{S_i} \frac{M_i}{m_i} y_{ij} = \sum_{i \in S}\sum_{j \in S_i} \frac{N}{n}\frac{M_i}{m_i} y_{ij}$$

En este tipo de muestreo las $\hat{t}_i$ son variables aleatorias. Por lo tanto, la varianza de $\hat{t}_\pi$ tiene 2 componentes:

- La variabilidad entre unidades primarias
- La variabilidad de las unidades secundarias dentro de las unidades primarias.

Por lo tanto

$$\hat{V}(\hat{t}_\pi) = N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n} + \frac{N}{n}\sum_{i\in S}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{S_i^2}{m_i}$$

En muchos casos, $\frac{N}{n}$ será pequeño respecto a N^2 de tal manera que la contribución del segundo término será despreciable en comparación con la contribución del primer término.
La media de la población se estima de la siguiente manera:

$$\hat{\bar{y}} = \frac{\hat{t}_\pi}{k}$$

Su error estándar es $\widehat{EE}(\hat{\bar{y}}) = \frac{\widehat{EE}(\hat{t}_\pi)}{k}$

Si los tamaños de los conglomerados son muy diferentes entre sí, el primer componente de la varianza es grande, pues es afectado por las variaciones de tamaño de las unidades primarias (las M_i) y por las variaciones de las $\bar{y}_i$. Este caso se puede representar mediante el siguiente ejemplo:

Supóngase que se quiere estimar el número promedio de patas que tienen los cachorros que viven en las perreras de la ciudad, la ciudad cuenta con 2 únicas perreras. Las que llamaremos perrera A y perrera B las cuales alojan a 30 y 10 cachorros respectivamente.
Elegimos una perrera con probabilidad de ½ después seleccionamos la perrera, elegimos 2 cachorros al azar y usamos $\hat{\bar{y}}$ para estimar el promedio de patas.
Ahora supóngase que elegimos la perrera A. cada uno de los 2 cachorros en la muestra tienen 4 patas de tal manera que

$$\hat{t}_A = M_i\sum_{j\in S_i}\frac{y_{ij}}{m_i}$$

$$= \frac{30}{2}(4) + \frac{30}{2}(4) = 120$$

$$\hat{t}_\pi = \frac{N}{n}\sum_{i\in S}\boxed{\hat{t}_A} \rightarrow \hat{t}_i$$

$$\hat{t}_\pi = 2(120) = 240$$

Por lo tanto

$$\hat{\bar{y}} = \frac{\hat{t}_\pi}{k} = \frac{240}{40} = 6$$

Además si se hubiera tomado la perrera B

$$\hat{t}_B = M_i \sum_{j \in S_i} \frac{y_{ij}}{m_i} = 40$$

$$\hat{t}_\pi = 2\hat{t}_B = 2(40) = 80$$

$$\hat{\bar{y}} = {}^{80}/_{4} = 2$$

Pero es claro que esto es incorrecto, el estimador insesgado desde el punto de vista matemático es

$$\frac{(6+2)}{2} = 4$$

De tal manera que el promedio de todas las muestras posibles produce el número correcto; la mala calidad del estimador se refleja en la enorme varianza

$$\hat{V}(\hat{t}_\pi) = 2^2 \left(1 - \frac{1}{2}\right)\frac{S_t^2}{1} + \frac{2}{1}\sum_{i \in S}\left(1 - \frac{m_i}{M_i}\right) M_i^2 \frac{S_i^2}{m_i} = 6400$$

$$\sqrt{\hat{V}(\hat{t}_\pi)} = 80$$

En general el estimador insesgado del total de la población $(\hat{t}_\pi)$ es ineficiente si los tamaños de los conglomerados son distintos y si t_i es aproximadamente proporcional a M_i. La varianza de $\hat{t}_\pi$ depende de la varianza de t_i y dicha varianza puede ser grande si los M_i son distintos.
Por lo que para este caso este tipo de muestreo por conglomerados es ineficiente. Veremos el mismo ejemplo con el muestreo por conglomerados en 2 etapas probabilidad proporcional al tamaño en la primera y MAS en la segunda.

Capítulo VII. Muestreo con probabilidad proporcional al tamaño

En el muestreo con probabilidad proporcional al tamaño[10], se asume que se conocen los valores de una variable auxiliar para cada elemento de la población la cual llamaremos z_k la cual es conocida también como "medida de tamaño del k-ésimo elemento de la población". Se espera una reducción en la varianza si la variable z y la variable de interés "y" están altamente correlacionadas.

Bajo este tipo de muestreo, las probabilidades de inclusión varían de acuerdo al tamaño relativo de los elementos de la población. El tamaño relativo de la unidad k se mide por:

$$p_k = \frac{z_k}{\sum_{k \in U} z_k = T_z}$$

En general la variable z_k se puede usar en muestreo por conglomerados, siendo z_k el número de elementos del conglomerado k.

VII.I. Pasos para una muestra PPT-Sistemática

1. Definir el intervalo de muestreo q= $\frac{T_z}{n}$
2. Seleccionar aleatoriamente un número entre 1 y q. Sea q_o
3. Definir los siguientes totales acumulados: $G_k=\sum_{j=1}^{k} z_j \qquad k = 1,2, \dots, N \ (G_N = T_z)$
4. Los n números seleccionados para estar en la muestra son q_0, $q_0 + q$, $q_0 + 2q$, … , $q_0 + (n-1)q$
5. Los elementos que forman parte de la muestra son aquellos cuyos G_k es el número mayor o igual más cercano correspondiente número seleccionado en el paso 4.

Entonces la probabilidad de inclusión del k – ésimo elemento es

$$\pi_k = n\frac{z_k}{T_z} = n \cdot p_k$$

Algunos ejemplos de variables auxiliares Z:

- Muestreo de caja de algún producto (galletas por ejemplo), Z_k como número de galletas en la caja k
- En una muestra de tiendas de autoservicio, Z_k como el número de trabajadores en la tienda k.

[10] También abreviado como PPT por su primera letra de cada palabra.

VII.II. Muestreo por conglomerados en 2 etapas con PPT en la primera y MAS en la segunda.

La probabilidad de elegir una unidad primaria con el método PPT es

$$\pi_k = n\frac{z_k}{T_z}$$

Si dentro de cada unidad primaria, consideramos una muestra MAS de unidades secundarias se tiene:

$$\pi_{ij} = \pi_i \pi_{j|i} = n\frac{z_i}{T_i}\frac{m_i}{M_i}$$

Por lo tanto el estimador horvitz Thompson es:

$$\hat{t}_\pi = \sum_{k\in S}\frac{T_z}{nz_k}\hat{t}_k$$

$$= \sum_{k\in S}\frac{T_z}{nz_k}\sum_{j\in S_k}\frac{M_k}{m_k}y_{jk}$$

$$= \sum_{k\in S}\sum_{j\in S_k}\frac{T_z}{nz_k}\frac{M_k}{m_k}y_{jk}$$

$$= \sum_{k\in S}\sum_{j\in S_k} w_{jk}\, y_{jk}$$

Para estimar la varianza de H-T se puede considerar sólo la variabilidad entre las unidades primarias y muestreo con reemplazo, teniendo la fórmula

$$\hat{V}(\hat{t}_\pi) = \frac{1}{n}\sum_{k\in S}\frac{\left(\frac{\hat{t}_k}{p_k} - \hat{t}_\pi\right)^2}{n-1}$$

Teniendo en cuenta este muestreo podemos regresar al ejemplo de las perreras A y B. en caso de elegir la perrera A

$$\hat{t}_\pi = \frac{1}{n}\sum_{k\in S}\frac{\hat{t}_k}{\frac{z_k}{T_z}} = \frac{1}{1}\left[\frac{\hat{t}_1}{\frac{30}{40}}\right]$$

$$\hat{t}_1 = \sum_{k\in S}\frac{M_1}{m_1}y_1 = \frac{(30)4}{2} + \frac{(30)4}{2} = 120$$

$$\hat{t}_\pi = 160$$

Y por lo tanto $\bar{y}_k = \frac{160}{4} = 4$

Si en cambio se toma la perrera B

$$\hat{t}_\pi = \frac{1}{1}\left[\frac{\hat{t}_2}{\frac{10}{40}}\right]$$

$$\hat{t}_2 = \frac{(10)4}{2} + \frac{(10)4}{2} = 40 \qquad \hat{t}_\pi = 160$$

Además la varianza del estimador $\hat{\bar{y}}_s$ es cero.

Conclusión

Los diseños muestrales mostrados y explicados en este trabajo no son los únicos ya que existen muchos más y mezclas entre ellos, se puede combinar por ejemplo conglomerados con muestreo estratificado.

Se mostró que el muestreo aleatorio simple es el más sencillo de todos pero es la base para la mayoría de diseños más complejos. Estimadores como Horvitz – Thompson se reducen a una forma muy simple e intuitiva aplicando este diseño provocando su facilidad de aplicación en la vida real.

El haber investigado más a fondo el tema de muestreo para la generación de esta tesina me mostró que el muestreo cada vez es está tomando más fuerza en la estadística, ya que cada vez tiene más demostraciones teóricas que le otorga más veracidad a los resultados del mismo. Muchos estudios cada vez requieren más esta herramienta para poder probar su impacto en el mercado como ejemplo. Beneficiando así a el actuario pues se abren más ofertas laborales en esta rama.

Pude observar también que muchos otros trabajos y análisis tienen como base una muestra para su población de estudio, y también comprendí mejor la importancia que tiene el generar una buena muestra ya que puede provocar sesgos en resultados y como consecuencia conclusiones falsas.

Por otro lado y en lo personal también le vi utilidad al muestreo incluso cuando se cuenta con toda la población. Por ejemplo en mi vida laboral pude entender este hecho ya que hay veces que la población total son millones y millones de registros en una base, por lo que poner a trabajar un modelo Multivariado con una cantidad de este tamaño puede provocar mucho tiempo de procesamiento. Al usar muestras representativas pude ver de manera significativa la reducción de tiempo para la operación y generación de resultados de los modelos.

El trabajo presentado aquí tiene la ventaja de que cubre de manera óptima el temario de Muestreo que se imparte en la carrera de Actuaría en la FES Acatlán, por lo que creo que se cumplió el objetivo de ser otra herramienta para ayudar al estudiante.

10.- Bibliografía.

Sharon L. Lohr. Muestreo: Diseño y análisis. Editorial Thomson.

César Pérez López. Muestreo estadístico: conceptos y problemas resueltos. Editorial Pearson-Prentice Hall, 2005.

Hugo Andrés Gutiérrez Rojas. Estrategias de muestreo. Diseño de encuestas y estimación de parámetros. Universidad Santo Tomás. Bogotá Colombia 2009.

Carl-Erik Särndal, Bengt Swensson, Jan Wretman. Model assisted suvery sampling. Editorial Springer. 2003.

Francisco Sánchez Villareal. Introducción al muestreo probabilístico para encuestas. [Archivo electrónico PDF]. Febrero 2007.

Risto Lehtonen y Erkki Pahkinen. Practical methods for design and analysis of complex surverys (second edition). Editorial John Wiley & Sons. Inglaterra 2004.

Patricia Romero Mares. Muestreo [material gráfico proyectable]. 76 diapositivas.

Hanwen Zhang & Hugo Andrés Gutiérrez Rojas. Teoría Estadística: aplicaciones y métodos. Universidad Santo Tomás. Bogotá Colombia. 2010.

Cassel, C. M., J. K. Wretman, and C. E. Särndal. Foundations of Inference in Survey Sampling. J. Wiley. New York. 1997.