Universidad Central de Venezuela
Facultad de Ciencias Económicas y Sociales
Escuela de Estadística y Ciencias Actuariales

ANÁLISIS MUESTRAL

Harold David Martín-Caro Malavé
Profesor de la Escuela de Estadística y
Ciencias Actuariales
Universidad Central de Venezuela

2006

Universidad Central de Venezuela
Facultad de Ciencias Económicas y Sociales
Escuela de Estadística y Ciencias Actuariales

ANÁLISIS MUESTRAL

Harold David Martín-Caro Malavé
Profesor de la Escuela de Estadística y
Ciencias Actuariales
Universidad Central de Venezuela

Trabajo presentado ante la ilustre Universidad Central de Venezuela para ascender a la categoría de Profesor Agregado

2006

Para el desarrollo de éste trabajo agradezco la formación académica y profesional a Luis Montero, Luis Armando Rodríguez, Guillermo Ramírez y Adelmo Fernández, así como el apoyo de mi esposa y mis padres.

A mi familia, especialmente a mi hermana María Fernanda, mi luz. A la memoria de mis abuelas Julia y Cristina.

Archivado por: Arayan
Nombres y Apellidos
Fecha: 14/11/07 Folio Nº 355

UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE CIENCIAS ECONOMICAS Y SOCIALES
ESCUELA DE ESTADISTICA Y CIENCIAS ACTUARIALES

ACTA

Quienes suscriben, Guillermo Ramirez, Alberto Camardiel e Irene Gurrea, miembros del Jurado designado al efecto por el Consejo de la Facultad de Ciencias Económicas y Sociales y por el Consejo de Desarrollo Científico y Humanístico de la Universidad Central de Venezuela, para examinar el Trabajo de Ascenso presentado por el Profesor **Harold David Martín-Caro Malavé**, cédula de identidad Nº 6.461.840, bajo el título "ANALISIS MUESTRAL" a los fines de su ascenso en el escalafón docente universitario a la categoría de AGREGADO, dejan constancia de lo siguiente:

1.- Leido como fue dicho Trabajo por cada uno de los miembros del Jurado, se fijó el día 4 de Julio de dos mil siete, a las 11am para que el autor lo defendiera en forma pública, lo que éste hizo en el Salón de Reuniones del Consejo de Escuela, mediante un resumen oral de su contenido, luego de lo cual respondió a las preguntas que le fueron formuladas, todo ello conforme a lo dispuesto en el Artículo 96 del Reglamento del Personal Docente y de Investigación de la Universidad Central de Venezuela.

2.- Finalizada la defensa del Trabajo de Ascenso, el Jurado decidió por UNANIMIDAD, de acuerdo con el Artículo 97 del Reglamento citado, el ADMITIRLO, por considerar sin hacerse solidario de las ideas expuestas por el autor, que se trata de un trabajo personal que significa un aporte a la materia por cuanto presenta un enfoque alternativo para la obtención de estimadores y sus varianzas correspondientes en diferentes diseños muestrales, constituyendo una contribución novedosa para el estudio de la teoría del muestreo, todo de conformidad con lo pautado en los Artículos 77 y siguientes del Reglamento de Personal Docente y de Investigación de la Universidad Central de Venezuela.

En fe de lo cual se levanta la presenta Acta, en Caracas a los 4 días del mes de Julio de dos mil siete, dejándose también constancia de que conforme al Artículo 93 del Reglamento citado, actuó como Coordinador del Jurado el Profesor Guillermo Ramirez.

Por el Consejo de la Facultad
Guillermo Ramirez (Coordinador)
Profesor Titular
C.I. 3.609.750

A. Camardiel
Por el Consejo de la Facultad
Alberto Camardiel
Profesor Titular
C.I. 2.135.649

Por el Consejo de Desarrollo
Científico y Humanístico
Irene Gurrea
Profesora Agregada
C.I. 2.993.540

INDICE

INDICE DE TABLAS

INDICE DE FIGURAS

PRÓLOGO

El uso del muestreo se ha expandido cada vez con mayor fuerza, incluso en disciplinas donde tiempo atrás era impensable usarlas, como pueden ser los casos de los censos de población. La preocupación que por el muestreo han tenido diferentes profesionales, ha ido en aumento, bien sean estadísticos, profesionales relacionados y hasta cualquier profesional en general; incluso gente común se preocupa por conocer, de manera general, algunos aspectos del muestreo.

Sin embargo, no todos consiguen aprender o profundizar en su búsqueda, quedando sin los conocimientos necesarios para poder enfrentar un proyecto específico, dando lugar a estudios faltos de credibilidad.

Particularmente, tengo la fortuna de haber trabajado en el Programa Censal correspondiente al XII Censo General de Población y Vivienda de Venezuela, que se levantó en Octubre de 1990; como Jefe de la Unidad de Muestreo. Las responsabilidades de dicha Unidad consistían en diseñar todas las muestras que se incluían en el Programa, dentro de las cuales resaltan el diseño de la muestra para el levantamiento del cuestionario ampliado, en todas sus fases, desde la recolección, hasta la expansión y cálculo de errores; también cabe destacar el diseño del Plan de Control de Calidad de la Codificación de los cuestionarios.

También he tenido la oportunidad de dictar clases de muestreo en la Escuela de Estadística de la Universidad Central de Venezuela, además de otros estudios e investigaciones sobre el tema.

Durante mi experiencia, he podido observar un tabú en cuanto al muestreo, por quienes no lo conocen y por quienes lo conocen poco. El muestreo, como muchas disciplinas, requiere de un estudio y práctica, nadie se hará muestrista leyendo un libro o tomando un curso, además de eso, tendrá que practicarlo, ejercerlo, equivocarse y corregirse, hasta llegar a dominarlo. Claro está que mucha gente, que no quiere hacerse muestrista, también requiere de cierto dominio sobre su terminología y algunas técnicas.

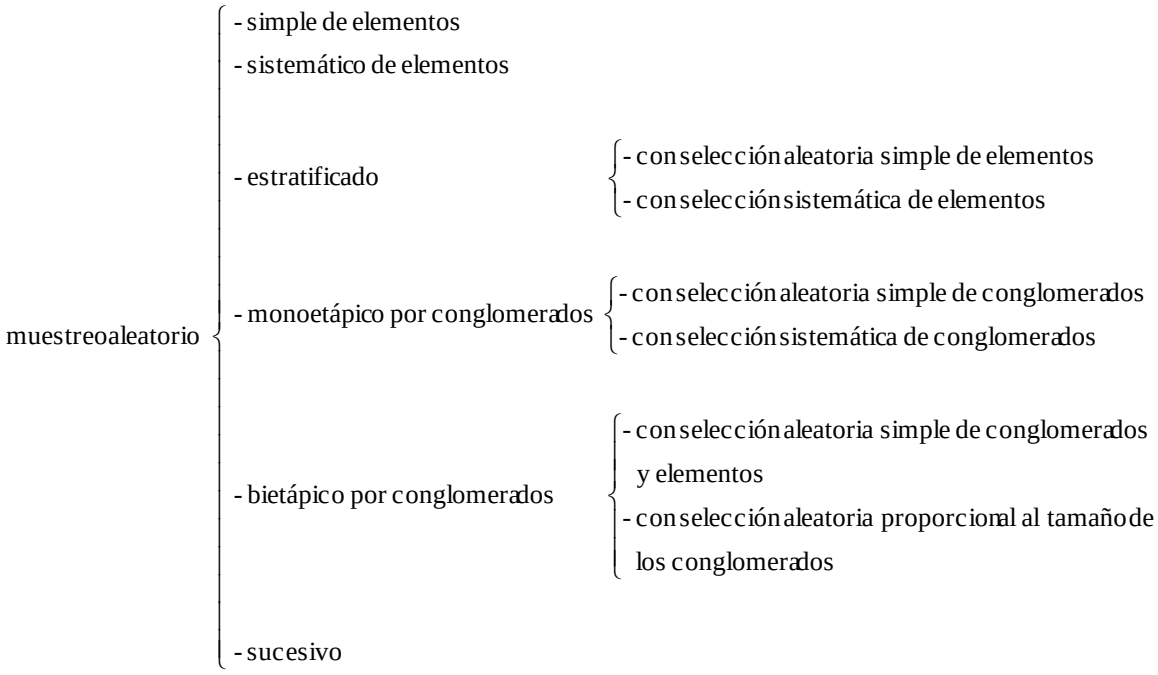
Este trabajo encontrará utilidad en varios públicos; i) los que se inician, bien para aquellos que continuarán profundizando, como para los que quieren sólo un enfoque e idea general; ii) los usuarios intermedios, que pueden diseñar o participar en el diseño de muestras pequeñas, y necesitan ciertas fórmulas de estimadores y varianzas; y iii) los estudiosos del muestreo, aquellos que el muestreo ocupa parte importante de sus trabajos y necesitan conocerlo a fondo. No sustituirá a textos clásicos como los de Yates; Deming; Cochran; Hansen, Hurwitz y Madow, Sukhatme, Sampford, Kish, Murthy, Raj, Cansado, Azorín y Sánchez-Crespo; pero a través del enfoque planteado, puede ayudar a entenderlos y servir de complemento.

Igualmente intenta hacer comprender a todos aquellos interesados en el uso del muestreo, su importancia, sus principios y sus axiomas, primero para conocerlo más a fondo, segundo para saber interpretarlo y tercero para utilizarlo.

En el capítulo 1, se trata de los principios, fundamentos y filosofía del muestreo, así como los conceptos y términos que se usarán a lo largo del mismo. Este primer capítulo está dirigido a cualquier persona que desee o necesite de algunas nociones elementales de concepción del muestreo, y por supuesto, para los más experimentados. En los capítulos siguientes se trabaja cada diseño muestral, y se pueden dividir en una primera parte en la que se deducen los estimadores, partiendo de su forma más básica, aplicando los principios mencionados en el primer capítulo. La misma se orienta a las personas conocedoras como soporte de sus conocimientos, así como a todos los interesados, y sólo se requiere conocimiento matemático básico, más que todo para entender los símbolos utilizados, en cambio se necesita un poco más de sentido común, en lo que respecta a construcción de indicadores, valores poblacionales y estimadores.

En la segunda parte de cada capítulo, se deduce y desarrolla la varianza de cada estimador. Esta es la parte más complicada del trabajo, y está dirigida a los conocedores o interesados con conocimiento más avanzado de matemáticas. Para el que no esté interesado, puede tomar sólo las fórmulas finales de dichos desarrollos.

Se tratarán los diseños muestrales bajo la estructura del siguiente cuadro sinóptico:



Además se incluye el estudio de subpoblaciones y de los estimadores indirectos, como son de razón, de regresión lineal y de razón al tamaño, éste último sólo para el caso de conglomerados.

Espero que este esfuerzo sirva para comprender mejor el muestreo y así hacer un mejor uso de él, con el objetivo de obtener mejores estimaciones, así como por el simple hecho de usarlo responsablemente.

Harold D. Martín-Caro M.

1.- INTRODUCCIÓN AL MUESTREO

Hace miles de años se realizaron los primeros censos de población. Para recaudación de impuestos se requería saber cuántas personas trabajaban; cuando había una epidemia, se hacían conteos de personas afectadas; se dice que Herodes hizo un conteo de niños en sus dominios, aunque con fines perversos. Hoy en día es cada vez más necesario realizar estos estudios. Los gobiernos los necesitan para conocer la realidad social, económica, industrial, comercial, agrícola y educativa de su población, y para elaborar, orientar y hacer seguimiento de sus programas y políticas. Las empresas los necesitan para conocer la aceptación de sus productos, sus ventas, las preferencias de los consumidores, entre otras. Sin embargo pareciera cumplirse una suerte de "ley de insatisfacción de información", además bien justificada, que a medida que se tiene más información, más se necesita para seguir profundizando en el problema.

De manera que los estudios e investigaciones se han hecho, además de necesarios, más complejos. Esta necesidad ha traído como consecuencia el surgimiento y desarrollo de una serie de técnicas y disciplinas, entre las cuales se encuentra el muestreo.

1.1.- Breve historia del muestreo

Los primeros usos del muestreo fueron de manera empírica e informal, como por ejemplo cuando se quería diezmar a una unidad militar como castigo, se seleccionaban aleatoriamente soldados. En la reunión de Berna del Instituto Internacional de Estadística de 1895, Kiaer, Director de la Oficina Central de Estadística de Noruega, presentó su trabajo "Observations et experiences concernant les denombrements representatifs", donde se usaba el término "método representativo" en contraposición a la "investigación exhaustiva", en lo que se reconoce como el primer trabajo donde se trata el muestreo de una manera seria. Además de dicho trabajo, Cheysson y otros proponían el "método monográfico", que investigaba detalladamente un fenómeno único, seleccionando "unidades típicas".

El método de Kiaer fue rechazado en aquella reunión, pero en su reunión de Berlín (1903), recomendó que se aplicase y fuese sometido a observación. En 1924 es nombrada una comisión para el estudio de estos métodos, y estuvo integrada por Jensen, Bowley, Gini, March, Stuart y Zicek, pero es en la reunión de Roma (1926), cuando se reconocen las grandes ventajas del método.

Muchos estadísticos opusieron resistencia al método, por temor a que las muestras pudieran sustituir a los censos, como George Von Mayr que decía "Pas de calcul là ou l'observation peut être faite" (*no se pueden calcular las observaciones*).

Cabe destacar los trabajos de Carroll D. Wright, fundador del hoy denominado Bureau of Labor Statistics en Massachussetts; el de Arthur Bowley, de la Escuela de economía de Londres; los de Fredholm y Edin en Suecia y los de A. Kaufmann y A. Chuprov, entre otros. Igualmente, es importante destacar la importancia que tuvo la publicación de la primera tabla de números aleatorios para la obtención de muestras probabilísticas, desarrollada por Tippet en 1927.

En 1934, el profesor Jerzy Neyman publicó en la Royal Statistical Society de Londres, lo que puede considerarse como el primer trabajo científico sobre el muestreo de poblaciones finitas, en el que hace una crítica al diseño y resultados obtenidos por Gini y Galvani en 1929, sobre la muestra intencional (purposive sample).

El artículo de Neyman dejó establecido, sin lugar a dudas, que la selección aleatoria era la base de una teoría científica que permitía predecir la validez de las estimaciones muestrales. El contenido del trabajo sobre la selección de las unidades de muestreo, los métodos de estimación y la utilización de información complementaria para formar estratos, son ejemplo de un nuevo enfoque, que culmina con la afijación óptima, qu consiste en fijar los tamaños de muestra en cada estrato, para minimizar la varianza para un tamaño fijo de muestra total. En 1935, Yates y Zacopanay desarrollan un procedimiento más general que permite minimizar la varianza para un costo fijo.

En 1938 Neyman desarrolla la técnica del muestreo doble o bifásico, utilizada cuando no se conocen los tamaños de los estratos, ni pueden ubicarse a los elementos en ellos.

En 1942, Hansen y Hurwitz son los primeros en utilizar la palabra "cluster" (conglomerado) para designar un grupo de elementos que constituye una unidad de muestreo. Una modalidad de este método es el denominado muestreo por áreas, en el que la población en estudio se divide en áreas, de forma que cada elemento pertenece a una y sólo una de las unidades de muestreo o áreas. Hansen y Hurwitz introdujeron un esquema de muestreo con reposición y probabilidades desiguales de selección.

Los métodos de muestreo por conglomerados y áreas fueron desarrollados ante la imposibilidad de disponer de una lista de elementos que permitiera su selección como unidades de muestreo, así como por consideraciones de costo.

A finales de los años 40, surge el submuestreo, como respuesta al muestreo por conglomerados cuando estos, llamados también unidades primarias o de primera etapa, son muy grandes. Algunas publicaciones sobre este aspecto fueron "A Chapter in Population Sampling" (1950) del Bureau of the Census, y otras de Gray y Corlett (1950), Yates (1950) y Jebe (1952).

Propuestas de estimadores para diversas características poblacionales y sus varianzas para el muestreo polietápico pueden verse en Sukhatme (1950, 1953), Sukhatme y Panse (1951), Sukhatme y Narain (1952) y Thionet (1953).

La técnica de selección sistemática, con arranque aleatorio, fue utilizada por Bowley en el campo económico, pero la posibilidad de que existiesen correlaciones internas no fue considerada hasta 1942 por Osborne. Los primeros investigadores de la teoría del muestreo sistemático fueron W.G. y L.H. Madow, en 1944, quienes demostraron la igualdad en media de las varianzas correspondientes al muestreo sistemático y al muestreo aleatorio simple; este importante resultado condujo al uso intensivo del muestreo sistemático como una técnica conveniente de selección, más que como un método de muestreo. También demostraron que para poblaciones con tendencia lineal el muestreo sistemático es más preciso que el aleatorio simple, pero menos que el estratificado. L.H. Madow (1946), Cochran (1946), Yates (1946, 1948 y 1949), Hansen y Hurwitz (1949) y W.G. Madow (1949), Das (1950), Buckland (1951) y Iachan (1982), continuaron la elaboración de la teoría del muestreo sistemático.

A principios de los 50, se desarrollan los métodos de estimación indirectos, de razón y de regresión, que en general, son más precisos que los directos. La dificultad más teórica que práctica, de estos métodos radica en que no existen expresiones exactas del sesgo y la varianza, pero para muestras grandes, el sesgo es despreciable y las expresiones aproximadas de las varianzas son adecuadas.

Lahiri (1951) propuso un esquema en el que el estimador de razón es insesgado. También se debe a Lahiri el método de selección que evita enumerar todas las posibles muestras encontrando sus tamaños totales y acumulados.

Midzuno (1950) y Sen (1952) propusieron independientemente un procedimiento para seleccionar una muestra con probabilidad proporcional al tamaño, que consiste en elegir la primera unidad con probabilidad proporcional al tamaño (de la variable auxiliar) y las restantes unidades sin reemplazamiento y con probabilidades iguales.

Hansen, Hurwitz y Madow (1953) plantean en 1953, el método de estimación por diferencias, como alternativa al método de regresión. Este método se basa en la diferencia entre el valor real de una característica y el de la información complementaria disponible para la misma unidad. Hartley y Ross (1954) desarrollaron un estimador insesgado de la razón, partiendo de la media de razones (x_i/y_i) y un factor de corrección.

Sanchez-Crespo (1978) desarrolló un estimador general que resulta insesgado utilizando muestreo con selección con probabilidad proporcional al tamaño.

Horvitz y Thompson desarrollaron una técnica general de muestreo en la que las unidades primarias son elegidas sin reposición y con probabilidades de selección desiguales. Durbin (1953), Yates y Grundy (1953) propusieron otros estimadores lineales.

En 1957, Keyfitz proprcionó métodos simples para el cálculo de aproximaciones a las varianzas estimadas para estimadores del tipo razón, cuando se seleccionan dos elementos en cada estrato.

En los años 40 se inician dos líneas de evolución para el tratamiento de los errores ajenos al muestreo. La primera se dirige al tratamiento de errores específicos y la segunda al desarrollo de una teoría para el tratamiento del error total. Este análisis es probablemente una de las etapas más importantes en el desarrollo del muestreo desde el artículo de Neyman en 1934.

Según Murthy, Mahalanobis ya indicó la importancia de evaluar los errores ajenos al muestreo en 1938 y desarrolló su técnica de las submuestras interpenetrantes (1946), de gran utilidad para conocer la contribución de los entrevistadores a la variabilidad del estimador.

De gran importancia en los años 40 fueron las nuevas técnicas para el tratamiento de la falta de respuesta, como las de Hansen y Hurwitz (1946) y Politz y Simmons (1949), entre otras. El método de Hansen y Hurwitz consiste en tomar una muestra de las unidades que no contestaron. El de Politz y Simmons se basa en una sóla visita en k períodos de tiempo similares que se suponen elegidos aleatoriamente. Al entrevistado se le pregunta en cuantos de los k-1 períodos restantes hubiese sido encontrado en la casa y los resultados se ponderan de acuerdo a este dato. Por supuesto, quedan excluidas las personas que nunca están en la casa.

El problema de la falta de respuesta también fue estudiado por Birnbaum y Sirken (1950), que determinaron el tamaño de la muestra necesario para minimizar la raíz del error cuadrático medio. Deming (1953) consideró la repetición de visita como el procedimiento más eficiente para reducir el error cuadrático medio e hizo un intento para determinar el número óptimo de visitas a la misma unidad informante. Durbin (1954) tuvo en cuenta el costo de las visitas repetidas y sus consecuencias. Finkner (1952) continuó el estudio de un posible ajuste de los datos para paliar el efecto del sesgo proveniente de la falta de respuesta. Dalenius (1955) consideró la posibilidad de obtener datos de las unidades que no respondieron, cuando aún se está levantando la encuesta.

Al finalizar los años 40, se inicia la época en la que se consolida el muestreo de poblaciones finitas, con la publicación de los textos de Yates (1949), Deming (1950), Cochran (1953), Hansen, Hurwitz y Madow (1953) y Sukhatme (1954), y luego los de Sampford (1962), Kish (1965), Murthy (1967) y Raj (1968). El primer texto en

español se debe a E. Cansado (1950).

En los 70 se efectuaron numerosos desarrollos teóricos y prácticos, como por ejemplo el diseño de planes de control de calidad basados en la inspección de una muestra, las encuestas continuas, combinaciones de datos de encuestas por muestreo y datos censales, procedimientos para el cálculo de varianzas y estimaciones de estas, como para estimadores del tipo razón, de los grupos aleatorios y del muestreo sistemático, entre otros.

Se comienza a notar una separación entre teóricos y prácticos, porque algunos de los primeros critican a los segundos de no prestar suficiente atención a los últimos desarrollos.

Se comienza a utilizar el muestreo, no sólo en grandes encuestas sobre fuerza de trabajo, producción agrícola, industrial, sino en investigaciones de mercado, en mediciones de popularidad de políticos o en intención del voto.

Hoy en día se ha extendido el uso del muestreo en casi todos los ámbitos, y como todo, esto tiene su parte positiva y negativa. El muestreo es aceptado por todos como un conjunto de técnicas que permiten investigar una gran cantidad de situaciones en diferentes aspectos y espacios de la vida de las personas, gobiernos y empresas, entre otros, ya no hay discusión sobre su validez y aceptación. Sin embargo, esta expansión ha traído consigo un uso indiscriminado del muestreo, y es muy común encontrar estudios en los que no se cuidan aspectos importantes, así como otros, más osados, en los que tratan de disfrazar esta falta con medidas sobre el error y la confiabilidad, que confunden a los menos duchos, pero que suelen ser los que utilizan los resultados.

Deben cuidarse estos detalles, ya que se puede llegar a resultados falsos, y al final puede significar una pérdida de confianza en los métodos, en el muestreo y en los estadísticos. Por tal motivo, se hace un llamado al uso prudente del muestreo, y se hacen votos porque este trabajo sirva para orientar a los estudiosos en esa dirección. [1;351-361]

1.2.- Importancia del muestreo - ventajas y desventajas

Desde su inicio, las técnicas de muestreo han ido evolucionando y ganando adeptos, de modo que su uso ha aumentado de una manera acelerada y cada vez son más las disciplinas que utilizan el muestreo como parte de sus investigaciones, y cada vez son más los que lo consideran una disciplina científica que aporta resultados de manera relativamente rápida, económica y confiable.

Cada vez son más los “Censos” que incluyen el muestreo tanto para la investigación de algunas variables, como

para el propio control de las operaciones censales, algo que resultaba ilógico y contradictorio hace 60 años. El Control Total de la Calidad, tan empleado por los japoneses y absorbido por muchas empresas de países occidentales, se basa en buena medida en el uso de planes de muestreo, para conocer la calidad de los productos y procesos.

Aunque las ventajas del muestreo sobre la enumeración completa son bien conocidas, es importante mencionarlas. Las mismas se pueden clasificar en varias categorías, mayor rentabilidad, mayor rapidez, menor costo y mejor medición de las variables. Es evidente que al investigar sólo una parte de la población tanto la recolección como el procesamiento de datos es más rápido, al igual que menos costoso. Se pueden hacer investigaciones de poblaciones muy grandes a través de muestras relativamente pequeñas, por ejemplo, la Encuesta de Hogares por Muestreo de Venezuela, que tiene como objetivo principal estimar la tasa de desempleo, toma una muestra de 100.000 personas para obtener información a nivel de cuatro dominios de estudio, pero si se quisiera obtener información a nivel nacional, se necesitarían sólo unas 20.000 personas, y 350.000 si se quiere obtener a nivel de entidad federal, tamaños éstos que de cualquier modo son mucho menos que los más de 20.000.000 de personas que conformarían una enumeración completa.

Sin embargo, el muestreo también tiene sus desventajas. Cuando se hacen investigaciones estadísticas de cualquier tipo, se cometen errores, que deben ser controlados y, en lo posible, medidos. En general, estos errores pueden ser clasificados en "errores muestrales" y "errores no muestrales". Los primeros se refieren al error cometido por no incluir en la muestra a todos los elementos de la población en estudio, y pueden ser medidos siempre que la muestra haya sido tomada de acuerdo con ciertas condiciones, materia ésta que se desarrollará en el presente trabajo. Obviamente este error no permite que la información pueda ser desagregada, ya sea de manera geográfica o atendiendo a otras variables de interés, al menos con altos niveles de confianza y precisión.

Los errores no muestrales son aquellos que se derivan de factores distintos al muestreo, y están conformados por:

- errores en la recolección, debidos a fallas en;
    - el instrumento de recolección
    - las operaciones de recolección
    - los entrevistadores o encargados de hacer la recolección propiamente dicha
    - el informante o fuente de información
    - ausencia o negativa del informante a responder
- errores en el tratamiento de la información, debido a fallas en:
    - la codificación de datos
    - la transcripción de datos
    - el procesamiento de datos

Estos errores se cometen en cualquier investigación, pero tienen el inconveniente de que en muchos casos resulta difícil medirlos.

De manera que tendría sentido contrastar los errores no muestrales asociados a la enumeración completa con la suma de los errores muestrales y no muestrales asociados a la investigación por muestreo, claro está, con un diseño muestral apropiado. También se pudieran comparar los costos, la facilidad y rapidez de la recolección y el procesamiento.

Como ya se mencionó, los errores no muestrales son difíciles de medir, pero se puede decir que tienden a aumentar cuando la cantidad de elementos a investigar aumenta, ya que resulta más complicado controlarlos; y pudiera suceder que los errores no muestrales de la enumeración completa sean mayores que la suma de los errores muestrales y no muestrales de la investigación por muestreo. En cuanto a los costos, suelen ser mayores en una enumeración completa. En lo que respecta a la celeridad de recolección y procesamiento, es más rápido en una investigación por muestreo, aunque a veces puede ser más complicado que en la enumeración completa, al menos el procesamiento, ya que hay que incluir los procesos de expansión y estimación de los errores muestrales, mientras que en una enumeración completa basta con hacer una agregación de los datos recabados.

Por otro lado, en ocasiones se hace un uso inadecuado y hasta inescrupuloso de las técnicas de muestreo, donde no se presentan la metodología de selección, los estimadores utilizados ni los errores cometidos; y uso de muestras no probabilísticas, niveles de desagregación altos con muestras sumamente pequeñas, entre otras cosas.

1.3.- Parámetros, valores poblacionales y estimadores

En este apartado se definen algunos términos importantes, como son **universo** y **población**. Con cierta frecuencia se confunden estos términos, incluso hay autores que los aluden indistintamente, como si se tratara un mismo concepto, de cualquier modo, aquí se hará una diferenciación.

Se denomina Universo a cualquier conjunto finito o infinito de elementos que constituyen objeto de un estudio, y Población al conjunto de medidas de la variable en estudio en cada uno de los elementos que conforman el Universo. Es decir, cada una de las variables en estudio genera una población que viene dada por el conjunto de valores que ella toma en los elementos que conforman el universo. Por ello se puede decir, cuando el universo tiene N elementos, que la población es de tamaño N. Por lo tanto, en un estudio puede haber varias poblaciones, tantas como variables se definan. [7;76]

En consecuencia, se deben definir con mucha claridad y precisión el universo y las variables a estudiar, para poder definir las diferentes poblaciones en un estudio. A continuación se muestra un cuadro con ejemplos de universo y la población, además de algunos indicadores.

Figura 1.1.- Ejemplos de Universo, Población e Indicadores

Estudio	Universo	Población	Indicadores
Encuesta de Población	Todas las personas mayores de 12 años	- Conjunto de edades de todas las personas mayores de 12 años - Conjunto de valores de la variable "Situación en la Fuerza de Trabajo" de todas las personas mayores de 12 años	Tasa de desempleo = $= \frac{\text{total personas} > 12 \text{ años desempleadas}}{\text{total personas económicamente activas}}$
	Todas las personas	- Conjunto de edades de todas las personas	Edad promedio = $= \frac{\sum \text{edades de todas las personas}}{\text{total de personas}}$
		- Conjunto de valores de la variable "sexo" (masc., fem.) todas las personas	Proporción de Mujeres = $= \frac{\text{total de mujeres}}{\text{total de personas}}$
Encuesta Agrícola	Todas las reses menores de 3 años	- Conjunto de edades de todas las reses menores de 3 años - Conjunto de valores de la variable "vacunada" Menores a 3 años	Proporción de reses menores a 3 años y vacunadas = $= \frac{\text{total reses} < 3 \text{ años y vacunadas}}{\text{total de reses} < 3 \text{ años}}$
Encuesta Agrícola	Todas las fincas	- Conjunto de valores de la variable "presencia de ganado porcino" en todas las fincas - Conjunto de valores de la variable "total de cabezas de ganado porcino" en todas las fincas	Proporción de fincas con ganado porcino = $= \frac{\text{total fincas con ganado porcino}}{\text{total de fincas}}$
Encuesta Escolar	Todas las personas Entre 3 y 25 años que Están o han estado inscritos en el sistema educativo	- Conjunto de edades de todas las personas que están o han estado inscritos en el sistema educativo - Conjunto de valores de la variable "situación actual en el sistema educativo", de todas las personas que están o han estado inscritos en el sistema educativo	Tasa de deserción escolar = $= \frac{\text{total personas entre 3 y 25 años que abandonaron los estudios en el último año}}{\text{total personas entre 3 y 25 años que están o han estado en el sistema educativo en el último año}}$

Otras definiciones encontradas denominan universo al conjunto de elementos objeto de estudio, y población como la parte o subconjunto de éste conformado por los elementos de los cuales realmente se puede obtener información, es decir, los que están incluidos en el marco muestral. El marco muestral es una lista de unidades de muestreo, que pueden ser elementos (personas, animales, objetos) o alguna agregación de ellos, como áreas geográficas, o cualquier otra agrupación, de donde se seleccionará la muestra. Dichas listas suelen tener problemas de omisiones o duplicaciones, entre otros. En líneas generales, la definición de universo coincide con la mostrada anteriormente, y entonces se denominará al subconjunto incluido en el marco muestral como "universo estadístico".

A pesar de ésta diferencia entre universo, universo estadístico y población, se suele utilizar el término *poblacional* para referirse a conceptos relativos al universo, como son los casos de subpoblaciones, para denominar a parte del universo -tema será tratado en el capítulo 5-, y "Valor Poblacional", y en este sentido, es importante definir los términos "Valor Verdadero" y "Valor Poblacional", ambos son mediciones que se hacen sobre todos los elementos del universo, la diferencia radica en que en el primero se supone que las observaciones están exentas de errores no muestrales, mientras que en el segundo, éstos si están presentes. Así, el promedio verdadero de una variable, considera a todos los elementos del universo, considerando que no se cometen errores, es realmente el "promedio verdadero"; el promedio poblacional, igualmente incluye a todos los elementos del universo, pero es el resultado de la medición, en la cual se comenten errores; igual ocurre con un total , una proporción o una tasa. [4;30]

A menudo algunos autores definen a un parámetro como un valor poblacional, sin embargo, aquí se trabajarán de forma separada, incluyendo la diferenciación hecha con valor verdadero.

Los parámetros están asociados a distribuciones de probabilidad y no son observables, sólo se pueden estimar. Los valores verdaderos se pueden determinar con exactitud si se observa el universo completo [3;160-161], sin embargo, finalmente, lo que se suele obtenerse es el valor poblacional.

Sea una población de tamaño N, $y_1, y_2, \ldots, y_N$, entonces, $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ es el promedio verdadero. Supóngase que dicha variable sigue una distribución Normal, de promedio μ y varianza σ^2; entonces $\bar{Y}$ es promedio verdadero, que es el valor esperado de *y* en el universo.

La idea fundamental del muestreo es obtener esas medidas o indicadores, pero basándose en el estudio de una parte del universo. A la parte del universo que se estudia se le denomina "muestra" y a las mediciones hechas sobre todos los elementos de la muestra se le denominará "Valor Muestral", sin embargo, cuando se utilizan dichos "Valores Muestrales" o una función de ellos para inferir sobre la población o sobre los "Valores Poblacionales", se le denomina "estimador", que, como su nombre lo indica, estimará o hará inferencias sobre el valor poblacional.

Es importante destacar la diferencia entre los términos "estimador" y "estimación", tan similares y confundidos con frecuencia. Se denomina estimador a la medida o indicador usado para hacer los cálculos basados en la muestra, y estimación al valor obtenido de aplicarlo. En otras palabras, estimador es una función y estimación es el valor de dicha función evaluada en una muestra, en última instancia, estimador es la fórmula y estimación el resultado.

Otro concepto fundamental es el de censo. Originalmente, y a través de muchos años, se le denominó censo al conteo o enumeración de toda una población, así los romanos y otras culturas realizaban sus censos. Luego se fue anexando la investigación de otras variables, aprovechando el desplazamiento de recursos para obtener mayor información de la población. Estos censos se fueron haciendo cada vez más necesarios, de manera que se fueron sistematizando y complementando con otras fuentes de información, como las encuestas por muestreo y registros administrativos, hasta llegar a los actuales censos de población, vivienda, económicos, agropecuarios, y otros. Siendo fieles a los orígenes, tradicionalmente se ha llamado censo a una enumeración completa o exhaustiva, es decir, a una investigación que incluye a todos los elementos de la población.

Sin embargo, en los tiempos actuales se puede denominar censo a "una encuesta destinada a medir las características estructurales de un conjunto de elementos" [5;1], donde una encuesta es "un procedimiento en el cual se recopilan, procesan y analizan datos acerca de las características de un conjunto de elementos (personas, animales, viviendas, objetos), mediante entrevistas o por observación, que pueden ser personales, por teléfono o por correo", u observación. "Es necesario señalar que para la recopilación de los datos, se debe realizar un estudio previo que incluye la determinación de las variables a investigar, el diseño del instrumento para la recolección (cuestionario), el diseño de la muestra (en caso de haberla), la definición de las operaciones de campo, entre otros aspectos" que forman parte de la etapa de planificación. [5; 1]

Entonces, un censo es una encuesta, y como tal puede hacerse por enumeración completa o a través de una muestra. Esta definición de censos no está en contraposición a las primeras investigaciones que se hicieron y que denominaban censos. Aquellas, de hecho, eran censos porque medían características estructurales de la población, además, el término encuesta es más reciente y aún más el uso del muestreo.

Retomando la idea previa, calcular la tasa de desempleo tomando en cuenta a todas las personas mayores de 12 años del país, es el valor poblacional, si se toma una muestra y se hacen los cálculos sobre ella, se obtendrá la “tasa muestral de desempleo” y que puede fungir también como "tasa estimada de desempleo" o "tasa de desempleo estimada" o "una estimación de la tasa de desempleo". Igualmente ocurre con el caso de las reses, si se consideran a todas las menores de 3 años del país, se obtiene la "proporción poblacional de reses vacunadas", si se toman en

cuenta sólo las seleccionadas en una muestra, se obtiene la "proporción muestral de reses vacunadas", que también será la "proporción estimada de reses vacunadas" o "una estimación de la proporción de reses vacunadas".

El valor verdadero es una constante, no varía; pero el estimador, depende de la muestra seleccionada, es decir, que para cada muestra se obtiene un valor (estimación) diferente. Por esta razón, los estimadores son variables, y cuando la muestra es aleatoria, son variables aleatorias, y como tal tienen sus varianzas.

1.4.- Muestras probabilísticas

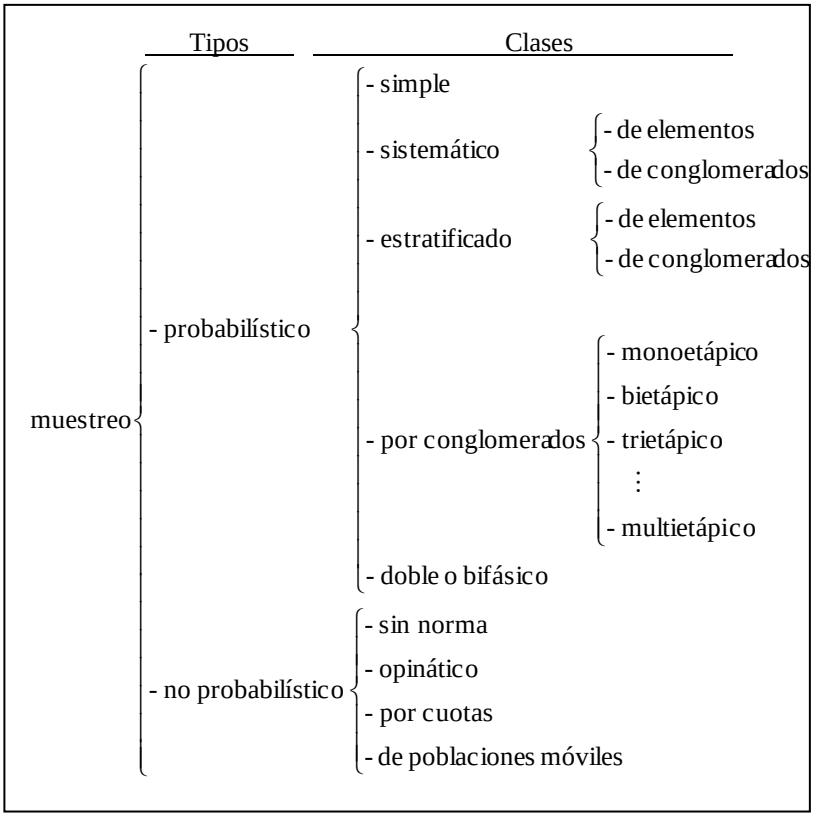
Los diseños muestrales se pueden dividir en dos grandes grupos,

i .- aleatorios o probabilísticos

ii.- no aleatorios o no probabilísticos

y ambos grupos se subdividen, de acuerdo a la técnica aplicada. A continuación se presenta un cuadro sinóptico de las divisiones del muestreo:

Figura 1.2.- Tipos y Clases de Muestreo



También suelen utilizarse los términos “muestreo de listas”, aplicado la selección de elementos, y “muestreo de

áreas”, más inclinado a la selección de conglomerados.

Una muestra es aleatoria o probabilística cuando todos los elementos de la población tienen una probabilidad de selección conocida y distinta de cero. Por supuesto, una muestra no probabilística es aquella que no cumple con estas dos condiciones. Dentro de estas últimas se enumeran el muestreo sin norma, que es aquel en el cual, como su nombre los indica, no se sigue ningún criterio para la selección de la muestra; el muestreo opinático se rige por el juicio u opinión del seleccionador, que a veces suele ser experto o conocedor del tema a investigar, por ejemplo, "las 20 personas que entren primero a este edificio", o “las 20 personas de mayor edad que trabajan en esta oficina”; luego se tiene el muestreo por cuotas, que si bien es no probabilístico, a diferencia de los anteriores, introduce elementos estadísticos, y consiste en seleccionar elementos de la población, pero cumpliendo con una cantidad de elementos de ciertas características, hasta completar un tamaño preestablecido de ellos; por ejemplo, se deben entrevistar 200 personas, cumpliendo las siguientes cuotas:

Tabla 1.1.- Cuotas de muestreo por sexo, según grupos edad

Grupos de Edad	Total	Hombres	Mujeres
Total	200	100	100
15 a 24 años	44	21	23
25 a 34 años	50	23	27
35 a 44 años	53	28	25
45 años y más	53	28	25

entonces se van entrevistando personas hasta satisfacer las cuotas establecidas, así, sí se consigue una persona que pertenece a una categoría cuya cuota ha sido cumplida, no se le hace la entrevista.

Mención aparte merece el muestreo para poblaciones móviles o silvestres, como peces o animales en su habitat natural, y su objetivo es estimar el tamaño del universo basado en técnicas de *captura-marca-recaptura*, es decir, se *capturan* elementos, se *marcan*, se regresan a su habitat, y se realiza otra *captura*, donde se identifican los elementos marcados de la primera muestra. Existen técnicas estadísticamente bien fundamentadas en este sentido, pero está clasificado como no probabilístico porque al ser su concepción y principal uso en poblaciones silvestres, la selección de las muestras difícilmente cumplirá criterios probabilísticos.

Las muestras no probabilísticas se utilizan mucho en pruebas donde lo importante no son las cifras propiamente, sino más bien el funcionamiento de instrumentos o procesos, como pueden ser el entendimiento de un cuestionario, el procesamiento de los datos, entre otras; por ejemplo, para saber si se capta el total de cuartos para dormir de una vivienda, se toman viviendas donde se conoce de antemano que hay tendencia al hacinamiento; o si se quiere saber si se incluyen todos los posibles materiales de las paredes, piso y techo de las viviendas, se toman una muestra por

cuotas de diferentes tipos de viviendas; o si se quieren probar unos programas computacionales de entrada y procesamiento de datos, se incluye elementos donde se sepa que va a actuar dicho software.

También son usadas muestras no probabilísticas cuando se requieren resultados rápidos, donde puede resultar complicado el diseño de una muestra probabilística. Sin embargo, y basados en este último argumento, se ha hecho un uso indiscriminado del muestreo no probabilístico, dejando, en muchos casos, mal parado al muestreo como técnica, y no a los responsables de este hecho.

En las muestras no probabilísticas no se puede calcular ninguna medida del error cometido, ya que una parte de las posibles muestras tienen probabilidad cero de ser seleccionadas.

En el caso, de las muestras aleatorias, si es calcular o estimar la varianza de los estimadores, aunque en oportunidades, dependiendo del diseño, puede ser complicado. Lógicamente, habrá diseños mejores que otros, y parte de esto es lo que se pretende presentar en este trabajo, mediante el uso de las medidas de precisión del estimador, como son el error cuadrático medio, la varianza, el error estándar, el coeficiente de variación y el sesgo.

Es frecuente escuchar términos como "muestra representativa", "muestras grandes o pequeñas", que es importante aclarar.

Cuando se toma una muestra, de la manera que sea, se toma con la finalidad de representar a toda la población, sin embargo, la idea generalizada que se tiene de una "muestra representativa" es el de una muestra que "represente bien" a toda la población, pero se acaba de ver que en las muestras probabilísticas todos los elementos tienen probabilidad conocida y distinta de cero de estar en la muestra, es decir, que toda la población está aleatoriamente representada. Existen diferentes diseños para seleccionar la muestra, así como diferentes estimadores de un mismo valor poblacional, lógicamente, habrá diseños y estimadores que se comporten mejor que otros, para un estudio en particular, es decir que se obtiene mayor precisión con unos diseños y/o estimadores que con otros. De manera que en lugar de hablar de "muestras representativas" se debe hablar de muestras aleatorias o probabilísticas, con tal o cual precisión. Incluso a veces se escucha el barbarismo de "una muestra más representativa que otra" para decir que "un diseño arroja resultados más precisos que otro".

En cuanto a los términos de una "muestra grande o pequeña", deben tener una referencia, ... *esta muestra es grande respecto a la anterior*, o ... *esta muestra es pequeña, tomando en cuenta el tamaño de la población*. Sin embargo, es preferible evitar estos términos, y hablar en términos de que se cumple con los objetivos propuestos, bien sean en términos de confiabilidad y desagregación de la información, entonces, una muestra será pequeña si por razones

del tamaño de muestra y no por el diseño en sí, no se logran dichos objetivos, y será verdaderamente grande si con un tamaño menor y el mismo diseño, se obtiene una precisión similar. Es decir, que para hablar de tamaños grandes o pequeños, se tiene que hablar de una *comparación*, en la que además, se hayan aislado todos los demás elementos que afectan la precisión de los estimadores.

1.5.- Mediciones de errores muestrales

La medida general para medir el error que se comete al no incluir a toda la población en el estudio, es el Error Cuadrático Medio del estimador, que es la suma de la varianza y el cuadrado del sesgo del estimador, es decir:

Sea $\hat{\theta}$ un estimador de θ, entonces el Error Cuadrático Medio de $\hat{\theta}$ es

$$ECM(\hat{\theta}) = V(\hat{\theta}) + \left[B(\hat{\theta})\right]^2$$

donde $B(\hat{\theta}) = E(\hat{\theta}) - \theta$

si el sesgo es cero, es decir, si el estimador es insesgado, entonces el Error Cuadrático Medio coincide con la varianza. El error de muestreo es medido por la raíz del Error Cuadrático Medio, de manera que si el estimador es insesgado, el error de muestreo es la desviación estándar del estimador.

El principio del muestreo es asignar a los elementos que no fueron seleccionados, el promedio de los elementos que fueron seleccionados. En algunos casos, de acuerdo con el diseño aplicado, este promedio será ponderado, en otros se fraccionará la población para obtener y aplicar promedios en cada fracción. Todos estos casos se verán a lo largo del trabajo.

Otro aspecto importante es el de diferenciar la varianza de la variable a estimar y la varianza del estimador. Para ilustrar esta diferencia se trabajará un ejemplo. En cierta área de asentamientos agrícolas se desea hacer una estimación de la producción de arroz para una determinada temporada. La población está compuesta por fincas. Sea y_i el total de kilos de arroz producidos en la finca *i-ésima* para la temporada en cuestión.

Se sabe que la varianza de una variable es el promedio de los cuadrados de sus desvíos respecto a su valor esperado. Entonces,

$$V(y)=S_y^2=S^2=\frac{\sum_{i=1}^{N}(y_i-\bar{Y})^2}{N} \qquad \text{y} \qquad S'^2_y=S'^2=\frac{\sum_{i=1}^{N}(y_i-\bar{Y})^2}{N-1}$$

son la varianza y cuasivarianza de la variable "*y"*, y miden la dispersión de la producción de arroz entre las fincas.

Por otro lado,

$$V\left(\hat{Y}\right)=S_{\hat{Y}}^2=\frac{\sum_{k=1}^{\binom{N}{n}}\left(\hat{Y}_k-Y\right)^2}{\binom{N}{n}} \qquad V(\hat{\bar{Y}})=V\left(\bar{y}\right)=S_{\bar{y}}^2=\frac{\sum_{k=1}^{\binom{N}{n}}\left(\bar{y}_k-\bar{Y}\right)^2}{\binom{N}{n}}$$

son las varianzas del estimador del total de arroz producido en todo el área y del estimador del promedio de arroz producido en cada finca del área respectivamente (si la muestra es aleatoria simple de fincas); donde $\binom{N}{n}$ indica el total de muestras posibles y $\hat{Y}_k$ la estimación del total, considerando la muestra *k-ésima*, con *k*=1,2,..., $\binom{N}{n}$. En el capítulo del muestreo aleatorio simple se explicará más en detalle, y se demostrará que

$$V(\hat{Y}) = S_{\hat{Y}}^2 = \frac{N^2(N-n)}{N \quad n}\frac{\sum_{i=1}^{N}\left(y_i-\bar{Y}\right)^2}{N-1} = N^2\frac{N-n}{N\,n}S'^2$$

$$V(\hat{\bar{Y}})=V(\bar{y}) = S_{\hat{\bar{Y}}}^2 = S_{\bar{y}}^2 = \frac{N-n}{N \quad n}\frac{\sum_{i=1}^{N}\left(y_i-\bar{Y}\right)^2}{N-1} = \frac{N-n}{N\,n}S'^2$$

Nótese que estas varianzas dependen de la cuasivarianza de la variable *"y"*. De manera que debe quedar bien en claro la diferencia entre $V(y), V(\bar{y}), V(\hat{Y})$.

A partir de la varianza se obtienen otras medidas de error, como son el error estándar y el coeficiente de variación, cuya expresión se muestra a continuación,

error estándar $ee(\hat{\theta})=\sqrt{V(\hat{\theta})}$

coeficiente de variación $CV(\hat{\theta})=\frac{ee(\hat{\theta})}{\theta}$

es decir, que la varianza y el error cuadrático medio muestran el error en términos de las unidades de medida de la propia variable elevadas al cuadrado, el error estándar y la raíz del error cuadrático medio lo expresan en términos de las unidades de medida de la propia variable, y el coeficiente de variación lo expresa en términos relativos. En

este sentido, el error estándar es más fácil de interpretar que la varianza, pero el coeficiente de variación tiene la ventaja sobre los otros, en que, por estar expresado en una escala única para cualquier medición de cualquier variable, es aún más fácil la interpretación y comparaciones.

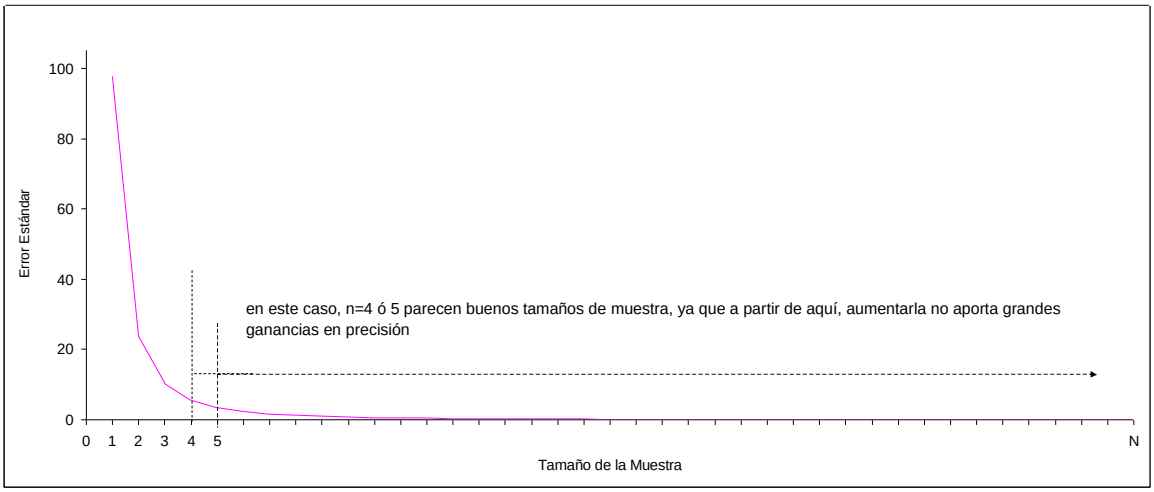
En el caso específico de los totales y promedios mostrados anteriormente, los errores estándar y coeficientes de variación quedarían como sigue,

$$ee(\hat{Y}) = \sqrt{N^2 \frac{N-n}{N\,n} S'^2} = \sqrt{N^2 \frac{N-n}{N\,n}}\, S' \qquad ; \qquad ee(\hat{\bar{Y}}) = \sqrt{\frac{N-n}{N\,n}}\, S'$$

$$CV(\hat{Y}) = \frac{\sqrt{N^2 \frac{N-n}{N\,n} S'^2}}{Y} = \frac{\sqrt{N^2 \frac{N-n}{N\,n}}\, S'}{Y} \qquad ; \qquad CV(\hat{\bar{Y}}) = \frac{\sqrt{\frac{N-n}{N\,n}}\, S'}{\bar{Y}}$$

En general, al aumentar el tamaño de la muestra, aumenta la precisión de los estimadores; esto ocurre cuando los estimadores que se utilizan son consistentes. Sin embargo, llega un punto donde aumentar el tamaño de la muestra no aporta grandes ganancias en precisión; utilizando un enfoque de precisión/costo, se puede decir que a partir de un tamaño de muestra, resulta muy costoso aumentar el mismo en relación con su aporte en la ganancia de precisión, en consecuencia, quizás se ha llegado a un tamaño de muestra adecuado. Este fenómeno puede verse en el siguiente gráfico.

Figura 1.3.- Precisión de un Estimador (error estándar) según tamaño de la muestra



1.6.- Principios del muestreo

Mucho se habla sobre el muestreo, y como se ha dicho, son muchos los que los usan como herramienta en sus trabajos, sin embargo, pareciera haber un cierto misterio en cuanto al conocimiento de las técnicas de muestreo, al punto de que cuando una persona conoce poco más de lo elemental, a veces se le considera un experto.

Aunque conocer en profundidad toda la teoría del muestreo es complejo, su enfoque del muestreo es sencillo. Se pueden enumerar dos principios donde recae básicamente todo el peso de la disciplina. Estos son:

Principio 1 **Una vez seleccionada una muestra, cualquier estimación de cualquier valor poblacional consiste en asignar el promedio de los valores de la variable en estudio, de los elementos o unidades que fueron seleccionados en la muestra, a los elementos o unidades que no fueron seleccionados**

Este primer principio está referido directamente a los estimadores y, como puede notarse, es sencillo de entender, aunque su construcción a veces puede resultar un poco complicada.

Cuando se aplican diseños sencillos, la construcción de los estimadores es evidente. A medida que los diseños se van complicando, se va haciendo más difícil entender o deducir un estimador; así se tiene que cuando la muestra es estratificada, un estimador consiste en asignar a cada elemento o unidad que no está en la muestra, el promedio de la variable en estudio de los elementos o unidades de su estrato, que sí están en la muestra. En otros casos el promedio a asignar es ponderado de acuerdo con pesos de cada unidad, que se calculan de diversas maneras.

Como ya se dijo, los estimadores son variables aleatorias y tienen su correspondiente varianza, que es una medida de su precisión. El segundo principio se refiere a la forma como se calcula esta varianza.

Principio 2 **La varianza de un estimador consiste en calcular la varianza de las estimaciones provenientes de todas las muestras posibles**

es decir, se toman todas las muestras posibles, de acuerdo con el diseño planteado, se halla la estimación proveniente de cada una ellas y se calcula la varianza, tomando como observaciones cada una de las estimaciones.

Claro que en la práctica este procedimiento resulta improcedente, ya que para valorar todas las muestras posibles habría que hacer una enumeración completa.

Sin embargo, el conocer la expresión matemática de la varianza es sumamente útil para analizar las condiciones bajo las cuales se puede controlar la precisión del método utilizado. Igualmente lo es para ayudar a definir la expresión de su estimador a partir de una sola muestra, sin necesidad de tomar todas las posibles muestras.

Estos procedimientos son más complejos que los involucrados en el principio anterior, y por lo tanto requieren de más conocimientos matemáticos. En el presente trabajo se obtendrán los estimadores y sus varianzas utilizando estos dos principios.

2.- MUESTREO ALEATORIO SIMPLE

El muestreo aleatorio simple, también conocido bajo el nombre de "muestreo irrestricto aleatorio", como es de esperar, es el más simple o básico de todos los diseños muestrales aleatorios, y consiste en tomar una muestra al azar de n elementos de un universo de N, sin restricciones de ningún tipo.

El muestreo aleatorio simple *(MAS)* tiene las siguientes características:

a) tamaño del universo = N
b) tamaño de la muestra = n
c) selección sucesiva de las unidades
d) selección sin reposición de las unidades seleccionadas
e) probabilidades iguales de selección en cada extracción, para las unidades que no han sido seleccionadas
f) las muestras que constan de las mismas unidades se consideran iguales

De manera que la probabilidad de seleccionar la unidad *i-ésima* en la *k-ésima* extracción es: $\frac{1}{N-k+1}$

donde $i=1,...,N$ y $k=1,...,n$.

Entonces, la probabilidad de seleccionar una muestra compuesta por las unidades

$$i,i+1,i+2,\ldots,i+(n-1)$$

es,

$$P(i,i+1,i+2,\ldots,i+(n-1))=\frac{1}{N}\ \frac{1}{N-1}\ \frac{1}{N-2}\ \cdots\ \frac{1}{N-(n-1)}=\frac{1}{N\,(N-1)(N-2)\ldots(N-(n-1))}=\frac{(N-n)!}{N!}$$

como las muestras que incluyen las mismas unidades pero en diferente orden, se consideran iguales, y se tienen *n!* permutaciones de las mismas unidades en la muestra, por lo tanto

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

que es la probabilidad de selección de una muestra posible, que no es otra cosa que *1* entre el total de muestras

posibles, lo que significa que las mismas son equiprobables. Es decir que existen $\binom{N}{n}$ muestras posibles.

2.1.- Valores poblacionales y estimadores - total y promedio

Sea un universo de *N* elementos, en el cual se quiere medir una variable *"y"* que se asocia a cada elemento, es decir,

y_i = el valor que toma la variable *"y"* en el elemento i-ésimo , $i=1,2,...,N$

Sean *Y* y $\bar{Y}$ el total y el promedio poblacional respectivamente,

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

Se toma una muestra aleatoria simple de tamaño *n*, que es $y_1, y_2, \ldots, y_n$, $n \leq N$, y se tienen el total y el promedio muestral, respectivamente,

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

Como *Y* se puede escribir de la siguiente manera, $Y = \sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} y_i$

entonces se define $\hat{Y}$ como un estimador de *Y*, $\hat{Y} = \sum_{i=1}^{n} y_i + \sum_{j=n+1}^{N} y'_j$

donde, y_i es el valor de i-ésimo elemento de la muestra, $i=1,...,n$

y'_j es un valor estimado de y_j , $j=n+1,...,N$, que aplicando el *principio 1*, viene dado por,

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

que es el promedio de los elementos que están en la muestra (promedio muestral), por lo tanto,

$$\hat{Y} = \sum_{i=1}^{n} y_i + \sum_{j=n+1}^{N} \bar{y} = \sum_{i=1}^{n} y_i + (N-n)\,\bar{y} = n\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) + (N-n)\,\bar{y} = n\,\bar{y} + (N-n)\,\bar{y} = N\,\bar{y} \qquad (2.1)$$

Análogamente, se hallará $\hat{\bar{Y}}$, el estimador de $\bar{Y}$,

$$\hat{\bar{Y}} = \frac{\sum_{i=1}^{n} y_i + \sum_{j=n+1}^{N} \bar{y}}{N} = \frac{\sum_{i=1}^{n} y_i + (N-n)\,\bar{y}}{N} = \frac{n\left(\dfrac{\sum_{i=1}^{n} y_i}{n}\right) + (N-n)\,\bar{y}}{N} = \frac{n\,\bar{y} + (N-n)\,\bar{y}}{N} = \frac{N\,\bar{y}}{N} = \bar{y} \qquad (2.2)$$

es decir, que los estimadores de los valores poblacionales son,

$$\hat{Y} = N\,\bar{y} \quad , \quad \hat{\bar{Y}} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

2.2.- Varianzas de los estimadores

Para el cálculo de la varianza, se debe obtener la estimación correspondiente a cada una de las muestras posibles y hallar la varianza de la distribución resultante. Se tienen $\binom{N}{n}$ muestras de tamaño *n*. A los totales muestrales se les denominará y_k, y se define por

$$y_k = \sum_{i=1}^{n} y_{ki} \quad , \quad y_{ki} = valor\,de\,la\ \text{var}\,iable\,"y"\,en\,el\,elemento\,i-ésimo\,de\,la\,muestra\,k$$

y a la media muestral, $\bar{y}_k = \dfrac{\sum_{i=1}^{n} y_{ki}}{n}$

La varianza del estimador de la media $V(\hat{\bar{Y}})$, viene dada por (*principio 2*):

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

que es igual a

$$V(\hat{\bar{Y}}) = \frac{\sum_{k=1}^{\binom{N}{n}} \bar{y}_k^2 - \binom{N}{n}\bar{Y}^2}{\binom{N}{n}} = \frac{\sum_{k=1}^{\binom{N}{n}} \bar{y}_k^2}{\binom{N}{n}} - \bar{Y}^2$$

descomponiendo $\sum_{k=1}^{\binom{N}{n}} \bar{y}_k^2$, se tiene que

$$\sum_{k=1}^{\binom{N}{n}} \bar{y}_k^2 = \sum_{k=1}^{\binom{N}{n}} \left(\frac{y_k}{n}\right)^2 = \frac{1}{n^2}\sum_{k=1}^{\binom{N}{n}} y_k^2 = \frac{1}{n^2}\left[\binom{N-1}{n-1}\sum_{k=1}^{N} y_i^2 + 2\binom{N-2}{n-2}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} y_i\, y_j\right]$$

por lo tanto,

$$V(\hat{\bar{Y}}) = \frac{\frac{1}{n^2}\left[\binom{N-1}{n-1}\sum_{k=1}^{N} y_i^2 + 2\binom{N-2}{n-2}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} y_i\, y_j\right]}{\binom{N}{n}} - \bar{Y}^2$$

$$= \frac{\frac{1}{n^2}\binom{N-1}{n-1}\sum_{k=1}^{N} y_i^2}{\binom{N}{n}} + \frac{\frac{2}{n^2}\binom{N-2}{n-2}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} y_i\, y_j}{\binom{N}{n}} - \bar{Y}^2$$

$$= \frac{1}{N\,n}\sum_{i=1}^{N} y_i^2 + \frac{2\,(n-1)}{N\,(N-1)\,n}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} y_i\, y_j - \bar{Y}^2$$

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

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

$$= \frac{1}{N\,n}\left[\left(\frac{n-1}{N-1}\right)\left(\sum_{i=1}^{N} y_i^2 + 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} y_i\, y_j\right) + \left(1-\frac{n-1}{N-1}\right)\sum_{i=1}^{N} y_i^2\right] - \bar{Y}^2 \qquad (2.3)$$

como

$$\sum_{i=1}^{N} y_i^2 + 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} y_i\, y_j = \left(\sum_{i=1}^{N} y_i\right)^2 = N^2\,\frac{\left(\sum_{i=1}^{N} y_i\right)^2}{N^2} = N^2\,\bar{Y}^2$$

entonces (2.3) queda,

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

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

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

$$V\left(\hat{\bar{Y}}\right) = \frac{(N-n)}{N}\frac{S'^2}{n} \quad \text{donde} \quad S'^2 = \frac{\sum_{i=1}^{N} y_i^2 - N\,\bar{Y}^2}{N-1} = \frac{\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2}{N-1} \tag{2.4}$$

Utilizando la relación $\hat{Y} = N\,\hat{\bar{Y}}$, se tiene que

$$V(\hat{Y}) = V(N\,\hat{\bar{Y}}) = N^2\,V(\bar{Y}) = \frac{N^2(N-n)}{N}\frac{S'^2}{n} \tag{2.5}$$

2.3.- Estimadores de las varianzas

El estimador de la varianza del estimador del promedio (*o varianza estimada del estimador del promedio*), resulta de sustituir los valores poblacionales que no se conocen, por sus respectivos valores muestrales,

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

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

Para el estudio del sesgo de este estimador, se desarrolla su esperanza matemática:

$$E\left[\hat{V}(\hat{\bar{Y}})\right] = E\left[\frac{N-n}{Nn}\frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}{n-1}\right] = \frac{N-n}{Nn(n-1)} \quad E\left[\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2\right] = \frac{N-n}{Nn(n-1)} \quad E\left[\sum_{i=1}^{n} y_i^2 - n\,\bar{y}^2\right] \tag{2.6}$$

desarrollando la esperanza de la suma,

$$E\left[\sum_{i=1}^{n} y_i^2 - n\,\bar{y}^2\right] = E\left[\sum_{i=1}^{n} y_i^2\right] - n\,E\left[\bar{y}^2\right]$$

así que de (2.6)

$$E\left[\hat{V}(\hat{\bar{Y}})\right] = \frac{N-n}{N\,n(n-1)}\left[E\left[\sum_{i=1}^{n} y_i^2\right] - n\,E\left[\bar{y}^2\right]\right] \tag{2.7}$$

como existen $\binom{N}{n}$ posibles muestras diferentes, cada una con la misma probabilidad de ser seleccionada, se tiene que, desarrollando el segundo término,

$$E\left[\bar{y}^2\right] = E\left[\left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^2\right] = \frac{1}{n^2}\,E\left[\left(\sum_{i=1}^{n} y_i\right)^2\right] = \frac{1}{n^2}\frac{1}{\binom{N}{n}}\sum_{k=1}^{\binom{N}{n}}\left(\sum_{i=1}^{n} y_{ki}\right)^2$$

$$= \frac{1}{n^2}\frac{1}{\binom{N}{n}}\sum_{k=1}^{\binom{N}{n}}\left(y_{k1}+y_{k2}+\ldots+y_{kn}\right)^2 = \frac{1}{n^2}\frac{1}{\binom{N}{n}}\left[\binom{N-1}{n-1}\sum_{i=1}^{N} y_i^2 + 2\binom{N-2}{n-2}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} y_i\,y_j\right]$$

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

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

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

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

y desarrollando el primer término

$$E\left[\sum_{i=1}^{n} y_i^2\right] = \frac{1}{\binom{N}{n}}\sum_{k=1}^{\binom{N}{n}}\left(y_{k1}^2 + y_{k2}^2 + \ldots + y_{kn}^2\right) = \frac{1}{\binom{N}{n}}\sum_{k=1}^{N}\binom{N-1}{n-1} y_i^2 = \frac{n}{N}\sum_{i=1}^{N} y_i^2$$

por lo tanto,

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

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

sustituyendo (2.7) , se tiene

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

luego, $\hat{V}(\hat{\bar{Y}})$ es un estimador insesgado de $V(\hat{\bar{Y}})$. Análogamente, $\hat{V}(\hat{Y})$ es un estimador insesgado de $V(\hat{Y})$.

Nótese que en cualquier caso, para calcular la varianza del estimador o para estimarla, es necesario conocer el tamaño de la población, *N*. Igualmente se necesita para estimar el total poblacional, no así para el promedio. El término $(N-n)/N$ es el corrector de poblaciones finitas, que se puede escribir como *1-f*, donde $f = n/N$, es la fracción de muestreo. A medida que *N* se hace grande con respecto a n, *f* converge a cero, en consecuencia *1-f* converge a 1; por tal motivo se le denomina corrector de poblaciones finitas, ya que no se utiliza en poblaciones "*infinitas*".

Nótese que si el universo es muy grande ($N \to \infty$), y además *f* converge a cero, se tiene que,

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

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

Igualmente,

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

2.4.- Proporciones

El enfoque de proporciones es un caso particular de lo expuesto hasta ahora, en toda su extensión, pero por su importancia y uso, se tratará en este apartado de forma exclusiva.

Sea un universo de *N* elementos del cual se tomará una muestra aleatoria simple de *n* de ellos, *n* < *N*. A diferencia de las variables supuestas hasta ahora, que son del tipo numérico, aquí se desea medir la presencia o no de una característica, para hallar la proporción o el total de elementos con la característica. Por ejemplo, se desea medir la

proporción o el total de mujeres, o de personas solteras, o de fincas que tienen sembrado café, o de empresas con más de 20 empleados, o de empresas con servicio telefónico, o viviendas con techo de material de desecho, entre otras muchas. Si bien puede ser tratado con todas las fórmulas de valores poblacionales, estimadores y varianzas planteadas hasta ahora, se tratará de una manera más específica.

Se define una variable *"a"* que indica la presencia o no de la característica, definida de la siguiente manera:

$$a_i = \begin{cases} 1 & \text{si el elemento } i-\acute{e}simo \text{ posee la característica en estudio} \\ 0 & \text{si el elemento } i-\acute{e}simo \text{ no posee la característica en estudio} \end{cases}$$

Luego se aplican las fórmulas conocidas, y se simplifican debido a las propiedades de la variable *"a"*.
Sea *"A"* el total poblacional, definido por,

$$A = \sum_{i=1}^{N} a_i$$

que mide el total de elementos con la característica, y *"P"* la proporción poblacional, definida por,

$$P = \frac{\sum_{i=1}^{N} a_i}{N} = \frac{A}{N}$$

que es un cociente entre el total de elementos que tienen la característica y el total de elementos, por ende, mide la proporción de elementos con la característica, y que al multiplicar por 100, se tiene el porcentaje. Nótese que no es otra cosa que el total y el promedio poblacional mostrados en el apartado 2.1, pero utilizando la variable *"a"*.

Sean los valores muestrales,

$$a = \sum_{i=1}^{n} a_i \qquad p = \frac{\sum_{i=1}^{n} a_i}{n} = \frac{a}{n}$$

el total y la proporción muestral respectivamente. Los estimadores del total y de la proporción poblacional respectivamente son,

$$\hat{A} = N\,p \qquad \hat{P} = \frac{\sum_{i=1}^{n} a_i}{n} = \frac{a}{n} = p$$

que resultan de sustituir y_i por a_i en

$$\hat{Y} = N\,\bar{y} \qquad \hat{\bar{Y}} = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}$$

Análogamente, para hallar sus varianzas, se sustituye y_i por a_i y P por $\bar{Y}$ en las fórmulas de $V(\hat{Y})$, $V(\hat{\bar{Y}})$ y se desarrollan,

$$V(\hat{A}) = \frac{N^2(N-n)}{N\,n}\,\frac{\sum_{i=1}^{N}(a_i - P)^2}{N-1} = \frac{N^2(N-n)}{N\,n}\,\frac{\sum_{i=1}^{N}a_i^2 - N\,P^2}{N-1}$$

pero como $a_i^2 = a_i$,

$$V(\hat{A}) = \frac{N^2(N-n)}{N\,n}\,\frac{\sum_{i=1}^{N} a_i - N\,P^2}{N-1} = \frac{N^2(N-n)}{N\,n}\,\frac{N\,P - N\,P^2}{N-1} = \frac{N^2(N-n)}{N\,n}\,\frac{N\,P\,Q}{N-1}$$

donde $Q=1-P$; y

$$V(\hat{P}) = V(p) = V\left(\frac{\hat{A}}{N}\right) = \frac{1}{N^2}V(\hat{A}) = \frac{(N-n)}{N\,n}\,\frac{N\,P\,Q}{N-1}$$

Aplicando el mismo procedimiento a las varianzas estimadas,

$$\hat{V}(\hat{A}) = \frac{N^2(N-n)}{N\,n}\,\frac{\sum_{i=1}^{n}(a_i - p)^2}{n-1} = \frac{N^2(N-n)}{N\,n}\,\frac{\sum_{i=1}^{n}a_i^2 - np^2}{n-1} = \frac{N^2(N-n)}{N\,n}\,\frac{npq}{n-1}$$
$$= \frac{N^2(N-n)}{N}\,\frac{pq}{n-1} = N^2(1-f)\frac{pq}{n-1}$$

$$\hat{V}(p) = \hat{V}\left(\frac{\hat{A}}{N}\right) = \frac{1}{N^2}\hat{V}(\hat{A}) = \frac{(N-n)}{N\,n}\,\frac{n\,p\,q}{n-1} = \frac{(N-n)}{N}\,\frac{p\,q}{n-1} = (1-f)\frac{p\,q}{n-1}$$ donde $q=1-p$.

Si el universo es muy grande ($N \to \infty$), y además f converge a cero, se tiene que,

$$V(\hat{A}) = \frac{N^2(1-f)}{n}\,\frac{N\,P\,Q}{N-1} = \frac{N^2\,P\,Q}{n}$$

ya que $(1-f) \to 1$ y como $(N-1) \to N$, entonces $\frac{N}{(N-1)} \to 1$.

Igualmente, $V(p) = \frac{P\,Q}{n}$

y las varianzas estimadas, $\hat{V}(\hat{A}) = \frac{N^2 p q}{n-1}$, $\hat{V}(p) = \frac{p q}{n-1}$.

A menudo se ve en publicaciones de censos y encuestas cuadros que cruzan dos o más variables, como por ejemplo el sexo, la edad, situación en la fuerza de trabajo y parentesco con el jefe del hogar. En el cuadro de la próxima página se observa un cruce de estas variables, donde Sexo tiene 2 categorías, Edad 8 categorías, Parentesco con el Jefe del Hogar 5 y Situación en la Fuerza de Trabajo 3; resultando un gran cuadro de 648 celdas, de las cuales 408 de totales (sombreadas), y sólo 240 no incluyen totales (figura 2.1)

En este caso, cada una de las celdas debe ser tratada como una característica, bajo el enfoque de proporciones. Por ejemplo, para la primera celda (excluyendo las de totales), se le asignará el valor 1 a una persona si es hombre, entre 12 y 14 años, que trabaja y es jefe de hogar; si no cumple con alguna de estas características tomará el valor cero, para esta celda y así sucesivamente. Lógicamente, toda persona de 12 años o más estará ubicada en sólo una celda (excluyendo las de totales), y para ella tomará el valor 1 y cero para todas y cada una del resto de las celdas de no totales.

Generalmente, en tabulados, el enfoque es de proporciones, aun cuando se obtengan totales. Nótese que incluso la edad, que es una variable numérica, es tratada como una variable cualitativa al cruzarse con otras, se puede tratar como una variable numérica, pero sólo para estimar la edad promedio, o la edad promedio de los hombres que trabajan y que son jefes de hogar (en el ejemplo).

Cabe destacar que en este cuadro se puede tratar la variable Situación en la Fuerza de Trabajo con más categorías (Trabaja, No Trabaja, Estudia, Pensionado/Jubilado, Buscando Trabajo por Primera Vez), e incluir la variable Área (urbana y rural), que aumentaría el total de celdas a 2916. Con esto se observa que a menudo se obtiene información compleja, y se deben estimar todas las celdas requeridas, por supuesto, con diferente precisión cada una.

Figura 2.1.- Total de Personas de 12 años o más, por Situación en la Fuerza de Trabajo y Parentesco con el Jefe del Hogar, según Sexo y Grupos de Edad

Sexo y Grupos de Edad	Total						Trabaja						No Trabaja						Inactivo					
	Total	Jefe de Hogar	Conyuge	Hijo	Otro Pariente	No Pariente	Total	Jefe de Hogar	Conyuge	Hijo	Otro Pariente	No Pariente	Total	Jefe de Hogar	Conyuge	Hijo	Otro Pariente	No Pariente	Total	Jefe de Hogar	Conyuge	Hijo	Otro Pariente	No Pariente
TOTAL																								
12 a 14 años																								
15 a 19 años																								
20 a 24 años																								
25 a 29 años																								
30 a 39 años																								
40 a 49 años																								
50 a 59 años																								
60 años y (+)																								
VARONES																								
12 a 14 años																								
15 a 19 años																								
20 a 24 años																								
25 a 29 años																								
30 a 39 años																								
40 a 49 años																								
50 a 59 años																								
60 años y (+)																								
HEMBRAS																								
12 a 14 años																								
15 a 19 años																								
20 a 24 años																								
25 a 29 años																								
30 a 39 años																								
40 a 49 años																								
50 a 59 años																								
60 años y (+)																								

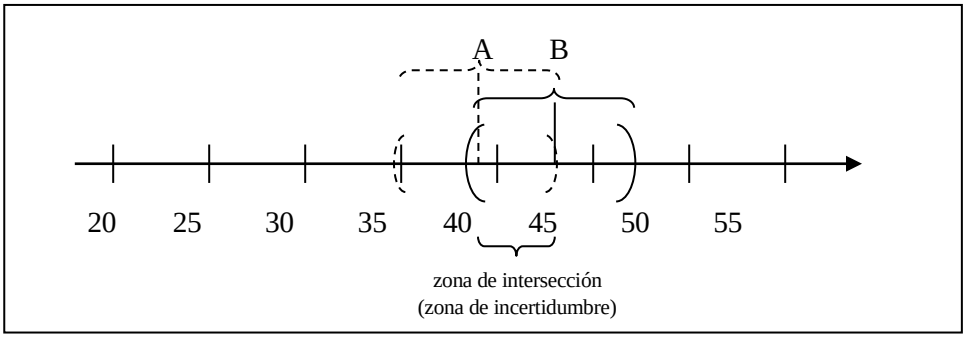
2.5.- Estimación por intervalos

Las estimaciones mostradas también son conocidas como “estimaciones puntuales”, y como se vio, están acompañadas de un error muestral, que abre un rango para dicho valor, con el cual aumenta la probabilidad de abarcar al valor poblacional, esto se hace construyendo un intervalo, que se denomina “intervalo de confianza” y es lo que se conoce como “estimación por intervalos”.

Se denomina intervalo de confianza al intervalo (o vector aleatorio) que con una cierta probabilidad, contiene al valor poblacional que se está estimando, y a dicha probabilidad “nivel de confianza” o simplemente “confianza, la misma se indica por 1-α y en algunas oportunidades se expresa en términos porcentuales 100(1-α)%.

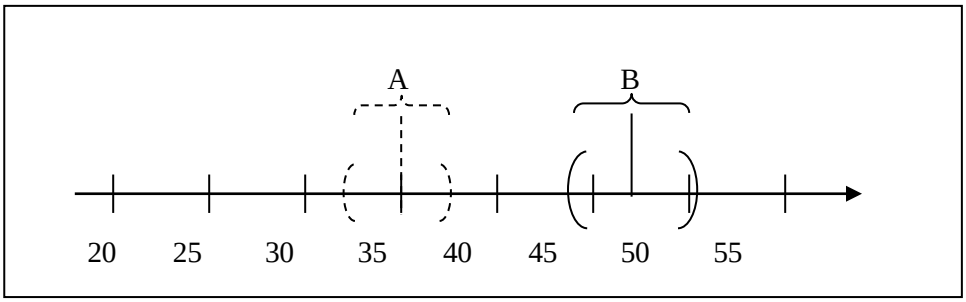
En muchas ocasiones es conveniente hacer estimaciones por intervalos. Por ejemplo, si se está comparando la preferencia sobre candidatos presidenciales, además de las estimaciones puntuales, es necesario analizar los comportamientos de ciertos rangos. Supóngase que sobre una muestra de 500 personas, la estimación de la preferencia del candidato A es del 39%, y la del candidato B es de 43%; gana el candidato B. Al construir los intervalos de confianza, se tiene que la preferencia del candidato A puede variar entre 34,7% y 43,3%, mientras que la del B está entre 38,7% y 47,3%, ambos con 0,95 de confianza. A claras se observa que se solapan los intervalos entre el 38,7% y el 43,3%, entonces, no se puede afirmar que el candidato B gana, porque existe una zona de incertidumbre, donde el candidato A puede estar ganándole al candidato B (figura 2.2).

Figura 2.2.- Intervalos de Confianza de mediciones electorales



Pero si en lugar de 500 personas, se entrevistan 1000 personas, y los resultados son del 35% y 47% respectivamente, los intervalos son (32% ; 38%) y (43,9% ; 50,1%). Puede apreciarse que no se interceptan los intervalos, esto indica que, con una confianza del 0,95 gana el candidato B (figura 2.3).

Figura 2.3.- Intervalos de Confianza de mediciones electorales



Generalmente se supone que las estimaciones $\hat{\bar{Y}}$ y $\hat{Y}$ se distribuyen de forma normal, cuyo valor esperado es el valor poblacional; esta aseveración se hace basado en el Teorema del Límite Central, que indica:

> Sea *y* una variable aleatoria, y $y_1, y_2, ..., y_N$ observaciones de dicha variable en un universo de *N* elementos, con promedio y varianzas finitos $\bar{Y}$ y S^2 respectivamente, y sea $\bar{y}$ el promedio de "*y*" en una muestra aleatoria de tamaño *n*. Entonces la distribución de
>
> $$\frac{\bar{y}-\bar{Y}}{V(\bar{y})}$$ converge a una distribución normal, *N*(0,1) cuando n $\rightarrow \infty$.
>
> Es decir, que $\bar{y}$ se distribuye como una normal $N(\bar{Y}, S^2)$. [9,155]

Basado en este resultado, la estructura del intervalo de confianza de $\hat{\theta}$, el estimador de θ, es la siguiente,

$$\left(\hat{\theta}-k_{\frac{\alpha}{2}}\sqrt{V(\hat{\theta})}\ ,\ \hat{\theta}+k_{1-\frac{\alpha}{2}}\sqrt{V(\hat{\theta})}\right)$$

donde *k* es el valor de una distribución de probabilidad, que en este caso es normal o *t*-student. Se usa la distribución *t*-student cuando no se conoce $V(\hat{\theta})$, entonces se usa el estimador $\hat{V}(\hat{\theta})$. Si se conoce $V(\hat{\theta})$ se usa la distribución normal. Sin embargo, cuando se desconoce $V(\hat{\theta})$, pero el tamaño de la muestra es grande, se usa la distribución normal, ya que, la distribución *t*-student converge a una distribución normal. Es decir,

i) si se conoce $V(\hat{\theta})$, $$\left(\hat{\theta}-z_{\frac{\alpha}{2}}\sqrt{V(\hat{\theta})}\ ,\ \hat{\theta}+z_{1-\frac{\alpha}{2}}\sqrt{V(\hat{\theta})}\right)$$

> donde *z* es el valor de la ordenada de la distribución normal *N*(0,1), cuando se deja a la izquierda de la distribución $\frac{\alpha}{2}$ y a la derecha $1-\frac{\alpha}{2}$ respectivamente, es decir, para un nivel de confianza de $(1-\alpha)$.

ii) si no se conoce $V(\hat{\theta})$, $\left(\hat{\theta} - t_{\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{\theta})} \quad , \quad \hat{\theta} + t_{1-\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{\theta})} \right)$

donde *t* es el valor de la ordenada de la distribución *t*-student, cuando se deja a la izquierda de la distribución $\frac{\alpha}{2}$ y a la derecha $1-\frac{\alpha}{2}$ respectivamente, es decir, para un nivel de confianza de $(1-\alpha)$.

iii) si no se conoce $V(\hat{\theta})$, pero *n* es grande, $\left(\hat{\theta} - z_{\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{\theta})} \ , \ \hat{\theta} + z_{1-\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{\theta})} \right)$

ya que, como se mencionó, en este caso $t_{\alpha/2}$ converge a $z_{\alpha/2}$, análogamente, $t_{1-\alpha/2}$ converge a $z_{1-\alpha/2}$, entonces es lo mismo usar la distribución normal que la *t*-student. Es importante aclarar que *n* se considera grande, según la mayoría de las tablas disponibles de la distribución *t*-student, para valores mayores a 120.

En otros términos,

Figura 2.4.- Distribución de probabilidad utilizada en intervalos de confianza del estimador del total y promedio

i)	si se conoce	$V(\hat{\theta})$	y *n* es grande	$\Rightarrow$ distribución normal
ii)	si se conoce	$V(\hat{\theta})$	y *n* es pequeña	$\Rightarrow$ distribución *t*-student
iii)	si no se conoce	$V(\hat{\theta})$	y *n* es grande	$\Rightarrow$ distribución normal
iv)	si no se conoce	$V(\hat{\theta})$	y *n* es pequeña	$\Rightarrow$ distribución *t*-student

3.- MUESTREO ALEATORIO SIMPLE CON REEMPLAZAMIENTO

El planteamiento de este diseño es muy similar al anterior, y su diferencia radica en que al seleccionar un elemento en la muestra, el mismo no es retirado del marco, sino que se mantiene y puede ser seleccionado nuevamente.

En el diseño anterior, la probabilidad de que el elemento *i-ésimo* sea seleccionado en la primera extracción es *1/N* , igual que en este diseño. La diferencia es a partir de la segunda extracción; mientras en el diseño en estudio la probabilidad se mantiene, porque se repone el elemento seleccionado, en el anterior es *1/(N-1)*. Así la probabilidad de seleccionar el elemento *i-ésimo* en la *h-ésima* selección, según el diseño en estudio y el anterior, respectivamente son:

$$\frac{1}{N} \quad , \quad \frac{1}{N-(h-1)} = \frac{1}{N-h+1}$$

3.1.- Estimadores

Curiosamente, los estimadores del total y promedio poblacional son exactamente iguales al caso anterior. Sean y_1, y_2, ... , y_N las observaciones de un universo de *N* elementos, y sean

$$Y = \sum_{i=1}^{N} y_i \quad , \quad \bar{Y} = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{Y}{N}$$

el total y el promedio poblacional; entonces, si $\hat{Y}_{cr}$ es el estimador de *Y*, se tiene que,

$$\hat{Y}_{cr} = \sum_{i=1}^{n} y_i + \sum_{j=n+1}^{N} \hat{y}_j \quad , \text{ donde} \quad \hat{y}_j = \frac{\sum_{i=1}^{n} y_j}{n} = \bar{y} \quad \forall\ j\ ;\ j = n+1, n+2, \ldots, N$$

por lo tanto, $\hat{Y}_{cr} = \sum_{i=1}^{n} y_i + (N-n)\bar{y} = n\,\bar{y} + (N-n)\bar{y} = N\,\bar{y}$

y como $\bar{Y} = \frac{Y}{N}$ entonces , $\hat{\bar{Y}}_{cr} = \bar{y}$

Para verificar si estos estimadores son insesgados se hallará la esperanza matemática. Anteriormente se tenían $\binom{N}{n}$ muestras posibles, ahora aumenta a N^n, ya que hay restitución. Entonces,

$$E(\hat{\bar{Y}}_{cr}) = \sum_{k=1}^{N^n} \hat{\bar{Y}}_k \, P_k = \frac{1}{N^n} \sum_{k=1}^{N^n} \hat{\bar{Y}}_k = \frac{1}{N^n} \sum_{k=1}^{N^n} \bar{y}_k = \frac{1}{N^n} \sum_{k=1}^{N^n} \left(\frac{\sum_{i=1}^{n} y_{ki}}{n} \right) = \frac{1}{n\,N^n} \sum_{k=1}^{N^n} \sum_{i=1}^{n} y_{ki}$$

$$= \frac{1}{n\,N^n} \sum_{i=1}^{N} n\,N^{n-1}\, y_i = \frac{1}{N} \sum_{i=1}^{N} y_i = \bar{Y}$$

por lo tanto, $\hat{\bar{Y}}_{cr} = \bar{y}$ es un estimador insesgado de $\bar{Y}$. Análogamente, $\hat{Y}_{cr}$ es un estimador insesgado de *Y*, ya que,

$$E(\hat{Y}_{cr}) = E(N\,\hat{\bar{Y}}_{cr}) = N\,E(\hat{\bar{Y}}_{cr}) = N\,\bar{Y} = Y$$

3.2.- Varianzas de los estimadores

Para determinar las varianzas, se procederá de la misma forma que en el capítulo anterior, aplicando el principio 2, esto es,

$$V(\hat{\bar{Y}}_{cr}) = \frac{\sum_{k=1}^{N^n} (\bar{y}_k - \bar{Y})^2}{N^n} = \frac{\sum_{k=1}^{N^n} (\bar{y}_k^2 - 2\,\bar{Y}\,\bar{y}_k + \bar{Y}^2)}{N^n} = \frac{\sum_{k=1}^{N^n} \bar{y}_k^2 - 2\,\bar{Y} \sum_{k=1}^{N^n} \bar{y}_k + \sum_{k=1}^{N^n} \bar{Y}^2}{N^n}$$

$$= \frac{\sum_{k=1}^{N^n} \bar{y}_k^2 - 2\,N^n\,\bar{Y} + N\,\bar{Y}^2}{N^n} = \frac{\sum_{k=1}^{N^n} \bar{y}_k^2}{N^n} - \bar{Y}^2$$

desarrollando el numerador del primer sumando,

$$\sum_{k=1}^{N^n} \bar{y}_k^2 = \sum_{k=1}^{N^n} \left(\frac{y_k}{n} \right)^2 = \frac{1}{n^2} \sum_{k=1}^{N^n} y_k^2 = \frac{1}{n^2} \sum_{k=1}^{N^n} \left(\sum_{i=1}^{n} y_i \right)^2$$

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

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

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

entonces,

$$V(\hat{\bar{Y}}_{cr}) = \frac{\frac{N^{n-2}}{n}\left[N\sum_{i=1}^{N} y_i^2 + (n-1)\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j\right]}{N^n} - \bar{Y}^2 = \frac{N\sum_{i=1}^{N} y_i^2 + (n-1)\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j}{n\,N^2} - \bar{Y}^2$$

$$= \frac{N\sum_{i=1}^{N} y_i^2 + (n-1)\sum_{i=1}^{N} y_i y_i + 2(n-1)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} y_i y_j}{n\,N^2} - \bar{Y}^2$$

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

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

donde,

$$S_|^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{Y})^2}{N} = \frac{\sum_{i=1}^{N} y_i^2 - N\bar{Y}^2}{N}$$

por lo tanto,

$$V(\hat{\bar{Y}}_{cr}) = \frac{S_|^2}{n} \qquad (3.1)$$

veáse que si *N* es grande y n/N es pequeño, la varianza del muestreo aleatorio simple sin restitución (2.4) converge a (3.1), es decir,

$$\text{Si} \quad N \to \infty \quad \text{y} \quad \frac{\text{n}}{\text{N}} \to 0 \quad \Rightarrow \quad V(\hat{\bar{Y}}) \to V(\hat{\bar{Y}}_{cr})$$

prueba.

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

pero como $f = n/N$ tiende a *0* (cero), entonces *(1-f)* tiende a 1, y resulta

$$V(\hat{\bar{Y}}) = \frac{1}{n}\frac{\sum_{i=1}^{N}(y_i - \bar{Y})^2}{N-1}$$

y como *N* tiende a infinito, entonces $S^2 \to S_|^2$, por lo tanto

$$V(\hat{\bar{Y}}) = \frac{1}{n}\frac{\sum_{i=1}^{N}(y_i - Y)^2}{N} = \frac{S_|^2}{n} = V(\hat{\bar{Y}}_{cr})$$

Igual que en el muestreo aleatorio simple sin restitución,

$$V(\hat{Y}_{cr}) = N^2\, V(\hat{\bar{Y}}_{cr}) = \frac{N^2}{n} S_|^2 \qquad (3.2)$$

Igual que para $V(\hat{\bar{Y}})$, si *N* es grande y n/N es pequeño, la varianza del muestreo aleatorio simple sin restitución (2.5) converge a (3.2), es decir,

$$\text{Si} \quad N \to \infty \quad \text{y} \quad \frac{\text{n}}{\text{N}} \to 0 \quad \Rightarrow \quad V(\hat{Y}) \to V(\hat{Y}_{cr})$$

Las estimaciones de las varianzas vienen dadas por,

$$\hat{V}(\hat{\bar{Y}}_{cr}) = \frac{1}{n}\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n} = \frac{s_|^2}{n}$$

$$\hat{V}(\hat{Y}_{cr}) = N^2\, \hat{V}(\hat{\bar{Y}}_{cr}) = \frac{N^2}{n} s_|^2$$

Los valores esperados de las varianzas estimadas son,

$$E\left[\hat{V}(\hat{\bar{Y}}_{cr})\right] = E\left[\frac{s_|^2}{n}\right] = \frac{1}{n}E\left[s_|^2\right] = \frac{1}{n}E\left[\frac{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}{n}\right] = \frac{1}{n}E\left[\frac{\sum_{i=1}^{n} y_i^2}{n} - \bar{y}^2\right] = \frac{1}{n^2}E\left[\sum_{i=1}^{n} y_i^2\right] - \frac{1}{n}E\left[\bar{y}^2\right]$$

desarrollando el segundo término, sabiendo que existen N^n muestras posibles,

$$E\left[\bar{y}^2\right] = E\left[\left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^2\right] = \frac{1}{n^2}E\left[\left(\sum_{i=1}^{n} y_i\right)^2\right] = \frac{1}{n^2}\frac{1}{N^n}\sum_{k=1}^{N^n}\left(\sum_{i=1}^{n} y_{ki}\right)^2$$

3.3.- Proporciones

Como se ha comentado, el enfoque de proporciones es un caso particular del promedio y del total, cuando la variable en estudio está definida para tomar únicamente como valores ceros y unos. En este sentido, adaptando las fórmulas mostradas arriba para este caso, se tiene que,

$$a_i = \begin{cases} 1 & \text{si el elemento } i-\acute{e}simo \text{ posee la característica en estudio} \\ 0 & \text{si el elemento } i-\acute{e}simo \text{ no posee la característica en estudio} \end{cases}$$

$$A = \sum_{i=1}^{N} a_i \quad , \quad P = \frac{\sum_{i=1}^{N} a_i}{N} = \frac{A}{N}$$

entonces,

$$\hat{A}_{cr} = N\,p \quad , \quad \hat{P}_{cr} = p$$

y las varianzas y sus estimadores,

$$V(\hat{P}_{cr}) = \frac{P(1-P)}{n} = \frac{PQ}{n} \quad , \quad V(\hat{A}_{cr}) = N^2\,V(\hat{P}_{cr}) = \frac{N^2 PQ}{n}$$

$$\hat{V}(\hat{P}_{cr}) = \frac{p\,q}{n} \quad , \quad \hat{V}(\hat{A}_{cr}) = N^2\,\hat{V}(\hat{P}_{cr}) = \frac{N^2 p\,q}{n}$$

4.- MUESTREO ALEATORIO ESTRATIFICADO

A veces tomar una muestra aleatoria simple, con o sin reemplazamiento, puede no ser la mejor elección, ya que puede resultar una muestra con alta concentración de elementos de ciertas características y pocos de otras, que generen un resultado poco ajustado a la realidad. Por ejemplo, si se quiere estimar el ingreso promedio de los empleados de una compañía, puede obtenerse una muestra con mayoría de directivos sobre otro tipo de empleados, y como los directivos tienen mayor ingreso, el promedio queda sobrestimado; análogamente, si se obtiene una muestra con mayoría de obreros, el ingreso promedio queda subestimado. Lo mismo puede ocurrir si se está midiendo el rating de televisión entre varias emisoras, y quedan seleccionados más hogares de un sector social bajo que del resto de los sectores, en una proporción diferente a como ocurre en la población, o el caso contrario, o más hogares de descendencia de un determinado país, o de una región específica del país, que traería como posible consecuencia, una estimación del rating muy diferente a la real.

En estos casos, lo más apropiado es asignar a la muestra una proporción de cada uno de estos grupos que mejore la estimación, que en algunos casos puede ser asignar a la muestra la misma proporción que en la población.

El muestreo estratificado prevé estos casos, y consiste en dividir la población en sectores o "estratos", y entonces seleccionar una muestra independiente en cada uno de ellos. La muestra en cada estrato, en este capítulo, será aleatoria simple sin reemplazamiento, y la asignación puede variar, según lo que convenga, en términos de reducir la varianza, el costo o las facilidades de desplazamiento en el campo (levantamiento). En este capítulo se describen varios tipos de asignación o afijación.

Sea un universo de N elementos, que se divide en L grupos exhaustivos y mutuamente excluyentes, que se denominarán "estratos", con $N_1,N_2,N_3, \ldots ,N_L$ elementos respectivamente, además $N = N_1 + N_2 + N_3 + \cdots + N_L$. El diseño consiste en tomar una muestra aleatoria simple sin reemplazamiento de elementos en cada uno de los estratos, de manera que se tendrán L tamaños muestrales, que se denotarán por $n_1,n_2,n_3, \ldots ,n_L$, que indican el total de elementos en la muestra en cada estrato. La fracción de muestreo en el estrato *h-ésimo* es $f_h = \frac{n_h}{N_h}$, $h=1,2,3, \ldots ,L$. La muestra debe ser tomada independientemente en cada estrato, es decir, que la selección de elementos en un estrato no afecta la selección de elementos de ningún otro estrato.

Las razones principales para utilizar el muestreo estratificado son:

i. si se desea obtener estimaciones separadas para cada grupo o sector, entonces éstos deben ser tratados como estratos.

ii. por conveniencia administrativa, cuando es necesario separar el levantamiento de la información por determinadas zonas o áreas geográficas, puede ser conveniente tratarlas como estratos.

iii. Cuando estratificar da lugar a ganancias en la precisión de los estimadores. En este caso es conveniente que los estratos sean homogéneos internamente y heterogéneos entre ellos, de acuerdo a la(s) variable(s) a estimar. [2, 125]

4.1.- Total y promedio poblacional y muestral

Se denotará por y_{hi} al valor de la variable *"y"* en el *i-ésimo* elemento del estrato *h*, $h=1,...,L$; $i=1,...,N_h$, entonces,

$$Y_h = \sum_{i=1}^{N_h} y_{hi} \quad \text{es el total poblaciona l del estrato } h$$

$$\bar{Y}_h = \frac{Y_h}{N_h} \quad \text{es el promedio poblaciona l del estrato } h$$

$$Y = \sum_{h=1}^{L} Y_h \quad \text{es el total poblaciona l}$$

$$\bar{Y} = \frac{Y}{N} \quad \text{es el promedio poblaciona l}$$

los valores muestrales equivalentes son,

$$y_h = \sum_{i=1}^{n_h} y_{hi} \quad \text{es el total muestral del estrato } h$$

$$\bar{y}_h = \frac{y_h}{n_h} \quad \text{es el promedio muestral del estrato } h$$

$$y = \sum_{h=1}^{L} y_h \quad \text{es el total muestral}$$

$$\bar{y} = \frac{y}{n} = \frac{1}{n}\sum_{h=1}^{L} y_h = \frac{1}{n}\sum_{h=1}^{L} n_h \bar{y}_h = \sum_{h=1}^{L} w_h \bar{y}_h \quad \text{es el promedio muestral}$$

que no es más que el promedio de los promedios de los estratos, ponderados por el peso de cada estrato en la muestra $\left(w_h = \frac{n_h}{n} \right)$.

4.2.- Estimadores y varianzas

Como en cada estrato se toma una muestra aleatoria simple, se utilizan los mismos estimadores mostrados en el capítulo anterior, para cada estrato, es decir,

$$\hat{Y}_h = N_h \bar{y}_h \qquad \hat{\bar{Y}}_h = \bar{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h}$$

$$V(\hat{Y}_h) = \frac{N_h^2 (N_h - n_h)}{N_h n_h} S_h^2 \qquad V(\hat{\bar{Y}}_h) = \frac{N_h - n_h}{N_h n_h} S_h^2$$

$$\hat{V}(\hat{Y}_h) = \frac{N_h^2 (N_h - n_h)}{N_h n_h} s_h^2 \qquad \hat{V}(\hat{\bar{Y}}_h) = \frac{N_h - n_h}{N_h n_h} s_h^2$$

donde

$$S_h^2 = \frac{\sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2}{N_h - 1} \qquad y \qquad s_h^2 = \frac{\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2}{n_h - 1}$$

El estimador del total poblacional es:

$$\hat{Y}_{st} = \sum_{h=1}^{L} \hat{Y}_h = \sum_{h=1}^{L} N_h \bar{y}_h$$

donde *st* indica que la muestra es estratificada. Para estimar el promedio poblacional, se tiene,

$$\hat{\bar{Y}}_{st} = \frac{1}{N} \hat{Y}_{st} = \frac{1}{N} \sum_{h=1}^{L} N_h \bar{y}_h = \sum_{h=1}^{L} W_h \bar{y}_h$$

que es equivalente a tomar el promedio ponderado de los promedios de los estratos, ponderando cada estrato por $W_h = \frac{N_h}{N}$, que es el peso del estrato *h* en la población.

La varianza del estimador del total poblacional es:

$$V(\hat{Y}_{st}) = V\left(\sum_{h=1}^{L} N_h \bar{y}_h \right) = \sum_{h=1}^{L} V(N_h \bar{y}_h) + 2 \sum_{h=1}^{L-1} \sum_{k=h+1}^{L} COV(N_h \bar{y}_h, N_k \bar{y}_k)$$

pero como las muestras son independientes en cada estrato, entonces

$$COV(N_h \bar{y}_h, N_k \bar{y}_k) = 0 \qquad , \forall\, h \neq k$$

y se tiene que,

$$V(\hat{Y}_{st}) \;=\; \sum V(N_h\,\bar{y}_h) \;=\; \sum_{h=1}^{L} N_h^2\,V(\bar{y}_h) \;=\; \sum_{h=1}^{L} \frac{N_h\,(N_h-n_h)}{n_h} S_h^2$$

y $$V(\bar{Y}_{st}) = V\left(\frac{1}{N}\hat{Y}_{st}\right) = \frac{1}{N^2}V(\hat{Y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h(N_h-n_h)}{n_h}S_h^2 = \sum_{h=1}^{L} W_h^2\,\frac{(N_h-n_h)}{N_h}\,\frac{S_h^2}{n_h}$$

4.3.- Afijación de la muestra

La afijación de la muestra consiste en determinar los tamaños muestrales de cada estrato, es decir, fijar los n_h. Existen varias formas de hacerlo, y en gran medida dependerá del tipo de estudio, el tiempo y recursos económicos disponibles para hacerlo, la disponibilidad y condiciones del marco muestral, precisión de las estimaciones. Sin embargo, aquí se enumerarán las formas de hacerlo, para poder evaluar las ventajas y desventajas de los mismos, y así determinar, tomando en cuenta los factores mencionados, la afijación a efectuar.

En principio se conocen 4 formas de afijación, igual, proporcional, arbitraria y óptima o de Neyman.

Afijación Igual: se toman muestras aleatorias simples del mismo tamaño en cada estrato

$$n=n_1+n_2+\cdots+n_L \quad , \quad n_1=n_2=\cdots=n_L$$

$$\Rightarrow \; n_h=\frac{n}{L} \qquad , \; h=1,2,...,L \tag{4.1}$$

Afijación Proporcional: se toman muestras aleatorias simples en cada estrato, de manera de distribuir la muestra de tamaño n de acuerdo con el peso relativo que cada estrato toma en la población, según el total de elementos. Sea W_h el peso relativo del estrato h en la población, $W_h=N_h/N$, y w_h el peso relativo del estrato h en la muestra, $w_h=n_h/n$; se tiene que

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

entonces

$$n_h = n\frac{N_h}{N} \qquad , \; \forall h \; , \; h=1,2,...,L \tag{4.2}$$

nótese que

$$n_1+n_2+\cdots+n_L = n\frac{N_1}{N}+n\frac{N_2}{N}+\cdots+n\frac{N_L}{N}=\frac{n}{N}\left(N_1+N_2+\cdots+N_L\right)=\frac{n}{N}N = n$$

Si los tamaños de los estratos *(N_h)* son iguales, las afijaciones proporcional e igual coinciden.

Afijación Arbitraria: consiste en asignar un tamaño de muestra a cada estrato, que no responde a ningún criterio pre-establecido

Afijación de Neyman: esta afijación debe satisfacer la condición que $V(\bar{y}_{st})$ sea menor que la correspondiente a cualquier otra afijación. El nombre se debe a su autor Jerzy Neyman (1934), aunque más tarde se encontró un artículo de A. A. Tschuprow (o Chuprov) que data de 1923, con algunas fórmulas aplicadas por Neyman. Para hallar los tamaños muestrales se debe minimizar (*min*) la varianza del estimador, sujeto a (*sa*) que $n_1+n_2+...+n_L = n$, esto es,

$$min \left[\frac{1}{N^2} \sum_{h=1}^{L} N_h (N_h - n_h) \frac{S_h^2}{n_h} \right]$$

$$sa \quad \sum_{h=1}^{L} n_h \quad = \quad n$$

Cabe destacar, que aunque el tamaño de muestra es un valor entero, es decir, no es continuo, por facilidad, se hará la abstracción de considerarlo como tal. Aplicando el método de los multiplicadores de Lagrange, se tiene la siguiente ecuación,

$$L = \frac{1}{N^2} \left[\sum_{h=1}^{L} N_h (N_h - n_h) \frac{S_h^2}{n_h} \right] - \lambda \left[\sum_{h=1}^{L} n_h - n \right]$$

$$= \frac{1}{N^2} \left[\sum_{h=1}^{L} \frac{N_h^2 S_h^2}{n_h} - N_h^2 S_h^2 \right] - \lambda \left[\sum_{h=1}^{L} n_h - n \right]$$

y derivando,

$$\frac{\partial L}{\partial n_1} = -\frac{N_1^2 S_1^2}{N^2 n_1^2} - \lambda = 0 \qquad \Rightarrow \; n_1 = \frac{N_1 S_1}{N} \sqrt{-\frac{1}{\lambda}} \qquad (1)$$

$$\vdots$$

$$\frac{\partial L}{\partial n_L} = -\frac{N_L^2 S_L^2}{N^2 n_L^2} - \lambda = 0 \qquad \Rightarrow\ n_L = \frac{N_L S_L}{N}\sqrt{-\frac{1}{\lambda}} \qquad \text{(L)}$$

$$\frac{\partial L}{\partial \lambda} = n - n_1 - n_2 - \cdots - n_L = 0 \qquad \text{(L+1)}$$

sustituyendo $n_1, \ldots, n_L$ hallados en las ecuaciones $1,\ldots,L$, en la ecuación $(L+1)$,

$$n - \frac{N_1 S_1}{N}\sqrt{-\frac{1}{\lambda}} - \frac{N_2 S_2}{N}\sqrt{-\frac{1}{\lambda}} - \ldots - \frac{N_L S_L}{N}\sqrt{-\frac{1}{\lambda}} = 0$$

$$n - \left(\frac{1}{N}\right)\left(\sqrt{-\frac{1}{\lambda}}\right)\left[N_1 S_1 + N_2 S_2 + \ldots + N_L S_L\right] = 0$$

$$\lambda = -\left[\frac{N_1 S_1 + N_2 S_2 + \ldots + N_L S_L}{N \quad n}\right]^2 = -\left[\frac{\sum_{h=1}^{L} N_h S_h}{N \quad n}\right]^2$$

sustituyendo este resultado en la ecuación *h-ésima*,

$$n_h = \frac{N_h S_h}{N}\sqrt{-\frac{1}{-\left[\frac{\sum_{h=1}^{L} N_h S_h}{N n}\right]^2}} = n\frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} \qquad (4.3)$$

Puede observarse que es una afijación proporcional al peso relativo del estrato, medido en términos del total de elementos y dispersión de la variable a medir (y), simultáneamente, es decir que no sólo toma en cuenta el tamaño, sino también la variabilidad del estrato; entonces, si dos estratos tienen el mismo tamaño, el tamaño de la muestra será mayor para aquel que sea más heterogéneo, mientras que la afijación proporcional asigna tamaños de muestras iguales.

Sin embargo, si las varianzas de los estratos (S_h^2) son iguales, entonces la afijación de Neyman coincide con la afijación proporcional (4.2).

Afijación

Optima: este es un caso generalizado de la afijación de Neyman, que considera los costos, y que se debe a Yates y Zacopanay (1935). Consiste en minimizar la misma varianza del caso anterior, pero sujeto a un costo fijo, aunque también se puede hacer al contrario, minimizar el costo sujeto a una precisión o varianza fija, a continuación se presentan ambos procedimientos para hallar n_h. Antes se debe definir la variable *c* como el costo del levantamiento de la información, y

$$c = c_0 + c_1 n_1 + c_2 n_2 + \cdots + c_L n_L$$

la función de costo, en su forma más simple, donde c_0 es el costo fijo, y c_h es el costo de incluir a un elemento en la muestra en el estrato *h, h=1,2,...,L*.

i) minimizar la varianza sujeta a un costo fijo

$$\min \quad \left[\frac{1}{N^2} \sum_{h=1}^{L} N_h \, (N_h - n_h) \frac{S_h^2}{n_h} \right]$$

$$sa \quad c_0 + \sum_{h=1}^{L} c_h \, n_h = c$$

aplicando el método de los multiplicadores de Lagrange, se tiene la siguiente ecuación,

$$L = \frac{1}{N^2} \left[\sum_{h=1}^{L} N_h (N_h - n_h) \frac{S_h^2}{n_h} \right] - \lambda \left[c_0 + \sum_{h=1}^{L} c_h n_h - c \right]$$

$$= \frac{1}{N^2} \left[\sum_{h=1}^{L} N_h \left(\frac{N_h}{n_h} - 1 \right) S_h^2 \right] - \lambda \left[c_0 + \sum_{h=1}^{L} c_h n_h - c \right]$$

y derivando,

$$\frac{\partial L}{\partial n_1} = -\frac{N_1^2\, S_1^2}{N^2\, n_1^2} - \lambda c_1 = 0 \qquad \Rightarrow \quad n_1 = \frac{N_1 S_1}{N}\sqrt{-\frac{1}{\lambda c_1}} \qquad (1)$$

$$\vdots \qquad\qquad\qquad \vdots$$

$$\vdots \qquad\qquad\qquad \vdots$$

$$\frac{\partial L}{\partial n_L} = -\frac{N_L^2\, S_L^2}{N^2\, n_L^2} - \lambda c_L = 0 \qquad \Rightarrow \quad n_L = \frac{N_L S_L}{N}\sqrt{-\frac{1}{\lambda c_L}} \qquad (L)$$

$$\frac{\partial L}{\partial \lambda} = c - c_0 - c_1 n_1 - c_2 n_2 - \ldots - c_L n_L = 0 \qquad (L+1)$$

sustituyendo en la ecuación (L+1), los valores de n_1, ... , n_L hallados en las ecuaciones *1,...,L*,

$$c - c_0 - c_1 \frac{N_1\, S_1}{N}\sqrt{-\frac{1}{\lambda c_1}} - c_2 \frac{N_2\, S_2}{N}\sqrt{-\frac{1}{\lambda c_2}} - \ldots - c_L \frac{N_L\, S_L}{N}\sqrt{-\frac{1}{\lambda c_L}} = 0$$

$$c - c_0 - \left(\frac{N_1 S_1}{N}\frac{\sqrt{c_1}}{\sqrt{-\lambda}}\right) - \left(\frac{N_2 S_2}{N}\frac{\sqrt{c_2}}{\sqrt{-\lambda}}\right) - \cdots - \left(\frac{N_L S_L}{N}\frac{\sqrt{c_L}}{\sqrt{-\lambda}}\right) = 0$$

$$(c - c_0)\, N\sqrt{-\lambda} = N_1\, S_1\, \sqrt{c_1} + N_2\, S_2 \sqrt{c_2} + \ldots + N_L\, S_L\, \sqrt{c_L}$$

$$\lambda = -\left[\frac{N_1 S_1\sqrt{c_1} + N_2 S_2\sqrt{c_2} + \ldots + N_L S_L\sqrt{c_L}}{(c - c_0)N}\right]^2 = -\left[\frac{\sum_{h=1}^{L} N_h S_h \sqrt{c_h}}{(c - c_0)N}\right]^2$$

sustituyendo este resultado en la ecuación *h-ésima*,

$$n_h = \frac{N_h S_h}{N}\sqrt{-\frac{1}{-\left[\frac{\sum_{h=1}^{L} N_h S_h \sqrt{c_h}}{(c-c_0)N}\right]^2}\, c_h} = \frac{(c-c_0)\left(\frac{N_h S_h}{\sqrt{c_h}}\right)}{\sum_{h=1}^{L} N_h S_h \sqrt{c_h}}$$

pero se tiene n_h en función del costo fijo, de los costos variables, de las desviaciones y del tamaño de los estratos, pero no en función del tamaño de la muestra, *n*. En tal caso habría que calcular los n_h y luego sumarlos para obtener *n*. Esto puede ser un poco tedioso y a veces hasta impreciso, ya que si al calcular *n* por la suma de los n_h, se obtiene un tamaño

de muestra más grande de lo previsto, habría que ir modificando los n_h para cuadrar a n. Por ello sería conveniente tener n_h en función de los costos, desviaciones y tamaños de los estratos, pero también del tamaño muestral total, n. Obsérvese el siguiente artificio,

$$n = \sum_{h=1}^{L} n_h = \frac{(c - c_0)\left(\sum_{h=1}^{L} \frac{N_h S_h}{\sqrt{c_h}}\right)}{\sum_{h=1}^{L} N_h S_h \sqrt{c_h}}$$

entonces,

$$\frac{n_h}{n} = w_h = \frac{\left(\frac{N_h S_h}{\sqrt{c_h}}\right)}{\sum_{h=1}^{L}\left(\frac{N_h S_h}{\sqrt{c_h}}\right)}$$

y de aquí se puede escribir n_h en función de n,

$$n_h = n \frac{\left(\frac{N_h S_h}{\sqrt{c_h}}\right)}{\sum_{h=1}^{L}\left(\frac{N_h S_h}{\sqrt{c_h}}\right)} \qquad (4.4)$$

Si se aplica el otro planteamiento,

ii) minimizar el costo sujeto a una varianza fija

$$\min \quad c_o + \sum_{h=1}^{L} c_h n_h$$

$$sa \quad \frac{1}{N^2} \sum_{h=1}^{L} N_h (N_h - n_h) \frac{S_h^2}{n_h} = V$$

donde V es la varianza deseada, ya que se tiene que fijar un nivel de precisión. La ecuación de Lagrange es,

$$L = c_0 + \sum_{h=1}^{L} c_h \, n_h - \lambda \left[\frac{1}{N^2} \left(\sum_{h=1}^{L} N_h \, (N_h - n_h) \frac{S_h^2}{n_h} \right) - V \right]$$

$$= c_0 + \sum_{h=1}^{L} c_h \, n_h - \frac{\lambda}{N^2} \left[\sum_{h=1}^{L} N_h \, S_h^2 \left(\frac{N_h}{n_h} - 1 \right) - V N^2 \right]$$

y derivando,

$$\frac{\partial L}{\partial n_1} = c_1 + \lambda \, \frac{N_1^2 \, S_1^2}{N^2 \, n_1^2} = 0 \qquad \Rightarrow \quad n_1 = \frac{N_1 S_1}{N} \sqrt{-\frac{\lambda}{c_1}} \qquad (1)$$

$$\vdots \qquad\qquad \vdots$$

$$\vdots \qquad\qquad \vdots$$

$$\frac{\partial L}{\partial n_L} = c_L + \lambda \, \frac{N_L^2 \, S_L^2}{N^2 \, n_L^2} = 0 \qquad \Rightarrow \quad n_L = \frac{N_L S_L}{N} \sqrt{-\frac{\lambda}{c_L}} \qquad (L)$$

$$\frac{\partial L}{\partial \lambda} = -\frac{1}{N^2} \left[\sum_{h=1}^{L} N_h \, S_h^2 \left(\frac{N_h}{n_h} - 1 \right) \right] + V = 0 \qquad (L+1)$$

sustituyendo en la ecuación (L+1), los valores de n_1, ... , n_L hallados en las ecuaciones *1,...,L*,

$$-\frac{1}{N^2}\left[\sum_{h=1}^{L} N_h S_h^2\left(\frac{N_h}{\frac{N_h S_h}{N}\sqrt{\frac{-\lambda}{c_h}}}-1\right)-V N^2\right]=0$$

entonces,

$$\sum_{h=1}^{L}\frac{N_h S_h N\sqrt{c_h}}{\sqrt{-\lambda}}-N_h S_h^2=V N^2$$

$$\lambda=-\frac{N^2\left[\sum_{h=1}^{L} N_h S_h\sqrt{c_h}\right]^2}{\left[V N^2+\sum_{h=1}^{L} N_h S_h^2\right]^2}$$

sustituyendo este resultado en la ecuación *h-ésima*,

$$n_h=\frac{N_h S_h}{\sqrt{c_h}}\frac{\sum_{h=1}^{L} N_h S_h\sqrt{c_h}}{V N^2+\sum_{h=1}^{L} N_h S_h^2}$$

aplicando el mismo artificio del caso anterior para obtener n en función de n_h,

$$n=\sum_{h=1}^{L}\left(\frac{N_h S_h}{\sqrt{c_h}}\right)\frac{\sum_{h=1}^{L} N_h S_h\sqrt{c_h}}{V N^2+\sum_{h=1}^{L} N_h S_h^2}$$

entonces,

$$\frac{n_h}{n}=w_h=\frac{\left(\frac{N_h S_h}{\sqrt{c_h}}\right)}{\sum_{h=1}^{L}\left(\frac{N_h S_h}{\sqrt{c_h}}\right)}$$

y de aquí se puede escribir n_h en función de n,

$$n_h = n \frac{\left(\frac{N_h S_h}{\sqrt{c_h}} \right)}{\sum_{h=1}^{L} \left(\frac{N_h S_h}{\sqrt{c_h}} \right)} \qquad (4.5)$$

que es el mismo resultado al que se llegó en (4.4).

La afijación óptima le asignará mayor tamaño muestral a los estratos que sean:

- más grandes
- más dispersos
- más económicos (la captura de información)

Nótese que si se toma la fórmula de n_h en la afijación óptima, pero con un costo fijo en cada estrato, es decir $c_h = c_k$, $h \neq k$, con $h,k = 1,2,..,L$; entonces $c = c_0 + Lc_1$, y se obtiene la fórmula de la afijación de Neyman vista en (4.3). Por esto se dijo que la afijación óptima es un caso general de la afijación de Neyman.

5.- ESTUDIO DE SUBPOBLACIONES

A menudo se requiere hacer estudios sobre un grupo específico del universo, del cual no se dispone un marco exclusivo o detallado, sino de uno que incluye tanto a esta población como a otras, sin que sea factible identificar a los elementos de interés a partir de este. Por ejemplo, se desea hacer un estudio sobre los hogares con niños en edad escolar, para medir su nivel socioeconómico, así como el total de hogares con niños que no asisten a la escuela. Lo más seguro, en el mejor de los casos, es que se disponga de un marco que consista en una lista de hogares, con su respectiva ubicación geográfica.

En general, se desea hacer un estudio sobre un grupo o subconjunto del universo, contenido en el universo estadístico, pero que no puede identificarse a priori en el marco muestral. A dicho grupo se le conoce como "subpoblación."

Otro caso es cuando se desea hacer estudios similares sobre varias poblaciones, y se requiere recolectar la misma información para cada población, y sería muy costoso hacer selecciones independientes para cada una, aun en el caso de que se disponga del marco específico. Tal es el caso de una encuesta que captara el total de hombres entre 15 y 19 años que trabajan, hombres entre 20 y 24 años, igualmente para mujeres. Suponiendo que se tiene un marco de personas, diferenciadas por sexo y edad, con su respectiva ubicación geográfica, se hace costoso e improcedente hacer cuatro muestras independientes, en lugar de seleccionar hogares e incluir a todas las personas de esos hogares.

En otros casos, se requiere estimar la proporción de hogares cuyo jefe sea mujer, con niños y que además sea pobre; lo más probable es encontrar un marco que incluya a todos los hogares, y además el sólo hecho de medir la pobreza es algo complejo.

De cualquier modo, el muestreo es complicado, ya que no se sabría si un elemento pertenece o no a la subpoblación específica sino hasta después de que ha sido seleccionado.

A efectos de ilustrar el procedimiento, se tomará inicialmente el muestreo aleatorio simple. Sea N el total de elementos en el universo estadístico, y N_i el total de elementos en la subpoblación i-ésima, $i=1,2,...,H$; se toma una muestra de n elementos, donde n_i pertenecen a la subpoblación i-ésima. Se define y_{ij} como el valor de la variable "y" arrojado por el elemento j-ésimo de la subpoblación i-ésima, donde $j=1,2,...,N_i$.

Se tienen los siguientes valores poblacionales:

$$Y_i = \sum_{j=1}^{N_i} y_{ij}$$ total de la subpoblación *i-ésima*

$$\overline{Y}_i = \frac{\sum_{j=1}^{N_i} y_{ij}}{N_i}$$ promedio de la subpoblación *i-ésima*

$$S_i^2 = \frac{\sum_{j=1}^{N_i} \left(y_{ij} - \overline{Y}_i \right)^2}{N_i - 1}$$ varianza de la variable *"y"* en la subpoblación *i-ésima*

y como es una muestra aleatoria simple, sus respectivos estimadores son:

$$\hat{Y}_i = N_i \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i} = N_i \overline{y}_i$$ estimador del total de la subpoblación *i-ésima* (5.1)

$$\hat{\overline{Y}}_i = \overline{y}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$$ estimador del promedio de la subpoblación *i-ésima* (5.2)

$$\hat{S}_i^2 = s_i^2 = \frac{\sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2}{n_i - 1}$$ varianza muestral de la variable *"y"* en la subpoblación *i-ésima* (5.3)

y las varianzas de los estimadores son los siguientes:

$$V(\hat{Y}_i) = \frac{N_i^2 (N_i - n_i)}{N_i} \frac{S_i^2}{n_i} = \frac{N_i^2 (1 - f_i)}{n_i} S_i^2 \quad , \quad V(\hat{\overline{Y}}_i) = V(\overline{y}_i) = \frac{N_i - n_i}{N_i} \frac{S_i^2}{n_i} = \frac{1 - f_i}{n_i} S_i^2$$

y las varianzas estimadas

$$\hat{V}(\hat{Y}_i) = \frac{N_i^2 (N_i - n_i)}{N_i} \frac{s_i^2}{n_i} = \frac{N_i^2 (1 - f_i)}{n_i} s_i^2 \quad , \quad \hat{V}(\hat{\overline{Y}}_i) = \hat{V}(\overline{y}_i) = \frac{N_i - n_i}{N_i} \frac{s_i^2}{n_i} = \frac{1 - f_i}{n_i} s_i^2$$

En muchos casos no se conoce N_i, entonces los estimadores y, por supuesto, sus varianzas sufren algunas modificaciones. El procedimiento para hallar los estimadores, bien sea del total o del promedio, es similar al de proporciones, en el sentido de que se le asigna el valor *0* (cero) a los elementos que no pertenecen a la subpoblación

en cuestión y si pertenece, toma el valor que ya tiene, es decir,

$$y'_{ij} = \begin{cases} y_{ij} & \text{si la subpoblación en estudio es la i–ésima} \\ 0 & \text{si la subpoblación en estudio no es la i–ésima} \end{cases}$$

Para determinar el estimador del total de la subpoblación *i*-ésima, se aplica el estimador conocido para el total en el muestreo aleatorio simple, que es,

$$\hat{Y} = N\frac{\sum_{k=1}^{H}\sum_{j=1}^{n_i} y'_{kj}}{n} = \frac{N}{n}\left[\sum_{j=1}^{n_i} y'_{ij} + \sum_{\substack{k=1 \\ k\neq i}}^{H}\sum_{j=1}^{n_k} y'_{kj}\right] = \frac{N}{n}\left[\sum_{j=1}^{n_i} y'_{ij} + 0\right] = \frac{N}{n}\sum_{j=1}^{n_i} y_{ij} = \hat{Y}_i \qquad (5.4)$$

en otras palabras, se sustituye el factor de expansión de $\frac{N_i}{n_i}$ por $\frac{N}{n}$. De igual forma se calcula la varianza,

$$\hat{S}'^2 = s'^2 = \frac{\sum_{k=1}^{H}\sum_{j=1}^{n_k}(y'_{kj} - \bar{y}')^2}{n-1} = \frac{\sum_{j=1}^{n_i}(y'_{ij} - \bar{y}')^2 + \sum_{\substack{k=1 \\ k\neq i}}^{H}\sum_{j=1}^{n_k}(y'_{kj} - \bar{y}')^2}{n-1} = \frac{\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}'\right)^2 + (n - n_i)\,\bar{y}'^2}{n-1} = \frac{\sum_{j=1}^{n_i} y_{ij}^2 - n\,\bar{y}'^2}{n-1} \qquad (5.5)$$

donde, $\bar{y}' = s'^2 = \frac{\sum_{k=1}^{H}\sum_{j=1}^{n_k} y'_{kj}}{n} = \frac{\sum_{j=1}^{n_i} y'_{ij} + \sum_{\substack{k=1 \\ k\neq i}}^{H}\sum_{j=1}^{n_k} y'_{kj}}{n} = \frac{\sum_{j=1}^{n_i} y'_{ij}}{n} = \frac{\sum_{j=1}^{n_i} y_{ij}}{n}$

Nótese que para estimar el promedio no hace falta conocer N_i, entonces, para el estimador del promedio se mantiene la expresión (5.2). La fórmula para hallar la varianza estimada es la misma utilizada en el muestreo aleatorio simple,

$$\hat{V}(\hat{Y}) = v(\hat{Y}) = N^2(N-n)\frac{s'^2}{n}$$

Como ilustración, supóngase que se hace una encuesta para investigar ciertas características sobre los estudiantes de una universidad. El marco muestral está compuesto por una lista de 4375 estudiantes de toda la universidad y se toma una muestra aleatoria simple de tamaño 125, de los cuales 28 estudian administración, 35 ingeniería, 18 medicina, 12 derecho, 9 economía y 23 sociología. Sea y la siguiente variable

y_{ij} = matrícula pagada a través de todos sus estudios, por el estudiante j-ésimo de la carrera i-ésima

Se obtiene el siguiente cuadro,

Tabla 5.1 – Tamaños de muestra y valores muestrales

Carrera	n_i	y_i	$\sum_{i=1}^{n_i} y_{ij}^2$
Total	125	1.286.158	16.789.753.206
Administración	28	245.881	3.139.138.023
Ingeniería	35	353.897	4.247.781.979
Medicina	18	214.291	2.867.810.289
Derecho	12	178.529	2.840.802.275
Economía	9	100.616	1.336.507.834
Sociología	23	192.944	2.357.712.806

Se desea estimar el total cancelado por carrera, por concepto de matrícula, y además el promedio cancelado por cada estudiante, según la carrera. Para estimar el promedio, se toma el promedio muestral, y como no se conocen los N_i, para estimar el total de bolívares cancelados por todos los estudiantes, según la carrera, se utiliza la fórmula (5.4), y se obtienen los siguientes resultados (opción A):

Tabla 5.2 – Estimaciones de la opción A

Carrera	promedio	Total Estimado (A)
Total	10.289,26	45.015.530,00
Administración	8.781,46	8.605.835,00
Ingeniería	10.111,34	12.386.395,00
Medicina	11.905,06	7.500.185,00
Derecho	14.877,42	6.248.515,00
Economía	11.179,56	3.521.560,00
Sociología	8.388,87	6.753.040,00

Supóngase que las coordinaciones de administración y economía tienen registrados el total de estudiantes de sus respectivas carreras, que son 1.013 de administración y 526 de economía, de las demás se desconoce el número. Se recalculan las estimaciones del total de bolívares cancelados por todos los estudiantes, según la carrera; para administración y economía se utiliza la fórmula (5.1), y para el resto, se aplica la fórmula (5.4), pero sólo con los estudiantes de esas carreras, es decir, ahora se tendrá N=2836 y n=88 (opción B).

Tabla 5.3 – Tamaños de poblacionales y muestrales, estimaciones opción B

Carrera	N_i	n_i	Total Estimado (B)
Total	4.375	125	45.058.780,86
Administración	1.013	28	8.895.623,32
Ingeniería		35	11.405.135,14
Medicina		18	6.906.014,50
Derecho		12	5.753.502,77
Economía	526	9	5.880.446,22
Sociología		23	6.218.058,91

Luego de una ardua búsqueda, se encontró el total de estudiantes de cada carrera; entonces se procede a recalcular las estimaciones de las carreras para las cuales no se disponía de esta cifra, aplicando la fórmula (5.1). Para administración y economía las estimaciones resultan igual a las anteriores (opción C),

Tabla 5.3 – Tamaños de poblacionales y muestrales, estimaciones opción C

Carrera	N_i	n_i	Total Estimado (C)
Total	4375	125	46.292.560,47
Administración	1013	28	8.895.623,32
Ingeniería	1426	35	14.418.774,91
Medicina	456	18	5.428.705,33
Derecho	565	12	8.405.740,42
Economía	526	9	5.880.446,22
Sociología	389	23	3.263.270,26

Obsérvese que, salvo para las carreras mencionadas, las estimaciones son diferentes. Es de suponer que las mostradas en el último cuadro son las mejores, ya que se dispone del total de estudiantes de cada carrera, y en las otras no, sin embargo, habría que revisar sus respectivas medidas de precisión. De cualquier modo, está última estimación dio más trabajo que las anteriores, en el sentido de encontrar los datos necesarios. A continuación se muestran las varianzas, errores estándar y coeficientes de variación para cada caso, a efectos de compararlos.

Tabla 5.4 – Varianzas estimadas de los estimadores del total, por opción según carrera

Carrera	Varianzas Estimadas		
	(A)	(B)	(C)
Total	4.265.926.872.946,29	4.265.926.872.946,29	4.265.926.872.946,29
Administración	3.185.503.195.077,90	1.293.375.434.539,39	1.293.375.434.539,39
Ingeniería	3.893.695.937.996,29	2.875.230.168.538,69	1.115.811.007.943,13
Medicina	2.999.526.029.019,03	2.388.065.562.448,35	206.688.474.421,62
Derecho	3.101.943.063.189,19	2.523.073.240.726,96	437.309.276.923,77
Economía	1.506.116.780.281,13	799.455.364.381,33	799.455.364.381,33
Sociología	2.471.041.848.392,42	1.969.377.157.307,56	207.970.360.358,51

Tabla 5.5 – Errores estándar y coeficientes de variación estimados de los estimadores del total, por opción según carrera

Carrera	Errores Estándar Estimados			Coeficientes de Variación Estimados		
	(A)	(B)	(C)	(A)	(B)	(C)
Total	2.065.412,03	2.065.412,03	2.065.412,03	0,0459	0,0458	0,0446
Administración	1.784.797,80	1.137.266,65	1.137.266,65	0,2074	0,1278	0,1278
Ingeniería	1.973.245,03	1.695.650,37	1.056.319,56	0,1593	0,1487	0,0733
Medicina	1.731.913,98	1.545.336,71	454.630,04	0,2309	0,2238	0,0837
Derecho	1.761.233,39	1.588.418,47	661.293,64	0,2819	0,2761	0,0787
Economía	1.227.239,50	894.122,68	894.122,68	0,3485	0,1521	0,1521
Sociología	1.571.954,79	1.403.344,99	456.037,67	0,2328	0,2257	0,1397

Nótese que, como se esperaba, tienen mejor precisión las estimaciones hechas en última instancia (C), ya que se conocían todos los N_i, además, puede observarse que las varianzas y demás medidas de precisión para administración y economía en las estimaciones (B) y (C) son iguales, porque, como se aclaró, son los mismos estimadores. Resulta curioso que si bien las estimaciones de los totales de bolívares cancelados por todos los estudiantes sin tomar en cuenta las carreras, son diferentes, las varianzas y errores estándar son iguales, no así con los coeficientes de variación. Estos últimos son diferentes, debido a que los totales son diferentes, ya que $cv(\beta)=ee(\beta)/E(\beta)$.

Al estimar la suma de los totales de las subpoblaciones, se utiliza la fórmula (5.1), donde N_i y n_i son N y n respectivamente, ya que los elementos de la subpoblación son todos los elementos de la población. Esto hace que las fórmulas (5.1) y (5.4) resulten iguales para la suma de los totales, y además resulten igual a (2.1), que es la fórmula del estimador del total en el muestreo aleatorio simple. Por este hecho, las varianzas y errores estándar son iguales. Lo extraño, después de esta explicación, es que las estimaciones de los totales sean diferentes. Resulta que en este ejemplo, al igual que ocurre en la mayoría de los estudios, el total mostrado en cada caso, no es producto de aplicar la citada fórmula, sino que es la suma de los totales para cada subpoblación, y esto se hace para que la presentación de los resultados luzca con mayor coherencia.

Existen estudios más complejos que este ejemplo, como cuando se hacen encuestas para determinar el desempleo y caracterizar la fuerza de trabajo, o cuando se toma una muestra para levantar información adicional, aunque simultánea, al Censo de Población, o en encuestas agropecuarias, donde se investigan muchas variables -en ocasiones puede variar entre 30 y 50-, entonces se trata de buscar una o varias variables con alta correlación o asociación positiva (si se trata de variables cualitativas) con las variables a estimar, y que además se conozcan sus totales poblacionales, para que funjan como variables auxiliares.

6.- TAMAÑO DE LA MUESTRA

El cálculo del tamaño de la muestra es un tema que preocupa a todos los que tienen que ver con investigaciones por muestreo. A menudo se oyen expresiones como "cuál es el tamaño de la muestra para estimar el coeficiente de penetración de este producto", o “para estimar el total de hogares con niños que no asisten a la escuela”, y mucha gente no sabe darle respuesta a estas situaciones. Algunos osados sacan a colación fórmulas aprendidas rápidamente, que en algunas circunstancias no bastan, y que además, de la manera en que la usan, siempre obtienen el mismo tamaño de muestra para estimar cualquier valor poblacional, a cualquier nivel de desagregación de la variable misma o geográficamente.

Como anécdota, recuerdo a un grupo de personas de una compañía, que cada vez que calculaban el tamaño de la muestra para cualquier investigación, obtenían que era 400, independientemente el producto o servicio. Pero no sólo eso, sino que pretendían dar información para diferentes clases sociales o tamaños de empresas (según fueran productos residenciales o empresariales, respectivamente), y para unas 550 áreas geográficas aproximadamente. Claro, se excusaban en decir que lo más importante era obtener únicamente dicho coeficiente y que el comportamiento entre áreas era homogéneo, en dichos casos falso desde cualquier punto de vista.

Como estos, muchos casos, y algunas personas menos osadas y conscientes de sus limitaciones en el tema, dicen que parece un enigma, un misterio el determinar tamaño de muestra adecuado para un estudio, esto sin incluir además, el diseño muestral apropiado.

Lo que debe hacerse es tomar en cuenta la información disponible, así como la precisión deseada, los recursos financieros, humanos y materiales disponibles, los niveles de desagregación deseados y la premura en la disponibilidad de los datos, para abordar la investigación con la seriedad del caso. Los usuarios de la información siempre requieren una gran cantidad de variables, al mayor nivel de detalle y en el menor tiempo posible; algo que resulta incompatible en la mayoría de los casos, a menos que ya se disponga de la información, es decir, que no haya que realizar la investigación. En general, a medida que aumenta el nivel de detalle deseado, debe aumentar el tamaño de la muestra, sin embargo, tamaños de muestras grandes, requieren mayores recursos y tiempo de recolección, tratamiento y procesamiento que tamaños pequeños. Este último detalle, pudiera reducirse si se disponen de los recursos económicos y humanos necesarios para acelerar los procesos, pero esto siempre tiene un límite.

Como puede apreciarse, para determinar el tamaño de la muestra deben considerarse varios factores, Cochran

[2;103-105], lo explica de una manera muy didáctica, a través de un ejemplo y luego una enumeración de pasos. Aquí se tratará de hacer algo similar.

Siempre se debe tomar en cuenta los siguientes aspectos:

1.- Nivel pretendido de desagregación de la(s) variable(s) - (cobertura vertical)
2.- Nivel pretendido de desagregación geográfica. - (cobertura horizontal)
3.- Precisión deseada de las estimaciones.
4.- Recursos disponibles.
5.- Tiempo disponible.

Los tres primeros responden a criterios analíticos, los dos últimos a facilidades disponibles. La precisión puede expresarse en términos del error cuadrático medio, la varianza, el error estándar, el coeficiente de variación y el sesgo. Como se indicó en el capítulo 1, el error cuadrático medio es la suma de la varianza y el sesgo, para estimadores insesgados el error cuadrático medio coincide con la varianza, por tal motivo el sesgo no es muy utilizado para fijar una precisión a priori en función de determinar tamaños de muestra. A su vez, el error estándar y el coeficiente de variación son funciones de la varianza, pero expresados en unidades más fáciles de trabajar.

Azorín [1;68] y Félix Seijas [7;143], utilizan el término "error máximo admisible" (*EMA*), definido como el máximo error dispuesto a aceptar, y consiste en el término que se le suma y se le resta al estimador para determinar el intervalo de confianza.

En líneas generales, el enfoque es considerar alguna medida de precisión y su nivel deseado, utilizar su fórmula, de acuerdo con el diseño muestral establecido, determinar o asumir la varianza de la variable a estudiar (puede ser estimada de estudios anteriores o poblaciones similares), y entonces despejar el tamaño de la muestra. En otros casos el tamaño de muestra puede estar determinado exclusivamente por el tiempo o los recursos disponibles.

En este orden, consideremos la varianza del estimador del promedio,

$$V(\bar{y}) = \frac{N-n}{N}\frac{S^2}{n} \le V$$ siendo V un valor de la varianza fijada a priori, como cota superior de la varianza del estimador

Despejando n se tiene que, $$n \ge \frac{N S^2}{N V + S^2}$$

Replicando el procedimiento, de acuerdo a la medida utilizada para fijar la precisión a priori, y al valor poblacional a estimar se tiene el siguiente cuadro,

Tabla 6.1 – Tamaños de muestras por medida utilizada, según clase de muestreo y valor poblacional a estimar

Diseño Muestral	Valor Poblacional a Estimar	Medida utilizada para fijar la precisión			
		Varianza	Error Estándar	Coeficiente de Variación	Error Máximo Admisible
m.a.s. sin reemplazamiento	Y	$n \geq \frac{N^2 S^2}{V + N S^2}$	$n \geq \frac{N^2 S^2}{E^2 + N S^2}$	$n \geq \frac{N^2 S^2}{C^2 Y^2 + N S^2}$	$n \geq \frac{N^2 k^2_{1-\frac{\alpha}{2}} S^2}{M^2 + N k^2_{1-\frac{\alpha}{2}} S^2}$
m.a.s. sin reemplazamiento	$\overline{Y}$	$n \geq \frac{N S^2}{N V + S^2}$	$n \geq \frac{N S^2}{N E^2 + S^2}$	$n \geq \frac{N S^2}{N C^2 \overline{Y}^2 + S^2}$	$n \geq \frac{N k^2_{1-\frac{\alpha}{2}} S^2}{N M^2 + k^2_{1-\frac{\alpha}{2}} S^2}$
m.a.s. sin reemplazamiento	A	$n \geq \frac{N^3 PQ}{(N-1)V + N^2 PQ}$	$n \geq \frac{N^3 PQ}{(N-1)E^2 + N^2 PQ}$	$n \geq \frac{N^3 Q}{(N-1)C^2 P + N^2 Q}$	$n \geq \frac{N^3 k^2_{1-\frac{\alpha}{2}} PQ}{(N-1)M^2 + N^2 k^2_{1-\frac{\alpha}{2}} PQ}$
m.a.s. sin reemplazamiento	P	$n \geq \frac{N PQ}{(N-1)V + PQ}$	$n \geq \frac{N PQ}{(N-1)E^2 + PQ}$	$n \geq \frac{N Q}{(N-1)C^2 P + Q}$	$n \geq \frac{N^2 k^2_{1-\frac{\alpha}{2}} PQ}{(N-1)M^2 + k^2_{1-\frac{\alpha}{2}} PQ}$
m.a.s. con reemplazamiento	Y	$n \geq \frac{N^2 S_{\vert}^2}{V}$	$n \geq \frac{N^2 S_{\vert}^2}{E^2}$	$n \geq \frac{N^2 S_{\vert}^2}{C^2 Y^2}$	$n \geq \frac{N^2 k^2_{1-\frac{\alpha}{2}} S_{\vert}^2}{M^2}$
m.a.s. con reemplazamiento	$\overline{Y}$	$n \geq \frac{S_{\vert}^2}{V}$	$n \geq \frac{S_{\vert}^2}{E^2}$	$n \geq \frac{S_{\vert}^2}{C^2 \overline{Y}^2}$	$n \geq \frac{k^2_{1-\frac{\alpha}{2}} S_{\vert}^2}{M^2}$
m.a.s. con reemplazamiento	A	$n \geq \frac{N^2 PQ}{V}$	$n \geq \frac{N^2 PQ}{E^2}$	$n \geq \frac{N^2 Q}{C^2 P}$	$n \geq \frac{N^2 k^2_{1-\frac{\alpha}{2}} PQ}{M^2}$
m.a.s. con reemplazamiento	P	$n \geq \frac{PQ}{V}$	$n \geq \frac{PQ}{E^2}$	$n \geq \frac{Q}{C^2 P}$	$n \geq \frac{k^2_{1-\frac{\alpha}{2}} PQ}{M^2}$

Donde V, E, C y M son las varianzas, errores estándar, coeficientes de variación y errores máximos admisibles fijados con anterioridad como cota superior de precisión a efectos del cálculo del tamaño de la muestra, y k es el valor de *t-student* o de la distribución normal.

Para mostrar el siguiente razonamiento se tomará la fórmula del tamaño de muestra para estimar *Y*, en muestreo aleatorio simple sin reemplazamiento, utilizando como medida de precisión la varianza, que es

$$n \geq \frac{N^2 S^2}{V + N S^2}$$ pero multiplicando y dividiendo por *V* se tiene que $$n \geq \frac{\frac{N^2 S^2}{V}}{1 + \frac{N S^2}{V}} = \frac{n_0}{1 + \frac{n_0}{N}}$$ donde $$n_0 = \frac{N^2 S^2}{V}$$

Luego, si se desconoce S^2, se puede tomar una muestra inicial o piloto, aleatoria simple sin reemplazamiento de tamaño n_0, estimar S^2 con dicha muestra (s^2) y el tamaño muestral sería,

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

donde se puede deducir n_0 de *n*, es decir, el tamaño de muestra faltante será $n-n_0$.

Este procedimiento se puede aplicar a todos los casos, y los n_0 se presentan en la siguiente tabla,

Tabla 6.2 – Tamaños de muestra inicial o piloto por medida utilizada, según clase de muestreo y valor poblacional a estimar

Diseño Muestral	Valor Poblacional a Estimar	Medida utilizada para fijar la precisión			
		Varianza	Error Estándar	Coeficiente de Variación	Error Máximo Admisible
m.a.s. sin reemplazamiento	Y	$n_0 = \frac{N^2 S^2}{V}$	$n_0 = \frac{N^2 S^2}{E^2}$	$n_0 = \frac{N^2 S^2}{C^2 Y^2}$	$n_0 = \frac{N^2 k^2_{1-\frac{\alpha}{2}} S^2}{M^2}$
m.a.s. sin reemplazamiento	$\bar{Y}$	$n_0 = \frac{S^2}{V}$	$n_0 = \frac{S^2}{E^2}$	$n_0 = \frac{S^2}{C^2 \bar{Y}^2}$	$n_0 = \frac{k^2_{1-\frac{\alpha}{2}} S^2}{M^2}$
m.a.s. sin reemplazamiento	A	$n_0 = \frac{N^3 PQ}{(N-1)V}$	$n_0 = \frac{N^3 PQ}{(N-1)E^2}$	$n_0 = \frac{NQ}{(N-1)C^2 P}$	$n_0 = \frac{N^3 k^2_{1-\frac{\alpha}{2}} PQ}{(N-1)M^2}$
m.a.s. sin reemplazamiento	P	$n_0 = \frac{NPQ}{(N-1)V}$	$n_0 = \frac{NPQ}{(N-1)E^2}$	$n_0 = \frac{NQ}{(N-1)C^2 P}$	$n_0 = \frac{N k^2_{1-\frac{\alpha}{2}} PQ}{(N-1)M^2}$

Donde *V*, *E*, *C* y *M* son respectivamente las varianzas, errores estándar, coeficientes de variación y errores máximos admisibles fijados con anterioridad como cota superior de precisión a efectos del cálculo del tamaño de la muestra, y *k* es el valor de *t-student* o de la distribución normal.

Nótese que si *N* es suficientemente grande $(N \to \infty)$ todos los n_0 coinciden con los *n* respectivos del MAS con reemplazamiento.

Cuando la muestra es estratificada, el cálculo del tamaño de muestra se puede dividir en dos, de acuerdo con la razón por la cual se toma este diseño. La decisión sobre el uso del muestreo aleatorio estratificado (con selección simple o sistemática de elementos) se debe a una de las siguientes consideraciones:

i.- para reducir la varianza de los estimadores (si el diseño realmente lo permite)
ii.- para generar/obtener estimaciones en cada estrato

Para el primer caso, se debe calcular primero *n* (el tamaño total de la muestra) y luego hallar los n_h (el tamaño de muestra para cada estrato) según la afijación seleccionada. Las fórmulas para calcular *n* se muestran en la siguiente tabla.

Tabla 6.3 – Tamaños de muestra en el muestreo estratificado, por estimador, según afijación utilizada

Afijación	Estimación de $\overline{Y}$	Estimación de Y	Estimación de P	Estimación de A
General	$n = \dfrac{\sum_{h=1}^{L} \dfrac{W_h^2 S_h^2}{w_h}}{\hat{V}(\overline{Y}_{st}) + \left(\dfrac{\sum_{h=1}^{L} W_h S_h^2}{N} \right)}$	$n = \dfrac{\sum_{h=1}^{L} \dfrac{N_h^2 S_h^2}{w_h}}{\hat{V}(Y_{st}) + \sum_{h=1}^{L} N_h S_h^2}$	$n = \dfrac{\sum_{h=1}^{L} \dfrac{W_h^2 p_h q_h}{w_h}}{V(p_{st}) + \dfrac{\sum_{h=1}^{L} W_h p_h q_h}{N}}$	$n = \dfrac{\sum_{h=1}^{L} \dfrac{N_h^2 p_h q_h}{w_h}}{\hat{V}(A_{st}) + \sum_{h=1}^{L} N_h p_h q_h}$
Igual	$n = \dfrac{L\sum_{h=1}^{L} W_h^2 S_h^2}{\hat{V}(\overline{Y}_{st}) + \left(\dfrac{\sum_{h=1}^{L} W_h S_h^2}{N} \right)}$	$n = \dfrac{L\sum_{h=1}^{L} N_h^2 S_h^2}{\hat{V}(Y_{st}) + \sum_{h=1}^{L} N_h S_h^2}$	$n = \dfrac{L\sum_{h=1}^{L} W_h^2 p_h q_h}{V(p_{st}) + \dfrac{\sum_{h=1}^{L} W_h p_h q_h}{N}}$	$n = \dfrac{L\sum_{h=1}^{L} N_h^2 p_h q_h}{\hat{V}(A_{st}) + \sum_{h=1}^{L} N_h p_h q_h}$
Proporcional	$n = \dfrac{\sum_{h=1}^{L} W_h S_h^2}{\hat{V}(\overline{Y}_{st}) + \left(\dfrac{\sum_{h=1}^{L} W_h S_h^2}{N} \right)}$	$n = \dfrac{\sum_{h=1}^{L} W_h S_h^2}{\hat{V}(\overline{Y}_{st}) + \left(\dfrac{\sum_{h=1}^{L} W_h S_h^2}{N} \right)}$	$n = \dfrac{\sum_{h=1}^{L} W_h p_h q_h}{V(p_{st}) + \dfrac{\sum_{h=1}^{L} W_h p_h q_h}{N}}$	$n = \dfrac{\sum_{h=1}^{L} N_h p_h q_h}{\hat{V}(A_{st}) + \sum_{h=1}^{L} N_h p_h q_h}$
Optima	$n = \dfrac{\left(\sum_{h=1}^{L} W_h S_h \right)^2}{\hat{V}(\overline{Y}_{st}) + \left(\dfrac{\sum_{h=1}^{L} W_h S_h^2}{N} \right)}$	$n = \dfrac{\left(\sum_{h=1}^{L} N_h S_h \right)^2}{\hat{V}(Y_{st}) + \sum_{h=1}^{L} N_h S_h^2}$	$n = \dfrac{\left(\sum_{h=1}^{L} W_h \sqrt{p_h q_h} \right)^2}{V(p_{st}) + \dfrac{\sum_{h=1}^{L} W_h p_h q_h}{N}}$	$n = \dfrac{\left(\sum_{h=1}^{L} N_h \sqrt{p_h q_h} \right)^2}{\hat{V}(A_{st}) + \sum_{h=1}^{L} N_h p_h q_h}$

Si se desea obtener estimaciones en cada estrato, entonces se calculan los directamente los n_h de acuerdo con las fórmulas del muestreo aleatorio simple mostradas en las tablas 6.1 y 6.2, y luego se totalizan para obtener *n*.

Si la varianza (S^2) es muy alta, es porque existe gran heterogeneidad, y se debe incluir a muchos elementos en la

muestra; si la varianza es muy baja, es porque hay gran homogeneidad en la población, y el tamaño de muestra puede ser pequeño. En todo caso, el tamaño de muestra (n) debe variar entre *2* y N; se debe evitar $n=1$, ya que no se puede calcular la varianza muestral (s^2)

De manera que sí al realizar los cálculos del tamaño de muestra, resulta que $n<0$, *0* ó *1*, se debe hacer $n=2$, si por otra parte, $n>N$, se hace $n=N$.

Cabe destacar que ningún cálculo garantiza una precisión fijada de antemano, ya que la verdadera varianza, y demás medidas, se desconocen en la gran mayoría de los casos, por lo tanto, lo que se hace es estimarlas a partir de la muestra tomada, sin embargo, se pueden obtener tamaños de muestra que permitan obtener estimadores con una precisión aproximada. Por lo tanto, aquellos estudios que usan como precisión el coeficiente de variación determinado a priori, no sólo está mal calculado, sino que es inapropiado.

Claro está que esta precisión va asociada a un estimador y a un nivel de desagregación determinado. Se puede determinar un tamaño de muestra que permita obtener estimaciones del ingreso promedio por vivienda de un barrio, con un coeficiente de variación (*cv*) del 5%, pero al desagregarlo por manzana, seguramente el *cv* va a aumentar en cada una, y por supuesto, en las manzanas con una observación no se pueden dar estimaciones. También se podría obtener el ingreso promedio por vivienda, según el tipo de vivienda, pero con un error mayor que al darlo agregado.

Se considerarán algunos casos para ilustrar una secuencia de pasos, sin que esto implique un orden pre-establecido.

Supóngase que un agrónomo desea probar un insecticida en plantas de aguacate, y habla con un estadístico para que le indique el tamaño de una muestra aleatoria simple que le permita obtener las estimaciones necesarias.

Como respuesta, el estadístico le hace las siguientes preguntas:

¿Qué información necesita?
¿A qué nivel necesita desagregar la(s) variable(s)?
¿A qué nivel de desagregación geográfica desea las estimaciones?
¿A qué nivel de precisión desea las estimaciones?

El agrónomo respondió diciendo que quería estimar la proporción de plantas sanadas por el insecticida, y la

proporción de las que dieron aguacates con sabor diferente al usual, a causa del producto, según si la planta había sido curada o no. También dijo que el estudio lo haría en una plantación de 5000 matas de aguacate, de las cuales aproximadamente 1500 estaban enfermas; el procedimiento sería seleccionar una muestra de plantas enfermas, aplicar el insecticida y observar, luego de un tiempo prudencial, cuáles de ellas han sanado. En cuanto a la precisión, se conforma con un error de (+/-) 5%. Es decir que si la proporción estimada de efectividad es de 0,6 entonces la verdadera proporción estaría entre 0,55 y 0,65, con una confianza de 0,95.

Esta medida del error dado por el agrónomo, se refiere al "error máximo admisible". Se sabe que si la muestra es aleatoria simple y el tamaño de muestra es grande, la proporción muestral se distribuye como una normal, pero como no se conoce la verdadera varianza, el intervalo de confianza sería:

$$(p-t_{(n-1;0,05)}\, ee(p) \, , \, p+t_{(n-1;0,95)}\, ee(p))$$

pero el valor de *t-student* depende del tamaño de la muestra y no se tiene. Al buscar en la tabla, se nota que está cerca de 2, se toma este valor, entonces, se tiene que el intervalo es:

$$(p-2ee(p) \, , \, p+2ee(p))$$

donde $2ee(p) = 0,05$, entonces $ee(p) = 0,025$. El error estándar de p es la raíz cuadrada de la varianza de p, que es

$$V(p) = \frac{N-n}{n} \frac{PQ}{N-1}$$

es decir, que se tiene que

$$ee(p) = \sqrt{\frac{N-n}{n} \frac{PQ}{N-1}} \leq \frac{0,05}{2} = 0,025$$

despejando n,

$$n \geq \frac{N\,P\,Q}{(0,025)^2\,(N-1)+P\,Q}$$

donde $P=1-Q$. Entonces se tiene que el tamaño de la muestra es una función de la precisión, del tamaño de la población y del valor poblacional a estimar. Pudiera parecer una contradicción, tener que conocer el valor poblacional a estimar, si precisamente eso es lo que se desconoce, y si se conociera, no se tomaría una muestra; pero no se requiere sino una aproximación, una idea, para poder hallar n.

Se puede tomar $P=0,5$, para maximizar el producto PQ, y por ende, maximizar el tamaño mínimo de la muestra, sin embargo, este método puede aumentar el tamaño de la muestra más de lo necesario, y desperdiciar recursos, por

ello se debe aplicar este criterio cuando realmente no se tiene idea alguna del comportamiento de *P*.

Al consultarle al agrónomo sobre un valor aproximado de *P*, este revisó un informe del producto que indica que el porcentaje de efectividad es del 90%, y que en estudios previos se determinó que afecta el sabor del 15% de los frutos. Al calcular *n* para ambas estimaciones, se obtiene,

para la efectividad $$n \geq \frac{1500(0,9)(0,1)}{(0,025)^2\ (1499)+(0,9)(0,1)} = 131,5$$

para alteración de sabor $$n \geq \frac{1500(0,15)(0,85)}{(0,025)^2\ (1499)+(0,15)(0,85)} = 179,7$$

Se toma el valor entero más próximo por la derecha, es decir, la parte entera más 1. Entonces, para la estimación de la efectividad, hace falta una muestra de 132 matas enfermas, y para la estimación de la proporción de sabor afectado se necesitan 180 matas enfermas. Como es un sólo estudio, se tomará una sóla muestra que debe ser de 180 plantas enfermas para satisfacer ambos requerimientos de precisión. El agrónomo indica que estos tamaños son muy grandes, que le resulta costoso por la cantidad de insecticida a utilizar, además que va a tardar más de lo esperado, a menos que contrate más empleados que lo estipulado, y se encarece más el estudio. Así que propone bajar la precisión al 7%. Se obtiene,

para la efectividad $$n \geq \frac{1500(0,9)(0,1)}{(0,035)^2\ (1499)+(0,9)(0,1)} = 70,08$$

para alteración de sabor $$n \geq \frac{1500(0,15)(0,85)}{(0,035)^2\ (1499)+(0,15)(0,85)} = 97,39$$

Es decir, que la muestra debe ser de 98, y el investigador queda satisfecho con este tamaño. Aquí se tienen dos opciones,

- el investigador tiene o construye un marco con las plantas enfermas, de las cuales selecciona 98 aleatoriamente
- no se tiene dicho marco y se selecciona una muestra aleatoria de más de 98 plantas, esperando encontrar aproximadamente 98 enfermas. Esto es, la proporción de plantas enfermas es 0,3, entonces se seleccionan 327 plantas, esperando que el 30% (98) estén enfermas

Otro procedimiento, muy utilizado, es recorrer la plantación y tomar las primeras 98 plantas enfermas, pero el mismo no es aleatorio, y en consecuencia se descartará como opción válida.

Veáse el siguiente ejemplo. Una empresa de lácteos desea hacer un "store audit" en una ciudad, para medir las ventas de los clientes de uno de sus quesos, claro que aprovecharán para medir las de otros productos. La empresa

tiene dividida la ciudad en tres zonas, y ellos quieren conocer el comportamiento del mercado en cada zona, y además diferenciada según sean pequeños establecimientos (panaderías, charcuterías y pequeños abastos) y grandes establecimientos (supermercados, automercados y grandes abastos). El siguiente cuadro muestra el total de establecimientos,

Tabla 6.4 - Total de establecimientos por tipo, según zona

	Total	Pequeños	Grandes
Total	1.065	803	262
Zona 1	435	330	105
Zona 2	372	289	83
Zona 3	258	184	74

Se quiere que la estimación tenga un error no mayor al +/- 10%, pero no se conoce comportamiento alguno de las ventas, por lo cual se decide tomar una muestra piloto, fijada arbitrariamente en 10 establecimientos pequeños y 10 grandes, y se obtuvieron los siguientes resultados:

Tabla 6.5 – Ventas de establecimientos de la muestra piloto y valores muestrales

Establecimiento No.	Pequeños Establecimientos		Grandes Establecimientos	
	Ventas (Kg)	Zona	Ventas (Kg)	Zona
1	10.3	1	203.2	1
2	25	2	351.2	3
3	28.1	1	233.4	2
4	31.5	3	275.1	1
5	15	2	384	2
6	18.5	1	309.5	1
7	23	3	193.8	3
8	25	2	301.7	2
9	21.5	3	285.7	3
10	29.6	1	215.9	1

Valores Muestrales

Total Muestral	227,50	2.753,50
Promedio	22,75	275,35
Varianza	39,70	3.667,59
Cuasivarianza	44,11	4.075,10
Desviación	6,30	60,56
Cuasidesviación	6,64	63,84

como la estimación por intervalo debe ser, $\left(\bar{y} - (0{,}10)\bar{y}\ ,\ \bar{y} + (0{,}10)\bar{y}\right)$

y el intervalo de confianza es, $\left(\bar{y} - 2\,ee(\bar{y})\ ,\ \bar{y} + 2\,ee(\bar{y})\right)$

entonces, $2\,ee(\bar{y}) \leq (\,0{,}10\,)\,\bar{y}$

es decir, $ee(\bar{y}) \leq (0{,}05)\,\bar{y}$

despejando *n* se tiene,

$$n \geq \frac{N\ S^2}{N\,(0{,}05)^2\ \bar{Y}^2 + S^2}$$

Entonces se toma S^2 y $\bar{Y}$ de la muestra piloto, suponiendo que en las tres zonas, la dispersión de las compras del queso en los establecimientos pequeños es similar, al igual que con los grandes. Se sustituye *N*, S^2 y $\bar{Y}$ para cada zona y tipo de establecimiento, de manera de hallar un tamaño de muestra para cada cruce, funcionando como estratos. Los resultados son:

Tabla 6.6 - Tamaños de Muestra Calculados

	Pequeños Establecimientos	Grandes Establecimientos
Zona 1	30,898244	17,845498
Zona 2	30,493194	17,076233
Zona 3	28,761436	16,659380

y aproximando al entero superior,

Tabla 6.7 - Tamaños de Muestra Definitivos

	Total	Pequeños Establecimientos	Grandes Establecimientos
Total	144	91	53
Zona 1	49	31	18
Zona 2	49	31	18
Zona 3	46	29	17

Pero como se tomó una muestra piloto, esos elementos se deducen del tamaño muestral, entonces, la cantidad de nuevos elementos a seleccionar son:

Tabla 6.8 - Total de Nuevos Establecimientos a Entrevistar

	Total	Pequeños Establecimientos	Grandes Establecimientos
Total	134	91	43
Zona 1	41	27	14
Zona 2	53	38	15
Zona 3	40	26	14

Es importante aclarar que para la dispersión de la variable “ventas de queso”, si bien se hizo una diferenciación por tipo de establecimiento, se consideró que dentro de ellos su distribución era igual en cada zona. Si se toma una

muestra piloto en cada estrato los cálculos para el tamaño de muestra serán más confiables.

Ahora se verá un ejemplo más complejo. Supóngase que el Ministerio de Desarrollo Social (MDS) de un país necesita elaborar un plan de distribución de alimentos gratuitos a la población más necesitada. Desea conocer el perfil de los hogares, de acuerdo con su conformación, el tipo de la vivienda que habitan, su ingreso, la clase socioeconómica a la cual pertenecen, el total de niños y su asistencia escolar, el empleo y el nivel educativo de los adultos, entre otras.

Al investigar varias variables, el cálculo del tamaño de muestra es más complejo, y se debe determinar la o las variables sobre las cuales se efectuarán dichos cálculos, ya que hacerlo sobre todas sería engorroso. Luego de varias conversaciones con el Instituto Nacional de Estadística (INE), se determinó que utilizarían la tabla 6.9 para el cálculo del tamaño de la muestra, debido a su importancia, al comportamiento heterogéneo de algunas variables y a la disponibilidad de dichas variables en estudios anteriores, como el más reciente Censo de Población que levantó el INE hace 5 años. Dicho cuadro forma parte del plan de tabulaciones del estudio, y se requiere para todas y cada una de las Entidades Federales.

Tabla 6.9 - Total de Hogares por Tenencia de Automóvil y Tipo de Hogar, según Tenencia y Tipo de Vivienda (MAQUETA)

Tenencia y Tipo de Vivienda	TOTAL			Con Vehículo			Sin Vehículo		
	Total	Uni-personal	Multi-personal	Total	Uni-personal	Multi-personal	Total	Uni-personal	Multi-personal
Total									
Quinta									
Casa-Quinta									
Casa									
Apartamento									
Rancho									
Vivienda Propia									
Quinta									
Casa-Quinta									
Casa									
Apartamento									
Rancho									
Vivienda Alquilada									
Quinta									
Casa-Quinta									
Casa									
Apartamento									
Rancho									

Se tienen 162 celdas, 122 de las cuales son de totales (sombreadas). Cada indica la presencia de una o más características, y cada una de las celdas sombreadas, indica una subpoblación.

Cada hogar está ubicado en una de las 40 celdas no sombreadas, y en otras 15 celdas de totales. Nótese que al excluir las celdas sombreadas se está en presencia de estimación de totales bajo el enfoque de proporciones.

En principio se acuerda que los coeficientes de variación (cv) de las estimaciones deben estar entre 0,10 y 0,15, incluso pueden llegar hasta 0,2 en algunas celdas consideradas por los usuarios no tan importantes o con poca población.

La varianza del estimador del total en subpoblaciones, cuando se conoce N_i (tamaño de la subpoblación) es,

$$V(\hat{A}_i) = \frac{N_i^2 (N_i - n_i)}{N_i \; n_i} \; \frac{N_i P_i Q_i}{N_i - 1}$$

y el coeficiente de variación es,

$$CV(\hat{A}_i) = \frac{\sqrt{\frac{N_i^2 (N_i - n_i)}{N_i \; n_i} \; \frac{N_i P_i Q_i}{N_i - 1}}}{A_i} = \sqrt{\frac{(N_i - n_i)}{(N_i - 1) n_i} \; \frac{N_i - A_i}{A_i}} = \sqrt{\frac{(N_i - n_i) Q_i}{(N_i - 1) n_i P_i}}$$

Para hacer una prueba, se seleccionó una Entidad Federal de 347.530 hogares, y los valores poblacionales -de acuerdo al censo más reciente- se muestran en la tabla 6.10. Se aplica la fórmula donde las columnas 4 y 7 (referentes a totales), indican los N_i y las columnas 5, 6 y 8, 9 las respectivas características. Cuando se le aplica la fórmula a una celda, esa es la característica, así que si el hogar está allí, la variable toma el valor *1* (uno), si no, toma el valor *0* (cero). Al aplicarle la fórmula al archivo respectivo, con *cv*=0,10, se obtuvieron los resultados que se muestran la tabla 6.11.

Tabla 6.10 - Total de Hogares por Tenencia de Automóvil y Tipo de Hogar, según Tenencia y Tipo de Vivienda
(Censo Anterior – Entidad Federal seleccionada)

Tenencia y Tipo de Vivienda	TOTAL			Con Vehículo			Sin Vehículo		
	Total (1)	Uni-personal (2)	Multi-personal (3)	Total (4)	Uni-personal (5)	Multi-personal (6)	Total (7)	Uni-personal (8)	Multi-personal (9)
Total	347.530	21.266	326.264	86.292	5.131	81.161	261.238	16.135	245.103
Quinta	5.132	281	4.851	4.534	246	4.288	598	35	563
Casa-Quinta	22.917	842	22.075	11.564	421	11.143	11.353	421	10.932
Casa	216.069	12.233	203.836	49.316	2.777	46.539	166.753	9.456	157.297
Apartamento	41.091	3.305	37.786	19.754	1.582	18.172	21.337	1.723	19.614
Rancho	62.321	4.605	57.716	1.124	105	1.019	61.197	4.500	56.697
Vivienda Propia	268.618	16.169	252.449	64.640	3.831	60.809	203.978	12.338	191.640
Quinta	3.621	211	3.410	3.199	176	3.023	422	35	387
Casa-Quinta	16.907	632	16.275	8.506	316	8.190	8.401	316	8.085
Casa	171.427	9.631	161.796	39.157	2.179	36.978	132.270	7.452	124.818
Apartamento	26.785	2.215	24.570	12.900	1.090	11.810	13.885	1.125	12.760
Rancho	49.878	3.480	46.398	878	70	808	49.000	3.410	45.590
Vivienda Alquilada	78.912	5.097	73.815	21.652	1.300	20.352	57.260	3.797	53.463
Quinta	1.511	70	1.441	1.335	70	1.265	176	0	176
Casa-Quinta	6.010	210	5.800	3.058	105	2.953	2.952	105	2.847
Casa	44.642	2.602	42.040	10.159	598	9.561	34.483	2.004	32.479
Apartamento	14.306	1.090	13.216	6.854	492	6.362	7.452	598	6.854
Rancho	12.443	1.125	11.318	246	35	211	12.197	1.090	11.107

Tabla 6.11 – Tamaños de muestra calculados para el (hogares) por Tenencia de Automóvil y Tipo de Hogar, según Tenencia y Tipo de Vivienda
(Tamaños Muestrales con cv=0,10 - Entidad Federal seleccionada)

Tenencia y Tipo de Vivienda	TOTAL			Con Vehículo			Sin Vehículo		
	Total	Uni-personal	Multi-personal	Total	Uni-personal	Multi-personal	Total	Uni-personal	Multi-personal
Total									
Quinta									
Casa-Quinta									
Casa									
Apartamento									
Rancho									
Vivienda Propia									
Quinta					176	6		35	9
Casa-Quinta					316	4		316	4
Casa					1.627	6		1.655	6
Apartamento					1.000	10		1.049	9
Rancho					70	9		1.302	8
Vivienda Alquilada									
Quinta					70	6		0	1
Casa-Quinta					105	4		105	4
Casa					598	7		1.548	7
Apartamento					492	8		598	9
Rancho					35	16		941	10

Los tamaños resultantes para las columnas de hogares unipersonales son sumamente altos, de hecho en algunas celdas $n_i=N_i$. Como se trata de subpoblaciones, no se pueden identificar dichos hogares a priori, por lo tanto se debe calcular el tamaño de muestra (n) que garantice el total de elementos calculados en cada celda, que en este caso será $n=N$=347.530. Se flexibilizan los requerimientos de precisión y se llega a un tamaño de muestra de 12.100 hogares, y se espera que la misma esté distribuida como se muestra en la tabla 6.12, de acuerdo a la distribución poblacional.

En la tabla 6.13 se muestran los coeficientes de variación esperados para cada una de las celdas. Nótese que las celdas correspondientes a hogares unipersonales, tienen coeficientes muy altos, de forma acentuada en las celdas blancas, esto indica que no se deben generar estas estimaciones, o hay que manejarlas con cuidado; afortunadamente, los totales poblacionales no son tan altos, por lo tanto, las desviaciones en términos porcentuales pueden ser grandes, pero no necesariamente en términos absolutos.

Tabla 6.12 – Distribución esperada de la muestra (hogares) por Tenencia de Automóvil y Tipo de Hogar, según Tenencia y Tipo de Vivienda
(Tamaños Muestrales definitivos - Entidad Federal seleccionada)

Tenencia y Tipo de Vivienda	TOTAL			Con Vehículo			Sin Vehículo		
	Total	Uni-personal	Multi-personal	Total	Uni-personal	Multi-personal	Total	Uni-personal	Multi-personal
Total	12.177	743	11.434	3.022	178	2.844	9.155	565	8.590
Quinta	179	9	170	158	8	150	21	1	20
Casa-Quinta	804	30	774	406	15	391	398	15	383
Casa	7.572	428	7.144	1.728	97	1.631	5.844	331	5.513
Apartamento	1.439	115	1.324	692	55	637	747	60	687
Rancho	2.183	161	2.022	38	3	35	2.145	158	1.987
Vivienda Propia	9.413	565	8.848	2.264	133	2.131	7.149	432	6.717
Quinta	127	7	120	112	6	106	15	1	14
Casa-Quinta	592	22	570	298	11	287	294	11	283
Casa	6.008	337	5.671	1.372	76	1.296	4.636	261	4.375
Apartamento	938	77	861	452	38	414	486	39	447
Rancho	1.748	122	1.626	30	2	28	1.718	120	1.598
Vivienda Alquilada	2.764	178	2.586	758	45	713	2.006	133	1.873
Quinta	52	2	50	46	2	44	6	0	6
Casa-Quinta	212	8	204	108	4	104	104	4	100
Casa	1.564	91	1.473	356	21	335	1.208	70	1.138
Apartamento	501	38	463	240	17	223	261	21	240
Rancho	435	39	396	8	1	7	427	38	389

Tabla 6.13 – Coeficientes de Variación Esperados de Hogares por Tenencia de Automóvil y Tipo de Hogar, según Tenencia y Tipo de Vivienda, para los tamaños muestrales de la tabla 6.12 - Entidad Federal seleccionada

Tenencia y Tipo de Vivienda	TOTAL			Con Vehículo			Sin Vehículo		
	Total	Uni-personal	Multi-personal	Total	Uni-personal	Multi-personal	Total	Uni-personal	Multi-personal
Total	0,00	0,03	0,00	0,02	0,07	0,00	0,01	0,04	0,00
Quinta	0,07	0,31	0,02	0,03	0,33	0,02	0,20	0,86	0,05
Casa-Quinta	0,03	0,18	0,01	0,03	0,25	0,01	0,03	0,25	0,01
Casa	0,01	0,05	0,00	0,02	0,10	0,01	0,01	0,05	0,00
Apartamento	0,02	0,09	0,01	0,03	0,13	0,01	0,02	0,12	0,01
Rancho	0,02	0,07	0,01	0,16	0,50	0,05	0,00	0,08	0,01
Vivienda Propia	0,00	0,04	0,00	0,02	0,08	0,01	0,01	0,05	0,00
Quinta	0,09	0,35	0,02	0,03	0,38	0,02	0,24	0,84	0,08
Casa-Quinta	0,04	0,20	0,01	0,04	0,29	0,01	0,04	0,29	0,01
Casa	0,01	0,05	0,00	0,02	0,11	0,01	0,01	0,06	0,00
Apartamento	0,03	0,11	0,01	0,03	0,15	0,01	0,03	0,15	0,01
Rancho	0,02	0,09	0,01	0,18	0,61	0,05	0,00	0,09	0,01
Vivienda Alquilada	0,02	0,07	0,00	0,03	0,14	0,01	0,01	0,08	0,01
Quinta	0,13	0,62	0,03	0,05	0,62	0,03	0,38	-	0,00
Casa-Quinta	0,07	0,35	0,01	0,07	0,50	0,02	0,07	0,50	0,02
Casa	0,02	0,10	0,01	0,05	0,21	0,01	0,01	0,11	0,01
Apartamento	0,04	0,15	0,01	0,05	0,23	0,02	0,04	0,21	0,02
Rancho	0,04	0,15	0,01	0,33	0,86	0,14	0,01	0,15	0,01

7.- ESTIMADORES INDIRECTOS

En cualquier caso es conveniente mejorar las estimaciones, y en muchas oportunidades se puede lograr utilizando los métodos indirectos. Estos consisten en aprovechar la información auxiliar, básica o complementaria, relativa a una variable correlacionada con la que se estudia, entendiéndose por auxiliar información proveniente de estudios anteriores -encuestas o censos-, de registros administrativos o de estudios de poblaciones diferentes, pero altamente correlacionada con la variable a investigar.

Los estimadores indirectos generalmente son usados para mejorar la precisión de las estimaciones, y tienen la característica de utilizar adicionalmente otra variable, diferente a la variable a estimar, y que se denomina "variable auxiliar". De esta variable se necesita conocer los valores de los elementos de la muestra, igual que de la variable en estudio, pero además, ciertos valores poblacionales, como el total y el promedio. La precisión de las estimaciones mejorará en la medida que la variable auxiliar esté más correlacionada, en algunos casos positivamente, con la variable en estudio.

Estos estimadores son muy utilizados porque permiten ajustar los resultados a los totales o promedios de otras variables conocidas, dándole mayor consistencia a los resultados; es decir, para investigar la situación en la fuerza de trabajo, una variable auxiliar puede ser los tamaños poblacionales por edad, obtenidos por proyecciones de población, así los resultados siempre mostrarán dichos totales por edad.

Estos estimadores tienen la particularidad de ser sesgados, pero suele suceder que al tener tamaños de muestra grandes, el sesgo se hace despreciable, logrando, en algunas oportunidades, un error cuadrático medio menor al del estimador directo.

Los estimadores indirectos conocidos son los de razón y los de regresión. En los puntos siguientes se tratarán cada uno de ellos.

7.1.- Estimadores de razón

Sean $y_1, y_2, \ldots, y_N$ y $x_1, x_2, \ldots, x_N$ los valores de las variables y , x para cada uno de los N elementos de la población. Se denomina "razón de y a x" o simplemente "razón" y se identifica por R, a

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

y su estimador es,

$$\hat{R} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{y}{x} = \frac{\overline{y}}{\overline{x}} = \frac{\hat{Y}}{\hat{X}}$$

entonces, los estimadores de razón del total y el promedio respectivamente son,

$$\hat{Y}_R = \hat{R}\ X = \frac{y}{x}\ X$$

$$\hat{\overline{Y}}_R = \overline{y}_R = \hat{R}\ \overline{X} = \frac{y}{x}\ \overline{X} = \frac{1}{N}\ \hat{Y}_R$$

7.1.1.- Esperanza de los estimadores de razón

En principio debe recalcarse que en los estimadores de razón, tanto el numerador como el denominador son variables aleatorias, por lo tanto los cálculos de la esperanza y la varianza son complejos, y se usarán procedimientos diferentes a los mostrados en los capítulos anteriores, específicamente se utilizará un procedimiento mostrado por Sukhatme [8;141].

Se definirán

$$\begin{cases} y_i = \overline{Y} + e_i \\ x_i = \overline{X} + e'_i \end{cases}$$ donde e_i , e'_i son los desvíos de y_i , x_i respecto de $\overline{Y}$, $\overline{X}$ respectivamente.

de manera que,

$$\overline{y} = \frac{\sum_{i=1}^{n} (\overline{Y} + e_i)}{n} = \frac{\sum_{i=1}^{n} \overline{Y} + \sum_{i=1}^{n} e_i}{n} = \frac{n\ \overline{Y}}{n} + \frac{\sum_{i=1}^{n} e_i}{n} = \overline{Y} + \overline{e}$$

siendo $\overline{e}$ la media muestral de los desvíos de *y*. Entonces,

$$E(\overline{y}) = E(\overline{Y} + \overline{e}) = E(\overline{Y}) + E(\overline{e}) = \overline{Y} + E(\overline{e}) \qquad (7.1)$$

desarrollando el segundo miembro,

$$E(\bar{e}) = \frac{\sum_{k=1}^{\binom{N}{n}} \frac{e_k}{n}}{\binom{N}{n}} = \frac{\sum_{k=1}^{\binom{N}{n}} \frac{1}{n} \sum_{i=1}^{n} e_{ki}}{\binom{N}{n}} = \frac{\binom{N-1}{n-1} \sum_{i=1}^{N} e_i}{n \binom{N}{n}} = \frac{\sum_{i=1}^{N} e_i}{N} = \bar{E} = 0$$

donde $\bar{E}$ es el promedio poblacional de los desvíos de *y*. Es decir que la esperanza del promedio muestral de los desvíos, es el promedio poblacional de los desvíos, que a su vez es cero. Entonces, sustituyendo este resultado en (7.1), se tiene que,

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

que es un estimador insesgado, tal como ya se conoce. Lo que se hizo fue verificar que esta manera de definir a y_i es correcta. Ahora se hallará la varianza de e.

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

desarrollando el primer término,

$$E(\bar{e}^2) = \frac{\sum_{k=1}^{\binom{N}{n}} \left(\frac{e_k}{n}\right)^2}{\binom{N}{n}} = \frac{\sum_{k=1}^{\binom{N}{n}} e_k^2}{n^2 \binom{N}{n}} = \frac{1}{n^2 \binom{N}{n}} \left[\binom{N-1}{n-1} \sum_{i=1}^{N} e_i^2 + 2 \binom{N-2}{n-2} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} e_i \, e_j \right]$$

$$= \frac{\binom{N-1}{n-1} \sum_{i=1}^{N} e_i^2}{n^2 \binom{N}{n}} + 2 \frac{\binom{N-2}{n-2} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} e_i \, e_j}{n^2 \binom{N}{n}} = \frac{\sum_{i=1}^{N} e_i^2}{n \, N} + \frac{2\,(n-1)}{N-1} \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} e_i \, e_j}{n \, N}$$

$$= \frac{1}{n \, N} \left[\sum_{i=1}^{N} e_i^2 + \frac{2\,(n-1)}{N-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} e_i \, e_j \right]$$

$$= \frac{1}{n \, N} \left[\sum_{i=1}^{N} e_i^2 + 2 \frac{n-1}{N-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} e_i \, e_j + \frac{n-1}{N-1} \sum_{i=1}^{N} e_i^2 - \frac{n-1}{N-1} \sum_{i=1}^{N} e_i^2 \right]$$

$$= \frac{1}{n \, N} \left[\left(\frac{n-1}{N-1} \right) \left(\sum_{i=1}^{N} e_i^2 + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} e_i \, e_j \right) + \left(1 - \frac{n-1}{N-1} \right) \sum_{i=1}^{N} e_i^2 \right]$$

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

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

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

$$= \frac{1}{nN}\left[\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2 + \frac{n-1}{N-1}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} y_i\, y_j - \frac{n-1}{N-1}\bar{Y}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} y_j - \frac{n-1}{N-1}\bar{Y}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} y_i + \frac{n-1}{N-1}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\bar{Y}^2\right] \quad (7.2)$$

pero como,

$$\sum_{i=1}^{N}\ \sum_{\substack{j=1\\ j\neq i}}^{N} y_i\ y_j = \sum_{i=1}^{N}\ \sum_{j=1}^{N} y_i\ y_j\ -\ \sum_{i=1}^{N} y_i^2 \qquad (7.3)$$

y

$$\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} y_j = (N-1)\sum_{j=1}^{N} y_j = N(N-1)\bar{Y} \quad ; \quad \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} y_i = (N-1)\sum_{i=1}^{N} y_i = N\,(N-1)\bar{Y} \qquad (7.4)$$

además,

$$\sum_{i=1}^{N}\ \sum_{\substack{j=1\\ j\neq i}}^{N} \bar{Y}^2 = N\,(N-1)\ \bar{Y}^2 \qquad (7.5)$$

sustituyendo (7.3),(7.4) y (7.5) en (7.2)

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

descomponiendo,

$$\sum_{i=1}^{N}\ \sum_{j=1}^{N} y_i\ y_j = \sum_{i=1}^{N} y_i\ \sum_{j=1}^{N} y_j = N^2\,\bar{Y}^2$$

entonces,

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

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

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

por lo tanto,

$$E(\bar{e}^2) = \frac{N-n}{N\ n}\ \frac{\sum_{i=1}^{N}(y_i - \bar{Y})^2}{N-1} = \frac{1-f}{n}\ S_y^2 \qquad (7.6)$$

es decir que,

$$V(\bar{e}) = E(\bar{e}^2) = V(\bar{y}) = \frac{1-f}{n}\ S_y^2$$

De la misma manera,

$$\bar{x} = \bar{X} + \bar{e}$$
$$E(\bar{e}) = \bar{E} = 0 \qquad \Rightarrow \qquad E(\bar{x}) = \bar{X}$$
$$V(\bar{e}) = E(\bar{e}^2) = V(\bar{x}) = \frac{1-f}{n}S_x^2 \qquad (7.7)$$

Desarrollando $\hat{R}$,

$$\hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{\bar{Y} + \bar{e}}{\bar{X} + \bar{e}'} = \frac{\bar{Y}\left(1 + \frac{\bar{e}}{\bar{Y}}\right)}{\bar{X}\left(1 + \frac{\bar{e}'}{\bar{X}}\right)}$$

por lo tanto,

$$E(\hat{R}) = E\left[\frac{\bar{Y}\left(1 + \frac{\bar{e}}{\bar{Y}}\right)}{\bar{X}\left(1 + \frac{\bar{e}'}{\bar{X}}\right)}\right] = \frac{\bar{Y}}{\bar{X}}E\left[\frac{1 + \frac{\bar{e}}{\bar{Y}}}{1 + \frac{\bar{e}'}{\bar{X}}}\right] = R\,E\left[\frac{1 + \frac{\bar{e}}{\bar{Y}}}{1 + \frac{\bar{e}'}{\bar{X}}}\right]$$

para resolver la esperanza, se desarrollará la serie de Taylor del denominador. La serie de Taylor de una función *f(x)* respecto de x_0 es,

$$\sum_{k=0}^{\infty} \frac{f^{k}(x_0)}{k!}(x - x_0)^k$$

donde $f^{k}(x_0)$ es la derivada de orden *k* de la función *f*, evaluada en el punto x_0. En este caso la función es,

$$f(\bar{e}') = \left(1 + \frac{\bar{e}'}{\bar{X}}\right)^{-1} = \frac{1}{1 + \frac{\bar{e}'}{\bar{X}}} = \frac{1}{\frac{\bar{X} + \bar{e}'}{\bar{X}}} = \frac{\bar{X}}{\bar{X} + \bar{e}'}$$

y se debe hallar la serie de Taylor de $f(\bar{e}')$ respecto de cero. Al hallar $f'(\bar{e}')$, $f''(\bar{e}')$, $f''(\bar{e}')$, ... , se llega a la expresión genérica

$$f^{k}(\bar{e}') = (-1)^k \frac{k!\ \ \bar{X}}{(\bar{X} + \bar{e}')^{k+1}}$$

y evaluando en cero,

$$f^{k}(0) = (-1)^k \frac{k!}{\bar{X}^k}$$

entonces la serie es,

$$\sum_{k=0}^{\infty} (-1)^k \frac{k!}{\bar{X}^k} \tag{7.8}$$

entonces, $$\sum_{k=0}^{\infty} (-1)^k \frac{k!}{\bar{X}^k} \frac{(\bar{e}'-0)^k}{k!} = \sum_{k=0}^{\infty} (-1)^k \frac{(\bar{e}')^k}{\bar{X}^k}$$

luego,

$$E(\hat{R}) = R\,E\left[\left(1 + \frac{\bar{e}}{\bar{Y}}\right)\sum_{k=0}^{\infty}(-1)^k \frac{\bar{e}'^k}{\bar{X}^k}\right] = R\,E\left[\left(1 + \frac{\bar{e}}{\bar{Y}}\right)\left(1 - \frac{\bar{e}'}{\bar{X}} + \frac{\bar{e}'^2}{\bar{X}^2} - \frac{\bar{e}'^3}{\bar{X}^3} + \cdots + (-1)^k \frac{\bar{e}'^k}{\bar{X}^k} + \cdots\right)\right]$$

$$= R\,E\left[1 + \frac{\bar{e}}{\bar{Y}} - \frac{\bar{e}'}{\bar{X}} - \frac{\bar{e}\,\bar{e}'}{\bar{Y}\,\bar{X}} + \frac{\bar{e}'^2}{\bar{X}^2} + \frac{\bar{e}\,\bar{e}'^2}{\bar{Y}\,\bar{X}^2} - \frac{\bar{e}'^3}{\bar{X}^3} - \frac{\bar{e}\,\bar{e}'^3}{\bar{Y}\,\bar{X}^3} + \frac{\bar{e}'^4}{\bar{X}^4} + \frac{\bar{e}\,\bar{e}'^4}{\bar{Y}\,\bar{X}^4} - \frac{\bar{e}'^5}{\bar{X}^5} - \frac{\bar{e}\,\bar{e}'^5}{\bar{Y}\,\bar{X}^5} + \cdots\right]$$

truncando la serie, suponiendo que

$$\frac{\bar{e}'^k}{\bar{X}^k} \quad \forall k > 2 \qquad \wedge \qquad \frac{\bar{e}\,\bar{e}'^k}{\bar{Y}\,\bar{X}^k} \quad \forall k > 1$$

no aportan mucho y se pueden despreciar,

$$E(\hat{R}) = R\,E\left[1+\frac{\bar{e}}{\bar{Y}}-\frac{\bar{e}'}{\bar{X}}-\frac{\bar{e}\,\bar{e}'}{\bar{Y}\,\bar{X}}+\frac{\bar{e}'^{2}}{\bar{X}^{2}}\right] = R\left[1+\frac{1}{\bar{Y}}E(\bar{e})-\frac{1}{\bar{X}}E(\bar{e}')-\frac{1}{\bar{Y}\,\bar{X}}E(\bar{e}\,\bar{e}')+\frac{1}{\bar{X}^{2}}E(\bar{e}'^{2})\right]$$

$$= R\left[1-\frac{1}{\bar{Y}\,\bar{X}}E(\bar{e}\,\bar{e}')+\frac{1}{\bar{X}^{2}}\frac{N-n}{N}\frac{S_x^2}{n}\right] \qquad (7.9)$$

ahora se hallará $E(\bar{e}\,\bar{e}')$;

$$E(\bar{e}\,\bar{e}') = E\left[\left(\frac{\sum_{i=1}^{n}e_i}{n}\right)\left(\frac{\sum_{i=1}^{n}e'_i}{n}\right)\right] = \frac{1}{n^2}E\left[\left(\sum_{i=1}^{n}e_i\right)\left(\sum_{i=1}^{n}e'_i\right)\right] = \frac{1}{n^2}E\left[\sum_{i=1}^{n}e_i e'_i+\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}e_i e'_j\right]$$

$$= \frac{1}{n^2}\frac{1}{\binom{N}{n}}\left[\sum_{k=1}^{\binom{N}{n}}\sum_{i=1}^{n}e_{ki}\,e'_{ki}+\sum_{k=1}^{\binom{N}{n}}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}e_{ki}\,e'_{kj}\right] = \frac{1}{n^2}\frac{1}{\binom{N}{n}}\left[\binom{N-1}{n-1}\sum_{i=1}^{N}e_i e'_i+\binom{N-2}{n-2}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}e_i e'_j\right]$$

$$= \frac{1}{n^2}\left[\frac{n}{N}\sum_{i=1}^{N}e_i e'_i+\frac{n(n-1)}{N(N-1)}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}e_i e'_j\right]$$

$$= \frac{1}{n\,N}\sum_{i=1}^{N}(y_i-\bar{Y})(x_i-\bar{X})+\frac{n-1}{n\,N\,(N-1)}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}(y_i-\bar{Y})(x_j-\bar{X}) \qquad (7.10)$$

desarrollando la doble sumatoria,

$$\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}(y_i-\bar{Y})(x_j-\bar{X}) = \sum_{i=1}^{N}\sum_{j=1}^{N}(y_i-\bar{Y})(x_j-\bar{X})-\sum_{i=1}^{N}(y_i-\bar{Y})(x_i-\bar{X})$$

$$= \sum_{i=1}^{N}(y_i-\bar{Y})\sum_{j=1}^{N}(x_j-\bar{X})-\sum_{i=1}^{N}(y_i-\bar{Y})(x_i-\bar{X})$$

sustituyendo este resultado en (10), se tiene que,

$$E(\bar{e}\,\bar{e}') = \frac{1}{n\,N}\sum_{i=1}^{N}(y_i-\bar{Y})(x_i-\bar{X})+\frac{n-1}{n\,N\,(N-1)}\sum_{i=1}^{N}(y_i-\bar{Y})\sum_{j=1}^{N}(x_j-\bar{X})-\frac{n-1}{n\,N\,(N-1)}\sum_{i=1}^{N}(y_i-\bar{Y})(x_i-\bar{X})$$

pero como,

$$\sum_{i=1}^{N}(y_i-\bar{Y})=0 \qquad y \qquad \sum_{j=1}^{N}(x_j-\bar{X})=0$$

$$E(\bar{e}\,\bar{e}')=\frac{N-n}{n\,N\,(N-1)}\sum_{i=1}^{N}(y_i-\bar{Y})(x_i-\bar{X})=\frac{N-n}{N}\frac{S_{xy}}{n} \tag{7.11}$$

y al sustituir (7.11) en (7.9), se obtiene,

$$E(\hat{R})=R\left[1-\frac{1}{\bar{Y}\,\bar{X}}\frac{N-n}{N}\frac{S_{xy}}{n}+\frac{1}{\bar{X}^2}\frac{N-n}{N}\frac{S_x^2}{n}\right]=R\left[1+\frac{N-n}{N\,n}\left(\frac{S_x^2}{\bar{X}^2}-\frac{S_{xy}}{\bar{X}\,\bar{Y}}\right)\right] \tag{7.12}$$

que evidencia que $\hat{R}$ no es un estimador insesgado de R, y además su sesgo es,

$$B(\hat{R})=R-E(\hat{R})=R-R\left[1+\frac{N-n}{N\,n}\left(\frac{S_x^2}{\bar{X}^2}-\frac{S_{xy}}{\bar{X}\,\bar{Y}}\right)\right]=R\frac{N-n}{N\,n}\left(\frac{S_{xy}}{\bar{X}\bar{Y}}-\frac{S_x^2}{\bar{X}^2}\right) \tag{7.13}$$

7.1.2.- Error cuadrático medio de los estimadores

Ahora se hallará el error cuadrático medio, que en este caso será diferente de la varianza, por existir sesgo.

$$ECM(\hat{R})=V(\hat{R})+\left(B(\hat{R})\right)^2$$

Debe desarrollarse la expresión de la varianza de $\hat{R}$,

$$V(\hat{R})=E\left[\left(\hat{R}-E(\hat{R})\right)^2\right]=E(\hat{R}^2)-\left[E(\hat{R})\right]^2=E(\hat{R}^2)-R^2\left[1+\frac{N-n}{N\,n}\left(\frac{S_x^2}{\bar{X}^2}-\frac{S_{xy}}{\bar{X}\,\bar{Y}}\right)\right]^2$$

$$V(\hat{R})=E\left[\left(\hat{R}-E(\hat{R})\right)^2\right]=E(\hat{R}^2)-\left[E(\hat{R})\right]^2=E(\hat{R}^2)-R^2\left[1+\frac{N-n}{N\,n}\left(\frac{S_x^2}{\bar{X}^2}-\frac{S_{xy}}{\bar{X}\,\bar{Y}}\right)\right]^2$$

$$=E(\hat{R}^2)-R^2\left[1+2\left(\frac{N-n}{N\,n}\right)\left(\frac{S_x^2}{\bar{X}^2}-\frac{S_{xy}}{\bar{X}\,\bar{Y}}\right)+\left(\frac{N-n}{N\,n}\right)^2\left(\frac{S_x^2}{\bar{X}^2}-\frac{S_{xy}}{\bar{X}\,\bar{Y}}\right)^2\right]$$

$$=E(\hat{R}^2)-R^2-2R^2\left(\frac{N-n}{N\,n}\right)\frac{S_x^2}{\bar{X}^2}+2R^2\left(\frac{N-n}{N\,n}\right)\frac{S_{xy}}{\bar{X}\,\bar{Y}}-R^2\left(\frac{N-n}{N\,n}\right)^2\left(\frac{S_x^2}{\bar{X}^2}-\frac{S_{xy}}{\bar{X}\,\bar{Y}}\right)^2 \tag{7.14}$$

Al igual que para $E(\hat{R})$, para hallar $E(\hat{R}^2)$ se desarrollará la serie de Taylor del denominador de la siguiente expresión,

$$E(\hat{R}^2) = R^2\ E\left[\frac{\left(1+\frac{\bar{e}}{\bar{Y}}\right)^2}{\left(1+\frac{\bar{e}'}{\bar{X}}\right)^2}\right] \tag{7.15}$$

es decir,

$$\left(1+\frac{\bar{e}'}{\bar{X}}\right)^{-2} = \frac{1}{\left(1+\frac{\bar{e}'}{\bar{X}}\right)^2} = \frac{1}{\left(1+\frac{\bar{e}'}{\bar{X}}\right)^2} = \frac{1}{\left(\frac{\bar{X}+\bar{e}'}{\bar{X}}\right)^2} = \frac{\bar{X}^2}{\left(\bar{X}+\bar{e}'\right)^2}$$

entonces, se tiene que,

$$f^k(\bar{e}') = (-1)^k\ \frac{(k+1)!\,\bar{X}^2}{(\bar{X}+\bar{e}')^{k+2}} \qquad \text{y evaluando en cero,} \qquad f^k(0) = (-1)^k\ \frac{(k+1)!}{\bar{X}^k}$$

$$\sum_{k=0}^{\infty}(-1)^k\ \frac{(k+1)\,\bar{e}^k}{\bar{X}^k}$$

y la Serie de Taylor de $\left(1+\frac{\bar{e}'}{\bar{X}}\right)^{-2}$ respecto de cero es, $\sum_{k=0}^{\infty}(-1)^k\ \frac{(k+1)\,\bar{e}'^k}{\bar{X}^k}$

que sustituyendo en (7.15)

$$E(\hat{R}^2) = R^2\ E\left[\left(1+\frac{\bar{e}}{\bar{Y}}\right)^2\left(\sum_{k=0}^{\infty}(-1)^k\ \frac{(k+1)\,\bar{e}'^k}{\bar{X}^k}\right)\right] = R^2\ E\left[\left(1+2\frac{\bar{e}}{\bar{Y}}+\frac{\bar{e}^2}{\bar{Y}^2}\right)^2\left(\sum_{k=0}^{\infty}(-1)^k\ \frac{(k+1)\,\bar{e}'^k}{\bar{X}^k}\right)\right]$$

$$= R^2\ E\left[1+2\frac{\bar{e}}{\bar{Y}}+\frac{\bar{e}^2}{\bar{Y}^2}-2\frac{\bar{e}'}{\bar{X}}-4\frac{\bar{e}}{\bar{Y}}\frac{\bar{e}'}{\bar{X}}-2\frac{\bar{e}^2}{\bar{Y}^2}\frac{\bar{e}'}{\bar{X}}+3\frac{\bar{e}'^2}{\bar{X}^2}+6\frac{\bar{e}}{\bar{Y}}\frac{\bar{e}'^2}{\bar{X}^2}+3\frac{\bar{e}^2}{\bar{Y}^2}\frac{\bar{e}'^2}{\bar{X}^2}+\cdots\cdots\right]$$

Truncando la serie en las potencias de 2, se tiene,

$$E(\hat{R}^2) \approx R^2\ E\left[1+2\frac{\bar{e}}{\bar{Y}}+\frac{\bar{e}^2}{\bar{Y}^2}-2\frac{\bar{e}'}{\bar{X}}-4\frac{\bar{e}}{\bar{Y}}\frac{\bar{e}'}{\bar{X}}+3\frac{\bar{e}'^2}{\bar{X}^2}\right]$$

$$\approx R^2\left[1+\frac{2}{\bar{Y}}E(e)+\frac{1}{\bar{Y}^2}E(e^2)-\frac{2}{\bar{X}}E(e')-\frac{4}{\bar{Y}\,\bar{X}}E(ee')+\frac{3}{\bar{X}^2}E(e'^2)\right]$$

como $E(\bar{e}) = 0$ y $E(\bar{e}') = 0$,

$$E(\hat{R}^2) \approx R^2\left[1+\frac{1}{\bar{Y}^2}E(e^2)-\frac{4}{\bar{Y}\,\bar{X}}E(ee')+\frac{3}{\bar{X}^2}E(e'^2)\right]$$

$$\approx R^2\left[1+\frac{N-n}{Nn}\frac{S_y^2}{\bar{Y}^2}-4\frac{N-n}{Nn}\frac{S_{xy}}{\bar{X}\,\bar{Y}}+3\frac{N-n}{Nn}\frac{S_x^2}{\bar{X}^2}\right] \qquad (7.16)$$

Sustituyendo (7.16) en (7.14)

$$V(\hat{R})\approx R^2+R^2\frac{N-n}{Nn}\frac{S_y^2}{\bar{Y}^2}-4R^2\frac{N-n}{Nn}\frac{S_{xy}}{\bar{X}\bar{Y}}+3R^2\frac{N-n}{Nn}\frac{S_x^2}{\bar{X}^2}-R^2-2R^2\frac{N-n}{Nn}\frac{S_x^2}{\bar{X}^2}+2R^2\frac{N-n}{Nn}\frac{S_{xy}}{\bar{X}\bar{Y}}-R^2\left(\frac{N-n}{Nn}\right)^2\left(\frac{S_x^2}{\bar{X}^2}-\frac{S_{xy}}{\bar{X}\bar{Y}}\right)^2$$

$$\approx R^2\frac{N-n}{Nn}\frac{S_y^2}{\bar{Y}^2}+R^2\frac{N-n}{Nn}\frac{S_x^2}{\bar{X}^2}-2R^2\frac{N-n}{Nn}\frac{S_{xy}}{\bar{X}\bar{Y}}-R^2\left(\frac{N-n}{Nn}\right)^2\left(\frac{S_x^2}{\bar{X}^2}-\frac{S_{xy}}{\bar{X}\bar{Y}}\right)^2$$

entonces, el error cuadrático medio de $\hat{R}$ es,

$$EMC(\hat{R})=V(\hat{R})+\left[B(\hat{R})\right]^2\approx R^2\frac{N-n}{Nn}\frac{S_y^2}{\bar{Y}^2}+R^2\frac{N-n}{Nn}\frac{S_x^2}{\bar{X}^2}-2R^2\frac{N-n}{Nn}\frac{S_{xy}}{\bar{X}\bar{Y}}-R^2\left(\frac{N-n}{Nn}\right)^2\left(\frac{S_x^2}{\bar{X}^2}-\frac{S_{xy}}{\bar{X}\bar{Y}}\right)^2+\left[R\left(\frac{N-n}{Nn}\right)\left(\frac{S_{xy}}{\bar{X}\bar{Y}}-\frac{S_x^2}{\bar{X}^2}\right)\right]^2$$

$$\approx R^2\frac{N-n}{Nn}\frac{S_y^2}{\bar{Y}^2}+R^2\frac{N-n}{Nn}\frac{S_x^2}{\bar{X}^2}-2R^2\frac{N-n}{Nn}\frac{S_{xy}}{\bar{X}\bar{Y}}-R^2\left(\frac{N-n}{Nn}\right)^2\left(\frac{S_x^2}{\bar{X}^2}-\frac{S_{xy}}{\bar{X}\bar{Y}}\right)^2+R^2\left(\frac{N-n}{Nn}\right)^2\left(\frac{S_{xy}}{\bar{X}\bar{Y}}-\frac{S_x^2}{\bar{X}^2}\right)^2$$

y como

$$\left(\frac{S_x^2}{\bar{X}^2}-\frac{S_{xy}}{\bar{X}\bar{Y}}\right)^2=\left(\frac{S_{xy}}{\bar{X}\bar{Y}}-\frac{S_x^2}{\bar{X}^2}\right)^2$$

se tiene que

$$EMC(\hat{R})\approx R^2\frac{N-n}{Nn}\frac{S_y^2}{\bar{Y}^2}+R^2\frac{N-n}{Nn}\frac{S_x^2}{\bar{X}^2}-2R^2\frac{N-n}{Nn}\frac{S_{xy}}{\bar{X}\bar{Y}}=\frac{1-f}{n}\frac{1}{\bar{X}^2}\left[S_y^2+R^2S_x^2-2RS_{xy}\right] \qquad (7.17)$$

Luego, la Esperanza, Varianza, Sesgo y Error Cuadrático Medio para el estimador de la Razón son,

$$E(\hat{R}) = R\left[1 + \frac{1-f}{n}\left(\frac{S_x^2}{\overline{X}^2} - \frac{S_{xy}}{\overline{X}\,\overline{Y}}\right)\right]$$

$$V(\hat{R}) \approx R^2\left(\frac{1-f}{n}\right)\left[\frac{S_y^2}{\overline{Y}^2} + \frac{S_x^2}{\overline{X}^2} - 2\frac{S_{xy}}{\overline{X}\,\overline{Y}} - \left(\frac{1-f}{n}\right)\left(\frac{S_x^2}{\overline{X}^2} - \frac{S_{xy}}{\overline{X}\,\overline{Y}}\right)^2\right]$$

$$B(\hat{R}) = R\left(\frac{1-f}{n}\right)\left(\frac{S_{xy}}{\overline{X}\,\overline{Y}} - \frac{S_x^2}{\overline{X}^2}\right)$$

$$ECM(\hat{R}) \approx R^2\left(\frac{1-f}{n}\right)\left(\frac{S_y^2}{\overline{Y}^2} + \frac{S_x^2}{\overline{X}^2} - 2\frac{S_{xy}}{\overline{X}\,\overline{Y}}\right) = \frac{1}{\overline{X}^2}\left(\frac{1-f}{n}\right)\left(S_y^2 + R^2 S_x^2 - 2RS_{xy}\right)$$

Igualmente, la Esperanza, Varianza, Sesgo y Error Cuadrático Medio para el estimador del Promedio, del tipo Razón son,

$$E(\hat{\overline{Y}}_R) = \overline{Y}\left[1 + \frac{1-f}{n}\left(\frac{S_x^2}{\overline{X}^2} - \frac{S_{xy}}{\overline{X}\,\overline{Y}}\right)\right] \tag{7.18}$$

$$V(\hat{\overline{Y}}_R) \approx \overline{Y}^2\left(\frac{1-f}{n}\right)\left[\frac{S_y^2}{\overline{Y}^2} + \frac{S_x^2}{\overline{X}^2} - 2\frac{S_{xy}}{\overline{X}\,\overline{Y}} - \left(\frac{1-f}{n}\right)\left(\frac{S_x^2}{\overline{X}^2} - \frac{S_{xy}}{\overline{X}\,\overline{Y}}\right)^2\right]$$

$$B(\hat{\overline{Y}}_R) = \overline{Y}\left(\frac{1-f}{n}\right)\left(\frac{S_{xy}}{\overline{X}\,\overline{Y}} - \frac{S_x^2}{\overline{X}^2}\right) \tag{7.19}$$

$$ECM(\hat{\overline{Y}}_R) \approx \overline{Y}^2\left(\frac{1-f}{n}\right)\left(\frac{S_y^2}{\overline{Y}^2} + \frac{S_x^2}{\overline{X}^2} - 2\frac{S_{xy}}{\overline{X}\,\overline{Y}}\right) = \left(\frac{1-f}{n}\right)\left(S_y^2 + R^2 S_x^2 - 2RS_{xy}\right) \tag{7.20}$$

Finalmente, la Esperanza, Varianza, Sesgo y Error Cuadrático Medio para el estimador del Total, del tipo Razón son,

$$E(\hat{Y}_R) = Y\left[1 + \frac{1-f}{n}\left(\frac{S_x^2}{\overline{X}^2} - \frac{S_{xy}}{\overline{X}\,\overline{Y}}\right)\right] \tag{7.21}$$

$$V(\hat{Y}_R) \approx Y^2\left(\frac{1-f}{n}\right)\left[\frac{S_y^2}{\overline{Y}^2} + \frac{S_x^2}{\overline{X}^2} - 2\frac{S_{xy}}{\overline{X}\,\overline{Y}} - \left(\frac{1-f}{n}\right)\left(\frac{S_x^2}{\overline{X}^2} - \frac{S_{xy}}{\overline{X}\,\overline{Y}}\right)^2\right]$$

$$B(\hat{Y}_R) = Y\left(\frac{1-f}{n}\right)\left(\frac{S_{xy}}{\overline{X}\,\overline{Y}} - \frac{S_x^2}{\overline{X}^2}\right) \tag{7.22}$$

$$ECM(\hat{Y}_R) \approx Y^2\left(\frac{1-f}{n}\right)\left(\frac{S_y^2}{\overline{Y}^2} + \frac{S_x^2}{\overline{X}^2} - 2\frac{S_{xy}}{\overline{X}\,\overline{Y}}\right) = N^2\left(\frac{1-f}{n}\right)\left(S_y^2 + R^2 S_x^2 - 2RS_{xy}\right) \tag{7.23}$$

P.V. Sukhatme [8;145-147] realizó una mejor aproximación a la varianza de $\hat{R}$ desarrollando las series de Taylor hasta las potencias de orden 4, y llegó a que el sesgo de $\hat{R}$ $(B(\hat{R}))$ es

$$B'(\hat{R}) = B(\hat{R})\left[1 + \frac{3}{n}\frac{S_x^2}{\overline{X}^2}\right]$$

donde $B(\hat{R})$ es el sesgo mostrado en la ecuación (7.13), es decir, truncando las Series de Taylor en las potencias de orden 2. Nótese que si *n* es grande, ese incremento se hace despreciable, ocurriendo lo mismo con la varianza; incluso con el propio sesgo.

Nótese que las expresiones $ECM(\hat{R})$, $ECM(\hat{Y}_R)$ y $ECM(\hat{\overline{Y}}_R)$ están en función del término

$$\left[S_y^2 + R^2 S_x^2 - 2RS_{xy}\right]$$

entonces, se desarrollará éste, para llegar a una expresión equivalente de las varianzas, que será de utilidad más adelante.

$$\begin{aligned}
\left[S_y^2 + R^2 S_x^2 - 2RS_{xy}\right] &= \frac{\sum_{i=1}^{N}(y_i-\overline{Y})^2}{N-1} + R^2\frac{\sum_{i=1}^{N}(x_i-\overline{X})^2}{N-1} - 2R\frac{\sum_{i=1}^{N}(x_i-\overline{X})(y_i-\overline{Y})}{N-1} \\
&= \frac{1}{N-1}\left[\sum_{i=1}^{N}(y_i-\overline{Y})^2 + R^2\sum_{i=1}^{N}(x_i-\overline{X})^2 - 2R\sum_{i=1}^{N}(x_i-\overline{X})(y_i-\overline{Y})\right] \\
&= \frac{1}{N-1}\left[\sum_{i=1}^{N}y_i^2 - N\overline{Y}^2 + R^2\sum_{i=1}^{N}x_i^2 - R^2N\overline{X}^2 - 2R\sum_{i=1}^{N}(x_iy_i - \overline{X}y_y - x_i\overline{Y} + \overline{X}\overline{Y})\right] \\
&= \frac{1}{N-1}\left[\sum_{i=1}^{N}y_i^2 - N\overline{Y}^2 + R^2\sum_{i=1}^{N}x_i^2 - R^2N\overline{X}^2 - 2R\sum_{i=1}^{N}x_iy_i + 2\overline{R}Y + \overline{X}\overline{Y})\right] \\
&= \frac{1}{N-1}\left[\sum_{i=1}^{N}(y_i^2 - 2Rx_iy_i + R^2x_i^2) - N(\overline{Y}^2 - 2R\overline{X}\,\overline{Y} + R^2\overline{X}^2)\right] \\
&= \frac{1}{N-1}\left[\sum_{i=1}^{N}(y_i - Rx_i)^2 - N(\overline{Y} - R\overline{X})^2\right] \\
&= \frac{1}{N-1}\left[\sum_{i=1}^{N}(y_i - Rx_i)^2 - N(\overline{Y} - R\overline{X})^2 + N(\overline{Y} - R\overline{X})^2 - N(\overline{Y} - R\overline{X})^2\right] \\
&= \frac{1}{N-1}\left[\sum_{i=1}^{N}(y_i - Rx_i)^2 - 2N(\overline{Y} - R\overline{X})^2 + N(\overline{Y} - R\overline{X})^2\right]
\end{aligned}$$

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

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

$$= \frac{1}{N-1}\sum_{i=1}^{N}\left[(y_i - Rx_i) - (\bar{Y} - R\bar{X})\right]^2 \qquad (7.24)$$

pero,

$$\bar{Y} - R\bar{X} = \bar{Y} - \left(\frac{\bar{Y}}{\bar{X}}\right)\bar{X} = 0 \qquad (7.25)$$

entonces,

$$S_y^2 - 2RS_{xy} + R^2 S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - Rx_i)^2 \qquad (7.26)$$

por lo tanto, sustituyendo (7.26) en (7.17), (7.20) y (7.23),

$$ECM(\hat{R}) = \frac{1-f}{n}\frac{1}{\bar{X}^2}\frac{\sum_{i=1}^{N}(y_i - Rx_i)^2}{N-1} \qquad (7.27)$$

$$ECM(\hat{\bar{Y}}_R) = \frac{1-f}{n}\frac{\sum_{i=1}^{N}(y_i - Rx_i)^2}{N-1} \qquad (7.28)$$

$$ECM(\hat{Y}_R) = \frac{N^2(1-f)}{n}\frac{\sum_{i=1}^{N}(y_i - Rx_i)^2}{N-1} \qquad (7.29)$$

7.1.3.- Estimadores de los errores cuadráticos medios de los estimadores

Para estimar los errores cuadráticos medios de los estimadores, bastaría con sustituir en la fórmula de la varianza, los valores poblacionales por los estimados, es decir, las cuasivarianzas, las razones, los totales y los promedios. Esto es,

$$E\hat{C}M(\hat{R}) \approx \hat{R}^2\left(\frac{1-f}{n}\right)\left(\frac{s_y^2}{\bar{y}^2}+\frac{s_x^2}{\bar{x}^2}-2\frac{s_{xy}}{\bar{x}\,\bar{y}}\right)=\frac{1}{x^2}\left(\frac{1-f}{n}\right)\left(s_y^2+\hat{R}^2 s_x^2-2\hat{R}s_{xy}\right)$$

$$E\hat{C}M(\hat{\bar{Y}}_R) \approx y^2\left(\frac{1-f}{n}\right)\left(\frac{s_y^2}{y^2}+\frac{s_x^2}{x^2}-2\frac{s_{xy}}{x\,y}\right)=\left(\frac{1-f}{n}\right)\left(s_y^2+\hat{R}^2 s_x^2-2\hat{R}s_{xy}\right)$$

$$E\hat{C}M(\hat{Y}_R) \approx \hat{Y}^2\left(\frac{1-f}{n}\right)\left(\frac{s_y^2}{y^2}+\frac{s_x^2}{x^2}-2\frac{s_{xy}}{x\,y}\right)=N^2\left(\frac{1-f}{n}\right)\left(s_y^2+\hat{R}^2 s_x^2-2\hat{R}s_{xy}\right)$$

7.1.4.- <u>Estimadores de razón en el muestreo estratificado</u>

Si la muestra es estratificada, se tiene, en principio, un estimador para cada estrato, es decir,

$$R_h=\frac{\sum_{i=1}^{N} y_{hi}}{\sum_{i=1}^{N} x_{hi}}=\frac{Y_h}{X_h}=\frac{\bar{Y}_h}{\bar{X}_h}$$

y su estimador es,

$$\hat{R}_h=\frac{\sum_{i=1}^{n} y_{hi}}{\sum_{i=1}^{n} x_{hi}}=\frac{y_h}{x_h}=\frac{\bar{y}_h}{\bar{x}_h}=\frac{\hat{Y}_h}{\hat{X}_h}$$

entonces, los estimadores de razón del total y el promedio respectivamente son,

$$\hat{Y}_{R_h}=\hat{R}_h\;X_h=\frac{y_h}{x_h}\;X_h$$

$$\hat{\bar{Y}}_{R_h}=\bar{y}_{R_h}=\hat{R}_h\;\bar{X}_h=\frac{y_h}{x_h}\;\bar{X}_h=\frac{1}{N_h}\;\hat{Y}_{R_h}$$

y sus esperanzas, varianzas, sesgos y errores cuadráticos medios son:

para el estimador de la Razón son,

$$E(\hat{R}_h) = R_h \left[1 + \frac{1-f_h}{n_h} \left(\frac{S_{x_h}^2}{\overline{X}_h^2} - \frac{S_{xy_h}}{\overline{X}_h \overline{Y}_h} \right) \right]$$

$$V(\hat{R}_h) \approx R_h^2 \left(\frac{1-f_h}{n_h} \right) \left[\frac{S_{y_h}^2}{\overline{Y}_h^2} + \frac{S_{x_h}^2}{\overline{X}_h^2} - 2\frac{S_{xy_h}}{\overline{X}_h \overline{Y}_h} - \left(\frac{1-f_h}{n_h} \right) \left(\frac{S_{x_h}^2}{\overline{X}_h^2} - \frac{S_{xy_h}}{\overline{X}_h \overline{Y}_h} \right)^2 \right]$$

$$B(\hat{R}_h) = R_h \left(\frac{1-f_h}{n_h} \right) \left(\frac{S_{xy_h}}{\overline{X}_h \overline{Y}_h} - \frac{S_{x_h}^2}{\overline{X}_h^2} \right)$$

$$ECM(\hat{R}_h) \approx R_h^2 \left(\frac{1-f_h}{n_h} \right) \left(\frac{S_{y_h}^2}{\overline{Y}_h^2} + \frac{S_{x_h}^2}{\overline{X}_h^2} - 2\frac{S_{xy_h}}{\overline{X}_h \overline{Y}_h} \right) = \frac{1}{\overline{X}_h^2} \left(\frac{1-f_h}{n_h} \right) \left(S_{y_h}^2 + R_h^2 S_{x_h}^2 - 2R_h S_{xy_h} \right)$$

para el estimador del Promedio,

$$E(\hat{\overline{Y}}_{R_h}) = \overline{Y}_h \left[1 + \frac{1-f_h}{n_h} \left(\frac{S_{x_h}^2}{\overline{X}_h^2} - \frac{S_{xy_h}}{\overline{X}_h \overline{Y}_h} \right) \right]$$

$$V(\hat{\overline{Y}}_{R_h}) \approx \overline{Y}_h^2 \left(\frac{1-f_h}{n_h} \right) \left[\frac{S_{y_h}^2}{\overline{Y}_h^2} + \frac{S_{x_h}^2}{\overline{X}_h^2} - 2\frac{S_{xy_h}}{\overline{X}_h \overline{Y}_h} - \left(\frac{1-f_h}{n_h} \right) \left(\frac{S_{x_h}^2}{\overline{X}_h^2} - \frac{S_{xy_h}}{\overline{X}_h \overline{Y}_h} \right)^2 \right]$$

$$B(\hat{\overline{Y}}_{R_h}) = \overline{Y}_h \left(\frac{1-f_h}{n_h} \right) \left(\frac{S_{xy_h}}{\overline{X}_h \overline{Y}_h} - \frac{S_{x_h}^2}{\overline{X}_h^2} \right)$$

$$ECM(\hat{\overline{Y}}_{R_h}) \approx \overline{Y}_h^2 \left(\frac{1-f_h}{n_h} \right) \left(\frac{S_{y_h}^2}{\overline{Y}_h^2} + \frac{S_{x_h}^2}{\overline{X}_h^2} - 2\frac{S_{xy_h}}{\overline{X}_h \overline{Y}_h} \right) = \left(\frac{1-f_h}{n_h} \right) \left(S_{y_h}^2 + R_h^2 S_{x_h}^2 - 2R_h S_{xy_h} \right)$$

para el estimador del Total,

$$E(\hat{Y}_{R_h}) = Y_h \left[1 + \frac{1-f_h}{n_h} \left(\frac{S_{x_h}^2}{\overline{X}_h^2} - \frac{S_{xy_h}}{\overline{X}_h \overline{Y}_h} \right) \right]$$

$$V(\hat{Y}_{R_h}) \approx Y_h^2 \left(\frac{1-f_h}{n_h} \right) \left[\frac{S_{y_h}^2}{\overline{Y}_h^2} + \frac{S_{x_h}^2}{\overline{X}_h^2} - 2\frac{S_{xy_h}}{\overline{X}_h \overline{Y}_h} - \left(\frac{1-f_h}{n_h} \right) \left(\frac{S_{x_h}^2}{\overline{X}_h^2} - \frac{S_{xy_h}}{\overline{X}_h \overline{Y}_h} \right)^2 \right]$$

$$B(\hat{Y}_{R_h}) = Y_h \left(\frac{1-f_h}{n_h} \right) \left(\frac{S_{xy_h}}{\overline{X}_h \overline{Y}_h} - \frac{S_{x_h}^2}{\overline{X}_h^2} \right)$$

$$ECM(\hat{Y}_{R_h}) \approx Y_h^2 \left(\frac{1-f_h}{n_h} \right) \left(\frac{S_{y_h}^2}{\overline{Y}_h^2} + \frac{S_{x_h}^2}{\overline{X}_h^2} - 2\frac{S_{xy_h}}{\overline{X}_h \overline{Y}_h} \right) = N_h^2 \left(\frac{1-f_h}{n_h} \right) \left(S_{y_h}^2 + R_h^2 S_{x_h}^2 - 2R_h S_{xy_h} \right)$$

Aplicando el razonamiento de Muestreo Estratificado,

$$\hat{Y}_R = \sum_{h=1}^{L} \hat{Y}_{R_h} = \sum_{h=1}^{L} \hat{R}_h X_h = \sum_{h=1}^{L} \frac{y_h}{x_h} X_h \quad ; \quad \hat{\bar{Y}}_R = \frac{1}{N}\hat{Y}_R = \frac{1}{N}\sum_{h=1}^{L} \hat{Y}_{R_h} = \frac{1}{N}\sum_{h=1}^{L} N_h \hat{\bar{Y}}_{R_h} = \sum_{h=1}^{L} W_h \hat{\bar{Y}}_{R_h}$$

A estos estimadores se les denominarán "Estimadores de Razón Separado" del total y del promedio respectivamente, y se denotarán por $\hat{Y}_{Rs}$, $\hat{\bar{Y}}_{Rs}$. Es decir,

$$\hat{Y}_{Rs} = \sum_{h=1}^{L} \hat{Y}_{R_h} = \sum_{h=1}^{L} \hat{R}_h X_h = \sum_{h=1}^{L} \frac{y_h}{x_h} X_h \quad ; \quad \hat{\bar{Y}}_{Rs} = \frac{1}{N}\hat{Y}_{Rs} = \frac{1}{N}\sum_{h=1}^{L} \hat{Y}_{R_h} = \frac{1}{N}\sum_{h=1}^{L} N_h \hat{\bar{Y}}_{R_h} = \sum_{h=1}^{L} W_h \hat{\bar{Y}}_{R_h}$$

7.1.4.1.- Estimadores de razón separado

Como se indicó en el apartado anterior, los Estimadores de Razón Separado del total y del promedio, respectivamente son,

$$\hat{Y}_{Rs} = \sum_{h=1}^{L} \hat{Y}_{R_h} = \sum_{h=1}^{L} \hat{R}_h X_h = \sum_{h=1}^{L} \frac{y_h}{x_h} X_h \quad ; \quad \hat{\bar{Y}}_{Rs} = \frac{1}{N}\hat{Y}_{Rs} = \frac{1}{N}\sum_{h=1}^{L} \hat{Y}_{R_h} = \frac{1}{N}\sum_{h=1}^{L} N_h \hat{\bar{Y}}_{R_h} = \sum_{h=1}^{L} W_h \hat{\bar{Y}}_{R_h}$$

y se denominan así porque utilizan una Razón para cada estrato, dicho de otra forma, trabajan con la razón separada por estrato.

Para hallar la esperanza se tiene que,

$$E(\hat{Y}_{Rs}) = E\left[\sum_{h=1}^{L} \hat{Y}_{R_h}\right] = \sum_{h=1}^{L} E(\hat{Y}_{R_h}) = \sum_{h=1}^{L} Y_h \left[1 + \frac{1-f_h}{n_h}\left(\frac{S_{x_h}^2}{\bar{X}_h^2} - \frac{S_{xy_h}}{\bar{X}_h \bar{Y}_h}\right)\right] = Y + \sum_{h=1}^{L} Y_h \left(\frac{1-f_h}{n_h}\right)\left(\frac{S_{x_h}^2}{\bar{X}_h^2} - \frac{S_{xy_h}}{\bar{X}_h \bar{Y}_h}\right)$$

por lo tanto,

$$B(\hat{Y}_{Rs}) = \sum_{h=1}^{L} Y_h \left(\frac{1-f_h}{n_h}\right)\left(\frac{S_{xy_h}}{\bar{X}_h \bar{Y}_h} - \frac{S_{x_h}^2}{\bar{X}_h^2}\right)$$

En el caso de la varianza, se tiene que,

$$V(\hat{Y}_{Rs}) = V\left(\sum_{h=1}^{L} \hat{Y}_{R_h}\right) = \sum_{h=1}^{L} V(\hat{Y}_{R_h}) + 2\sum_{h=1}^{L-1}\sum_{k=h+1}^{L} COV(\hat{Y}_{R_h}, \hat{Y}_{R_h}) = \sum_{h=1}^{L} V(\hat{Y}_{R_h})$$

ya que las muestras en los estratos son independientes, entonces,

$$V(\hat{Y}_{Rs}) \approx \sum_{h=1}^{L} Y_h^2 \left(\frac{1-f_h}{n_h}\right)\left[\frac{S_{y_h}^2}{\overline{Y}_h^2}+\frac{S_{x_h}^2}{\overline{X}_h^2}-2\frac{S_{xy_h}}{\overline{X}_h\,\overline{Y}_h}-\left(\frac{1-f_h}{n_h}\right)\left(\frac{S_{x_h}^2}{\overline{X}_h^2}-\frac{S_{xy_h}}{\overline{X}_h\,\overline{Y}_h}\right)^2\right]$$

y el error cuadrático medio,

$$ECM(\hat{Y}_{Rs}) = V(\hat{Y}_{Rs}) + \left[B(\hat{Y}_{Rs})\right]^2$$

$$ECM(\hat{Y}_{Rs}) \approx \sum_{h=1}^{L} Y_h^2\left(\frac{1-f_h}{n_h}\right)\left(\frac{S_{y_h}^2}{\overline{Y}_h^2}+\frac{S_{x_h}^2}{\overline{X}_h^2}-\frac{2S_{xy_h}}{\overline{X}_h\overline{Y}_h}\right)+\sum_{h=1}^{L}\sum_{\substack{k=1\\k\neq h}}^{L} Y_hY_k\left(\frac{1-f_h}{n_h}\right)\left(\frac{1-f_k}{n_k}\right)\left(\frac{S_{xy_h}}{\overline{X}_h\,\overline{Y}_h}-\frac{S_{x_h}^2}{\overline{X}_h^2}\right)\left(\frac{S_{xy_k}}{\overline{X}_k\,\overline{Y}_k}-\frac{S_{x_k}^2}{\overline{X}_k^2}\right)$$

$$\approx \sum_{h=1}^{L} N_h^2\left(\frac{1-f_h}{n_h}\right)\left(S_{y_h}^2+R_h^2S_{x_h}^2-2R_hS_{xy_h}\right)+\sum_{h=1}^{L}\sum_{\substack{k=1\\k\neq h}}^{L} Y_hY_k\left(\frac{1-f_h}{n_h}\right)\left(\frac{1-f_k}{n_k}\right)\left(\frac{S_{xy_h}}{\overline{X}_h\,\overline{Y}_h}-\frac{S_{x_h}^2}{\overline{X}_h^2}\right)\left(\frac{S_{xy_k}}{\overline{X}_k\,\overline{Y}_k}-\frac{S_{x_k}^2}{\overline{X}_k^2}\right)$$

$$\approx \sum_{h=1}^{L} N_h^2\left(\frac{1-f_h}{n_h}\right)\left(S_{y_h}^2+R_h^2S_{x_h}^2-2R_hS_{xy_h}\right)+2\sum_{h=1}^{L-1}\sum_{k=h+1}^{L} Y_hY_k\left(\frac{1-f_h}{n_h}\right)\left(\frac{1-f_k}{n_k}\right)\left(\frac{S_{xy_h}}{\overline{X}_h\,\overline{Y}_h}-\frac{S_{x_h}^2}{\overline{X}_h^2}\right)\left(\frac{S_{xy_k}}{\overline{X}_k\,\overline{Y}_k}-\frac{S_{x_k}^2}{\overline{X}_k^2}\right)$$

De manera que,

$$E(\hat{Y}_{Rs}) = \sum_{h=1}^{L} Y_h\left[1+\frac{1-f_h}{n_h}\left(\frac{S_{x_h}^2}{\overline{X}_h^2}-\frac{S_{xy_h}}{\overline{X}_h\,\overline{Y}_h}\right)\right] \tag{7.30}$$

$$B(\hat{Y}_{Rs}) = \sum_{h=1}^{L} Y_h\left(\frac{1-f_h}{n_h}\right)\left(\frac{S_{xy_h}}{\overline{X}_h\,\overline{Y}_h}-\frac{S_{x_h}^2}{\overline{X}_h^2}\right) \tag{7.31}$$

$$V(\hat{Y}_{Rs}) \approx \sum_{h=1}^{L} Y_h^2\left(\frac{1-f_h}{n_h}\right)\left[\frac{S_{y_h}^2}{\overline{Y}_h^2}+\frac{S_{x_h}^2}{\overline{X}_h^2}-2\frac{S_{xy_h}}{\overline{X}_h\,\overline{Y}_h}-\left(\frac{1-f_h}{n_h}\right)\left(\frac{S_{x_h}^2}{\overline{X}_h^2}-\frac{S_{xy_h}}{\overline{X}_h\,\overline{Y}_h}\right)^2\right] \tag{7.32}$$

$$ECM(\hat{Y}_{Rs}) \approx \sum_{h=1}^{L} N_h^2\left(\frac{1-f_h}{n_h}\right)\left(S_{y_h}^2+R_h^2S_{x_h}^2-2R_hS_{xy_h}\right)+$$
$$+2\sum_{h=1}^{L-1}\sum_{k=h+1}^{L} Y_hY_k\left(\frac{1-f_h}{n_h}\right)\left(\frac{1-f_k}{n_k}\right)\left(\frac{S_{xy_h}}{\overline{X}_h\,\overline{Y}_h}-\frac{S_{x_h}^2}{\overline{X}_h^2}\right)\left(\frac{S_{xy_k}}{\overline{X}_k\,\overline{Y}_k}-\frac{S_{x_k}^2}{\overline{X}_k^2}\right) \tag{7.33}$$

$$\hat{ECM}(\hat{Y}_{Rs}) \approx \sum_{h=1}^{L} N_h^2\left(\frac{1-f_h}{n_h}\right)\left(s_{y_h}^2+\hat{R}_h^2s_{x_h}^2-2\hat{R}_hs_{xy_h}\right)+$$
$$+2\sum_{h=1}^{L-1}\sum_{k=h+1}^{L} Y_hY_k\left(\frac{1-f_h}{n_h}\right)\left(\frac{1-f_k}{n_k}\right)\left(\frac{s_{xy_h}}{\overline{x}_h\,\overline{y}_h}-\frac{s_{x_h}^2}{\overline{x}_h^2}\right)\left(\frac{s_{xy_k}}{\overline{x}_h\,\overline{y}_h}-\frac{s_{x_k}^2}{\overline{x}_h^2}\right) \tag{7.34}$$

y para el estimador del Promedio,

$$E(\hat{\bar{Y}}_{Rs}) = \sum_{h=1}^{L} W_h \bar{Y}_h \left[1 + \frac{1-f_h}{n_h} \left(\frac{S_{x_h}^2}{\bar{X}_h^2} - \frac{S_{xy_h}}{\bar{X}_h \bar{Y}_h} \right) \right]$$

$$B(\hat{\bar{Y}}_{Rs}) = \sum_{h=1}^{L} W_h \bar{Y}_h \left(\frac{1-f_h}{n_h} \right) \left(\frac{S_{xy_h}}{\bar{X}_h \bar{Y}_h} - \frac{S_{x_h}^2}{\bar{X}_h^2} \right)$$

$$V(\hat{\bar{Y}}_{Rs}) \approx \sum_{h=1}^{L} W_h^2 \bar{Y}_h^2 \left(\frac{1-f_h}{n_h} \right) \left[\frac{S_{y_h}^2}{\bar{Y}_h^2} + \frac{S_{x_h}^2}{\bar{X}_h^2} - 2\frac{S_{xy_h}}{\bar{X}_h \bar{Y}_h} - \left(\frac{1-f_h}{n_h} \right) \left(\frac{S_{x_h}^2}{\bar{X}_h^2} - \frac{S_{xy_h}}{\bar{X}_h \bar{Y}_h} \right)^2 \right]$$

$$ECM(\hat{\bar{Y}}_{Rs}) \approx \sum_{h=1}^{L} W_h^2 \left(\frac{1-f_h}{n_h} \right) \left(S_{y_h}^2 + R_h^2 S_{x_h}^2 - 2R_h S_{xy_h} \right) + 2 \sum_{h=1}^{L-1} \sum_{k=h+1}^{L} \bar{Y}_h \bar{Y}_k \left(\frac{1-f_h}{n_h} \right) \left(\frac{1-f_k}{n_k} \right) \left(\frac{S_{xy_h}}{\bar{X}_h \bar{Y}_h} - \frac{S_{x_h}^2}{\bar{X}_h^2} \right) \left(\frac{S_{xy_k}}{\bar{X}_k \bar{Y}_k} - \frac{S_{x_k}^2}{\bar{X}_k^2} \right)$$

$$\hat{ECM}(\hat{\bar{Y}}_{Rs}) \approx \sum_{h=1}^{L} W_h^2 \left(\frac{1-f_h}{n_h} \right) \left(s_{y_h}^2 + \hat{R}_h^2 s_{x_h}^2 - 2\hat{R}_h s_{xy_h} \right) + 2 \sum_{h=1}^{L-1} \sum_{k=h+1}^{L} \bar{y}_h \bar{y}_k \left(\frac{1-f_h}{n_h} \right) \left(\frac{1-f_k}{n_k} \right) \left(\frac{s_{xy_h}}{\bar{x}_h \bar{y}_h} - \frac{s_{x_h}^2}{\bar{x}_h^2} \right) \left(\frac{s_{xy_k}}{\bar{x}_h \bar{y}_h} - \frac{s_{x_k}^2}{\bar{x}_h^2} \right)$$

El estimador de la Razón es,

$$\hat{R}_s = \frac{\hat{\bar{Y}}_{Rs}}{\bar{X}} = \frac{1}{\bar{X}} \sum_{h=1}^{L} \hat{R}_h X_h = \sum_{h=1}^{L} \left(\frac{X_h}{X} \right) \hat{R}_h \qquad \text{o bien}$$

$$\hat{R}_s = \frac{\hat{\bar{Y}}_{Rs}}{\bar{X}} = \frac{1}{N\bar{X}} \sum_{h=1}^{L} N_h \hat{\bar{Y}}_{R_h} = \frac{1}{X} \sum_{h=1}^{L} N_h \hat{R}_h \bar{X}_h = \frac{1}{X} \sum_{h=1}^{L} X_h \hat{R}_h = \sum_{h=1}^{L} \left(\frac{X_h}{X} \right) \hat{R}_h \quad \text{, llegando al mismo resultado.}$$

De manera que,

$$E(\hat{R}_s) = \frac{1}{\bar{X}} \sum_{h=1}^{L} W_h \bar{Y}_h \left[1 + \frac{1-f_h}{n_h} \left(\frac{S_{x_h}^2}{\bar{X}_h^2} - \frac{S_{xy_h}}{\bar{X}_h \bar{Y}_h} \right) \right]$$

$$B(\hat{R}_s) = \frac{1}{\bar{X}} \sum_{h=1}^{L} W_h \bar{Y}_h \left(\frac{1-f_h}{n_h} \right) \left(\frac{S_{xy_h}}{\bar{X}_h \bar{Y}_h} - \frac{S_{x_h}^2}{\bar{X}_h^2} \right)$$

$$V(\hat{R}_s) \approx \frac{1}{\bar{X}^2} \sum_{h=1}^{L} W_h^2 \bar{Y}_h^2 \left(\frac{1-f_h}{n_h} \right) \left[\frac{S_{y_h}^2}{\bar{Y}_h^2} + \frac{S_{x_h}^2}{\bar{X}_h^2} - 2\frac{S_{xy_h}}{\bar{X}_h \bar{Y}_h} - \left(\frac{1-f_h}{n_h} \right) \left(\frac{S_{x_h}^2}{\bar{X}_h^2} - \frac{S_{xy_h}}{\bar{X}_h \bar{Y}_h} \right)^2 \right]$$

$$ECM(\hat{R}_s) \approx \frac{1}{\bar{X}^2} \sum_{h=1}^{L} W_h^2 \left(\frac{1-f_h}{n_h} \right) \left(S_{y_h}^2 + R_h^2 S_{x_h}^2 - 2R_h S_{xy_h} \right) + 2 \sum_{h=1}^{L-1} \sum_{k=h+1}^{L} \bar{Y}_h \bar{Y}_k \left(\frac{1-f_h}{n_h} \right) \left(\frac{1-f_k}{n_k} \right) \left(\frac{S_{xy_h}}{\bar{X}_h \bar{Y}_h} - \frac{S_{x_h}^2}{\bar{X}_h^2} \right) \left(\frac{S_{xy_k}}{\bar{X}_k \bar{Y}_k} - \frac{S_{x_k}^2}{\bar{X}_k^2} \right)$$

$$\hat{ECM}(\hat{R}_s) \approx \frac{1}{\bar{X}^2} \sum_{h=1}^{L} W_h^2 \left(\frac{1-f_h}{n_h} \right) \left(s_{y_h}^2 + \hat{R}_h^2 s_{x_h}^2 - 2\hat{R}_h s_{xy_h} \right) + 2 \sum_{h=1}^{L-1} \sum_{k=h+1}^{L} \bar{y}_h \bar{y}_k \left(\frac{1-f_h}{n_h} \right) \left(\frac{1-f_k}{n_k} \right) \left(\frac{s_{xy_h}}{x_h \bar{y}_h} - \frac{s_{x_h}^2}{\bar{x}_h^2} \right) \left(\frac{s_{xy_k}}{x_h \bar{y}_h} - \frac{s_{x_k}^2}{\bar{x}_h^2} \right)$$

En la bibliografía clásica, se muestra el error cuadrático medio como la varianza, y se desprecia el segundo término en todos los casos -estimador del promedio, total y razón-, es decir, la doble sumatoria.

7.1.4.2.- Estimadores de razón combinado

El uso de estimadores de razón y/o del tipo razón en el muestreo estratificado, contempla dos casos, en el primero (estimadores de razón separada), visto anteriormente, se consideran las razones de cada estrato "separadamente", en el segundo se utiliza una única razón para toda la población, "combinando", de alguna manera, la información de cada uno de los estratos, y están basados en el hecho de que las razones en cada estrato son muy similares, de manera que no se justifica trabajar con diferentes razones para cada estrato si la ganancia en precisión no es sustancial, o visto de otra manera, no se sacrifica mucha precisión por simplificar el procedimiento.

Cabe destacar que para aplicar los estimadores de razón separado, se requieren tamaños de muestra suficientemente grandes en cada estrato que permitan estimaciones confiables de las razones R_h. En el caso de los estimadores de razón combinado, no se requieren tamaños de muestra grandes en cada estrato, sino un tamaño de muestra total suficientemente grande para estimar la razón combinada; esto lleva a la conclusión de que generalmente se requieren tamaños de muestra más grandes para aplicar razones separadas que combinada.

Por otra parte, si las razones en cada estrato son iguales, tendría el mismo efecto utilizarlas de forma separada, que una razón común. Realmente esto es muy difícil que ocurra, pero si pueden ser parecidas, y en ese caso se puede usar, no una razón común estrictamente hablando, pero si una razón combinada, esto es,

$$R_1 = R_2 = \ldots = R_L = R_c \qquad \text{que no es más que,}$$

$$\frac{\bar{Y}_1}{\bar{X}_1} = \frac{\bar{Y}_2}{\bar{X}_2} = \ldots = \frac{\bar{Y}_L}{\bar{X}_L} = \frac{\bar{Y}}{\bar{X}} \qquad \text{ó} \qquad \frac{Y_1}{X_1} = \frac{Y_2}{X_2} = \ldots = \frac{Y_L}{X_L} = \frac{Y}{X}$$

es decir, $$R_c = \frac{\bar{Y}}{\bar{X}} = \frac{Y}{X}$$

Para estimar R_c se tienen dos opciones,

i) usar la razón de los promedios, $$\hat{R}_c = \frac{\bar{y}_{st}}{\bar{x}_{st}} = \frac{\hat{\bar{Y}}_{st}}{\hat{\bar{X}}_{st}} = \frac{\hat{Y}_{st}}{\hat{X}_{st}} = \frac{\sum_{h=1}^{L} W_h y_h}{\sum_{h=1}^{L} W_h x_h}$$

ii) usar el promedio de las razones, $$\hat{R}_{cst} = \sum_{h=1}^{L} W_h \hat{R}_h = \sum_{h=1}^{L} W_h \left(\frac{\bar{y}_h}{\bar{x}_h} \right) = \hat{R}_{st}$$

El primero, que se tratará a continuación, se denomina “estimador de Razón Combinado”, y es ampliamente conocido y trabajado en muchos textos, el segundo se denominará “estimador de Razón Combinado Estratificado”, suele tener menor precisión y la estructura de la varianza y el sesgo son bastante más complicadas que el anterior, motivos por los cuales no son trabajados ni conocidos.

Los Estimadores de Razón Combinado del total y del promedio, respectivamente son,

$$\hat{Y}_{Rc} = \frac{\bar{y}_{st}}{\bar{x}_{st}} X \qquad ; \qquad \hat{\bar{Y}}_{Rc} = \frac{\bar{y}_{st}}{\bar{x}_{st}} \bar{X}$$

a continuación se procederá a desarrollar la esperanza, sesgo, varianza y error cuadrático medio de $\hat{Y}_{Rc}$, y luego para $\hat{Y}_{Rc}$.

$$E\left(\hat{Y}_{Rc}\right) = E\left(\frac{\hat{Y}_{st}}{\hat{X}_{st}} X\right) = X\, E\left(\frac{\hat{Y}_{st}}{\hat{X}_{st}}\right)$$

haciendo $y_{hi} = \bar{Y}_h + e_h$, se tiene,

$$\bar{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_{hi}} = \frac{\sum_{i=1}^{n_h} \bar{Y}_h}{n_{hi}} + \frac{\sum_{i=1}^{n_h} e_h}{n_{hi}} = \bar{Y}_h + \bar{e}_h$$

$$E\left(\bar{y}_h\right) = \bar{Y}_h + E\left(\bar{e}_h\right) = \bar{Y}_h + \bar{E}_h = \bar{Y}_h \quad , \; V\left(\bar{y}_h\right) = \frac{N_h - n_h}{N_h} \frac{S_{y_h}^2}{n_h}$$

$$\hat{\bar{Y}}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h = \sum_{h=1}^{L} W_h \bar{Y}_h + \sum_{h=1}^{L} W_h \bar{e}_h = \bar{Y} + \bar{e}$$

$$E\left(\hat{\bar{Y}}_{st}\right) = E\left(\bar{Y} + \bar{e}\right) = \bar{Y} + E\left(\bar{e}\right) = \bar{Y} + \bar{E} = \bar{Y}$$

análogamente para $x_{hi} = \bar{X}_h + e'_h$

entonces,

$$\frac{\hat{Y}_{st}}{\hat{X}_{st}} = \frac{\left(\bar{Y} + \bar{e}\right)}{\left(\bar{X} + \bar{e}'\right)} = \frac{\bar{Y}\left(1 + \frac{\bar{e}}{\bar{Y}}\right)}{\bar{X}\left(1 + \frac{\bar{e}'}{\bar{X}}\right)} \qquad \text{luego,} \quad E\left(\frac{\hat{Y}_{st}}{\hat{X}_{st}}\right) = \frac{\bar{Y}}{\bar{X}} E\left[\frac{\left(1 + \frac{\bar{e}}{\bar{Y}}\right)}{\left(1 + \frac{\bar{e}'}{\bar{X}}\right)}\right]$$

aplicando el mismo procedimiento desarrollado en la sección 7.1.1. (Esperanza de los Estimadores de Razón), se tiene,

$$E\left(\frac{\hat{Y}_{st}}{\hat{X}_{st}}\right) = \frac{\bar{Y}}{\bar{X}}\left[1 + \frac{1}{\bar{Y}} E\left(\bar{e}\right) - \frac{1}{\bar{X}} E\left(\bar{e}'\right) - \frac{1}{\bar{Y}\bar{X}} E\left(\bar{e}\bar{e}'\right) + \frac{1}{\bar{X}^2} E\left(\bar{e}'^2\right)\right]$$

donde, $E(\bar{e})=0$; $E(\bar{e}')=0$; $E(\bar{e}\bar{e}')=\sum_{h=1}^{L}W_h^2\left(\frac{1-f_h}{n_h}\right)S_{xy_h}$; $E(\bar{e}'^2)=\sum_{h=1}^{L}W_h^2\left(\frac{1-f_h}{n_h}\right)S_{x_h}^2$

entonces,

$$E\left(\frac{\hat{\bar{Y}}_{st}}{\hat{\bar{X}}_{st}}\right)=\frac{\bar{Y}}{\bar{X}}\left[1-\frac{1}{\bar{Y}\bar{X}}\sum_{h=1}^{L}W_h^2\left(\frac{1-f_h}{n_h}\right)S_{xy_h}+\frac{1}{\bar{X}^2}\sum_{h=1}^{L}W_h^2\left(\frac{1-f_h}{n_h}\right)S_{x_h}^2\right]$$

$$=\frac{\bar{Y}}{\bar{X}}-\frac{1}{\bar{X}^2}\sum_{h=1}^{L}W_h^2\left(\frac{1-f_h}{n_h}\right)S_{xy_h}+\frac{\bar{Y}}{\bar{X}^3}\sum_{h=1}^{L}W_h^2\left(\frac{1-f_h}{n_h}\right)S_{x_h}^2$$

$$=\frac{\bar{Y}}{\bar{X}}-\frac{1}{\bar{X}^2}\sum_{h=1}^{L}W_h^2\left(\frac{1-f_h}{n_h}\right)\left[S_{xy_h}-\frac{\bar{Y}}{\bar{X}}S_{x_h}^2\right]$$

luego,

$$E(\hat{\bar{Y}}_{Rc})=\bar{Y}-\frac{1}{\bar{X}}\sum_{h=1}^{L}W_h^2\left(\frac{1-f_h}{n_h}\right)\left[S_{xy_h}-\frac{\bar{Y}}{\bar{X}}S_{x_h}^2\right]$$

por lo tanto,

$$B(\hat{\bar{Y}}_{Rc})=\bar{Y}-E(\hat{\bar{Y}}_{Rc})=\frac{1}{\bar{X}}\sum_{h=1}^{L}W_h^2\left(\frac{1-f_h}{n_h}\right)\left[S_{xy_h}-\frac{\bar{Y}}{\bar{X}}S_{x_h}^2\right]$$

Para el desarrollo de la varianza,

$$V(\hat{\bar{Y}}_{Rc})=V\left(\frac{\hat{\bar{Y}}_{st}}{\hat{\bar{X}}_{st}}\bar{X}\right)=\bar{X}^2V\left(\frac{\hat{\bar{Y}}_{st}}{\hat{\bar{X}}_{st}}\right)$$

y

$$V\left(\frac{\hat{\bar{Y}}_{st}}{\hat{\bar{X}}_{st}}\right)=E\left[\left(\frac{\hat{\bar{Y}}_{st}}{\hat{\bar{X}}_{st}}-E\left(\frac{\hat{\bar{Y}}_{st}}{\hat{\bar{X}}_{st}}\right)\right)^2\right]=E\left[\left(\frac{\hat{\bar{Y}}_{st}}{\hat{\bar{X}}_{st}}\right)^2-2\left(\frac{\hat{\bar{Y}}_{st}}{\hat{\bar{X}}_{st}}\right)E\left(\frac{\hat{\bar{Y}}_{st}}{\hat{\bar{X}}_{st}}\right)+\left(E\left(\frac{\hat{\bar{Y}}_{st}}{\hat{\bar{X}}_{st}}\right)\right)^2\right]=E\left[\left(\frac{\hat{\bar{Y}}_{st}}{\hat{\bar{X}}_{st}}\right)^2\right]-\left[E\left(\frac{\hat{\bar{Y}}_{st}}{\hat{\bar{X}}_{st}}\right)\right]^2$$

(7.35)

Aplicando el procedimiento desarrollado en la sección 7.1.2. (Error Cuadrático Medio de los Estimadores de Razón), a partir de la ecuación 7.15 hasta la 7.16, se tiene que,

$$E\left[\left(\frac{\hat{\bar{Y}}_{st}}{\hat{\bar{X}}_{st}}\right)^2\right]=E\left[\frac{\bar{Y}^2\left(1+\frac{\bar{e}}{\bar{Y}}\right)^2}{\bar{X}^2\left(1+\frac{\bar{e}'}{\bar{X}}\right)^2}\right]=\frac{\bar{Y}^2}{\bar{X}^2}E\left[\frac{\left(1+\frac{\bar{e}}{\bar{Y}}\right)^2}{\left(1+\frac{\bar{e}'}{\bar{X}}\right)^2}\right]=\frac{\bar{Y}^2}{\bar{X}^2}\left[1+\frac{1}{\bar{Y}^2}E(\bar{e}^2)-\frac{4}{\bar{Y}\bar{X}}E(\bar{e}\bar{e}')+\frac{3}{\bar{X}^2}E(\bar{e}'^2)\right]$$

y como $E(\bar{e}\bar{e}') = \sum_{h=1}^{L} W_h^2 \left(\frac{1-f_h}{n_h}\right) S_{xy_h}$; $E(\bar{e}^2) = \sum_{h=1}^{L} W_h^2 \left(\frac{1-f_h}{n_h}\right) S_{y_h}^2$; $E(\bar{e}'^2) = \sum_{h=1}^{L} W_h^2 \left(\frac{1-f_h}{n_h}\right) S_{x_h}^2$

entonces,

$$E\left[\left(\frac{\hat{\bar{Y}}_{st}}{\hat{\bar{X}}_{st}}\right)^2\right] = \frac{\bar{Y}^2}{\bar{X}^2}\left[1 + \frac{1}{\bar{Y}^2}\sum_{h=1}^{L} W_h^2\left(\frac{1-f_h}{n_h}\right)S_{y_h}^2 - \frac{4}{\bar{Y}\bar{X}}\sum_{h=1}^{L} W_h^2\left(\frac{1-f_h}{n_h}\right)S_{xy_h} + \frac{3}{\bar{X}^2}\sum_{h=1}^{L} W_h^2\left(\frac{1-f_h}{n_h}\right)S_{x_h}^2\right]$$

luego,

$$V\left(\hat{\bar{Y}}_{Rc}\right) = \bar{X}^2\left[E\left[\left(\frac{\hat{\bar{Y}}_{st}}{\hat{\bar{X}}_{st}}\right)^2\right] - \left[E\left(\frac{\hat{\bar{Y}}_{st}}{\hat{\bar{X}}_{st}}\right)\right]^2\right]$$

$$\approx \bar{Y}^2 + \sum_{h=1}^{L} W_h^2\left(\frac{1-f_h}{n_h}\right)S_{y_h}^2 - 4\frac{\bar{Y}}{\bar{X}}\sum_{h=1}^{L} W_h^2\left(\frac{1-f_h}{n_h}\right)S_{xy_h} + 3\frac{\bar{Y}^2}{\bar{X}^2}\sum_{h=1}^{L} W_h^2\left(\frac{1-f_h}{n_h}\right)S_{x_h}^2 -$$

$$-\left[\bar{Y}^2 + \frac{1}{\bar{X}^2}\left(\sum_{h=1}^{L} W_h^2\left(\frac{1-f_h}{n_h}\right)\left(S_{xy_h} - \frac{\bar{Y}}{\bar{X}}S_{x_h}^2\right)\right)^2 - 2\frac{\bar{Y}^2}{\bar{X}^2}\sum_{h=1}^{L} W_h^2\left(\frac{1-f_h}{n_h}\right)\left(S_{xy_h} - \frac{\bar{Y}}{\bar{X}}S_{x_h}^2\right)\right]$$

$$\approx \sum_{h=1}^{L} W_h^2\left(\frac{1-f_h}{n_h}\right)S_{y_h}^2 - 2\frac{\bar{Y}}{\bar{X}}\sum_{h=1}^{L} W_h^2\left(\frac{1-f_h}{n_h}\right)S_{xy_h} + \frac{\bar{Y}^2}{\bar{X}^2}\sum_{h=1}^{L} W_h^2\left(\frac{1-f_h}{n_h}\right)S_{x_h}^2 - \frac{1}{\bar{X}^2}\left(\sum_{h=1}^{L} W_h^2\left(\frac{1-f_h}{n_h}\right)\left(S_{xy_h} - \frac{\bar{Y}}{\bar{X}}S_{x_h}^2\right)\right)^2$$

$$\approx \sum_{h=1}^{L} W_h^2\left(\frac{1-f_h}{n_h}\right)\left(S_{y_h}^2 - 2R_c S_{xy_h} + R_c^2 S_{x_h}^2\right) - \frac{1}{\bar{X}^2}\left(\sum_{h=1}^{L} W_h^2\left(\frac{1-f_h}{n_h}\right)\left(S_{xy_h} - \frac{\bar{Y}}{\bar{X}}S_{x_h}^2\right)\right)^2$$

finalmente,

$$ECM\left(\hat{\bar{Y}}_{Rc}\right) = V\left(\hat{\bar{Y}}_{Rc}\right) + \left[B\left(\hat{\bar{Y}}_{Rc}\right)\right]^2 = \sum_{h=1}^{L} W_h^2\left(\frac{1-f_h}{n_h}\right)\left(S_{y_h}^2 - 2R_c S_{xy_h} + R_c^2 S_{x_h}^2\right)$$

Resumiendo se tiene,

$$E\left(\hat{\bar{Y}}_{Rc}\right) = \bar{Y} - \frac{1}{\bar{X}}\sum_{h=1}^{L} W_h^2\left(\frac{1-f_h}{n_h}\right)\left[S_{xy_h} - \frac{\bar{Y}}{\bar{X}}S_{x_h}^2\right] = \bar{Y} + \frac{1}{\bar{X}}\sum_{h=1}^{L} W_h^2\left(\frac{1-f_h}{n_h}\right)\left[R_c S_{x_h}^2 - S_{xy_h} - \frac{\bar{Y}}{\bar{X}}S_{x_h}^2\right]$$

$$B\left(\hat{\bar{Y}}_{Rc}\right) = \frac{1}{\bar{X}}\sum_{h=1}^{L} W_h^2\left(\frac{1-f_h}{n_h}\right)\left[S_{xy_h} - \frac{\bar{Y}}{\bar{X}}S_{x_h}^2\right] = \frac{1}{\bar{X}}\sum_{h=1}^{L} W_h^2\left(\frac{1-f_h}{n_h}\right)\left[S_{xy_h} - R_c S_{x_h}^2\right]$$

$$V(\hat{\bar{Y}}_{R_c}) \approx \sum_{h=1}^{L} W_h^2\left(\frac{1-f_h}{n_h}\right)\left(S_{y_h}^2 - 2R_c S_{xy_h} + R_c^2 S_{x_h}^2\right) - \frac{1}{\bar{X}^2}\left(\sum_{h=1}^{L} W_h^2\left(\frac{1-f_h}{n_h}\right)\left(S_{xy_h} - \frac{\bar{Y}}{\bar{X}}S_{x_h}^2\right)\right)^2$$

$$ECM(\hat{\bar{Y}}_{R_c}) \approx \sum_{h=1}^{L} W_h^2 \left(\frac{1-f_h}{n_h}\right)\left(S_{y_h}^2 - 2R_c S_{xy_h} + R_c^2 S_{x_h}^2\right)$$

$$\hat{ECM}(\hat{\bar{Y}}_{R_c}) \approx \sum_{h=1}^{L} W_h^2 \left(\frac{1-f_h}{n_h}\right)\left(s_{y_h}^2 - 2\hat{R}_c s_{xy_h} + \hat{R}_c^2 s_{x_h}^2\right)$$

y para el total

$$E\left(\hat{Y}_{Rc}\right) = Y - \frac{1}{X}\sum_{h=1}^{L} N_h^2 \left(\frac{1-f_h}{n_h}\right)\left[S_{xy_h} - \frac{\bar{Y}}{\bar{X}} S_{x_h}^2\right] = Y + \frac{1}{X}\sum_{h=1}^{L} N_h^2 \left(\frac{1-f_h}{n_h}\right)\left[R_c S_{x_h}^2 - S_{xy_h} - \frac{\bar{Y}}{\bar{X}} S_{x_h}^2\right]$$

$$B\left(\hat{Y}_{Rc}\right) = \frac{1}{X}\sum_{h=1}^{L} N_h^2 \left(\frac{1-f_h}{n_h}\right)\left[S_{xy_h} - \frac{\bar{Y}}{\bar{X}} S_{x_h}^2\right] = \frac{1}{X}\sum_{h=1}^{L} N_h^2 \left(\frac{1-f_h}{n_h}\right)\left[S_{xy_h} - R_c S_{x_h}^2\right]$$

$$V(\hat{Y}_{R_c}) \approx \sum_{h=1}^{L} N_h^2 \left(\frac{1-f_h}{n_h}\right)\left(S_{y_h}^2 - 2R_c S_{xy_h} + R_c^2 S_{x_h}^2\right) - \frac{1}{X^2}\left(\sum_{h=1}^{L} N_h^2 \left(\frac{1-f_h}{n_h}\right)\left(S_{xy_h} - \frac{\bar{Y}}{\bar{X}} S_{x_h}^2\right)\right)^2$$

$$ECM(\hat{Y}_{R_c}) \approx \sum_{h=1}^{L} N_h^2 \left(\frac{1-f_h}{n_h}\right)\left(S_{y_h}^2 - 2R_c S_{xy_h} + R_c^2 S_{x_h}^2\right)$$

$$\hat{ECM}(\hat{Y}_{R_c}) \approx \sum_{h=1}^{L} N_h^2 \left(\frac{1-f_h}{n_h}\right)\left(s_{y_h}^2 - 2\hat{R}_c s_{xy_h} + \hat{R}_c^2 s_{x_h}^2\right)$$

El estimador de la Razón sería,

$$\hat{R}_c = \frac{\bar{y}_{st}}{\bar{x}_{st}} = \frac{\hat{Y}_{Rc}}{X}$$

y sus respectivas esperanza, sesgo, varianza y error cuadrático medio son,

$$E\left(\hat{R}_c\right) = R - \frac{1}{X^2}\sum_{h=1}^{L} N_h^2 \left(\frac{1-f_h}{n_h}\right)\left[S_{xy_h} - \frac{\bar{Y}}{\bar{X}} S_{x_h}^2\right] = R + \frac{1}{X^2}\sum_{h=1}^{L} N_h^2 \left(\frac{1-f_h}{n_h}\right)\left[R_c S_{x_h}^2 - S_{xy_h} - \frac{\bar{Y}}{\bar{X}} S_{x_h}^2\right]$$

$$B\left(\hat{R}_c\right) = \frac{1}{X^2}\sum_{h=1}^{L} N_h^2 \left(\frac{1-f_h}{n_h}\right)\left[S_{xy_h} - \frac{\bar{Y}}{\bar{X}} S_{x_h}^2\right] = \frac{1}{X^2}\sum_{h=1}^{L} N_h^2 \left(\frac{1-f_h}{n_h}\right)\left[S_{xy_h} - R_c S_{x_h}^2\right]$$

$$V\left(\hat{R}_c\right) \approx \sum_{h=1}^{L} N_h^2 \left(\frac{1-f_h}{n_h}\right)\left(S_{y_h}^2 - 2R_c S_{xy_h} + R_c^2 S_{x_h}^2\right) - \frac{1}{X^4}\left(\sum_{h=1}^{L} N_h^2 \left(\frac{1-f_h}{n_h}\right)\left(S_{xy_h} - \frac{\bar{Y}}{\bar{X}} S_{x_h}^2\right)\right)^2$$

$$ECM\left(\hat{R}_c\right) \approx \frac{1}{X^2}\sum_{h=1}^{L} N_h^2 \left(\frac{1-f_h}{n_h}\right)\left(S_{y_h}^2 - 2R_c S_{xy_h} + R_c^2 S_{x_h}^2\right)$$

$$\hat{ECM}\left(\hat{R}_c\right) \approx \frac{1}{X^2}\sum_{h=1}^{L} N_h^2 \left(\frac{1-f_h}{n_h}\right)\left(s_{y_h}^2 - 2\hat{R}_c s_{xy_h} + \hat{R}_c^2 s_{x_h}^2\right)$$

Ahora se desarrollarán los estimadores del tipo razón combinado estratificado, donde la razón combinada se estima como el promedio ponderado de las razones de cada estrato.

$$\hat{R}_{cst} = \sum_{h=1}^{L} W_h \hat{R}_h = \sum_{h=1}^{L} W_h \left(\frac{\bar{y}_h}{\bar{x}_h} \right) \quad \Rightarrow \quad \hat{Y}_{Rcst} = \hat{R}_{cst} X \quad , \quad \hat{\bar{Y}}_{Rcst} = \hat{R}_{cst} \bar{X} = \bar{X} \sum_{h=1}^{L} W_h \left(\frac{\bar{y}_h}{\bar{x}_h} \right)$$

Se desarrollará primero el del promedio, y luego el del total y la razón.

$$E\left[\hat{\bar{Y}}_{Rcst} \right] = E\left[\bar{X} \sum_{h=1}^{L} W_h \left(\frac{\bar{y}_h}{\bar{x}_h} \right) \right] = \bar{X} E\left[\sum_{h=1}^{L} W_h \left(\frac{\bar{y}_h}{\bar{x}_h} \right) \right] = \bar{X} \sum_{h=1}^{L} E\left[W_h \left(\frac{\bar{y}_h}{\bar{x}_h} \right) \right] = \bar{X} \sum_{h=1}^{L} E\left[W_h \hat{R}_h \right]$$

aplicando las mismas transformaciones y procedimientos de la sección 7.1.1 (Esperanza de los Estimadores de Razón), se tiene que,

$$\hat{R}_h = \frac{\bar{Y}_h \left(1 + \frac{\bar{e}_h}{\bar{y}_h} \right)}{\bar{X}_h \left(1 + \frac{\bar{e}'_h}{\bar{x}_h} \right)}$$

entonces, $E\left[W_h \hat{R}_h \right] = E\left[\frac{\bar{Y}_h \left(1 + \frac{\bar{e}_h}{\bar{y}_h} \right)}{\bar{X}_h \left(1 + \frac{\bar{e}'_h}{\bar{x}_h} \right)} \right] = \frac{N_h}{N} \frac{\bar{Y}_h}{\bar{X}_h} E\left[\frac{\left(1 + \frac{\bar{e}_h}{\bar{y}_h} \right)}{\left(1 + \frac{\bar{e}'_h}{\bar{x}_h} \right)} \right] = \frac{N_h}{N} \frac{\bar{Y}_h}{\bar{X}_h} \left[1 + \frac{N_h - n_h}{N_h n_h} \left(\frac{S^2_{x_h}}{\bar{X}_h^2} - \frac{S_{xy_h}}{\bar{X}_h \bar{Y}_h} \right) \right]$

por lo tanto,

$E\left[\hat{\bar{Y}}_{Rcst} \right] = \bar{X} \sum_{h=1}^{L} W_h R_h \left[1 + \frac{N_h - n_h}{N_h n_h} \left(\frac{S^2_{x_h}}{\bar{X}_h^2} - \frac{S_{xy_h}}{\bar{X}_h \bar{Y}_h} \right) \right]$, que es un estimador sesgado, y su sesgo es,

$$B\left[\hat{\bar{Y}}_{Rcst} \right] = \bar{Y} - \bar{X} \sum_{h=1}^{L} W_h R_h \left[1 + \frac{N_h - n_h}{N_h n_h} \left(\frac{S^2_{x_h}}{\bar{X}_h^2} - \frac{S_{xy_h}}{\bar{X}_h \bar{Y}_h} \right) \right]$$

Ahora se desarrollará la varianza,

$$V\left[\hat{\bar{Y}}_{Rcst} \right] = E\left[\left(\hat{\bar{Y}}_{Rcst} - E\left(\hat{\bar{Y}}_{Rcst} \right) \right)^2 \right] = E\left[\left(\hat{\bar{Y}}_{Rcst} \right)^2 \right] - \left[E\left(\hat{\bar{Y}}_{Rcst} \right) \right]^2$$

se desarrollará el primer término,

$$E\left[\left(\hat{\bar{Y}}_{Rcst} \right)^2 \right] = E\left[\left(\bar{X} \sum_{h=1}^{L} W_h \left(\frac{\bar{y}_h}{\bar{x}_h} \right) \right)^2 \right] = \bar{X}^2 E\left[\left(\sum_{h=1}^{L} W_h \left(\frac{\bar{y}_h}{\bar{x}_h} \right) \right)^2 \right] = \bar{X}^2 E\left[\sum_{h=1}^{L} W_h^2 \left(\frac{\bar{y}_h}{\bar{x}_h} \right)^2 + 2 \sum_{h=1}^{L-1} \sum_{k=h+1}^{L} W_h W_k \left(\frac{\bar{y}_h}{\bar{x}_h} \right) \left(\frac{\bar{y}_k}{\bar{x}_k} \right) \right]$$

$$= \bar{X}^2 \left[\sum_{h=1}^{L} W_h^2 E\left[\left(\frac{\bar{y}_h}{\bar{x}_h} \right)^2 \right] + 2 \sum_{h=1}^{L-1} \sum_{k=h+1}^{L} W_h W_k E\left[\left(\frac{\bar{y}_h}{\bar{x}_h} \right) \left(\frac{\bar{y}_k}{\bar{x}_k} \right) \right] \right]$$

desarrollando las esperanzas, y aplicando el mismo procedimiento de 7.1.2 (Error Cuadrático Medio de los Estimadores de Razón), por el cual se llegó a (7.16), ahora aplicado al estrato *h*,

$$E\left[\left(\frac{\bar{y}_h}{\bar{x}_h}\right)^2\right]=\left(\frac{\bar{Y}_h}{\bar{X}_h}\right)^2\left[1+\frac{N_h-n_h}{N_h n_h}\frac{S_{y_h}^2}{\bar{Y}_h^2}-4\frac{N_h-n_h}{N_h n_h}\frac{S_{xy_h}}{\bar{X}_h\bar{Y}_h}+3\frac{N_h-n_h}{N_h n_h}\frac{S_{x_h}^2}{\bar{X}_h^2}\right]$$

y

$$E\left[\left(\frac{y_h}{x_h}\right)\left(\frac{y_k}{x_k}\right)\right]=E\left[\frac{\bar{Y}_h\left(1+\frac{\bar{e}_h}{\bar{Y}_h}\right)\bar{Y}_k\left(1+\frac{\bar{e}_k}{\bar{Y}_k}\right)}{\bar{X}_h\left(1+\frac{\bar{e}'_h}{\bar{X}_h}\right)\bar{X}_k\left(1+\frac{\bar{e}'_k}{\bar{X}_k}\right)}\right]=\left(\frac{\bar{Y}_h}{\bar{X}_h}\right)\left(\frac{\bar{Y}_k}{\bar{X}_k}\right)E\left[\frac{\left(1+\frac{\bar{e}_h}{\bar{Y}_h}\right)\left(1+\frac{\bar{e}_k}{\bar{Y}_k}\right)}{\left(1+\frac{\bar{e}'_h}{\bar{X}_h}\right)\left(1+\frac{\bar{e}'_k}{\bar{X}_k}\right)}\right]$$

desarrollando las Series de Taylor en el denominador, se tiene que,

$$E\left[\frac{\left(1+\frac{\bar{e}_h}{\bar{Y}_h}\right)\left(1+\frac{\bar{e}_k}{\bar{Y}_k}\right)}{\left(1+\frac{\bar{e}'_h}{\bar{X}_h}\right)\left(1+\frac{\bar{e}'_k}{\bar{X}_k}\right)}\right]=E\left[\left(1+\frac{\bar{e}_h}{\bar{Y}_h}\right)\left(1+\frac{\bar{e}_k}{\bar{Y}_k}\right)\left(\sum_{j=0}^{\infty}(-1)^j\frac{\bar{e}'^j_h}{\bar{X}^j_h}\right)\left(\sum_{j=0}^{\infty}(-1)^j\frac{\bar{e}'^j_k}{\bar{X}^j_k}\right)\right]$$

$$=E\left[\left(1+\frac{\bar{e}_h}{\bar{Y}_h}\right)\left(1+\frac{\bar{e}_k}{\bar{Y}_k}\right)\left(1-\frac{\bar{e}'_h}{\bar{X}_h}+\frac{\bar{e}'^2_h}{\bar{X}^2_h}-\frac{\bar{e}'^3_h}{\bar{X}^3_h}+\cdots+(-1)^j\frac{\bar{e}'^j_h}{\bar{X}^j_h}+\cdots\right)\left(1-\frac{\bar{e}'_k}{\bar{X}_k}+\frac{\bar{e}'^2_k}{\bar{X}^2_k}-\frac{\bar{e}'^3_k}{\bar{X}^3_k}+\cdots+(-1)^j\frac{\bar{e}'^j_k}{\bar{X}^j_k}+\cdots\right)\right]$$

truncando las series en potencias de 2º. grado,

$$\approx E\left[\left(1+\frac{\bar{e}_h}{\bar{Y}_h}+\frac{\bar{e}_k}{\bar{Y}_k}+\frac{\bar{e}_h}{\bar{Y}_h}\frac{\bar{e}_k}{\bar{Y}_k}\right)\left(1-\frac{\bar{e}'_h}{\bar{X}_h}+\frac{\bar{e}'^2_h}{\bar{X}^2_h}\right)\left(1-\frac{\bar{e}'_k}{\bar{X}_k}+\frac{\bar{e}'^2_k}{\bar{X}^2_k}\right)\right]$$

$$\begin{aligned}\approx E\Bigg[&1+\frac{\bar{e}_h}{\bar{Y}_h}+\frac{\bar{e}_k}{\bar{Y}_k}+\frac{\bar{e}_h}{\bar{Y}_h}\frac{\bar{e}_k}{\bar{Y}_k}-\frac{\bar{e}'_h}{\bar{X}_h}-\frac{\bar{e}_h\bar{e}'_h}{\bar{Y}_h\bar{X}_h}-\frac{\bar{e}_k\bar{e}'_h}{\bar{Y}_k\bar{X}_h}-\frac{\bar{e}_h\bar{e}_k\bar{e}'_h}{\bar{Y}_h\bar{Y}_k\bar{X}_h}+\frac{\bar{e}'^2_h}{\bar{X}^2_h}+\frac{\bar{e}_h\bar{e}'^2_h}{\bar{Y}_h\bar{X}^2_h}+\frac{\bar{e}_k\bar{e}'^2_h}{\bar{Y}_k\bar{X}^2_h}+\frac{\bar{e}_h\bar{e}_k\bar{e}'^2_h}{\bar{Y}_h\bar{Y}_k\bar{X}^2_h}-\\&-\frac{\bar{e}'_k}{\bar{X}_k}-\frac{\bar{e}_h\bar{e}'_k}{\bar{Y}_h\bar{X}_k}-\frac{\bar{e}_k\bar{e}'_k}{\bar{Y}_k\bar{X}_k}-\frac{\bar{e}_h\bar{e}_k\bar{e}'_k}{\bar{Y}_h\bar{Y}_k\bar{X}_k}+\frac{\bar{e}'_h\bar{e}'_k}{\bar{X}_h\bar{X}_k}+\frac{\bar{e}_h\bar{e}'_h\bar{e}'_k}{\bar{Y}_h\bar{X}_h\bar{X}_k}+\frac{\bar{e}_k\bar{e}'_h\bar{e}'_k}{\bar{Y}_k\bar{X}_h\bar{X}_k}+\frac{\bar{e}_h\bar{e}_k\bar{e}'_h\bar{e}'_k}{\bar{Y}_h\bar{Y}_k\bar{X}_h\bar{X}_k}-\\&-\frac{\bar{e}'^2_h\bar{e}'_k}{\bar{X}^2_h\bar{X}_k}-\frac{\bar{e}_h\bar{e}'^2_h\bar{e}'_k}{\bar{Y}_h\bar{X}^2_h\bar{X}_k}-\frac{\bar{e}_h\bar{e}_k\bar{e}'^2_h\bar{e}'_k}{\bar{Y}_h\bar{Y}_k\bar{X}^2_h\bar{X}_k}-\frac{\bar{e}_k\bar{e}'^2_h\bar{e}'_k}{\bar{Y}_k\bar{X}^2_h\bar{X}_k}+\frac{\bar{e}'^2_k}{\bar{X}^2_k}+\frac{\bar{e}_h\bar{e}'^2_k}{\bar{Y}_h\bar{X}^2_k}+\frac{\bar{e}_k\bar{e}'^2_k}{\bar{Y}_k\bar{X}^2_k}+\frac{\bar{e}_h\bar{e}_k\bar{e}'^2_k}{\bar{Y}_h\bar{Y}_k\bar{X}^2_k}-\\&-\frac{\bar{e}'_h\bar{e}'^2_k}{\bar{X}_h\bar{X}^2_k}-\frac{\bar{e}_h\bar{e}'_h\bar{e}'^2_k}{\bar{Y}_h\bar{X}_h\bar{X}^2_k}-\frac{\bar{e}_k\bar{e}'_h\bar{e}'^2_k}{\bar{Y}_k\bar{X}_h\bar{X}^2_k}-\frac{\bar{e}_h\bar{e}_k\bar{e}'_h\bar{e}'^2_k}{\bar{Y}_h\bar{Y}_k\bar{X}_h\bar{X}^2_k}+\frac{\bar{e}'^2_h\bar{e}'^2_k}{\bar{X}^2_h\bar{X}^2_k}+\frac{\bar{e}_h\bar{e}'^2_h\bar{e}'^2_k}{\bar{Y}_h\bar{X}^2_h\bar{X}^2_k}+\frac{\bar{e}_k\bar{e}'^2_h\bar{e}'^2_k}{\bar{Y}_k\bar{X}^2_h\bar{X}^2_k}+\frac{\bar{e}_h\bar{e}_k\bar{e}'^2_h\bar{e}'^2_k}{\bar{Y}_h\bar{Y}_k\bar{X}^2_h\bar{X}^2_k}\Bigg]\end{aligned}$$

como se ha mencionado anteriormente, $E(\bar{e}_h)=E(\bar{e}_k)=E(\bar{e}'_h)=E(\bar{e}'_k)=0$. Se revisarán los otros miembros,

$$E(\bar{e}_h\bar{e}_k)=E\left[\left(\frac{\sum_{i=1}^{n_h}e_{hi}}{n_h}\right)\left(\frac{\sum_{i=1}^{n_k}e_{ki}}{n_k}\right)\right]=\frac{1}{n_hn_k}E\left[\left(\sum_{i=1}^{n_h}e_{hi}\right)\left(\sum_{i=1}^{n_k}e_{ki}\right)\right]=\frac{1}{n_hn_k\binom{N_h}{n_h}\binom{N_k}{n_k}}\binom{N_h-1}{n_h-1}\binom{N_k-1}{n_k-1}\sum_{i=1}^{N_h}\sum_{i=1}^{N_k}e_{hi}e_{ki}$$

$$=\frac{1}{N_hN_k}\sum_{i=1}^{N_h}\sum_{i=1}^{N_k}e_{hi}e_{ki}=\frac{\sum_{i=1}^{N_h}e_{hi}}{N_h}\frac{\sum_{i=1}^{N_k}e_{ki}}{N_k}=\bar{E}_h\,\bar{E}_k=0$$

igualmente, $E(\bar{e}_k\bar{e}'_h)=\dfrac{\sum_{i=1}^{N_k}e_{ki}}{N_k}\dfrac{\sum_{i=1}^{N_h}e'_{hi}}{N_h}=\bar{E}_k\,\bar{E}'_h=0$, y análogamente, $E(\bar{e}_h\bar{e}'_k)=E(\bar{e}'_h\,\bar{e}'_k)=0$.

Esto ocurre porque los estratos son independientes. Así,

$$E(\bar{e}_h\bar{e}_k\bar{e}'_h)=E(\bar{e}_h\bar{e}'_h)E(\bar{e}_k)=\left(\frac{1-f_h}{n_h}S_{xy_h}\right)0=0 \quad , \text{ y } E(\bar{e}_h\bar{e}_k\bar{e}'_k)=E(\bar{e}_h\bar{e}'_h\,\bar{e}'_k)=E(\bar{e}_k\bar{e}'_h\,\bar{e}'_k)=0.$$

$$E(\bar{e}_k\bar{e}'^2_h)=\left(\frac{1-f_h}{n_h}S^2_{x_h}\right)0=0 \quad , \text{ y } E(\bar{e}'^2_h\,\bar{e}'_k)=E(\bar{e}_h\bar{e}'^2_k)=E(\bar{e}'_h\,\bar{e}'^2_k)=0.$$

$$E(\bar{e}_h\bar{e}_k\bar{e}'^2_h)=E(\bar{e}_h\bar{e}'^2_h)E(\bar{e}_k)=0 \quad , \text{ y } E(\bar{e}_h\bar{e}'^2_h\,\bar{e}_k)=E(\bar{e}_h\bar{e}_k\bar{e}'^2_k)=E(\bar{e}_k\bar{e}'_h\,\bar{e}'^2_k)=0.$$

$$E(\bar{e}'^2_h\,\bar{e}'^2_k)=E(\bar{e}'^2_h)E(\bar{e}'^2_k)=\left(\frac{1-f_h}{n_h}S^2_{x_h}\right)\left(\frac{1-f_k}{n_k}S^2_{x_k}\right)$$

$$E(\bar{e}_k\bar{e}'^2_h\,\bar{e}'_k)=E(\bar{e}_k\bar{e}'_k)E(\bar{e}'^2_h)=\left(\frac{1-f_h}{n_h}S^2_{x_h}\right)\left(\frac{1-f_k}{n_k}S_{xy_k}\right)$$

$$E(\bar{e}_h\bar{e}'_h\,\bar{e}'^2_k)=E(\bar{e}_h\bar{e}'_h)E(\bar{e}'^2_k)=\left(\frac{1-f_h}{n_h}S_{xy_h}\right)\left(\frac{1-f_k}{n_k}S^2_{x_k}\right)$$

$$E(\bar{e}_h\bar{e}_k\bar{e}'_h\,\bar{e}'_k)=E(\bar{e}_h\bar{e}'_h)E(\bar{e}_k\bar{e}'_k)=\left(\frac{1-f_h}{n_h}S_{xy_h}\right)\left(\frac{1-f_k}{n_k}S_{xy_k}\right)$$

Los términos

$$\frac{\bar{e}_h\bar{e}'^2_h}{\bar{Y}_h\bar{X}^2_h},\frac{\bar{e}_k\bar{e}'^2_k}{\bar{Y}_k\bar{X}^2_k},\frac{\bar{e}_h\bar{e}_k\bar{e}'^2_k}{\bar{Y}_h\bar{Y}_k\bar{X}^2_k},\frac{\bar{e}_h\bar{e}_k\bar{e}'^2_h}{\bar{Y}_h\bar{Y}_k\bar{X}^2_h},\frac{\bar{e}_k\bar{e}'_h\,\bar{e}'^2_k}{\bar{Y}_k\bar{X}_h\bar{X}^2_k},\frac{\bar{e}_h\bar{e}_k\bar{e}'^2_k}{\bar{Y}_h\bar{Y}_k\bar{X}^2_k},\frac{\bar{e}_h\bar{e}_k\bar{e}'^2_h\,\bar{e}'_k}{\bar{Y}_h\bar{Y}_k\bar{X}^2_h\bar{X}_k},\frac{\bar{e}_h\bar{e}_k\bar{e}'_h\,\bar{e}'^2_k}{\bar{Y}_h\bar{Y}_k\bar{X}_h\bar{X}^2_k},\frac{\bar{e}_h\bar{e}'^2_h\,\bar{e}'^2_k}{\bar{Y}_h\bar{X}^2_h\bar{X}^2_k},\frac{\bar{e}_k\bar{e}'^2_h\,\bar{e}'^2_k}{\bar{Y}_k\bar{X}^2_h\bar{X}^2_k},\frac{\bar{e}_h\bar{e}_k\bar{e}'^2_h\,\bar{e}'^2_k}{\bar{Y}_h\bar{Y}_k\bar{X}^2_h\bar{X}^2_k}$$

se despreciarán, suponiendo que aportan poco a la serie, pero además, por facilidad en la consecución de la fórmula final.

Entonces,

$$E\left[\left(\frac{\bar{y}_h}{\bar{x}_h}\right)\left(\frac{\bar{y}_k}{\bar{x}_k}\right)\right] \approx \left(\frac{\bar{Y}_h}{\bar{X}_h}\right)\left(\frac{\bar{Y}_k}{\bar{X}_k}\right)E\left[1-\frac{\bar{e}_h\bar{e}'_h}{\bar{Y}_h\bar{X}_h}+\frac{\bar{e}'^2_h}{\bar{X}_h^2}-\frac{\bar{e}_k\bar{e}'_k}{\bar{Y}_k\bar{X}_k}+\frac{\bar{e}'^2_k}{\bar{X}_k^2}+\frac{\bar{e}_h\bar{e}_k\bar{e}'_h\bar{e}'_k}{\bar{Y}_h\bar{Y}_k\bar{X}_h\bar{X}_k}-\frac{\bar{e}_k\bar{e}'^2_h\bar{e}'_k}{\bar{Y}_k\bar{X}_h^2\bar{X}_k}-\frac{\bar{e}_h\bar{e}'_h\bar{e}'^2_k}{\bar{Y}_h\bar{X}_h\bar{X}_k^2}+\frac{\bar{e}'^2_h\bar{e}'^2_k}{\bar{X}_h^2\bar{X}_k^2}\right]$$

$$\approx \left(\frac{\bar{Y}_h}{\bar{X}_h}\right)\left(\frac{\bar{Y}_k}{\bar{X}_k}\right)\left[1-\left(\frac{1-f_h}{n_h}\right)\frac{S_{xy_h}}{\bar{Y}_h\bar{X}_h}+\left(\frac{1-f_h}{n_h}\right)\frac{S^2_{x_h}}{\bar{X}_h^2}-\left(\frac{1-f_k}{n_k}\right)\frac{S_{xy_k}}{\bar{Y}_k\bar{X}_k}+\left(\frac{1-f_k}{n_k}\right)\frac{S^2_{x_k}}{\bar{X}_k^2}+\right.$$

$$\left.+\left(\frac{1-f_h}{n_h}\right)^2\frac{S_{xy_h}S_{xy_k}}{\bar{Y}_h\bar{X}_h\bar{Y}_k\bar{X}_k}-\left(\frac{1-f_h}{n_h}\right)^2\frac{S^2_{x_h}S_{xy_k}}{\bar{X}_h^2\bar{Y}_k\bar{X}_k}-\left(\frac{1-f_h}{n_h}\right)^2\frac{S_{xy_h}S^2_{x_k}}{\bar{Y}_h\bar{X}_h\bar{X}_k^2}+\left(\frac{1-f_h}{n_h}\right)^2\frac{S^2_{x_h}S^2_{x_k}}{\bar{X}_h^2\bar{X}_k^2}\right]$$

$$\approx \left(\frac{\bar{Y}_h}{\bar{X}_h}\right)\left(\frac{\bar{Y}_k}{\bar{X}_k}\right)\left[\left(1+\left(\frac{1-f_h}{n_h}\right)\left(\frac{S^2_{x_h}}{\bar{X}_h^2}-\frac{S_{xy_h}}{\bar{Y}_h\bar{X}_h}\right)\right)\left(1+\left(\frac{1-f_k}{n_k}\right)\left(\frac{S^2_{x_k}}{\bar{X}_k^2}-\frac{S_{xy_k}}{\bar{Y}_k\bar{X}_k}\right)\right)\right]$$

Por lo tanto,

$$E\left[\left(\hat{\bar{Y}}_{Rcst}\right)^2\right] \approx \bar{X}^2\left[\sum_{h=1}^{L}W_h^2\left(\frac{\bar{Y}_h}{\bar{X}_h}\right)^2\left(1+\frac{N_h-n_h}{N_h\,n_h}\frac{S^2_{y_h}}{\bar{Y}_h^2}-4\frac{N_h-n_h}{N_h\,n_h}\frac{S_{xy_h}}{\bar{X}_h\,\bar{Y}_h}+3\frac{N_h-n_h}{N_h\,n_h}\frac{S^2_{x_h}}{\bar{X}_h^2}\right)+\right.$$

$$\left.+2\sum_{h=1}^{L-1}\sum_{k=h+1}^{L}W_hW_k\left(\frac{\bar{Y}_h}{\bar{X}_h}\right)\left(\frac{\bar{Y}_k}{\bar{X}_k}\right)\left(1+\left(\frac{1-f_h}{n_h}\right)\left(\frac{S^2_{x_h}}{\bar{X}_h^2}-\frac{S_{xy_h}}{\bar{Y}_h\bar{X}_h}\right)\right)\left(1+\left(\frac{1-f_k}{n_k}\right)\left(\frac{S^2_{x_k}}{\bar{X}_k^2}-\frac{S_{xy_k}}{\bar{Y}_k\bar{X}_k}\right)\right)\right]$$

Por tanto, la varianza es,

$$V\left[\hat{\bar{Y}}_{Rcst}\right] \approx \bar{X}^2\left[\sum_{h=1}^{L}W_h^2\left(\frac{\bar{Y}_h}{\bar{X}_h}\right)^2\left(1+\frac{N_h-n_h}{N_h\,n_h}\frac{S^2_{y_h}}{\bar{Y}_h^2}-4\frac{N_h-n_h}{N_h\,n_h}\frac{S_{xy_h}}{\bar{X}_h\,\bar{Y}_h}+3\frac{N_h-n_h}{N_h\,n_h}\frac{S^2_{x_h}}{\bar{X}_h^2}\right)+\right.$$

$$\left.+2\sum_{h=1}^{L-1}\sum_{k=h+1}^{L}W_hW_k\left(\frac{\bar{Y}_h}{\bar{X}_h}\right)\left(\frac{\bar{Y}_k}{\bar{X}_k}\right)\left(1+\left(\frac{1-f_h}{n_h}\right)\left(\frac{S^2_{x_h}}{\bar{X}_h^2}-\frac{S_{xy_h}}{\bar{Y}_h\bar{X}_h}\right)\right)\left(1+\left(\frac{1-f_k}{n_k}\right)\left(\frac{S^2_{x_k}}{\bar{X}_k^2}-\frac{S_{xy_k}}{\bar{Y}_k\bar{X}_k}\right)\right)\right]-$$

$$-\bar{X}^2\left[\sum_{h=1}^{L}W_h\,R_h\left(1+\frac{N_h-n_h}{N_h\,n_h}\left(\frac{S^2_{x_h}}{\bar{X}_h^2}-\frac{S_{xy_h}}{\bar{X}_h\,\bar{Y}_h}\right)\right)\right]^2$$

y el error cuadrático medio,

$$ECM\left[\hat{\bar{Y}}_{Rcst}\right] = V\left[\hat{\bar{Y}}_{Rcst}\right] + \left(B\left[\left(\hat{\bar{Y}}_{Rcst}\right)\right]\right)^2$$

$$\approx \bar{X}^2\left[\sum_{h=1}^{L} W_h^2\left(\frac{\bar{Y}_h}{\bar{X}_h}\right)^2\left(1+\frac{N_h-n_h}{N_h\,n_h}\frac{S_{y_h}^2}{\bar{Y}_h^2}-4\frac{N_h-n_h}{N_h\,n_h}\frac{S_{xy_h}}{\bar{X}_h\,\bar{Y}_h}+3\frac{N_h-n_h}{N_h\,n_h}\frac{S_{x_h}^2}{\bar{X}_h^2}\right)+\right.$$

$$\left.+2\sum_{h=1}^{L-1}\sum_{k=h+1}^{L} W_h W_k\left(\frac{\bar{Y}_h}{\bar{X}_h}\right)\left(\frac{\bar{Y}_k}{\bar{X}_k}\right)\left(1+\left(\frac{1-f_h}{n_h}\right)\left(\frac{S_{x_h}^2}{\bar{X}_h^2}-\frac{S_{xy_h}}{\bar{Y}_h\bar{X}_h}\right)\right)\left(1+\left(\frac{1-f_k}{n_k}\right)\left(\frac{S_{x_k}^2}{\bar{X}_k^2}-\frac{S_{xy_k}}{\bar{Y}_k\bar{X}_k}\right)\right)\right]-$$

$$-\bar{X}^2\left[\sum_{h=1}^{L} W_h\,R_h\left(1+\frac{N_h-n_h}{N_h\,n_h}\left(\frac{S_{x_h}^2}{\bar{X}_h^2}-\frac{S_{xy_h}}{\bar{X}_h\,\bar{Y}_h}\right)\right)\right]^2+\bar{Y}^2+$$

$$+\bar{X}^2\left[\sum_{h=1}^{L} W_h\,R_h\left(1+\frac{N_h-n_h}{N_h\,n_h}\left(\frac{S_{x_h}^2}{\bar{X}_h^2}-\frac{S_{xy_h}}{\bar{X}_h\,\bar{Y}_h}\right)\right)\right]^2-2\bar{X}\bar{Y}\sum_{h=1}^{L} W_h\,R_h\left(1+\frac{N_h-n_h}{N_h\,n_h}\left(\frac{S_{x_h}^2}{\bar{X}_h^2}-\frac{S_{xy_h}}{\bar{X}_h\,\bar{Y}_h}\right)\right)$$

$$\approx \bar{X}^2\left[\sum_{h=1}^{L} W_h^2\left(\frac{\bar{Y}_h}{\bar{X}_h}\right)^2\left(1+\frac{N_h-n_h}{N_h\,n_h}\frac{S_{y_h}^2}{\bar{Y}_h^2}-4\frac{N_h-n_h}{N_h\,n_h}\frac{S_{xy_h}}{\bar{X}_h\,\bar{Y}_h}+3\frac{N_h-n_h}{N_h\,n_h}\frac{S_{x_h}^2}{\bar{X}_h^2}\right)+\right.$$

$$\left.+2\sum_{h=1}^{L-1}\sum_{k=h+1}^{L} W_h W_k\left(\frac{\bar{Y}_h}{\bar{X}_h}\right)\left(\frac{\bar{Y}_k}{\bar{X}_k}\right)\left(1+\left(\frac{1-f_h}{n_h}\right)\left(\frac{S_{x_h}^2}{\bar{X}_h^2}-\frac{S_{xy_h}}{\bar{Y}_h X_h}\right)\right)\left(1+\left(\frac{1-f_k}{n_k}\right)\left(\frac{S_{x_k}^2}{\bar{X}_k^2}-\frac{S_{xy_k}}{\bar{Y}_k X_k}\right)\right)\right]+$$

$$+\bar{Y}^2-2\bar{X}\bar{Y}\sum_{h=!}^{L} W_h\,R_h\left(1+\frac{N_h-n_h}{N_h\,n_h}\left(\frac{S_{x_h}^2}{\bar{X}_h^2}-\frac{S_{xy_h}}{\bar{X}_h\,\bar{Y}_h}\right)\right)$$

Para finalizar, el estimador del Error Cuadrático Medio es,

$$\hat{ECM}\left[\hat{\bar{Y}}_{Rcst}\right] \approx \bar{X}^2\left[\sum_{h=1}^{L} W_h^2\left(\frac{\bar{y}_h}{\bar{x}_h}\right)^2\left(1+\frac{N_h-n_h}{N_h\,n_h}\frac{s_{y_h}^2}{\bar{y}_h^2}-4\frac{N_h-n_h}{N_h\,n_h}\frac{s_{xy_h}}{\bar{x}_h\,\bar{y}_h}+3\frac{N_h-n_h}{N_h\,n_h}\frac{s_{x_h}^2}{\bar{x}_h^2}\right)+\right.$$

$$+2\sum_{h=1}^{L-1}\sum_{k=h+1}^{L} W_h W_k\left(\frac{\bar{y}_h}{\bar{x}_h}\right)\left(\frac{\bar{y}_k}{\bar{x}_k}\right)\left(1+\left(\frac{1-f_h}{n_h}\right)\left(\frac{s_{x_h}^2}{\bar{x}_h^2}-\frac{s_{xy_h}}{\bar{y}_h\bar{x}_h}\right)\right)\left(1+\left(\frac{1-f_k}{n_k}\right)\left(\frac{s_{x_k}^2}{\bar{x}_k^2}-\frac{s_{xy_k}}{\bar{y}_k\bar{x}_k}\right)\right)+$$

$$\left.+\hat{R}_{cst}^2-2\hat{R}_{cst}\sum_{h=!}^{L} W_h\,\hat{R}_h\left(1+\frac{N_h-n_h}{N_h\,n_h}\left(\frac{s_{x_h}^2}{\bar{x}_h^2}-\frac{s_{xy_h}}{\bar{x}_h\,\bar{y}_h}\right)\right)\right]$$

Para el total se tiene,

$$\hat{Y}_{Rcst} = \hat{R}_{cst}\, X = N\, \hat{R}_{cst}\, \overline{X} = N\, \hat{\overline{Y}}_{Rcst}$$

$$E\left[\hat{Y}_{Rcst}\right] = X \sum_{h=1}^{L} W_h\, R_h \left[1 + \frac{N_h - n_h}{N_h\, n_h}\left(\frac{S^2_{x_h}}{\overline{X}^2_h} - \frac{S_{xy_h}}{\overline{X}_h\, \overline{Y}_h}\right)\right]$$

$$B\left[\hat{Y}_{Rcst}\right] = Y - X \sum_{h=1}^{L} W_h\, R_h \left[1 + \frac{N_h - n_h}{N_h\, n_h}\left(\frac{S^2_{x_h}}{\overline{X}^2_h} - \frac{S_{xy_h}}{\overline{X}_h\, \overline{Y}_h}\right)\right]$$

$$V\left[\hat{Y}_{Rcst}\right] \approx X^2 \left[\sum_{h=1}^{L} W_h^2 \left(\frac{\overline{Y}_h}{\overline{X}_h}\right)^2 \left(1 + \frac{N_h - n_h}{N_h\, n_h}\frac{S^2_{y_h}}{\overline{Y}^2_h} - 4\frac{N_h - n_h}{N_h\, n_h}\frac{S_{xy_h}}{\overline{X}_h\, \overline{Y}_h} + 3\frac{N_h - n_h}{N_h\, n_h}\frac{S^2_{x_h}}{\overline{X}^2_h}\right) + \right.$$

$$\left. + 2\sum_{h=1}^{L-1}\sum_{k=h+1}^{L} W_h W_k \left(\frac{\overline{Y}_h}{\overline{X}_h}\right)\left(\frac{\overline{Y}_k}{\overline{X}_k}\right)\left(1 + \left(\frac{1 - f_h}{n_h}\right)\left(\frac{S^2_{x_h}}{\overline{X}^2_h} - \frac{S_{xy_h}}{\overline{Y}_h \overline{X}_h}\right)\right)\left(1 + \left(\frac{1 - f_k}{n_k}\right)\left(\frac{S^2_{x_k}}{\overline{X}^2_k} - \frac{S_{xy_k}}{\overline{Y}_k \overline{X}_k}\right)\right)\right] -$$

$$- X^2 \left[\sum_{h=1}^{L} W_h\, R_h \left(1 + \frac{N_h - n_h}{N_h\, n_h}\left(\frac{S^2_{x_h}}{\overline{X}^2_h} - \frac{S_{xy_h}}{\overline{X}_h\, \overline{Y}_h}\right)\right)\right]^2$$

$$ECM\left[\hat{\overline{Y}}_{Rcst}\right] \approx X^2 \left[\sum_{h=1}^{L} W_h^2 \left(\frac{\overline{Y}_h}{\overline{X}_h}\right)^2 \left(1 + \frac{N_h - n_h}{N_h\, n_h}\frac{S^2_{y_h}}{\overline{Y}^2_h} - 4\frac{N_h - n_h}{N_h\, n_h}\frac{S_{xy_h}}{\overline{X}_h\, \overline{Y}_h} + 3\frac{N_h - n_h}{N_h\, n_h}\frac{S^2_{x_h}}{\overline{X}^2_h}\right) + \right.$$

$$\left. + 2\sum_{h=1}^{L-1}\sum_{k=h+1}^{L} W_h W_k \left(\frac{\overline{Y}_h}{\overline{X}_h}\right)\left(\frac{\overline{Y}_k}{\overline{X}_k}\right)\left(1 + \left(\frac{1 - f_h}{n_h}\right)\left(\frac{S^2_{x_h}}{\overline{X}^2_h} - \frac{S_{xy_h}}{\overline{Y}_h \overline{X}_h}\right)\right)\left(1 + \left(\frac{1 - f_k}{n_k}\right)\left(\frac{S^2_{x_k}}{\overline{X}^2_k} - \frac{S_{xy_k}}{\overline{Y}_k \overline{X}_k}\right)\right)\right] +$$

$$+ Y^2 - 2XY \sum_{h=!}^{L} W_h\, R_h \left(1 + \frac{N_h - n_h}{N_h\, n_h}\left(\frac{S^2_{x_h}}{\overline{X}^2_h} - \frac{S_{xy_h}}{\overline{X}_h\, \overline{Y}_h}\right)\right)$$

$$\hat{ECM}\left[\hat{Y}_{Rcst}\right] \approx X^2 \left[\sum_{h=1}^{L} W_h^2 \left(\frac{\overline{y}_h}{\overline{x}_h}\right)^2 \left(1 + \frac{N_h - n_h}{N_h\, n_h}\frac{s^2_{y_h}}{\overline{y}^2_h} - 4\frac{N_h - n_h}{N_h\, n_h}\frac{s_{xy_h}}{\overline{x}_h\, \overline{y}_h} + 3\frac{N_h - n_h}{N_h\, n_h}\frac{s^2_{x_h}}{\overline{x}^2_h}\right) + \right.$$

$$\left. + 2\sum_{h=1}^{L-1}\sum_{k=h+1}^{L} W_h W_k \left(\frac{\overline{y}_h}{\overline{x}_h}\right)\left(\frac{\overline{y}_k}{\overline{x}_k}\right)\left(1 + \left(\frac{1 - f_h}{n_h}\right)\left(\frac{s^2_{x_h}}{\overline{x}^2_h} - \frac{s_{xy_h}}{\overline{y}_h \overline{x}_h}\right)\right)\left(1 + \left(\frac{1 - f_k}{n_k}\right)\left(\frac{s^2_{x_k}}{\overline{x}^2_k} - \frac{s_{xy_k}}{\overline{y}_k \overline{x}_k}\right)\right) + \right.$$

$$\left. + \hat{R}^2_{cst} - 2\hat{R}_{cst} \sum_{h=!}^{L} W_h\, \hat{R}_h \left(1 + \frac{N_h - n_h}{N_h\, n_h}\left(\frac{s^2_{x_h}}{\overline{x}^2_h} - \frac{s_{xy_h}}{\overline{x}_h\, \overline{y}_h}\right)\right)\right]$$

Finalmente, para la razón,

$$\hat{R}_{cst} = \frac{\hat{\bar{Y}}_{Rcst}}{\bar{X}}$$

$$E\left[\hat{R}_{cst}\right] = \sum_{h=1}^{L} W_h \, R_h \left[1 + \frac{N_h - n_h}{N_h \, n_h}\left(\frac{S_{x_h}^2}{\bar{X}_h^2} - \frac{S_{xy_h}}{\bar{X}_h \, \bar{Y}_h}\right)\right]$$

$$B\left[\hat{R}_{cst}\right] = R - \sum_{h=1}^{L} W_h \, R_h \left[1 + \frac{N_h - n_h}{N_h \, n_h}\left(\frac{S_{x_h}^2}{\bar{X}_h^2} - \frac{S_{xy_h}}{\bar{X}_h \, \bar{Y}_h}\right)\right]$$

$$V\left[\hat{R}_{cst}\right] \approx \left[\sum_{h=1}^{L} W_h^2 \left(\frac{\bar{Y}_h}{\bar{X}_h}\right)^2 \left(1 + \frac{N_h - n_h}{N_h \, n_h}\frac{S_{y_h}^2}{\bar{Y}_h^2} - 4\frac{N_h - n_h}{N_h \, n_h}\frac{S_{xy_h}}{\bar{X}_h \, \bar{Y}_h} + 3\frac{N_h - n_h}{N_h \, n_h}\frac{S_{x_h}^2}{\bar{X}_h^2}\right) + \right.$$

$$+ 2\sum_{h=1}^{L-1}\sum_{k=h+1}^{L} W_h W_k \left(\frac{\bar{Y}_h}{\bar{X}_h}\right)\left(\frac{\bar{Y}_k}{\bar{X}_k}\right)\left(1 + \left(\frac{1 - f_h}{n_h}\right)\left(\frac{S_{x_h}^2}{\bar{X}_h^2} - \frac{S_{xy_h}}{\bar{Y}_h \bar{X}_h}\right)\right)\left(1 + \left(\frac{1 - f_k}{n_k}\right)\left(\frac{S_{x_k}^2}{\bar{X}_k^2} - \frac{S_{xy_k}}{\bar{Y}_k \bar{X}_k}\right)\right)\Bigg] -$$

$$\left. - \left[\sum_{h=1}^{L} W_h \, R_h \left(1 + \frac{N_h - n_h}{N_h \, n_h}\left(\frac{S_{x_h}^2}{\bar{X}_h^2} - \frac{S_{xy_h}}{\bar{X}_h \, \bar{Y}_h}\right)\right)\right]^2$$

$$ECM\left[\hat{R}_{cst}\right] \approx \left[\sum_{h=1}^{L} W_h^2 \left(\frac{\bar{Y}_h}{\bar{X}_h}\right)^2 \left(1 + \frac{N_h - n_h}{N_h \, n_h}\frac{S_{y_h}^2}{\bar{Y}_h^2} - 4\frac{N_h - n_h}{N_h \, n_h}\frac{S_{xy_h}}{\bar{X}_h \, \bar{Y}_h} + 3\frac{N_h - n_h}{N_h \, n_h}\frac{S_{x_h}^2}{\bar{X}_h^2}\right) + \right.$$

$$+ 2\sum_{h=1}^{L-1}\sum_{k=h+1}^{L} W_h W_k \left(\frac{\bar{Y}_h}{\bar{X}_h}\right)\left(\frac{\bar{Y}_k}{\bar{X}_k}\right)\left(1 + \left(\frac{1 - f_h}{n_h}\right)\left(\frac{S_{x_h}^2}{\bar{X}_h^2} - \frac{S_{xy_h}}{\bar{Y}_h \bar{X}_h}\right)\right)\left(1 + \left(\frac{1 - f_k}{n_k}\right)\left(\frac{S_{x_k}^2}{\bar{X}_k^2} - \frac{S_{xy_k}}{\bar{Y}_k \bar{X}_k}\right)\right)\Bigg] +$$

$$\left. + R^2 - 2R\sum_{h=!}^{L} W_h \, R_h \left(1 + \frac{N_h - n_h}{N_h \, n_h}\left(\frac{S_{x_h}^2}{\bar{X}_h^2} - \frac{S_{xy_h}}{\bar{X}_h \, \bar{Y}_h}\right)\right)\right.$$

$$E\hat{C}M\left[\hat{R}_{cst}\right] \approx \left[\sum_{h=1}^{L} W_h^2 \left(\frac{\bar{y}_h}{\bar{x}_h}\right)^2 \left(1 + \frac{N_h - n_h}{N_h \, n_h}\frac{s_{y_h}^2}{\bar{y}_h^2} - 4\frac{N_h - n_h}{N_h \, n_h}\frac{s_{xy_h}}{\bar{x}_h \, \bar{y}_h} + 3\frac{N_h - n_h}{N_h \, n_h}\frac{s_{x_h}^2}{\bar{x}_h^2}\right) + \right.$$

$$+ 2\sum_{h=1}^{L-1}\sum_{k=h+1}^{L} W_h W_k \left(\frac{\bar{y}_h}{\bar{x}_h}\right)\left(\frac{\bar{y}_k}{\bar{x}_k}\right)\left(1 + \left(\frac{1 - f_h}{n_h}\right)\left(\frac{s_{x_h}^2}{\bar{x}_h^2} - \frac{s_{xy_h}}{\bar{y}_h \bar{x}_h}\right)\right)\left(1 + \left(\frac{1 - f_k}{n_k}\right)\left(\frac{s_{x_k}^2}{\bar{x}_k^2} - \frac{s_{xy_k}}{\bar{y}_k \bar{x}_k}\right)\right) +$$

$$\left. + \hat{R}_{cst}^2 - 2\hat{R}_{cst}\sum_{h=!}^{L} W_h \, \hat{R}_h \left(1 + \frac{N_h - n_h}{N_h \, n_h}\left(\frac{s_{x_h}^2}{\bar{x}_h^2} - \frac{s_{xy_h}}{\bar{x}_h \, \bar{y}_h}\right)\right)\right]$$

7.2.- Estimadores de regresión lineal

Sea una población de *N* elementos, y sean $y_1, y_{2,\dots}, y_N$ y $x_1, x_{2,\dots}, x_N$ las observaciones de las variables *y*, *x* sobre cada uno de ellos.

Se puede expresar a la variable *"y"* como una función lineal de *"x"*, de la siguiente manera,

$$y_i = a + bx_i + e_i \tag{7.36}$$

donde *"a"* es el punto de corte de la recta en el eje de la variable *"y"*, *"b"* es la pendiente de la recta, y *"e_i"* es el desvío que tiene *"y_i"* respecto de la recta, además,

$$e_i \in \Re \quad \text{y} \quad \sum_{i=1}^{N} e_i = 0$$

Sea $y'_i = a + bx_i$ (7.37)

un estimador de y_i , y se cumple que $\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} y'_i$

ya que,

$$\sum_{i=}^{N} y_i = \sum_{i=1}^{N} (a + bx_i + e_i) = Na + b\sum_{i=1}^{N} x_i + \sum_{i=1}^{N} e_i = Na + b\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} (a + bx_i) = \sum_{i=1}^{N} y'_i$$

Si se aplica a (7.36) sumatoria hasta *N* y luego se divide entre *N*, se tiene,

$$Y = \sum_{i=}^{N} y_i = N\,a + b\sum_{i=1}^{N} x_i + \sum_{i=1}^{N} e_i = N\,a + b\sum_{i=1}^{N} x_i \qquad y \qquad \bar{Y} = a + b\bar{X} \tag{7.38}$$

si se toma una muestra de tamaño *n*, $n \leq N$, y se le aplica a (7.36) las mismas operaciones anteriores, pero hasta *n*, se tiene que,

$$\sum_{i=}^{n} y_i = n\,a + b\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} e_i \qquad y \qquad \bar{y} = a + b\bar{x} + \bar{e} \tag{7.39}$$

donde $\bar{e}$ es el promedio muestral de los desvíos y además es diferente de cero. Restando (7.39) de (7.38),

$$\bar{Y} - \bar{y} = b\bar{X} - b\bar{x} - \bar{e} \qquad \Rightarrow \qquad \bar{Y} = \bar{y} + b(\bar{X} - \bar{x}) - \bar{e}$$

si se hace $e'_i = -e_i$, entonces $$\overline{Y} = \overline{y} + b(\overline{X} - \overline{x}) + \overline{e}' \quad (7.40)$$

El valor de *"a"* se puede obtener igualando (7.40) a (7.38), $a = \overline{y} - b\overline{x} + \overline{e}'$

Si los datos se comportan como una recta, entonces $\overline{e} = \overline{e}' = 0$

de lo que puede decirse que un estimador del promedio poblacional es,

$$\hat{\overline{Y}}_{lr} = \overline{y} + b(\overline{X} - \overline{x}) \quad (7.41)$$

y será bueno en la medida que e_i converja a cero, es decir, cuando y_i se aproxime a la recta $a+bx_i$ .El estimador mostrado en (7.41) es el "estimador de regresión lineal del promedio poblacional", y el estimador del total viene dado por la expresión,

$$\hat{Y}_{lr} = N\hat{\overline{Y}}_{lr} = \hat{Y} + b(X - \hat{X}) \quad (7.42)$$

7.2.1.- Esperanza de los estimadores

Vease la esperanza del estimador de regresión lineal del promedio,

$$E(\hat{\overline{Y}}_{lr}) = E\left[\overline{y} + b(\overline{X} - \overline{x})\right] = E(\overline{y}) + \overline{X}\,E(b) - E(b\overline{x}) = E(\overline{y}) + E(\overline{x})E(b) - E(b\overline{x})$$

$$= E(\overline{y}) - \left[E(b\overline{x}) - E(\overline{x})\,E(b)\right] = \overline{Y} - COV(b, \overline{x}) \quad (7.43)$$

por lo tanto, no es un estimador insesgado, y su sesgo es

$$B(\hat{\overline{Y}}_{lr}) = \overline{Y} - E(\hat{\overline{Y}}_{lr}) = \overline{Y} - \overline{Y} + COV(b, \overline{x}) = COV(b, \overline{x}) \quad (7.44)$$

Al no ser un estimador insesgado, el error cuadrático medio es,

$$ECM(\hat{\overline{Y}}_{lr}) = V(\hat{\overline{Y}}_{lr}) + \left(B(\hat{\overline{Y}}_{lr})\right)^2$$

7.2.2.- Error cuadrático medio de los estimadores

como ya se demostró, $B(\hat{\overline{Y}}_{lr}) = COV(b, \overline{x}) = E(b\overline{x}) - \overline{X}E(b)$,

entonces,

$$\left(B(\hat{\overline{Y}}_{lr})\right)^2 = \left(E(b\overline{x})\right)^2 - 2\overline{X}\,E(b)\,E(b\overline{x}) + \overline{X}^2\left(E(b)\right)^2 \quad (7.45)$$

Se procederá a hallar la varianza del estimador del promedio.

$$V(\hat{\bar{Y}}_{lr}) = E\left[\hat{\bar{Y}}_{lr} - E(\hat{\bar{Y}}_{lr})\right]^2 = E\left(\hat{\bar{Y}}_{lr}^2\right) - \left[E(\hat{\bar{Y}}_{lr})\right]^2 = E(\hat{\bar{Y}}_{lr}^2) - \left[\bar{Y} - E(b\bar{x}) + \bar{X}E(b)\right]^2$$

$$= E(\hat{\bar{Y}}_{lr}^2) - \left[\bar{Y}^2 + \left(E(b\bar{x})\right)^2 + \bar{X}^2\left(E(b)\right)^2 - 2\bar{Y}E(b\bar{x}) + 2\bar{X}\bar{Y}E(b) - 2\bar{X}E(b)E(b\bar{x})\right]$$

$$= E(\hat{\bar{Y}}_{lr}^2) - \bar{Y}^2 - \left(E(b\bar{x})\right)^2 - \bar{X}^2\left(E(b)\right)^2 + 2\bar{Y}E(b\bar{x}) - 2\bar{X}\bar{Y}E(b) + 2\bar{X}E(b)E(b\bar{x}) \qquad (7.46)$$

pero,

$$E(\hat{\bar{Y}}_{lr}^2) = E\left[\left(\bar{y} + b(\bar{X} - \bar{x})\right)^2\right] = E\left[\bar{y}^2 + 2b\bar{y}(\bar{X} - \bar{x}) + b^2(\bar{X} - \bar{x})^2\right]$$

$$= E\left[\bar{y}^2 + 2b\bar{X}\bar{y} - 2b\bar{x}\bar{y} + b^2\bar{X}^2 - 2b^2\bar{X}\bar{x} + b^2\bar{x}^2\right]$$

$$= E(\bar{y}^2) + 2\bar{X}E(b\bar{y}) - 2E(b\bar{x}\bar{y}) + \bar{X}^2E(b^2) - 2\bar{X}E(b^2\bar{x}) + E(b^2\bar{x}^2)$$

que al sustituir en (7.46), se tiene que,

$$V(\hat{\bar{Y}}_{lr}) = E(\bar{y}^2) + 2\bar{X}E(b\bar{y}) - 2E(b\bar{x}\bar{y}) + \bar{X}^2E(b^2) - 2\bar{X}E(b^2\bar{x}) + E(b^2\bar{x}^2) - \bar{Y}^2 - \left(E(b\bar{x})\right)^2 -$$
$$- \bar{X}^2\left(E(b)\right)^2 + 2\bar{Y}E(b\bar{x}) - 2\bar{X}\bar{Y}E(b) + 2\bar{X}E(b)E(b\bar{x})$$
$$= \left[E(\bar{y}^2) - \bar{Y}^2\right] + E(b^2\bar{x}^2) - \left(E(b\bar{x})\right)^2 + \bar{X}^2E(b^2) - \bar{X}^2\left(E(b)\right)^2 + 2\bar{X}E(b\bar{y}) - 2E(b\bar{x}\bar{y}) -$$
$$- 2\bar{X}E(b^2\bar{x}) + 2\bar{Y}E(b\bar{x}) - 2\bar{X}\bar{Y}E(b) + 2\bar{X}E(b)E(b\bar{x})$$

$$= V(\bar{y}) + E(b^2\bar{x}^2) - \left(E(b\bar{x})\right)^2 + \bar{X}^2E(b^2) - \bar{X}^2\left(E(b)\right)^2 + 2\bar{X}E(b\bar{y}) - 2E(b\bar{x}\bar{y}) -$$
$$- 2\bar{X}E(b^2\bar{x}) + 2\bar{Y}E(b\bar{x}) - 2\bar{X}\bar{Y}E(b) + 2\bar{X}E(b)E(b\bar{x}) \qquad (7.47)$$

entonces el Error Cuadrático Medio es la suma de (7.45) y (7.47), es decir,

$$ECM(\hat{\bar{Y}}_{lr}) = V(\bar{y}) + E(b^2\bar{x}^2) + \bar{X}^2E(b^2) + 2\bar{X}E(b\bar{y}) - 2E(b\bar{x}\bar{y}) - 2\bar{X}E(b^2\bar{x}) + 2\bar{Y}E(b\bar{x}) - 2\bar{X}\bar{Y}E(b) \qquad (7.48)$$

Es de hacer notar que si $\bar{x} = \bar{X}$, entonces, por (7.40), $\bar{Y} = \bar{y} + \bar{e}'$, es decir que si $\bar{e}$ y $\bar{e}'$ están cerca de cero, $\bar{y}$ estará cerca de $\bar{Y}$.

En el caso del estimador del total se tiene que, como se había mostrado en (7.42),

$$\hat{Y}_{lr} = N\hat{\bar{Y}}_{lr} = \hat{Y} + b(X - \hat{X}) \qquad (7.49)$$

su respectiva esperanza, varianza, sesgo y Error Cuadrático Medio son:

$$E(\hat{Y}_{lr}) = E\left[\hat{Y} + b(X - \hat{X})\right] = E(\hat{Y}) + X\ E(b) - E(b\hat{X}) = E(\hat{Y}) + E(\hat{X})\ E(b) - E(b\hat{X})$$
$$= E(\hat{Y}) - \left[E(b\hat{X}) - E(\hat{X})\ E(b)\right] = Y - COV(b, \hat{X}) \qquad (7.50)$$

$$V(\hat{Y}_{lr}) = V(\hat{Y}) + E(b^2\hat{X}^2) - \left(E(b\hat{X})\right)^2 + X^2E(b^2) - X^2\left(E(b)\right)^2 + 2XE(b\hat{Y}) - 2E(b\hat{X}\hat{Y}) - \\ - 2XE(b^2\hat{X}) + 2YE(b\hat{X}) - 2XYE(b) + 2XE(b)E(b\hat{X}) \quad (7.51)$$

$$B(\hat{Y}_{lr}) = Y - E(\hat{Y}_{lr}) = Y - Y + COV(b,\hat{X}) = COV(b,\hat{X}) \quad (7.52)$$

$$ECM(\hat{Y}_{lr}) = V(\hat{Y}) + E(b^2\hat{X}^2) + X^2E(b^2) + 2XE(b\hat{Y}) - 2E(b\hat{X}\hat{Y}) - 2XE(b^2\hat{X}) + 2YE(b\hat{X}) - 2XYE(b) \quad (7.53)$$

Cabe destacar que en la mayoría de los casos *b* es una constante, considerando esta situación, se tiene que,

$COV(b,\bar{x}) = 0$, en consecuencia, $E(\hat{Y}_{lr}) = \bar{Y}$ y $B(\hat{Y}_{lr}) = 0$, es decir, que $\hat{Y}_{lr}$ es un estimador insesgado del promedio poblacional.

Su error cuadrático medio sería,

$$ECM(\hat{\bar{Y}}_{lr}) = V(\hat{\bar{Y}}_{lr}) = V(\bar{y}) + E(b^2\bar{x}^2) - \left(E(b\bar{x})\right)^2 + \bar{X}^2E(b^2) - \bar{X}^2\left(E(b)\right)^2 + 2\bar{X}E(b\bar{y}) - 2E(b\bar{x}\bar{y}) - \\ - 2\bar{X}E(b^2\bar{x}) + 2\bar{Y}E(b\bar{x}) - 2\bar{X}\bar{Y}E(b) + 2\bar{X}E(b)E(b\bar{x})$$

$$= V(\bar{y}) + b^2E(\bar{x}^2) - b^2\left(E(\bar{x})\right)^2 + b^2\bar{X}^2 - b^2\bar{X}^2 + 2b\bar{X}E(\bar{y}) - 2bE(\bar{x}\bar{y}) - 2b^2\bar{X}E(\bar{x}) + 2b\bar{Y}E(\bar{x}) - 2b\bar{X}\bar{Y} + 2b^2\bar{X}E(\bar{x})$$

$$= V(\bar{y}) + b^2E(\bar{x}^2) - b^2\left(E(\bar{x})\right)^2 + 2b\bar{X}\bar{Y} - 2bE(\bar{x}\bar{y}) - 2b^2\bar{X}^2 + 2b\bar{X}\bar{Y} - 2b\bar{X}\bar{Y} + 2b^2\bar{X}^2$$

$$= V(\bar{y}) + b^2E(\bar{x}^2) - b^2\left(E(\bar{x})\right)^2 + 2b\bar{X}\bar{Y} - 2bE(\bar{x}\bar{y})$$

$$= V(\bar{y}) + b^2\left[E(\bar{x}^2) - \left(E(\bar{x})\right)^2\right] - 2b\left[E(\bar{x}\bar{y}) - \bar{X}\bar{Y}\right]$$

$$= V(\bar{y}) + b^2V(\bar{x}) - 2bCOV(\bar{x}\bar{y})$$

como se sabe, $V(\bar{y}) = \frac{N-n}{N\,n}S_y^2$ y $V(\bar{x}) = \frac{N-n}{N\,n}S_x^2$

desarrollando $COV(\bar{x}\bar{y}) = \left[E(\bar{x}\bar{y}) - \bar{X}\bar{Y}\right]$,

$$E(\bar{x}\bar{y}) - \bar{X}\bar{Y} = E\left[\frac{\sum_{i=1}^{n}x_i}{n}\frac{\sum_{i=1}^{n}y_i}{n}\right] - \bar{X}\bar{Y} = \frac{1}{n^2}E\left[\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)\right] - \bar{X}\bar{Y}$$

$$= \frac{1}{n^2}\left[\frac{\sum_{k=1}^{\binom{N}{n}}\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{\binom{N}{n}}\right] - \bar{X}\bar{Y}$$

$$= \frac{1}{n^2}\left[\frac{\binom{N-1}{n-1}\sum_{i=1}^{N} x_i\, y_i + \binom{N-2}{n-2}\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} x_i\, y_i}{\binom{N}{n}}\right] - \overline{X}\,\overline{Y}$$

$$= \left[\frac{\binom{N-1}{n-1}\binom{N-2}{n-2}\sum_{i=1}^{N} x_i\ y_i + \binom{N-2}{n-2}\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} x_i\ y_i}{n^2\binom{N}{n}\binom{N-1}{n-1}\binom{N-2}{n-2}}\right] - \overline{\overline{XY}}$$

$$= \frac{1}{N\,n}\left[\sum_{i=1}^{N} x_i\ y_i + \left(\frac{n-1}{N-1}\right)\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} x_i\ y_i - Nn\overline{\overline{XY}}\right]$$

$$= \frac{1}{N\,n}\left[\sum_{i=1}^{N} x_i\, y_i + \left(\frac{n-1}{N-1}\right)\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} x_i\, y_i - Nn\overline{\overline{XY}} + \left(\frac{n-1}{N-1}\right)\sum_{i=1}^{N} x_i\, y_i - \left(\frac{n-1}{N-1}\right)\sum_{i=1}^{N} x_i\, y_i\right]$$

$$= \frac{1}{N\,n}\left[\left(1-\frac{n-1}{N-1}\right)\sum_{i=1}^{N} x_i\, y_i + \left(\frac{n-1}{N-1}\right)\left(\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} x_i\, y_i + \sum_{i=1}^{N} x_i\, y_i\right) - Nn\overline{\overline{XY}}\right]$$

$$= \frac{1}{N\,n}\left[\left(\frac{N-n}{N-1}\right)\sum_{i=1}^{N} x_i\, y_i + \left(\frac{n-1}{N-1}\right)\left(\sum_{i=1}^{N}\sum_{j=1}^{N} x_i\, y_i\right) - Nn\overline{\overline{XY}}\right]$$

$$= \frac{1}{N\,n}\left[\left(\frac{N-n}{N-1}\right)\sum_{i=1}^{N} x_i\, y_i + \left(\frac{n-1}{N-1}\right)N^2\,\overline{\overline{XY}} - Nn\overline{\overline{XY}}\right]$$

$$= \frac{1}{N\,n\,(N-1)}\left[(N-n)\sum_{i=1}^{N} x_i\ y_i + \left(N^2(n-1) - Nn\right)\overline{\overline{XY}}\right]$$

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

$$= \frac{(N-n)}{N\,n} S_{xy} \qquad (7.54)$$

donde, $S_{xy} = \dfrac{\sum_{i=1}^{N}(x_i - \overline{X})(y_i - \overline{Y})}{N-1} = \dfrac{\sum_{i=1}^{N} x_i\, y_i - \overline{\overline{XY}}}{N-1}$

De manera que

$$ECM(\hat{\bar{Y}}_{lr}) = V(\bar{y}) + b^2 V(\bar{x}) - 2b\left[E(\bar{x}\bar{y}) - \bar{X}\bar{Y}\right] = \frac{(N-n)}{N\,n}\left[S_y^2 + b^2 S_x^2 - 2bS_{xy}\right]$$

y para el total,

$$ECM(\hat{Y}_{lr}) = N^2\,\frac{(N-n)}{N\,n}\left[S_y^2 + b^2 S_x^2 - 2bS_{xy}\right] = \frac{N(N-n)}{n}\left[S_y^2 + b^2 S_x^2 - 2bS_{xy}\right]$$

7.2.3.- Estimadores de los errores cuadráticos medio de los estimadores

Para estimar los errores cuadráticos medios de los estimadores, bastaría con sustituir los valores poblacionales por los estimados, es decir, las cuasivarianzas, las razones, los totales y los promedios. Esto es,

$$\hat{ECM}(\hat{\bar{Y}}_{lr}) = \frac{(N-n)}{N\,n}\left[s_y^2 + b^2 s_x^2 - 2bs_{xy}\right]$$

y para el total,

$$\hat{ECM}(\hat{Y}_{lr}) = N^2\,\frac{(N-n)}{N\,n}\left[s_y^2 + b^2 s_x^2 - 2bs_{xy}\right] = \frac{N(N-n)}{n}\left[s_y^2 + b^2 s_x^2 - 2bs_{xy}\right]$$

De cualquier modo, a continuación se desarrollan algunos casos especiales, y en cada uno de ellos se mostrarán los errores cuadráticos medios y sus estimadores.

7.2.4.- Casos especiales de los estimadores de regresión lineal

A continuación se presentan algunos casos particulares del estimador de regresión lineal.

7.2.4.1.- Estimador insesgado

Cuando *b=0*, el estimador de regresión se transforma en el estimador insesgado.

$$\hat{\bar{Y}}_{lr} = \bar{y} + 0(\bar{X} - \bar{x}) = \bar{y}$$

Verficando la esperanza y el sesgo por (7.43)

$$E(\hat{\bar{Y}}_{lr})=\bar{Y}-COV(b\bar{x})=\bar{Y}-\left[E(b\bar{x})-E(\bar{x})E(b)\right]=\bar{Y}-\left[E(b\bar{x})-\bar{X}\,E(b)\right]$$

pero si *b=0*, entonces $E(b\bar{x})=E(0)=0$, y $E(b)=E(0)=0$, por lo tanto,

$$E(\hat{\bar{Y}}_{lr})=\bar{Y} \qquad \text{; y el sesgo} \qquad B(\hat{\bar{Y}}_{lr})=0$$

lo que verifica que es un estimador insesgado. Entonces, por (7.48),

$$ECM(\hat{\bar{Y}}_{lr})=V(\hat{\bar{Y}}_{lr})=V(\bar{y}) \qquad \text{; es decir,} \qquad ECM(\hat{\bar{Y}}_{lr})=\frac{N-n}{N}\frac{S_y^2}{n}$$

que es la varianza conocida del estimador insesgado del promedio, la varianza del promedio muestral. El error cuadrático medio del total es

$$ECM(\hat{\bar{Y}}_{lr})=V(\hat{Y}_{lr})=V(\hat{Y})=N^2\frac{N-n}{N}\frac{S_y^2}{n}=N(N-n)\frac{S_y^2}{n}$$

los respectivos estimadores también son conocidos,

$$\hat{ECM}(\hat{\bar{Y}}_{lr})=\hat{V}(\hat{\bar{Y}}_{lr})=\hat{V}(\bar{y})=\frac{N-n}{N}\frac{s_y^2}{n}$$

$$\hat{ECM}(\hat{\bar{Y}}_{lr})=\hat{V}(\hat{Y}_{lr})=\hat{V}(\hat{Y})=N^2\frac{N-n}{N}\frac{s_y^2}{n}=N(N-n)\frac{s_y^2}{n}$$

7.2.4.2.- Estimador de razón

Véase ahora qué ocurre si $b=\bar{y}/\bar{x}$. Sustituyendo en (7.41),

$$\hat{\bar{Y}}_{lr}=\bar{y}+b(\bar{X}-\bar{x})=\bar{y}+\frac{\bar{y}}{\bar{x}}(\bar{X}-\bar{x})=\bar{y}+\frac{\bar{y}}{\bar{x}}\bar{X}-\bar{y}=\frac{\bar{y}}{\bar{x}}\bar{X}$$

que no es más que el estimador del promedio del tipo razón, visto en la sección 7.1 (Estimadores de Razón) Ahora se verificará si su esperanza, sesgo, varianza y error cuadrático medio son iguales a los mostrados en las secciones 7.1.1 (Esperanza de los Estimadores de Razón) y 7.1.2. (Error Cuadrático Medio de los Estimadores de Razón)

Por (7.43) $$E(\hat{\bar{Y}}_{lr})=\bar{Y}-\left[E(b\bar{x})-\bar{X}\,E(b)\right]=\bar{Y}-\left[E\left(\left(\frac{\bar{y}}{\bar{x}}\right)\bar{x}\right)-\bar{X}\,E\left(\frac{\bar{y}}{\bar{x}}\right)\right]=\bar{Y}-E(\bar{y})+\bar{X}\,E\left(\frac{\bar{y}}{\bar{x}}\right)=\bar{X}\,E\left(\frac{\bar{y}}{\bar{x}}\right)$$

y por (7.12), $$E(\hat{\bar{Y}}_{lr})=\bar{X}\left(\frac{\bar{Y}}{\bar{X}}\right)\left[1+\frac{N-n}{N\ n}\left(\frac{S_x^2}{\bar{X}^2}-\frac{S_{xy}}{\bar{X}\bar{Y}}\right)\right]=\bar{Y}\left[1+\frac{N-n}{N\ n}\left(\frac{S_x^2}{\bar{X}^2}-\frac{S_{xy}}{\bar{X}\bar{Y}}\right)\right]$$

que es la esperanza de $\hat{\bar{Y}}_R$ mostrada en (7.18). Para hallar el error cuadrático medio, se sustituye $b=\bar{y}/\bar{x}$ en (7.48) y se obtiene lo siguiente,

$$\begin{aligned}ECM(\hat{\bar{Y}}_{lr}) &= V(\bar{y})+E(\bar{y}^2)+\bar{X}^2E\left(\frac{\bar{y}^2}{\bar{x}^2}\right)+2\bar{X}E\left(\frac{\bar{y}^2}{\bar{x}}\right)-2E(\bar{y}^2)-2\bar{X}E\left(\frac{\bar{y}^2}{\bar{x}}\right)+2\bar{Y}E(\bar{y})-2\bar{X}\bar{Y}E\left(\frac{\bar{y}}{\bar{x}}\right)\\ &= V(\bar{y})+E(\bar{y}^2)+2\bar{Y}E(\bar{y})-2\bar{X}\bar{Y}E\left(\frac{\bar{y}}{\bar{x}}\right)+\bar{X}^2E\left(\frac{\bar{y}^2}{\bar{x}^2}\right)\\ &= V(\bar{y})+E(\bar{y}^2)+2\left(E(\bar{y})\right)^2-2\bar{X}\bar{Y}E\left(\frac{\bar{y}}{\bar{x}}\right)+\bar{X}^2E\left(\frac{\bar{y}^2}{\bar{x}^2}\right)\\ &= V(\bar{y})-\left[E(\bar{y}^2)-\left(E(\bar{y})\right)^2\right]+\bar{Y}^2-2\bar{X}\bar{Y}E\left(\frac{\bar{y}}{\bar{x}}\right)+\bar{X}^2E\left(\frac{\bar{y}^2}{\bar{x}^2}\right)\\ &= V(\bar{y})-V(\bar{y})+\bar{Y}E(\bar{y})-2\bar{X}\bar{Y}E\left(\frac{\bar{y}}{\bar{x}}\right)+\bar{X}^2E\left(\frac{\bar{y}^2}{\bar{x}^2}\right)\end{aligned}$$

sustituyendo $E\left(\frac{\bar{y}}{\bar{x}}\right)$ y $E\left(\frac{\bar{y}^2}{\bar{x}^2}\right)$ por (7.9) y (7.13) respectivamente,

$$\begin{aligned}ECM(\hat{\bar{Y}}_{lr}) &= \bar{Y}^2-2\bar{X}\bar{Y}\frac{\bar{Y}}{\bar{X}}\left[1+\frac{N-n}{Nn}\left(\frac{S_x^2}{\bar{X}^2}-\frac{S_{xy}}{\bar{X}\bar{Y}}\right)\right]+\bar{X}^2\frac{\bar{Y}^2}{\bar{X}^2}\left[1+\frac{N-n}{Nn}\left(\frac{S_y^2}{\bar{Y}^2}-4\frac{S_{xy}}{\bar{X}\bar{Y}}+3\frac{S_x^2}{\bar{X}^2}\right)\right]\\ &= \bar{Y}^2-2\bar{Y}^2-2\frac{N-n}{Nn}\bar{Y}^2\frac{S_x^2}{\bar{X}^2}+2\frac{N-n}{Nn}\bar{Y}\frac{S_{xy}}{\bar{X}}+\bar{Y}^2+\frac{N-n}{Nn}S_y^2-4\frac{N-n}{Nn}\bar{Y}\frac{S_{xy}}{\bar{X}}+3\frac{N-n}{Nn}\bar{Y}^2\frac{S_x^2}{\bar{X}^2}\end{aligned}$$

$$ECM(\hat{\bar{Y}}_{lr})=\frac{N-n}{N\ n}\left[S_y^2-2\frac{\bar{Y}}{\bar{X}}S_{xy}+\left(\frac{\bar{Y}}{\bar{X}}\right)^2S_x^2\right] \qquad (7.54)$$

que es el error cuadrático medio de $\hat{\bar{Y}}_R$ mostrada en (7.20). Para el total se tiene que,

$$\hat{Y}_{lr}=\hat{Y}+b(X-\hat{X})=\hat{Y}+\frac{y}{x}(X-\hat{X})=\hat{Y}+\frac{y}{x}X-\hat{Y}=\frac{y}{x}X$$

$$E(\hat{Y}_{lr}) = \bar{Y}\left[1 + \frac{N-n}{N\ n}\left(\frac{S_x^2}{\bar{X}^2} - \frac{S_{xy}}{\bar{X}\bar{Y}}\right)\right]$$

$$ECM(\hat{Y}_{lr}) = N^2\,\frac{N-n}{N\ n}\left[S_y^2 - 2\frac{\bar{Y}}{\bar{X}}S_{xy} + \left(\frac{\bar{Y}}{\bar{X}}\right)^2 S_x^2\right]$$

Los estimadores de los errores cuadráticos medios son,

$$\hat{ECM}(\hat{Y}_{lr}) = \frac{(N-n)}{N\,n}\left[s_y^2 - 2\frac{y}{x}s_{xy} + \left(\frac{y}{x}\right)^2 s_x^2\right] \qquad \hat{ECM}(\hat{Y}_{lr}) = N(N-n)\frac{1}{n}\left[s_y^2 - 2\frac{y}{x}s_{xy} + \left(\frac{y}{x}\right)^2 s_x^2\right]$$

7.2.4.3.- Estimador por diferencia

Este es el caso cuando *b=1*, entonces se corrige el promedio muestral sólo por la diferencia entre los promedios poblacional y muestral de la variable auxiliar; esto es,

$$\hat{\bar{Y}}_{lr} = \bar{y} + (\bar{X} - \bar{x})$$

al sustituir b=1 en (7.43), se tiene,

$$E(\hat{\bar{Y}}_{lr}) = \bar{Y} + \left[E(\bar{X}) - E(x)E(1)\right] = \bar{Y}$$

por lo tanto es un estimador insesgado. Haciendo lo propio en (7.48),

$$\begin{aligned} ECM(\hat{\bar{Y}}_{lr}) = V(\hat{\bar{Y}}_{lr}) &= V(\bar{y}) + E(\bar{x}^2) + \bar{X}^2 + 2\bar{X}\bar{Y} - 2E(\bar{x}\bar{y}) - 2\bar{X}^2 + 2\bar{X}\bar{Y} - 2\bar{X}\bar{Y} \\ &= V(\bar{y}) + \left(E(\bar{x}^2) - \bar{X}^2\right) - 2\left(E(\bar{x}\bar{y}) - \bar{X}\bar{Y}\right) \\ &= V(\bar{y}) + V(\bar{x}) - 2COV(\bar{x},\bar{y}) \\ &= \frac{N-n}{N\,n}\left[S_y^2 + S_x^2 - 2S_{xy}\right] \end{aligned}$$

Para el caso del total se tiene que,

$$\hat{Y}_{lr} = N\bar{y} + N(\bar{X} - \bar{x}) = \hat{Y} + (X - \hat{X})$$
$$E(\hat{Y}_{lr}) = Y + \left[E(X) - E(\hat{X})\right] = Y$$

por lo tanto, también es un estimador insesgado.

$$ECM(\hat{Y}_{lr}) = V(\hat{Y}_{lr}) = N(N-n)\left[S_y^2 + S_x^2 - 2S_{xy}\right]$$

y su estimador,

$$\hat{ECM}(\hat{Y}_{lr}) = \hat{V}(\hat{Y}_{lr}) = N(N-n)\left[s_y^2 + s_x^2 - 2s_{xy}\right]$$

7.2.4.4.- Estimador de varianza mínima para *b* constante

Sea $b = b_0$ constante, entonces, sustituyendo en (7.43),

$$E(\hat{\bar{Y}}_{lr}) = E(\bar{y}) - \left[b_0 E(\bar{x}) - b_0 E(\bar{x})\right] = E(\bar{y}) = \bar{Y}$$

se tiene que es un estimador insesgado. Sustiyuyendo $b = b_0$ en (7.48),

$$\begin{aligned}
ECM(\hat{\bar{Y}}_{lr}) = V(\hat{\bar{Y}}_{lr}) &= V(\bar{y}) + b_0^2 E(\bar{x}^2) + b_0^2 \bar{X}^2 + 2b_0 \bar{X} E(\bar{y}) - 2b_0 E(\bar{x}\bar{y}) - 2b_0^2 \bar{X} E(\bar{x}) + 2b_0 \bar{Y} E(\bar{x}) - 2b_0 \bar{X}\bar{Y} \\
&= V(\bar{y}) + b_0^2 E(\bar{x}^2) + b_0^2 \bar{X}^2 - 2b_0^2 \bar{X} E(\bar{x}) - 2b_0 E(\bar{x}\bar{y}) + 2b_0 \bar{X} E(\bar{y}) + 2b_0 \bar{Y} E(\bar{x}) - 2b_0 \bar{X}\bar{Y} \\
&= V(\bar{y}) + b_0^2 E(\bar{x}^2) + b_0^2 \left(E(\bar{x})\right)^2 - 2b_0^2 \left(E(\bar{x})\right)^2 - 2b_0 E(\bar{x}\bar{y}) + 2b_0 E(\bar{x}) E(\bar{y}) + \\
&\quad + 2b_0 E(\bar{y}) E(\bar{x}) - 2b_0 E(\bar{x}) E(\bar{y}) \\
&= V(\bar{y}) + b_0^2 \left[E(\bar{x}^2) - \left(E(\bar{x})\right)^2\right] - 2b_0 \left[E(\bar{x}\bar{y}) - E(\bar{x}) E(\bar{y})\right] \\
&= V(\bar{y}) + b_0^2 V(\bar{x}) - 2b_0 COV(\bar{x}, \bar{y}) \\
&= \frac{N-n}{Nn}\left[S_y^2 + b_0^2 S_x^2 - 2b_0 S_{xy}\right]
\end{aligned}$$

Para minimizar la varianza, se deriva respecto de b_0 y se iguala a cero,

$$\frac{\partial V(\hat{\bar{Y}}_{lr})}{\partial b_0} = -2\frac{N-n}{Nn} S_{xy} + 2\frac{N-n}{Nn} b_0 S_x^2 = 2\frac{N-n}{Nn}\left[b_0 S_x^2 - S_{xy}\right] = 0$$

para que se cumpla la igualdad debe ocurrir uno de los siguientes casos,

$$\begin{cases} a)\ n = N \\ ó \\ b)\ b_0 S_x^2 - S_{xy} = 0 \end{cases}$$

no tiene sentido hacer *n=N*, ya que la investigación sería por enumeración completa y la varianza del estimador igual a cero, porque se tiene el parámetro. Tomando el caso (*b*) y despejando b_0,

$$b_0 = \frac{S_{xy}}{S_x^2} = \frac{\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^{N}(x_i - \bar{X})^2} = \frac{\sum_{i=1}^{N} x_i y_i - N\bar{X}\bar{Y}}{\sum_{i=1}^{N} x_i^2 - N\bar{X}^2}$$

y la segunda derivada es,

$\frac{\partial^2 V(\hat{\bar{Y}}_{lr})}{\partial^2 b_0} = 2\frac{N-n}{Nn}S_x^2 \geq 0$; por lo tanto, en $b_0 = \frac{S_{xy}}{S_x^2}$ se minimiza $V(\hat{\bar{Y}}_{lr})$

Es importante destacar que b_0, es el coeficiente utilizado en los estudios de regresión a través del método de los mínimos cuadrados. Por lo tanto,

$$ECM(\hat{\bar{Y}}_{lr}) = V(\hat{\bar{Y}}_{lr}) = \frac{N-n}{Nn}\left[S_y^2 + b_0^2 S_x^2 - 2b_0 S_{xy}\right] = \frac{N-n}{Nn}\left[S_y^2 + \left(\frac{S_{xy}^2}{S_x^4}\right)S_x^2 - 2\left(\frac{S_{xy}}{S_x^2}\right)S_{xy}\right] = \frac{N-n}{Nn}\left[S_y^2 - \frac{S_{xy}^2}{S_x^2}\right]$$

$$= \frac{N-n}{Nn}S_y^2\left[1 - \frac{S_{xy}^2}{S_x^2 S_y^2}\right] = \frac{N-n}{Nn}S_y^2\left[1-\rho_{xy}^2\right]$$

Para el total,

$$ECM(\hat{Y}_{lr}) = V(\hat{Y}_{lr}) = N^2\left(\frac{N-n}{Nn}\right)S_y^2\left[1 - \frac{S_{xy}^2}{S_x^2 S_y^2}\right] = N^2\left(\frac{N-n}{Nn}\right)S_y^2\left[1-\rho_{xy}^2\right] = N(N-n)S_y^2\left[1-\rho_{xy}^2\right]$$

Cuando no se dispone de los valores poblacionales, entonces,

$\hat{b}_0 = \frac{s_{xy}}{s_x^2}$ y entonces,

$$\hat{ECM}(\hat{\bar{Y}}_{lr}) = \hat{V}(\hat{\bar{Y}}_{lr}) = \frac{N-n}{Nn}s_y^2\left[1 - \frac{s_{xy}^2}{s_x^2 s_y^2}\right] = \frac{N-n}{Nn}s_y^2\left[1-\hat{\rho}_{xy}^2\right]$$

$$\hat{ECM}(\hat{Y}_{lr}) = \hat{V}(\hat{Y}_{lr}) = N(N-n)s_y^2\left[1-\hat{\rho}_{xy}^2\right]$$

donde $\hat{\rho}_{xy} = \frac{s_{xy}}{s_x s_y}$.

7.2.5.- <u>Estimadores de regresión en el muestreo estratificado</u>

Igual que para los estimadores de razón, los estimadores de regresión lineal pueden ser usados cuando la muestra es estratificada, considerando los mismos enfoques que entonces. En principio se verá que ocurre en cada estrato.

Supóngase que se tiene una muestra estratificada, entonces, igual que en los estimadores de Razón, al aplicar los

estimadores de Regresión Lineal en cada estrato se tiene que los estimadores de Regresión Lineal del promedio y total poblacional, respectivamente son:

$$\hat{\bar{Y}}_{lr_h} = \bar{y}_h + b_h(\bar{X}_h - \bar{x}_h)$$
$$\hat{Y}_{lr_h} = N_h\hat{\bar{Y}}_{lr_h} = \hat{Y}_h + b_h(X_h - \hat{X}_h)$$

y sus respectivas esperanza, sesgo, varianza y Error Cuadrático Medio son:

Para el estimador del promedio,

$$E(\hat{\bar{Y}}_{lr_h}) = \bar{Y}_h - COV(b_h\bar{x}_h)$$

$$B(\hat{\bar{Y}}_{lr_h}) = COV(b_h\bar{x}\ )$$

$$V(\hat{\bar{Y}}_{lr_h}) = V(\bar{y}_h) + E(b_h^2\bar{x}_h^2) - \left(E(b_h\bar{x}_h)\right)^2 + \bar{X}_h^2E(b_h^2) - \bar{X}_h^2\left(E(b_h)\right)^2 + 2\bar{X}_hE(b_h\bar{y}_h) - 2E(b_h\bar{x}_h\bar{y}_h) - 2\bar{X}_hE(b_h^2\bar{x}_h) + 2\bar{Y}_hE(b_h\bar{x}_h) - 2\bar{X}_h\bar{Y}_hE(b_h) + 2\bar{X}_hE(b_h)E(b_h\bar{x}_h)$$

$$ECM(\hat{\bar{Y}}_{lr_h}) = V(\bar{y}_h) + E(b_h^2\bar{x}_h^2) + \bar{X}_h^2E(b_h^2) + 2\bar{X}_h^2E(b_h\bar{y}_h) - 2E(b_h\bar{x}_h\bar{y}_h) - 2\bar{X}_hE(b_h^2\bar{x}_h) + 2\bar{Y}_hE(b_h\bar{x}_h) - 2\bar{X}_h\bar{Y}_hE(b_h)$$

para el estimador del total,

$$E(\hat{Y}_{lr_h}) = Y_h - COV(b_h, \hat{X}_h)$$
$$V(\hat{Y}_{lr_h}) = V(\hat{Y}_h) + E(b_h^2\hat{X}_h^2) - \left(E(b_h\hat{X}_h)\right)^2 + X_h^2E(b_h^2) - X_h^2\left(E(b_h)\right)^2 + 2X_hE(b_h\hat{Y}_h) - 2E(b_h\hat{X}_h\hat{Y}_h) - 2X_hE(b_h^2\hat{X}_h) + 2Y_hE(b_h\hat{X}_h) - 2X_hY_hE(b_h) + 2X_hE(b_h)E(b_h\hat{X}_h)$$

$$B(\hat{Y}_{lr_h}) = Y_h - E(\hat{Y}_{lr_h}) = Y_h - Y_h + COV(b_h, \hat{X}_h) = COV(b_h, \hat{X}_h)$$

$$ECM(\hat{Y}_{lr_h}) = V(\hat{Y}_h) + E(b_h^2\hat{X}_h^2) + X_h^2E(b_h^2) + 2X_hE(b_h\hat{Y}_h) - 2E(b_h\hat{X}_h\hat{Y}_h) - 2X_hE(b_h^2\hat{X}_h) + 2Y_hE(b_h\hat{X}_h) - 2X_hY_hE(b_h)$$

Si se considera b_h como una constante, los errores cuadráticos medios resultan,

$$ECM(\hat{\bar{Y}}_{lr_h}) = V(\bar{y}_h) + b_h^2V(\bar{x}_h) - 2b_h\left[E(\bar{x}_h\bar{y}_h) - \bar{X}_h\bar{Y}_h\right] = \frac{(N_h - n_h)}{N_h\, n_h}\left[S_{y_h}^2 + b_h^2S_{x_h}^2 - 2b_hS_{xy_h}\right]$$
$$ECM(\hat{Y}_{lr_h}) = N_h^2\,\frac{(N_h - n_h)}{N_h\, n_h}\left[S_{y_h}^2 + b^2S_{x_h}^2 - 2bS_{xy_h}\right] = \frac{N_h(N_h - n_h)}{n_h}\left[S_{y_h}^2 + b_h^2S_{x_h}^2 - 2b_hS_{xy_h}\right]$$

y sus estimadores,

$$E\hat{C}M(\hat{\bar{Y}}_{lr_h}) = \frac{(N_h - n_h)}{N_h\, n_h}\left[s_{y_h}^2 + b_h^2 s_{x_h}^2 - 2b_h s_{xy_h}\right]$$
$$E\hat{C}M(\hat{Y}_{lr_h}) = \frac{N_h(N_h - n_h)}{n_h}\left[S_{y_h}^2 + b_h^2 S_{x_h}^2 - 2b_h S_{xy_h}\right]$$

también se pueden aplicar los casos especiales de estimador insesgado, de razón, por diferencia y de varianza mínima para b_h constante, que resultan de aplicar los mismos procedimientos a cada estrato, o en términos generales, al estrato *h*.

Para hallar el estimador de promedio y del total poblacional, se tienen los mismos dos casos que en los estimadores de Razón, tratar cada estrato de manera separada o de forma combinada.

7.2.5.1.- Estimadores de regresión lineal separado

Aplicando el razonamiento de Muestreo Estratificado, considerando los valores de b_h de forma "separada" o diferenciada para cada estrato, se obtienen los "Estimadores de Regresión Lineal Separado" del total y del promedio respectivamente, y se denotarán por $\hat{Y}_{lrs}$, $\hat{\bar{Y}}_{lrs}$. Estos son:

$$\hat{Y}_{lrs} = \sum_{h=1}^{L}\hat{Y}_{lr_h} = \sum_{h=1}^{L}\hat{Y}_h + b_h(X_h - \hat{X}_h)$$
$$\hat{\bar{Y}}_{lrs} = \frac{1}{N}\hat{Y}_{lrs} = \frac{1}{N}\sum_{h=1}^{L}\hat{Y}_{lr_h} = \frac{1}{N}\sum_{h=1}^{L}N_h\hat{\bar{Y}}_{lr_h} = \sum_{h=1}^{L}W_h\left[\bar{y}_h + b_h(\bar{X}_h - \bar{x}_h)\right]$$

La esperanza de los estimadores de regresión lineal separado son los siguientes,

$$E(\hat{Y}_{lrs}) = E\left[\sum_{h=1}^{L}\hat{Y}_{lr_h}\right] = \sum_{h=1}^{L}E\left[\hat{Y}_{lr_h}\right] = \sum_{h=1}^{L}E\left[\hat{Y}_h - COV(b_h, \hat{X}_h)\right] = \sum_{h=1}^{L}E\left[\hat{Y}_h\right] - \sum_{h=1}^{L}E\left[COV(b_h, \hat{X}_h)\right]$$
$$= \sum_{h=1}^{L}\hat{Y}_h - \sum_{h=1}^{L}COV(b_h, \hat{X}_h) = Y - \sum_{h=1}^{L}COV(b_h, \hat{X}_h)$$

$$E(\hat{\bar{Y}}_{lrs}) = E\left[\sum_{h=1}^{L}W_h\left[\bar{y}_h + b_h(\bar{X}_h - \bar{x}_h)\right]\right] = \sum_{h=1}^{L}W_h\,E\left[\bar{y}_h + b_h(\bar{X}_h - \bar{x}_h)\right]$$

$$= \sum_{h=1}^{L} W_h \left[E(\bar{y}_h) + \bar{X}_h E(b_h) - E(b_h \bar{x}_h) \right] = \sum_{h=1}^{L} W_h \left[E(\bar{y}_h) + E(\bar{x}_h) E(b_h) - E(b_h \bar{x}_h) \right]$$

$$= \sum_{h=1}^{L} W_h \left[E(y_h) - \left(E(b_h x_h) - E(x_h) E(b_h) \right) \right] = \sum_{h=1}^{L} W_h \left[\bar{Y}_h - COV(b_h, x_h) \right]$$

$$= \sum_{h=1}^{L} W_h \bar{Y}_h - \sum_{h=1}^{L} W_h \, COV(b_h, x_h) = \bar{Y} - \sum_{h=1}^{L} W_h \, COV(b_h, x_h)$$

por lo tanto, no son estimadores insesgados, y sus sesgos son

$$B(\hat{\bar{Y}}_{lrs}) = \bar{Y} - E(\hat{\bar{Y}}_{lrs}) = \bar{Y} - \bar{Y} + \sum_{h=1}^{L} COV(b_h, \hat{X}_h) = \sum_{h=1}^{L} COV(b_h, \hat{X}_h)$$

$$B(\hat{\bar{Y}}_{lrs}) = \bar{Y} - E(\hat{\bar{Y}}_{lrs}) = \bar{Y} - \bar{Y} + \sum_{h=1}^{L} W_h \, COV(b_h, \bar{x}_h) = \sum_{h=1}^{L} W_h \, COV(b_h, \bar{x}_h)$$

Desarrollando las varianzas, se tiene que,

$$V(\hat{\bar{Y}}_{lrs}) = V\left(\sum_{h=1}^{L} \hat{\bar{Y}}_{lr_h} \right) = \sum_{h=1}^{L} V(\hat{\bar{Y}}_{lr_h}) + 2 \sum_{h=1}^{L-1} \sum_{k=h+1}^{L} COV(\hat{\bar{Y}}_{lr_h}, \hat{\bar{Y}}_{lr_h}) = \sum_{h=1}^{L} V(\hat{\bar{Y}}_{lr_h})$$

$$V(\hat{\bar{Y}}_{lrs}) = V\left(\sum_{h=1}^{L} W_h \hat{\bar{Y}}_{lr_h} \right) = \sum_{h=1}^{L} W_h^2 V(\hat{\bar{Y}}_{lr_h}) + 2 \sum_{h=1}^{L-1} \sum_{k=h+1}^{L} W_h W_k \, COV(\hat{\bar{Y}}_{lr_h}, \hat{\bar{Y}}_{lr_h}) = \sum_{h=1}^{L} W_h^2 V(\hat{\bar{Y}}_{lr_h})$$

ya que las muestras en los estratos son independientes, entonces,

$$ECM(\hat{\bar{Y}}_{lrs}) = \sum_{h=1}^{L} V(\hat{\bar{Y}}_{lr_h}) + \left(\sum_{h=1}^{L} COV(b_h, \hat{X}_h) \right)^2$$

$$ECM(\hat{\bar{\bar{Y}}}_{lrs}) = \sum_{h=1}^{L} W_h^2 V(\hat{\bar{\bar{Y}}}_{lr_h}) + \left(\sum_{h=1}^{L} W_h \, COV(b_h, \hat{\bar{\bar{X}}}_h) \right)^2$$

pero considerando b_h como una constante en cada estrato, los errores cuadráticos medios resultan,

$$ECM(\hat{\bar{Y}}_{lrs}) = \sum_{h=1}^{L} N_h^2 \frac{(N_h - n_h)}{N_h \, n_h} \left[S_{y_h}^2 + b_h^2 S_{x_h}^2 - 2 b_h S_{xy_h} \right] = \sum_{h=1}^{L} \frac{N_h (N_h - n_h)}{n_h} \left[S_{y_h}^2 + b_h^2 S_{x_h}^2 - 2 b_h S_{xy_h} \right]$$

$$ECM(\hat{\bar{\bar{Y}}}_{lrs}) = \sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h \, n_h} \left[S_{y_h}^2 + b_h^2 S_{x_h}^2 - 2 b_h S_{xy_h} \right]$$

y sus estimadores,

$$E\hat{C}M(\hat{Y}_{lrs}) = \sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{n_h}\left[s_{y_h}^2 + b_h^2 s_{x_h}^2 - 2b_h s_{xy_h}\right]$$

$$E\hat{C}M(\hat{\bar{Y}}_{lrs}) = \sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h\, n_h}\left[s_{y_h}^2 + b_h^2 s_{x_h}^2 - 2b_h s_{xy_h}\right]$$

Cabe destacar que aquí se presentan los mismos casos para los estimadores de regresión lineal en el muestreo aleatorio simple, estimador insesgado, de razón, por diferencia y de varianza mínima, incluso, debido a la independencia de los estratos, pueden presentarse combinaciones de ellos, es decir, tratamientos diferentes en cada estrato, por ejemplo que en algunos estratos se aplique el estimador por diferencia y en otros el de varianza mínima, aunque se debe decir que no es ni mucho menos lo más aplicado. A continuación se muestra la fórmula para el cálculo de *b* cuando se desea minimizar la varianza, que por supuesto, variará en cada estrato.

$$b_{h0} = \frac{S_{xy\,h}}{S_{x_h}^2} \qquad , \qquad \hat{b}_{h0} = \frac{s_{xy_h}}{s_{x_h}^2}$$

Esto no puede aplicarse en el caso siguiente, ya que, los estratos son tratados, a efectos del uso de b, de manera combinada, en consecuencia, el valor de b único,

7.2.5.2.- Estimadores de regresión lineal combinado

Igual que para el caso de los estimadores de razón, aquí se presenta la diferencia en el tratamiento de los estratos, a efectos del uso de los estimadores, sin embargo hay una ligera diferencia en el enfoque, y es que, en principio, los valores de b o b_h no son calculados, necesariamente, a partir de la relación entre las variables, como ocurre con la razón R ; únicamente se hace así cuando se quiere minimizar la varianza, que es cuando se calcula b_0 (ver sección 7.2.4.4.- Estimador de varianza mínima para *b* constante).

De manera que, para el caso general, manteniendo b , que en este caso es b_c , como una constante, los estimadores y sus respectivas esperanzas y varianzas son los siguientes,

$$\hat{\bar{Y}}_{lrc} = \sum_{h=1}^{L} W_h\left(\bar{y}_h + b_c(\bar{X}_h - \bar{x}_h)\right)$$

$$E(\hat{\bar{Y}}_{lrc}) = E\left[\sum_{h=1}^{L} W_h\left(\bar{y}_h + b_c(\bar{X}_h - \bar{x}_h)\right)\right] = \sum_{h=1}^{L} E\left[W_h\left(\bar{y}_h + b_c(\bar{X}_h - \bar{x}_h)\right)\right]$$

$$= \sum_{h=1}^{L}\left[W_h E(\bar{y}_h) + b_c W_h E(\bar{X}_h) - b_c W_h E(\bar{x}_h)\right] = \sum_{h=1}^{L}\left[W_h\bar{Y}_h + b_c W_h\bar{X}_h - b_c W_h\bar{X}_h\right]$$

$$= \sum_{h=1}^{L} W_h\bar{Y}_h = \bar{Y}$$

por lo tanto es un estimador insesgado.

$$ECM(\hat{\bar{Y}}_{lrc}) = V(\hat{\bar{Y}}_{lrc})$$

$$= \sum_{h=1}^{L} W_h^2\left[V(\bar{y}_h) + b_c^2 E(\bar{x}_h^2) + b_c^2\bar{X}_h^2 + 2b_c\bar{X}_h^2 E(\bar{y}_h) - 2b_c E(\bar{x}_h\bar{y}_h) - 2b_c^2\bar{X}_h E(\bar{x}_h) + 2b_c\bar{Y}_h E(\bar{x}_h) - 2b_c\bar{X}_h\bar{Y}_h\right]$$

$$= \sum_{h=1}^{L} W_h^2\left[V(y_h) + b_c^2 E(x_h^2) + b_c^2\left(E(x_h)\right)^2 - 2b_h^2\left(E(x_h)\right)^2 - 2b_c E(x_h y_h) + 2b_c E(x_h)E(y_h)\right.$$
$$\left. + 2b_c E(\bar{y}_h)E(\bar{x}_h) - 2b_c E(\bar{x}_h)E(\bar{y}_h)\right]$$

$$= \sum_{h=1}^{L} W_h^2\left[V(y_h) + b_c^2\left(E(x_h^2) - \left(E(x_h)\right)^2\right) - 2b_c\left(E(x_h y_h) - E(\bar{x}_h)E(y_h)\right)\right]$$

$$= \sum_{h=1}^{L} W_h^2\left[V(\bar{y}_h) + b_c^2 V(\bar{x}_h) - 2b_c COV(\bar{x}_h, \bar{y}_h)\right]$$

$$= \sum_{h=1}^{L} W_h^2\,\frac{N_h - n_h}{N_h\, n_h}\left[S_{y_h}^2 + b_c^2 S_{x_h}^2 - 2b_c S_{xy_h}\right]$$

ya que por analogía de (7.54) $COV(\bar{x}_h, \bar{y}_h) = \sum_{h=1}^{L}\frac{N_h - n_h}{N_h\, n_h} S_{xy_h}$

$$E\hat{C}M(\hat{\bar{Y}}_{lrc}) = \hat{V}(\hat{\bar{Y}}_{lrc}) = \sum_{h=1}^{L} W_h^2\,\frac{N_h - n_h}{N_h\, n_h}\left[s_{y_h}^2 + b_c^2 s_{x_h}^2 - 2b_c s_{xy_h}\right]$$

Para el total se tiene que,

$$\hat{Y}_{lrc} = \sum_{h=1}^{L} N_h\left(\bar{y}_h + b_c(\bar{X}_h - \bar{x}_h)\right)$$

$$E(\hat{\bar{Y}}_{lrc}) = E(N\hat{\bar{Y}}_{lrc}) = N\,E(\hat{\bar{Y}}_{lrc}) = N\bar{Y} = Y$$

$$ECM(\hat{Y}_{lrc}) = V(\hat{Y}_{lrc}) = \sum_{h=1}^{L}\frac{N_h\left(N_h - n_h\right)}{n_h}\left[S_{y_h}^2 + b_c^2 S_{x_h}^2 - 2b_c S_{xy_h}\right]$$

$$E\hat{C}M(\hat{Y}_{lrc}) = \hat{V}(\hat{Y}_{lrc}) = \sum_{h=1}^{L}\frac{N_h\left(N_h - n_h\right)}{n_h}\left[s_{y_h}^2 + b_c^2 s_{x_h}^2 - 2b_c s_{xy_h}\right]$$

Cuando se quiere minimizar la varianza, se deriva la varianza respecto de b_{c0} y se iguala a cero,

$$\frac{\partial V(\hat{\bar{Y}}_{lrc})}{\partial b_{c0}} = \frac{\partial\left[\sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h\, n_h}\left[S_{y_h}^2 + b_{c0}^2 S_{x_h}^2 - 2b_{c0} S_{xy_h}\right]\right]}{\partial b_{c0}} = 0$$

$$= \sum_{h=1}^{L} 2W_h^2 \frac{N_h - n_h}{N_h\, n_h} b_{c0} S_{x_h}^2 - 2W_h^2 \frac{N_h - n_h}{N_h\, n_h} S_{xy_h} = 0$$

$$= 2\sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h\, n_h} b_{c0} S_{x_h}^2 - 2\sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h\, n_h} S_{xy_h} = 0$$

$$= 2b_{c0}\sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h\, n_h} S_{x_h}^2 - 2\sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h\, n_h} S_{xy_h} = 0$$

entonces,

$$b_{c0}\sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h\, n_h} S_{x_h}^2 = \sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h\, n_h} S_{xy_h}$$

despejando b_{c0},

$$b_{c0} = \frac{\sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h\, n_h} S_{xy_h}}{\sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h\, n_h} S_{x_h}^2} = \frac{\sum_{h=1}^{L} W_h^2 \frac{1 - f_h}{n_h} S_{xy_h}}{\sum_{h=1}^{L} W_h^2 \frac{1 - f_h}{n_h} S_{x_h}^2}$$

y la segunda derivada es,

$$\frac{\partial^2 V(\hat{\bar{Y}}_{lrc})}{\partial^2 b_0} = \sum_{h=1}^{L} 2W_h^2 \frac{N_h - n_h}{N_h\, n_h} S_{x_h}^2 = 2\sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h\, n_h} S_{x_h}^2 > 0$$; por lo tanto, en b_{c0} se minimiza $V(\hat{\bar{Y}}_{lrc})$

En general resultan mejor los estimadores de regresión lineal que los de razón, sin embargo, si la relación entre las variables *y*, *x* es lineal y la recta de regresión pasa por el origen, tienden a igualarse la precisión de ambos estimadores, y como trabajar con los estimadores de razón resulta más cómodo debido a que el cálculo de *R* es más sencillo que el de *B*, puede decirse que si la relación entre las variables es una recta que pasa por el origen, es preferible utilizar el estimador de razón, pero si no pasa por el origen, es preferible utilizar el de regresión lineal.

Por otra parte, entre los estimadores separados y combinados, es preferible el uso de los primeros, pero cuando las razones (R_h) o los coeficientes B_h son muy parecidos entre los estratos, debido a la facilidad de cálculo de las

varianzas, es preferible utilizar los combinados; igual pasa si los tamaños de muestra en cada estrato no son los suficientemente grandes como para obtener estimaciones confiables de R_h o de B_h. Pero si por el contrario se requieren estimaciones del total o el promedio para cada estrato, se deben utilizar los separados, asegurándose de contar con tamaños de muestra relativamente grandes.

8.- MUESTREO ALEATORIO SISTEMÁTICO

El muestreo aleatorio sistemático o simplemente muestreo sistemático, es muy usado y tiene razonas para ello, es muy fácil de aplicar, basta con tomar un sólo número aleatorio para seleccionar una muestra, además, los elementos de la muestra se reparten por toda la población, de manera que, si el marco muestral responde a ciertos criterios de ordenamiento, o puede ser ordenado convenientemente, a través de éste método de selección se puede evitar concentraciones de elementos de un mismo tipo en la muestra, cosa que no es posible en el muestreo aleatorio simple. Igualmente, es más fácil hacer una supervisión sobre el levantamiento de la información que en el muestreo aleatorio simple, ya que es más fácil identificar los elementos seleccionados. Por otra parte, es a menudo más fácil no cometer errores con un muestreo sistemático que con el aleatorio simple, aunque se corre el riesgo de obtener resultados sesgados si en la población se dan periodicidades o rachas.

Al compararlo con el muestreo estratificado, el muestreo sistemático puede resultar más costoso, tanto el levantamiento como la supervisión; en cuanto a la precisión, ambos tienen grandes ventajas, y dependerá de la distribución de la población, usualmente son utilizados simultáneamente -muestreo aleatorio sistemático estratificado-, aspecto que se trabajará en el próximo capítulo.

Sea un universo de N elementos, del cual se extraerá una muestra de n de ellos. Sea $k=N/n$ el intervalo de selección. Se divide la población en n intervalos, cada uno de tamaño k, entonces se selecciona un número aleatorio entero r, entre 1 y k, y se selecciona el *r-ésimo* elemento de cada intervalo, de manera que hay k muestras posibles. Los elementos seleccionados son:

$$y_r, y_{r+k}, y_{r+2k}, y_{r+3k}, \quad \cdots \quad , y_{r+(n-1)k}$$

donde $r+(n-1)k < N$, y $r+(n-1)k = N$ si y sólo si $r=k$, para k entero. Cuando k no es entero, existen procedimientos alternos, que más adelante se tratarán.

8.1.- Valores poblacionales y estimadores - total y promedio

Sea y_{ij} la observación *j-ésima* de la muestra *i-ésima, j=1,...,n ; i=1,...,k* , es decir, el valor de *"y"* en el elemento *j-ésimo* de la muestra *i-ésima*, entonces,

$y_{syi} = \sum_{j=1}^{n} y_{ij}$ el total de la muestra *i-ésima*

$\bar{y}_{syi} = \frac{y_{syi}}{n}$ el promedio de la muestra *i-ésima*

$Y = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij} = \sum_{i=1}^{k} y_{syi}$ el total poblacional

$\bar{Y} = \frac{Y}{N}$ el promedio poblacional

como de los *k* totales y promedios muestrales, sólo uno es el seleccionado,

$$\hat{Y}_{sy} = y_{sy1} + \sum_{i=2}^{k} y'_{syi}$$

donde y'_{syi} es un valor estimado de los totales y_{syi}, que no están en la muestra, y consiste en el promedio de los totales que si están, que en este caso es sólo uno, por lo tanto $y'_{syi} = y_{sy1} \quad , \quad i = 2,...,k$. Entonces,

$$\hat{Y}_{sy} = y_{sy1} + (k-1)y_{sy1} = k\, y_{sy1} = \frac{N}{n} y_{sy1} = N\, \bar{y}_{sy1}$$

y como hay un sólo promedio en la muestra, que en este caso es y_1 , pero puede ser cualquier otro,

$$\hat{\bar{Y}}_{sy} = \frac{\hat{Y}_{sy}}{N} = \frac{N\bar{y}_{sy1}}{N} = \bar{y}_{sy1} = \bar{y}_{sy}$$

por lo tanto, $\hat{Y}_{sy} = N\, \bar{y}_{sy}$

y para el promedio, $\hat{\bar{Y}}_{sy} = \frac{\hat{Y}_{sy}}{N} = \frac{N\, \bar{y}_{sy}}{N} = \bar{y}_{sy}$

es decir, que el estimador del promedio poblacional, es el promedio muestral.

8.2.- Varianzas de los estimadores del promedio y el total

Las varianzas de los estimadores vienen expresadas por:

$$V(\hat{\bar{Y}}_{sy}) = V(\bar{y}_{sy}) = \frac{\sum_{i=1}^{k}\left(\bar{y}_{syi} - \bar{Y}\right)^2}{k}$$

$$V(\hat{Y}_{sy}) = \frac{\sum_{i=1}^{k}\left(\hat{Y}_{syi} - Y\right)^2}{k} = \frac{\sum_{i=1}^{k}\left(N\bar{y}_{syi} - N\bar{Y}\right)^2}{k} = \frac{N^2\sum_{i=1}^{k}\left(\bar{y}_{syi} - \bar{Y}\right)^2}{k} = N^2V(\hat{\bar{Y}}_{sy}) = N^2V(\bar{y}_{sy})$$

El muestreo sistemático tiene la ventaja que la muestra está distribuida por todo el marco, y si éste se puede arreglar y ordenar de acuerdo con la variable a investigar, el muestreo sistemático puede funcionar muy bien. Cuando se investigan varias variables, donde cada una puede tener un comportamiento parecido al de las otras, el muestreo sistemático puede ser una buena opción. Sin embargo, el muestreo sistemático presenta dos problemas, el primero de ellos se presenta cuando *k* no es entero, y el segundo es la estimación de las varianzas. El cálculo de las varianzas viene expresado en función de tomar las *k* muestras posibles, pero al hacer un estudio, se toma sólo una, es decir, "no es medible" [4;148].

Para ello se han estudiado varias maneras de estimar las varianzas de los estimadores, que resultan unos mejores que otros de acuerdo con la distribución de la población. Estas son:

i.- muestreo aleatorio simple; cuando la población está ordenada al azar,

$$\hat{V}(\hat{\bar{Y}}_{sy}) = \frac{1-f}{n}\sum_{i=1}^{n}(y_i - \bar{y}_{sy})^2$$

ii.- selecciones pareadas; cuando la población tiene un comportamiento que se asemeja a una estratificación implícita, donde los estratos estarían conformados, cada uno por dos intervalos completos. Se aplica la fórmula de muestreo estratificado donde,

$$N_h = 2k \;;\; n_h = 2 \;;\; w_h = \frac{n_h}{n} = \frac{2}{n} \;;\; W_h = \frac{N_h}{N} = \frac{2}{n} \;;\; L = \frac{n}{2}$$

luego,

$$\hat{V}(\hat{\bar{Y}}_{sy}) = \sum_{h=1}^{n/2} \left(\frac{2}{n}\right)^2 \frac{2k-2}{2(2k)} \sum_{i=1}^{2} (y_{hi} - \bar{y}_h)^2$$

$$= \frac{4}{n^2} \frac{2(k-1)}{4k} \frac{1}{2} \sum_{h=1}^{n/2} (y_{h1} - y_{h2})^2$$

$$= \frac{1-f}{n^2} \sum_{h=1}^{n/2} (y_{h1} - y_{h2})^2$$

iii.- diferencias sucesivas; se trata de una modificación del anterior, pero usando todas las diferencias entre elementos sucesivos, que son *(n-1)*. Adaptando la fórmula queda,

$$\hat{V}(\hat{\bar{Y}}_{sy}) = \frac{1-f}{2n(n-1)} \sum_{g=1}^{n-1} (y_g - y_{g+1})^2$$

en cada caso, para hallar la varianza estimada del estimador del total, se hace

$$V(\hat{Y}_{sy}) = N^2 \, V(\hat{\bar{Y}}_{sy})$$

Para la aplicación de estas tres fórmulas de estimación de las varianzas de los estimadores, se considera el comportamiento u orden de la población, y en vista de lo descrito, es conveniente en los siguientes casos

Comportamiento de la Población		Fórmula a Aplicar
Orden al Azar	=>	Muestreo Aleatorio Simple
Efecto de Estratificación	=>	Diferencias Pareadas
Tendencia	=>	Diferencias Sucesivas

Otro método para solventar el problema de la estimación de varianzas, y a la vez contrarrestar algún efecto de conglomeración o estratificación de la población, es el *"muestreo sistemático replicado"*.

8.3.- Muestreo sistemático replicado

Como se mencionó anteriormente, uno de los problemas del muestreo sistemático es la estimación de la varianza de los estimadores, ya que sólo se tiene una de las *k* posibles muestras; el muestreo sistemático replicado aborda este problema generando varias muestras sistemáticas o replicaciones, todas de igual tamaño, manteniendo el tamaño

original de la muestra, es decir, que la suma de los tamaños muestrales sea *n*, luego se aplica la teoría del muestreo aleatorio simple para determinar los estimadores y sus respectivas varianzas. [6;182]

Sea *l* el total de muestras replicadas que se tomarán, y *n'* el tamaño de la muestra de cada una de ellas, de manera que *ln'=n*. Entonces,

$$k' = \frac{N}{n'} = l\left(\frac{N}{n}\right) = l\ k \qquad \text{es el intervalo de selección.}$$

De cada muestra se tiene una estimación del promedio y del total, que son:

$$\bar{y}_{syr_g} = \frac{\sum_{i=1}^{n'} y_{syr_{gi}}}{n'} \quad ; \quad \hat{Y}_{syr_g} = N\ \bar{y}_{syr_g} \quad , \ g = 1,2,...,l$$

Para estimar el promedio y el total poblacional se hace,

$$\hat{\bar{Y}}_{syr} = \frac{\sum_{g=1}^{l} \bar{y}_{syr_g}}{l} \quad ; \quad \hat{Y}_{syr} = \frac{\sum_{g=1}^{l} \hat{Y}_{syr_g}}{l} = \frac{\sum_{g=1}^{l} N\ \bar{y}_{syr_g}}{l} = N\frac{\sum_{g=1}^{l} \bar{y}_{syr_g}}{l} = N\ \hat{\bar{Y}}_{syr}$$

De manera que se tiene dividida la población en *n'* intervalos, cada uno con *k'* elementos. De modo que existen *k'* muestras sistemáticas, de las cuales se seleccionarán *l* de ellas, es decir, se tomarán *l* muestras sistemáticas de un total de *k'*; la selección de las muestras se hará de manera aleatoria simple. De modo que se tienen

$$\binom{k'}{l} = \binom{kl}{l} \qquad \text{muestras posibles.}$$

Como es una muestra aleatoria simple de muestras sistemáticas, se usan las fórmulas del muestreo aleatorio simple para las varianzas, pero usando los valores aquí obtenidos, entonces,

$$V(\hat{\bar{Y}}_{syr}) = \frac{lk-l}{(lk)l}\frac{\sum_{g=1}^{k'}\left(\bar{y}_{syr_g}-\bar{Y}\right)^2}{k'-1} = \frac{k-1}{k\,l}\frac{\sum_{g=1}^{k'}\left(\bar{y}_{syr_g}-\bar{Y}\right)^2}{k'-1} = \frac{1-f}{l}\frac{\sum_{g=1}^{k'}\left(\bar{y}_{syr_g}-\bar{Y}\right)^2}{k'-1}$$

$$V(\hat{Y}_{syr}) = V(N\ \hat{\bar{Y}}_{syr}) = N^2\ V(\hat{\bar{Y}}_{syr}) = N^2\ \frac{k-1}{k\,l}\frac{\sum_{g=1}^{k'}\left(\bar{y}_{syr_g}-\bar{Y}\right)^2}{k'-1} = N^2\ \frac{1-f}{l}\frac{\sum_{g=1}^{k'}\left(\bar{y}_{syr_g}-\bar{Y}\right)^2}{k'-1}$$

y para las varianzas estimadas,

$$\hat{V}(\hat{\bar{Y}}_{syr}) = \frac{lk-l}{(lk)l}\frac{\sum_{g=1}^{l}\left(\bar{y}_{syr_g}-\bar{y}_{syr}\right)^2}{l-1} = \frac{k-1}{k\,l}\frac{\sum_{g=1}^{l}\left(\bar{y}_{syr_g}-\bar{y}_{syr}\right)^2}{l-1} = \frac{1-f}{l}\frac{\sum_{g=1}^{l}\left(\bar{y}_{syr_g}-\bar{y}_{syr}\right)^2}{l-1}$$

$$V(\hat{Y}_{syr}) = V(N\,\hat{\bar{Y}}_{syr}) = N^2\,V(\hat{\bar{Y}}_{syr}) = N^2\,\frac{k-1}{k\,l}\frac{\sum_{g=1}^{l}\left(\bar{y}_{syr_g}-\bar{y}_{syr}\right)^2}{l-1} = N^2\,\frac{1-f}{l}\frac{\sum_{g=1}^{l}\left(\bar{y}_{syr_g}-\bar{y}_{syr}\right)^2}{l-1}$$

Como puede observarse, esté método consiste en una muestra aleatoria simple de *l* muestras sistemáticas, de un total de *k'*, en consecuencia la estructura de los estimadores y varianzas responden a éste criterio, es decir,

8.4.- *k* no entero

En la mayoría de los casos resulta que al generar *k*, éste no es entero, y se presenta el problema de dividir el marco en *n* intervalos con un número no entero de elementos. Este hecho introduce una perturbación en la teoría del muestreo, que puede ser despreciable cuando el tamaño de la muestra es grande, aunque se espera que no sea grande en ningún caso. Existen varias formas de abordar este problema,

i.- tomar la parte entera de *k* ($[k]$), entonces resultarán unas muestras de tamaño *n* y otras de tamaño *n*+1. Se divide la población en *n+1* intervalos, los primeros *n* de $[k]$ elementos cada uno, y el último de $N-[k]n$; si $r=1,2,...,N-[k]n$, la muestra tendrá *n+1* elementos, y si $r=N-[k]n+1$, $N-[k]n+2,...,[k]$, la muestra tendrá *n* elementos.

ii.- sumar 1 a la parte entera de *k*, entonces el tamaño de muestra variará entre *n*-2, *n*-1 y *n* elementos.

Si $\frac{N}{[k]+1} < n-1 \Rightarrow$ resultan *n*-1 intervalos;

si $r = 1,2,...,N-([k]+1)(n-2)$ $\Rightarrow$ la muestra tendrá *n*-1 elementos

si $r = N-([k]+1)(n-2)+1, N-([k]+1)(n-2)+2,..., [k]+1$ $\Rightarrow$ la muestra tendrá *n*-2 elementos

Si $\frac{N}{[k]+1} = n-1 \Rightarrow$ resultan *n*-1 intervalos, y la muestra siempre será de *n*-1 elementos

Si $\frac{N}{[k]+1} > n-1 \Rightarrow$ resultan n intervalos;

si $r = 1,2,...,,...,N-([k]+1)(n-1)$ $\Rightarrow$ la muestra tendrá *n* elementos

si $r = N-([k]+1)(n-1)+1, N-([k]+1)(n-1)+2,..., [k]+1$ $\Rightarrow$ la muestra tendrá *n*-1 elementos

es decir, que aunque el intervalo de selección siempre sería igual ([*k*]+1), el tamaño de la muestra puede ser *n*-2, *n*-1 o *n*, lo que hace que el método resulte poco confiable en cuanto al control del tamaño de la muestra.

iii.- método del intervalo fraccional

sea $k^* = 10^m k$, y se selecciona un número aleatorio entero (r) entre *1* y *k**, entonces,

el 1er elemento en la muestra es el que corresponde al N° [$10^{-m} r$] en el marco muestral
el 2do elemento en la muestra es el que corresponde al N° [$10^{-m} r + k$] en el marco muestral
el 3er elemento en la muestra es el que corresponde al N° [$10^{-m} r + 2k$] en el marco muestral
⋮
el *n-ésimo* elemento en la muestra es el que corresponde al N° [$10^{-m} r + (n-1)k$] en el marco muestral

iv.- método circular

éste método fue sugerido por Lahiri en 1952, y considera el universo dispuesto en un círculo, se toma *k* como el entero más cercano a *N/n*, entonces se selecciona un número aleatorio *r* entre *1* y *N*, entonces, el 1er elemento es el *r*-ésimo, y luego se toma cada *k*-ésimo elemento a partir de allí, hasta completar los *n* elementos de la muestra. [2; 258]

Las opciones (iii) y (iv) son las que garantizan que todas las muestras sean de tamaño *n*, manteniendo la aleatoriedad de la selección y la equiprobabilidad de selección de todos los elementos de la lista. De cualquier modo, la opción (i) resulta muy práctica cuando la selección se hace en el terreno, ya que las instrucciones son más sencillas que para el resto.

9.- MUESTREO ALEATORIO SISTEMÁTICO ESTRATIFICADO

Este diseño consiste en principio, de un muestreo estratificado, es decir, se divide la población en L estratos, cada uno de tamaño N_h, $h=1,...,L$. De cada uno de ellos se toma una muestra sistemática de tamaño n_h.

9.1.- Total y promedio poblacional y muestral

Se denotará por y_{hji} al valor de la variable *"y"* en la *j-ésimo* elemento de la *i-ésima* muestra del estrato h, $h=1,...,L$; $i=1,...,k_h$; $j=1,...,n_h$, donde $k_h = \frac{N_h}{n_h}$, entonces,

$y_{sy_{hi}} = \sum_{j=1}^{n_h} y_{hij}$ el total de la muestra *i-ésima* del estrato h

$\bar{y}_{sy_{hi}} = \frac{y_{sy_{hi}}}{n_h}$ el promedio de la muestra *i-ésima* del estrato h

$Y_h = \sum_{i=1}^{k_h}\sum_{j=1}^{n_h} y_{hij} = \sum_{i=1}^{k_h} y_{sy_{hi}}$ el total poblacional del estrato h

$\bar{Y}_h = \frac{Y_h}{N_h}$ el promedio poblacional del estrato h

$Y = \sum_{h=1}^{L} Y_h$ el total poblacional

$\bar{Y} = \frac{Y}{N}$ el promedio poblacional

los valores muestrales equivalentes son,

$y_{sy_h} = \sum_{j=1}^{n_h} y_{hi}$ el total muestral del estrato h

$\bar{y}_{sy_h} = \frac{y_h}{n_h}$ el promedio muestral del estrato h

$y_{sy} = \sum_{h=1}^{L} y_h$ el total muestral

$\bar{y}_{sy} = \frac{y_{sy}}{n} = \sum_{h=1}^{L} w_h \bar{y}_{sy_h}$ el promedio muestral

donde $w_h = \frac{n_h}{N_h}$.

9.2.- Estimadores y varianzas

En cada estrato se toma una muestra aleatoria sistemática e independiente de elementos, entonces, los estimadores del total y del promedio del estrato *h*, respectivamente son,

$$\hat{Y}_{sy_h} = N_h\, \bar{y}_{sy_h} \qquad ; \qquad \hat{\bar{Y}}_{sy_h} = \bar{y}_{sy_h}$$

y los estimadores del total y del promedio poblacional,

$$\hat{Y}_{stsy} = \sum_{h=1}^{L} \hat{Y}_{sy_h} = \sum_{h=1}^{L} N_h\, \bar{y}_{sy_h}$$

$$\hat{\bar{Y}}_{stsy} = \frac{1}{N}\, \hat{Y}_{stsy} = \frac{1}{N} \sum_{h=1}^{L} N_h\, \bar{y}_{sy_h} = \sum_{h=1}^{L} W_h\, \bar{y}_{sy_h}$$

Como puede observarse, los valores poblacionales, muestrales y los estimadores, no difieren en gran medida de los del capítulo 4 (Muestreo Aleatorio Estratificado), pero no así las varianzas. Veamos las varianzas de los estimadores del total y del promedio de los estratos.

$$V(\hat{\bar{Y}}_{sy_h}) = \frac{\sum_{i=1}^{k_h} (y_{hi} - \bar{Y}_h)^2}{k_h} \qquad , \quad \text{donde} \quad k_h = \frac{N_h}{n_h}$$

$$V(\hat{Y}_{sy_h}) = N_h^2\, V(\hat{\bar{Y}}_{sy_h})$$

para estimar estas varianzas, se tienen los mismos problemas y tratamientos que en el muestreo sistemático, de tal manera que resulta:

i.- muestreo aleatorio simple; cuando la población en el estrato *h* tiene un orden al azar,

$$\hat{V}(\hat{\bar{Y}}_{sy_h}) = \frac{1 - f_h}{n_h} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$$

ii.- selecciones pareadas; cuando la población en el estrato h tiene un comportamiento donde los intervalos de selección se comportan como estratos.

$$\hat{V}(\hat{\bar{Y}}_{sy_h}) = \frac{1 - f_h}{n_h^2} \sum_{t=1}^{n_h/2} (y_{ht_1} - y_{ht_2})^2$$

iii.- diferencias sucesivas; cuando existe alguna tendencia en la población,

$$\hat{V}(\hat{\bar{Y}}_{sy_h}) = \frac{1 - f_h}{2\, n_h (n_h - 1)} \sum_{g=1}^{n_h - 1} \left(y_{hg} - y_{h(g+1)}\right)^2$$

iv.- aplicando muestreo sistemático replicado;

$$V(\hat{\bar{Y}}_{syr_h}) = \frac{1-f_h}{l_h} \frac{\sum_{r=1}^{k'_h} (y_{syr_{hg}} - \bar{Y}_h)^2}{k'_h - 1}$$

$$\hat{V}(\bar{Y}_{syr_h}) = \frac{1-f_h}{l_h} \frac{\sum_{g=1}^{l_h} (y_{syr_{hg}} - \bar{y}_{syr_h})^2}{l_h - 1}$$

en cada caso,

$$V(\hat{Y}_{sy_h}) = N^2 V(\hat{\bar{Y}}_{sy_h}) \qquad ; \qquad \hat{V}(\hat{Y}_{sy_h}) = N^2 \hat{V}(\hat{\bar{Y}}_{sy_h})$$

Las varianzas de los estimadores son:

$$V(\hat{Y}_{stsy}) = \sum_{h=1}^{L} V(N_h \hat{\bar{Y}}_{sy_h}) = \sum_{h=1}^{L} N_h^2 V(\hat{\bar{Y}}_{sy_h}) = \sum_{h=1}^{L} N_h^2 \frac{\sum_{i=1}^{k_h} (y_{hi} - \bar{Y}_h)^2}{k_h}$$

$$V(\hat{\bar{Y}}_{stsy}) = \frac{1}{N^2} V(\hat{Y}_{stsy}) = \frac{1}{N^2} \sum_{h=1}^{L} N_h^2 \frac{\sum_{i=1}^{k_h} (y_{hi} - \bar{Y}_h)^2}{k_h} = \sum_{h=1}^{L} W_h^2 \frac{\sum_{i=1}^{k_h} (y_{hi} - \bar{Y}_h)^2}{k_h}$$

y los estimadores de estas varianzas,

$$\hat{V}(\hat{Y}_{stsy}) = \sum_{h=1}^{L} \hat{V}(N_h \hat{\bar{Y}}_h) = \sum_{h=1}^{L} N_h^2 \hat{V}(\hat{\bar{Y}}_{sy_h})$$

$$\hat{V}(\hat{\bar{Y}}_{stsy}) = \frac{1}{N^2} \hat{V}(\hat{Y}_{stsy}) = \sum_{h=1}^{L} W_h^2 \hat{V}(\hat{\bar{Y}}_{sy_h})$$

Según las varianzas estimadas de $\hat{\bar{Y}}_{sy_h}$ que se enunciaron anteriormente, se tiene,

i.- muestreo aleatorio simple;

$$\hat{V}(\hat{Y}_{stsy}) = \sum_{h=1}^{L} \frac{N_h^2 (1-f_h)}{n_h} s_h^2$$

$$\hat{V}(\hat{Y}_{stsy}) = \sum_{h=1}^{L} \frac{W_h^2 (1-f_h)}{n_h} s_h^2$$

ii.- selecciones pareadas;

$$\hat{V}(\hat{Y}_{stsy}) = \sum_{h=1}^{L} \frac{N_h^2(1-f_h)}{n_h^2} \sum_{t=1}^{n_h/2} (y_{ht_1} - y_{ht_2})^2$$

$$\hat{V}(\hat{\bar{Y}}_{stsy}) = \sum_{h=1}^{L} \frac{W_h^2(1-f_h)}{n_h^2} \sum_{t=1}^{n_h/2} (y_{ht_1} - y_{ht_2})^2$$

iii.- diferencias sucesivas;

$$\hat{V}(\hat{Y}_{stsy}) = \sum_{h=1}^{L} \frac{N_h^2(1-f_h)}{2\,n_h\,(n_h-1)} \sum_{g=1}^{n_h} \left(y_{hg} - y_{h(g+1)}\right)^2$$

$$\hat{V}(\hat{Y}_{stsy}) = \sum_{h=1}^{L} \frac{W_h^2(1-f_h)}{2\,n_h\,(n_h-1)} \sum_{g=1}^{n_h} \left(y_{hg} - y_{h(g+1)}\right)^2$$

iv.- muestreo sistemático replicado;

$$\hat{V}(\hat{Y}_{stsyr}) = \sum_{h=1}^{L} \frac{N_h^2(1-f_h)}{l_h\,(l_h-1)} \sum_{r=1}^{l_h} \left(\bar{y}_{h_r} - \hat{\bar{Y}}_{syr_h}\right)^2$$

$$\hat{V}(\hat{\bar{Y}}_{stsyr}) = \sum_{h=1}^{L} \frac{W_h^2(1-f_h)}{l_h\,(l_h-1)} \sum_{r=1}^{l_h} \left(\bar{y}_{h_r} - \hat{\bar{Y}}_{syr_h}\right)^2$$

En los casos de estratos donde k_h no es entero, se aplican los mismos tratamientos mostrados en el capítulo anterior.

10.- MUESTREO MONOETÁPICO ALEATORIO DE CONGLOMERADOS

Sea un universo de tamaño M_0 que se divide en N partes, que se denominaran conglomerados, de los cuales se seleccionan n y se incluyen en la muestra todos los elementos de los conglomerados seleccionados. Se denotará por

y_{ij} = al valor asociado al elemento *j-ésimo* del conglomerado *i-ésimo*

M_i = al total de elementos del conglomerado *i-ésimo*

donde $i=1,...,N$; $j=1,...,M_i$

sean los siguientes valores poblacionales, relativos a la variable y,

$M_0 = \sum_{i=1}^{N} M_i$ es el total de elementos en el universo

$Y_i = \sum_{j=1}^{M_i} y_{ij}$ es el total poblacional en el conglomerado *i-ésimo*

$\overline{Y}_i = \frac{Y_i}{M_i}$ es el promedio poblacional en el conglomerado *i-ési*mo

$Y = \sum_{i=1}^{N} Y_i$ es el total poblacional

$\overline{Y} = \frac{Y}{N}$ es el promedio poblacional por conglomerado

$\overline{\overline{Y}} = \frac{Y}{M_0}$ es el promedio poblacional por elemento

y sean

$m_0 = \sum_{i=1}^{n} M_i$ el total de elementos en la muestra

$y = \sum_{i=1}^{n} Y_i$ el total muestral

$\overline{y} = \frac{y}{n}$ el promedio muestral por conglomerado

$\overline{\overline{y}} = \frac{y}{m_0}$ el promedio muestral por elemento

10.1.- Selección de conglomerados con iguales probabilidades

Se presentarán dos casos, el primero consiste en considerar que todos y cada uno de los conglomerados tienen la misma probabilidad de ser seleccionados, es decir, se selecciona una Muestra Aleatoria Simple de Conglomerados.

10.1.1.- Estimadores

Bajo el principio que se ha venido enfocando en este trabajo, de **"asignar a lo que no está en la muestra el promedio de lo que sí está"** (*principio 1*), para estimar el total poblacional, se trabaja con los totales de los conglomerados, esto es,

$$\hat{Y} = \sum_{i=1}^{n} Y_i + \sum_{j=n+1}^{N} Y'_j$$

donde Y'_i es la estimación del total de los conglomerados que no están incluidos en la muestra. A estos se les asignará el promedio de los conglomerados que si están en la muestra, es decir,

$$Y'_j = \frac{\sum_{i=1}^{n} Y_i}{n} = \bar{y} \qquad \forall\ j\ ,\ j = n+1,\ldots,N$$

por lo tanto, $Y'_{n+1} = Y'_{n+2} = \ldots = Y'_N = \bar{y}$. Entonces,

$$\hat{Y} = \sum_{i=1}^{n} Y_i + (N-n)\bar{y} = n\,\bar{y} + (N-n)\bar{y} = N\,\bar{y} = N\,\frac{\sum_{i=1}^{n} Y_i}{n} \qquad (10.1)$$

el estimador del promedio poblacional por conglomerados es,

$$\hat{\bar{Y}} = \frac{\hat{Y}}{N} = \frac{\sum_{i=1}^{n} Y_i}{n} = \bar{y} \qquad (10.2)$$

y el estimador del promedio poblacional por elemento es,

$$\hat{\bar{\bar{Y}}} = \frac{\hat{Y}}{M_0} = \frac{\hat{Y}}{\sum_{i=1}^{N} M_i} = \frac{N}{n}\,\frac{\sum_{i=1}^{n} Y_i}{M_0} \qquad (10.3)$$

pero si no se conoce el total de elementos M_0, se estima, y resulta lo siguiente,

$$\hat{\bar{\bar{Y}}} = \frac{\hat{Y}}{\hat{M}_0} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} M_i} = \frac{\frac{N}{n}\sum_{i=1}^{n} Y_i}{\frac{N}{n}\sum_{i=1}^{n} M_i} = \frac{y}{m_0} = \bar{\bar{y}}$$

El Principio 1 también puede aplicarse a los elementos, es decir, asignar a los elementos que no están en la muestra, el promedio de los que si están; así se tienen los "estimadores de razón al tamaño", que se desarrollan a continuación.

$$\bar{\bar{Y}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{M_i} y_{ij} + \sum_{k=n+1}^{N}\sum_{j=1}^{M_k} y_{kj}}{\sum_{i=1}^{N} M_i} \qquad \text{y su estimador} \qquad \hat{\bar{\bar{Y}}}_{RT} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{M_i} y_{ij} + \sum_{k=n+1}^{N}\sum_{j=1}^{M_k} y'_{kj}}{\sum_{i=1}^{N} M_i} \qquad (10.4)$$

si se considera que

$$y'_{kj} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{M_i} y_{ij}}{\sum_{i=1}^{n} M_i} = \bar{\bar{y}} \qquad \forall k, j \quad , k = n+1,\ldots,N \quad , j = 1,\ldots,M_i$$

que es el promedio muestral por elementos, entonces,

$$\hat{\bar{\bar{Y}}}_{RT} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{M_i} y_{ij} + \sum_{k=n+1}^{N} M_k \bar{\bar{y}}}{\sum_{i=1}^{N} M_i} = \frac{\frac{\left(\sum_{i=1}^{n} M_i\right)}{\left(\sum_{i=1}^{n} M_i\right)}\left(\sum_{i=1}^{n}\sum_{j=1}^{M_i} y_{ij}\right) + \sum_{k=n+1}^{N} M_k \bar{\bar{y}}}{\sum_{i=1}^{N} M_i} = \frac{\left(\sum_{i=1}^{n} M_i\right)\bar{\bar{y}} + \left(\sum_{k=n+1}^{N} M_k\right)\bar{\bar{y}}}{\sum_{i=1}^{N} M_i} = \left[\frac{\sum_{i=1}^{n} M_i + \sum_{k=n+1}^{N} M_k}{\sum_{i=1}^{N} M_i}\right]\bar{\bar{y}} = \bar{\bar{y}}$$

es decir que $\hat{\bar{\bar{Y}}}_{RT} = \bar{\bar{y}}$, y entonces,

$$\hat{Y}_{RT} = M_0 \hat{\bar{\bar{Y}}}_{RT} = M_0 \bar{\bar{y}} = M_0 \frac{\sum_{i=1}^{n}\sum_{j=1}^{M_i} y_{ij}}{\sum_{i=1}^{n} M_i} = \sum_{i=1}^{N} M_i \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} M_i} \qquad (10.5)$$

a estos estimadores se le denomina respectivamente **"estimador de razón al tamaño del promedio poblacional por elementos"** y **"estimador de razón al tamaño del total poblacional"**, ya que usan la variable *"tamaño del*

conglomerado" como variable auxiliar de un estimador de razón. Igualmente, el **"estimador de razón al tamaño del promedio poblacional por conglomerados"** es

$$\hat{\bar{Y}}_{RT} = \frac{\hat{Y}_{RT}}{N} = \frac{1}{N}\sum_{i=1}^{N} M_i \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} M_i} = \frac{M_0}{N}\bar{\bar{y}} = \frac{M_0}{N}\hat{\bar{\bar{Y}}}_{RT} \qquad (10.6)$$

10.1.2.- Varianzas de los estimadores

La varianza de $\hat{Y}$ resulta de tomar cada Y_i ,$i=1,...,N$, como una observación, de manera que, como es una muestra aleatoria simple de esas observaciones, se obtiene la misma fórmula del muestreo aleatorio simple,

$$V(\hat{Y}) = N^2 \frac{(N-n)}{N\,n} \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}{N-1} = N^2(1-f)\frac{S^2}{n} \qquad \text{donde} \quad S^2 = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}{N-1} \qquad (10.7)$$

y por ende,

$$V(\hat{\bar{Y}}) = V\left(\frac{\hat{Y}}{N}\right) = \frac{1}{N^2} V(\hat{Y}) = (1-f)\frac{S^2}{n} \qquad (10.8)$$

$$V(\hat{\bar{\bar{Y}}}) = V\left(\frac{\hat{Y}}{M_0}\right) = \frac{1}{M_0^2} = V(\hat{Y}) = \frac{N^2(1-f)}{M_0^2}\frac{S^2}{n} \qquad (10.9)$$

y las varianzas estimadas son:

$$\hat{V}(\hat{Y}) = \frac{N^2(N-n)}{N\,n} \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} = \frac{N^2(1-f)}{n} \frac{\sum_{i=1}^{n}(Y_i - \bar{y})^2}{n-1} = N^2(1-f)\frac{s^2}{n}$$

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

$$\hat{V}(\hat{\bar{\bar{Y}}}) = \hat{V}\left(\frac{\hat{Y}}{M_0}\right) = \frac{N^2(1-f)}{M_0^2}\frac{s^2}{n}$$

$$\text{donde} \qquad s^2 = \frac{\sum(Y_i - \bar{y})^2}{n-1}$$

Para el caso de los estimadores de razón al tamaño,

$$\hat{Y}_{RT} = \sum_{i=1}^{N} M_i \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} M_i} \quad , \quad \hat{\bar{Y}}_{RT} = \frac{\hat{Y}_{RT}}{N} \quad , \quad \hat{\bar{\bar{Y}}}_{RT} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} M_i} = \frac{\hat{Y}_{RT}}{\left(\sum_{i=1}^{N} M_i\right)} = \frac{\hat{Y}_{RT}}{M_0}$$

Nótese que es equivalente a utilizar la fórmula del estimador de razón desarrollados en el capítulo 7 (Estimadores Indirectos), pero aplicando el tamaño del conglomerado (M_i) como variable auxiliar, esto es,

$$\hat{Y}_{RT} = \frac{y}{x} X = \left(\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}\right) \sum_{i=1}^{N} x_i = \left(\frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} M_i}\right) \sum_{i=1}^{N} M_i$$

y sustituyendo x_i por M_i en las fórmulas de la esperanza, sesgo, varianza y error cuadrático medio, se tiene,

$$E(\hat{Y}_{RT}) = Y\left[1 + \frac{1-f}{n}\left(\frac{\sum_{i=1}^{N} M_i^2 - \frac{M_0^2}{N}}{\frac{N-1}{\left(\frac{M_0^2}{N^2}\right)}} - \frac{\sum_{i=1}^{N} M_i Y_i - M_0 \bar{Y}}{\frac{N-1}{\left(\frac{M_0}{N}\right)\left(\frac{Y}{N}\right)}}\right)\right] = Y\left[1 + \left(\frac{1-f}{n}\right)\left(\frac{N^2}{M_0^2 Y}\right)\left(\frac{Y\sum_{i=1}^{N} M_i^2 - M_0 \sum_{i=1}^{N} M_i Y_i}{N-1}\right)\right]$$

por lo tanto el sesgo resulta,

$$B(\hat{Y}_{RT}) = (-1)\left[\left(\frac{1-f}{n}\right)\left(\frac{N^2}{M_0^2}\right)\left(\frac{Y\sum_{i=1}^{N} M_i^2 - M_0 \sum_{i=1}^{N} M_i Y_i}{N-1}\right)\right]$$

$$ECM(\hat{Y}_{RT}) \approx N^2\left(\frac{1-f}{n}\right)\left[S_y^2 + \left(\frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} M_i}\right)^2 \left(\sum_{i=1}^{N} M_i^2 - N\frac{M_0^2}{N^2}\right) - 2\left(\frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} M_i}\right)\left(\sum_{i=1}^{N} M_i Y_i - N\bar{Y}\frac{M_0}{N}\right)\right]$$

$$\approx N^2\left(\frac{1-f}{n}\right)\left(\frac{1}{N-1}\right)\left[\left(\sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2\right) + \left(\frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} M_i}\right)^2 \left(\sum_{i=1}^{N} M_i^2 - \frac{M_0^2}{N}\right) - 2\left(\frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} M_i}\right)\left(\sum_{i=1}^{N} M_i Y_i - M_0\bar{Y}\right)\right]$$

$$\approx N^2\left(\frac{1-f}{n}\right)\left(\frac{1}{N-1}\right)\left[\left(\sum_{i=1}^{N}Y_i^2 - N\bar{Y}^2\right)+\bar{\bar{Y}}^2\left(\sum_{i=1}^{N}M_i^2-\frac{M_0^2}{N}\right)-2\bar{\bar{Y}}\left(\sum_{i=1}^{N}M_iY_i - M_0\bar{Y}\right)\right]$$

$$\approx N^2\left(\frac{1-f}{n}\right)\left(\frac{1}{N-1}\right)\left[\sum_{i=1}^{N}Y_i^2 - N\bar{Y}^2+\bar{\bar{Y}}^2\sum_{i=1}^{N}M_i^2-\frac{\bar{\bar{Y}}^2}{N}-2\bar{\bar{Y}}\sum_{i=1}^{N}M_iY_i+2M_0\bar{Y}\bar{\bar{Y}}\right]$$

pero

$N\bar{Y}^2 = N\frac{Y^2}{N^2}=\frac{Y^2}{N}$ entonces, $ECM(\hat{\bar{Y}}_{RT})\approx N^2\left(\frac{1-f}{n}\right)\left(\frac{1}{N-1}\right)\left[\sum_{i=1}^{N}Y_i^2-2\frac{\bar{\bar{Y}}^2}{N}+\bar{\bar{Y}}^2\sum_{i=1}^{N}M_i^2-2\bar{\bar{Y}}\sum_{i=1}^{N}M_iY_i+2M_0\bar{Y}\bar{\bar{Y}}\right]$

y $\frac{Y^2}{N}=\frac{M_0^2}{M_0^2}\frac{Y^2}{N}=M_0^2\frac{\bar{\bar{Y}}^2}{N}$ por lo tanto, $ECM(\hat{\bar{Y}}_{RT})\approx N^2\left(\frac{1-f}{n}\right)\left(\frac{1}{N-1}\right)\left[\sum_{i=1}^{N}Y_i^2+\bar{\bar{Y}}^2\sum_{i=1}^{N}M_i^2-2\bar{\bar{Y}}\sum_{i=1}^{N}M_iY_i\right]$

luego, $$ECM(\hat{\bar{Y}}_{RT})\approx N^2\left(\frac{1-f}{n}\right)\frac{\sum_{i=1}^{N}\left(Y_i-M_i\bar{\bar{Y}}\right)^2}{N-1} \tag{10.10}$$

y $$E\hat{C}M(\hat{\bar{Y}}_{RT})\approx N^2\left(\frac{1-f}{n}\right)\frac{\sum_{i=1}^{n}\left(Y_i-M_i\hat{\bar{\bar{Y}}}_{RT}\right)^2}{n-1} \tag{10.11}$$

Análogamente, para el promedio por conglomerado,

$$E(\hat{\bar{\bar{Y}}}_{RT})=\bar{Y}\left[1+\left(\frac{1-f}{n}\right)\left(\frac{N^2}{M_0^2Y}\right)\left(\frac{Y\sum_{i=1}^{N}M_i^2-M_0\sum_{i=1}^{N}M_iY_i}{N-1}\right)\right] \tag{10.12}$$

$$B(\hat{\bar{\bar{Y}}}_{RT})=(-1)\left[\left(\frac{1-f}{n}\right)\left(\frac{N}{M_0^2}\right)\left(\frac{Y\sum_{i=1}^{N}M_i^2-M_0\sum_{i=1}^{N}M_iY_i}{N-1}\right)\right] \tag{10.13}$$

$$ECM(\hat{\bar{\bar{Y}}}_{RT})\approx\left(\frac{1-f}{n}\right)\frac{\sum_{i=1}^{N}\left(Y_i-M_i\bar{\bar{Y}}\right)^2}{N-1} \tag{10.14}$$

$$E\hat{C}M(\hat{\bar{\bar{Y}}}_{RT})\approx\left(\frac{1-f}{n}\right)\frac{\sum_{i=1}^{n}\left(Y_i-M_i\hat{\bar{\bar{Y}}}_{RT}\right)^2}{n-1} \tag{10.15}$$

y para el promedio por elemento,

$$E(\hat{\bar{Y}}_{RT}) = \bar{\bar{Y}}\left[1+\left(\frac{1-f}{n}\right)\left(\frac{N^2}{M_0^2 Y}\right)\left(\frac{Y\sum_{i=1}^{N} M_i^2 - M_0\sum_{i=1}^{N} M_i Y_i}{N-1}\right)\right] \tag{10.16}$$

$$B(\hat{\bar{Y}}_{RT}) = (-1)\left[\left(\frac{1-f}{n}\right)\left(\frac{N^2}{M_0^3}\right)\left(\frac{Y\sum_{i=1}^{N} M_i^2 - M_0\sum_{i=1}^{N} M_i Y_i}{N-1}\right)\right] \tag{10.17}$$

$$ECM(\hat{\bar{Y}}_{RT}) \approx \frac{N^2}{M_0^2}\left(\frac{1-f}{n}\right)\frac{\sum_{i=1}^{N}\left(Y_i - M_i\bar{\bar{Y}}\right)^2}{N-1} \tag{10.18}$$

$$\hat{ECM}(\hat{\bar{Y}}_{RT}) \approx \frac{N^2}{M_0^2}\left(\frac{1-f}{n}\right)\frac{\sum_{i=1}^{n}\left(Y_i - M_i\hat{\bar{Y}}_{RT}\right)^2}{n-1} \tag{10.19}$$

10.2.- Selección de conglomerados con probabilidades desiguales

El segundo caso, que se estudiará a continuación, es cuando los conglomerados tienen diferente probabilidad de selección, es decir, basta con que uno de los conglomerados tenga diferente probabilidad de selección, para que se diferencia del caso estudiado en 10.1. Se pueden presentar diversos casos, el más sencillo es cuando se asigna la probabilidad de selección de cada conglomerado en función de su tamaño, es decir, que la probabilidad de selección coincide su tamaño relativo; entonces se dice que la probabilidad es proporcional al tamaño. También se pueden fijar las mismas en función de otros criterios diversos.

Para el primer caso se tiene que la probabilidad de selección es,

$$P(\text{selección del conglomerado i-ésimo}) = \frac{M_i}{M_0}$$

el estimador se denotará por $\hat{Y}_{PPS}$ (para el total), y la selección se puede hacer con o sin reemplazamiento, aquí se desarrollará la selección con reemplazamiento.

En principio se tiene que,

$$Y = \frac{1}{N}\left[M_0\left(\frac{Y_1}{M_1}\right) + M_0\left(\frac{Y_2}{M_2}\right) + \ldots + M_0\left(\frac{Y_N}{M_N}\right)\right] = \frac{M_0}{N}\sum_{i=1}^{N}\overline{Y}_i$$

entonces, $\hat{Y}_{PPS} = \frac{M_0}{N}\left[\sum_{i=1}^{n}\overline{Y}_i + \sum_{j=n+1}^{N}\overline{Y}_j'\right]$, aplicando el principio 1, $\overline{Y}_j' = \frac{\sum_{i=1}^{n} Y_i}{n} \quad \forall\ j\ ,\ j = n+1,\ldots,N$ siendo *n* el total de conglomerados en la muestra y $Y_1, Y_2, \ldots, Y_n$ los valores muestrales.

$$\hat{Y}_{PPS} = \frac{M_0}{N}\left[\sum_{i=1}^{n}\overline{Y}_i + (N-n)\frac{\sum_{i=1}^{n}\overline{Y}_i}{n}\right] = \frac{M_0}{N}\left[\left(\sum_{i=1}^{n}\overline{Y}_i\right)\left(1+\frac{N-n}{n}\right)\right] = \frac{M_0}{N}\left(\sum_{i=1}^{n}\overline{Y}_i\right)\left(\frac{n+N-n}{n}\right) = \frac{M_0}{n}\left(\sum_{i=1}^{n}\overline{Y}_i\right)$$

Sin embargo, a veces no se conoce con exactitud el tamaño de todos los conglomerados, o se tienen los tamaños de períodos anteriores, o simplemente se considera que es preferible utilizar una "medida de tamaño" que está altamente correlacionada con Y_i, distinta del total de elementos. Por ejemplo, en una encuesta de consumo de un producto alimenticio, el "tamaño" de un área residencial puede medirse en el total de viviendas, total de personas, total de supermercados y abastos. Para estos casos, se generalizará el planteamiento inicial, y a dicha medida de tamaño del conglomerado i-ésimo se denotará por M_i' y la probabilidad de selección será,

$$P(\text{selección del conglomerado i-ésimo}) = z_i = \frac{M_i'}{M_0'} \quad \text{donde} \quad M_0' = \sum_{i=1}^{N} M_i'$$

y un caso particular es cuando $M_i' = M_i$ para todo *i, i=1,2,...,N*.

Generalizando aún más, z_i puede ser cualquier probabilidad prefijada. Donde $z_i > 0 \ \forall i, i = 1,2,\ldots,N$ y $\sum_{i=1}^{N} z_i = 1$. Dichos estimadores se denotarán por $\hat{Y}_{PPZ}$ (para el total), e igual que para el caso anterior, la selección puede hacerse con o sin reemplazamiento. A continuación se desarrollará el primero de ellos, *selección de conglomerados con probabilidades diferentes con reemplazamiento*.

Sea

$$Y = \frac{1}{N}\left[\frac{Y_1}{z_1} + \frac{Y_2}{z_2} + \ldots + \frac{Y_N}{z_N}\right] = \frac{1}{N}\sum_{i=1}^{N}\frac{Y_i}{z_i}$$ y el estimador,

$$\hat{Y}_{PPZ} = \frac{1}{N}\left[\sum_{i=1}^{n}\frac{Y_i}{z_i} + \sum_{j=n+1}^{N}\frac{Y_j'}{z_i}\right]$$, aplicando el principio 1, $\overline{Y}_j' = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{z_i} \quad \forall\ j\ ,\ j = n+1,\ldots,N$

$$\hat{Y}_{PPZ} = \frac{1}{N}\left[\sum_{i=1}^{n}\frac{Y_i}{z_i} + \frac{1}{n}\sum_{j=n+1}^{N}\sum_{i=1}^{N}\frac{Y_i}{z_i}\right] = \frac{1}{N}\left[\sum_{i=1}^{n}\frac{Y_i}{z_i} + \frac{N-n}{n}\sum_{i=1}^{n}\frac{Y_i}{z_i}\right] = \frac{1}{N}\left(\sum_{i=1}^{n}\frac{Y_i}{z_i}\right)\left(\frac{n+N-n}{n}\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{z_i}$$

Nótese que si se sustituye z_i por M_i/M_0 se obtiene $\hat{Y}_{PPS}$, lo que confirma que es un caso particular.

Para hallar la esperanza,

$$E\left(\hat{Y}_{PPZ}\right) = E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{z_i}\right] = \frac{1}{n}E\left[\sum_{i=1}^{n}\frac{Y_i}{z_i}\right]$$, como la selección es con reemplazamiento, se tienen N^n muestras posibles, entonces,

$E\left(\hat{Y}_{PPZ}\right) = \frac{1}{n}\sum_{k=1}^{N^n}\sum_{i=1}^{n}P_k\frac{Y_i}{z_i}$, donde P_k es la probabilidad de selección de la *k-ésima* muestra , $k=1,2,...,N^n$

$$= \frac{1}{n}\sum_{i=1}^{N} n\frac{Y_i}{z_i}\left(z_i\sum_{j_1=1}^{N}\sum_{j_2=1}^{N}\cdots\sum_{j_{(n-1)}=1}^{N} z_{j_1}z_{j_2}\ldots z_{j_{(n-1)}}\right) = \frac{1}{n}\sum_{i=1}^{N} nY_i\left(\sum_{j_1=1}^{N}\sum_{j_2=1}^{N}\cdots\sum_{j_{(n-1)}=1}^{N} z_{j_1}z_{j_2}\ldots z_{j_{(n-1)}}\right)$$

$$= \sum_{i=1}^{N} Y_i\left(\sum_{j_1=1}^{N}\sum_{j_2=1}^{N}\cdots\sum_{j_{(n-1)}=1}^{N} z_{j_1}z_{j_2}\ldots z_{j_{(n-1)}}\right) = \sum_{i=1}^{N} Y_i\left(\left(\sum_{j_1=1}^{N} z_{j_1}\right)\left(\sum_{j_2=1}^{N} z_{j_2}\right)\ldots\left(\sum_{j_{(n-1)}=1}^{N} z_{j_{(n-1)}}\right)\right)$$

$$= \sum_{i=1}^{N} Y_i((1)(1)\ldots(1)) = \sum_{i=1}^{N} Y_i = Y$$

Por lo tanto $\hat{Y}_{PPZ}$ es un estimador insesgado de Y, y en consecuencia, $\hat{Y}_{PPS}$ también lo es. Ahora se procederá a determinar la varianza; para ello se utilizará la fórmula de la esperanza de los desvíos al cuadrado,

$$V\left(\hat{Y}_{PPZ}\right) = E\left[\left(\hat{Y}_{PPZ} - Y\right)^2\right] = E\left[\hat{Y}_{PPZ}^2 - 2Y\hat{Y}_{PPZ} + Y^2\right] = E\left(\hat{Y}_{PPZ}^2\right) - 2Y\,E\left(\hat{Y}_{PPZ}\right) + Y^2 = E\left(\hat{Y}_{PPZ}^2\right) - Y^2$$

donde,

$$E\left(\hat{Y}_{PPZ}^2\right) = \sum_{k=1}^{N^n}P_k\left[\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{z_i}\right]^2 = \frac{1}{n^2}\sum_{k=1}^{N^n}P_k\left[\sum_{i=1}^{n}\frac{Y_i}{z_i}\right]^2 = \frac{1}{n^2}\left[n\sum_{i=1}^{n}\frac{Y_i^2}{z_i} + n(n-1)Y^2\right] = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i^2}{z_i} + \frac{n-1}{n}Y^2 \quad (10.20)$$

Sustituyendo arriba,

$$V\left(\hat{Y}_{PPZ}\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i^2}{z_i} + \frac{n-1}{n}Y^2 - Y^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i^2}{z_i} - \frac{1}{n}Y^2 = \frac{1}{n}\left[\sum_{i=1}^{N}\frac{Y_i^2}{z_i} - Y^2\right] = \frac{1}{n}\left[\sum_{i=1}^{N}\frac{Y_i^2}{z_i} - Y^2 + Y^2 - Y^2\right]$$

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

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

es decir, $$V\left(\hat{Y}_{PPZ}\right) = \frac{1}{n}\sum_{i=1}^{N} z_i\left(\frac{Y_i}{z_i} - Y\right)^2$$

y un estimador insesgado de la varianza es, $$\hat{V}\left(\hat{Y}_{PPZ}\right) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{Y_i}{z_i} - \hat{Y}_{PPZ}\right)^2$$

véase,

$$E\left[\hat{V}\left(\hat{Y}_{PPZ}\right)\right] = \frac{1}{n(n-1)}E\left[\sum_{i=1}^{n}\left(\frac{Y_i}{z_i} - \hat{Y}_{PPZ}\right)^2\right] = \frac{1}{n(n-1)}E\left[\sum_{i=1}^{n}\left(\frac{Y_i}{z_i} - \hat{Y}_{PPZ} + Y - Y\right)^2\right]$$

$$= \frac{1}{n(n-1)}E\left[\sum_{i=1}^{n}\left(\left(\frac{Y_i}{z_i} - Y\right) - \left(\hat{Y}_{PPZ} - Y\right)\right)^2\right] = \frac{1}{n(n-1)}E\left[\sum_{i=1}^{n}\left(\frac{Y_i}{z_i} - Y\right)^2 - n\left(\hat{Y}_{PPZ} - Y\right)^2\right]$$

$$= \frac{1}{n(n-1)}E\left[\sum_{i=1}^{n}\left(\frac{Y_i}{z_i} - Y\right)^2\right] - \frac{1}{n-1}E\left[\left(\hat{Y}_{PPZ} - Y\right)^2\right]$$

$$= \frac{1}{n(n-1)}E\left[\sum_{i=1}^{n}\left(\frac{Y_i}{z_i} - Y\right)^2\right] - \frac{1}{n-1}\left[E\left(\hat{Y}_{PPZ}^2\right) - Y^2\right]$$

por (10.20), $$= \frac{1}{n(n-1)}E\left[\sum_{i=1}^{n}\left(\frac{Y_i}{z_i} - Y\right)^2\right] - \frac{1}{n-1}\left[\frac{1}{n}\sum_{I=1}^{N}\frac{Y_i^2}{z_i} + \frac{n-1}{n}Y^2 - Y^2\right] \qquad (10.21)$$

además,

$$E\left[\sum_{i=1}^{n}\left(\frac{Y_i}{z_i} - Y\right)^2\right] = E\left[\sum_{i=1}^{n}\left(\frac{Y_i}{z_i}\right)^2 - 2Y\sum_{i=1}^{n}\left(\frac{Y_i}{z_i}\right) + nY^2\right] = E\left[\sum_{i=1}^{n}\left(\frac{Y_i}{z_i}\right)^2 - nY^2\right] = E\left[\sum_{i=1}^{n}\left(\frac{Y_i}{z_i}\right)^2\right] - nY^2 \qquad (10.22)$$

y análogamente al desarrollo de $E\left(\hat{Y}_{PPZ}\right)$,

$$E\left[\sum_{i=1}^{n}\left(\frac{Y_i}{z_i}\right)^2\right] = \sum_{k=1}^{N^n}\sum_{i=1}^{n} P_k\left(\frac{Y_i}{z_i}\right)^2 = \sum_{i=1}^{N} n\frac{Y_i^2}{z_i^2}\left(z_i\sum_{j_1=1}^{N}\sum_{j_2=1}^{N}\cdots\sum_{j_{(n-1)}=1}^{N} z_{j_1}z_{j_2}\ldots z_{j_{(n-1)}}\right)$$

$$= \sum_{i=1}^{N} n\frac{Y_i^2}{z_i}\left(\sum_{j_1=1}^{N}\sum_{j_2=1}^{N}\cdots\sum_{j_2=1}^{N} z_{j_1}z_{j_2}\ldots z_{j_{(n-1)}}\right) = n\sum_{i=1}^{N}\frac{Y_i^2}{z_i}\left(\left(\sum_{j_1=1}^{N} z_{j_1}\right)\left(\sum_{j_2=1}^{N} z_{j_2}\right)\ldots\left(\sum_{j_{(n-1)}=1}^{N} z_{j_{(n-1)}}\right)\right)$$

$$= n\sum_{i=1}^{N}\frac{Y_i^2}{z_i}\left((1)(1)\ldots(1)\right) = n\sum_{i=1}^{N}\frac{Y_i^2}{z_i}$$

sustituyendo en (10.22) y luego en (10.21),

$$E\left[\hat{V}\left(\hat{Y}_{PPZ}\right)\right]=\frac{1}{n(n-1)}\left[n\sum_{i=1}^{N}\frac{Y_i^2}{z_i}-nY^2\right]-\frac{1}{n-1}\left[\frac{1}{n}\sum_{I=1}^{N}\frac{Y_i^2}{z_i}+\frac{n-1}{n}Y^2-Y^2\right]$$

$$=\frac{1}{n(n-1)}\left[n\sum_{i=1}^{N}\frac{Y_i^2}{z_i}-nY^2\right]-\frac{1}{n-1}\left[\frac{1}{n}\sum_{I=1}^{N}\frac{Y_i^2}{z_i}-\frac{1}{n}Y^2\right]$$

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

$$=V\left(\hat{Y}_{PPZ}\right)$$

por lo que se demuestra que es un estimador insesgado.

Si se sustituye $z_i=M_i\big/M_0$, se tiene que,

$$V\left(\hat{Y}_{PPS}\right)=\frac{M_0}{n}\sum_{i=1}^{N}M_i\left(\overline{Y}_i-\overline{\overline{Y}}\right)^2 \qquad \hat{V}\left(\hat{Y}_{PPS}\right)=\frac{M_0^2}{n(n-1)}\sum_{i=1}^{n}\left(\overline{Y}_i-\hat{\overline{\overline{Y}}}_{PPS}\right)^2$$

donde $\hat{\overline{\overline{Y}}}_{PPS}=\dfrac{\hat{Y}_{PPS}}{M_0}$, análogamente, $\hat{\overline{\overline{Y}}}_{PPZ}=\dfrac{\hat{Y}_{PPZ}}{M_0}$

además $\hat{\overline{Y}}_{PPS}=\dfrac{\hat{Y}_{PPS}}{N}$, $\hat{\overline{Y}}_{PPZ}=\dfrac{\hat{Y}_{PPZ}}{N}$

10.3.- Selección de conglomerados con probabilidades desiguales sin reemplazamiento

En algunas oportunidades se desea hacer la selección de los conglomerados sin reemplazamiento, es decir, los conglomerados seleccionados no se consideran para las siguientes extracciones.

Debido a que el desarrollo es diferente a la manera como se ha enfocado este trabajo, no se desarrollará este caso, simplemente se mostrarán el estimador y las varianzas, desarrollados por Horvitz y Thompson (1952).

Supóngase que el primer conglomerado seleccionado es el *i*-ésimo, y se selecciona con probabilidad z_i. Entonces, las probabilidades de selección de los *N-1* conglomerados restantes, en la segunda extracción, es $z_j\big/(1-z_i)$. Por lo tanto, la probabilidad total de seleccionar el *i*-ésimo conglomerado, en la primera o segunda extracción, es,

$$\pi_i = z_i + \sum_{\substack{j=1 \\ j\neq i}}^{N} \frac{z_j z_i}{(1-z_j)} = z_i \left(1 + \sum_{\substack{j=1 \\ j\neq i}}^{N} \frac{z_j}{(1-z_j)} \right) = z_i \left(1 + A - \frac{z_i}{(1-z_i)} \right)$$

donde $\pi_i = z_i + \sum_{\substack{j=1 \\ j\neq i}}^{N} \frac{z_j z_i}{(1-z_j)} = z_i \left(1 + \sum_{\substack{j=1 \\ j\neq i}}^{N} \frac{z_j}{(1-z_j)} \right) = z_i \left(1 + A - \frac{z_i}{(1-z_i)} \right)$

Para el caso general, donde n≥2, se considerará,

π_i = probabilidad de que la i-ésima unidad esté en la muestra

π_{ij} = probabilidad de que la i-ésima y j-ésima unidades estén ambas en la muestra

entonces, $\sum_{i=1}^{N} \pi_i = n$, $\sum_{i=1}^{N-1} \sum_{\substack{j=1 \\ j\neq i}}^{N} (n-1)\pi_i$, $\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \pi_{ij} = \frac{1}{2} n(n-1)$

El estimador de Horvitz-Thompson del total poblacional es,

$\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{Y_i}{\pi_i}$, que es un estimador insesgado de *Y*, cuya varianza es,

$$V(\hat{Y}_{HT}) = \sum_{i=1}^{N} \frac{(1-\pi_i)}{\pi_i} Y_i^2 + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} Y_i Y_j$$

En cuanto a las varianzas estimadas, se tienen dos, que son,

$\hat{V}_1(\hat{Y}_{HT}) = \sum_{i=1}^{n} \frac{(1-\pi_i)}{\pi_i^2} Y_i^2 + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j \pi_{ij}} Y_i Y_j$, siempre y cuando $\pi_{ij} > 0$, $\forall i, j = 1,2,...,N$. El segundo fue desarrollado por Yates y Grundy en 1953, y paralelamente por Sean, también en 1953, y es,

$$\hat{V}_2(\hat{Y}_{HT}) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2$$

Los términos *(π_{ij} - π_i π_j)* varían considerablemente, y a veces son negativos, en consecuencia, $\hat{V}_1(\hat{Y}_{HT})$ y $\hat{V}_2(\hat{Y}_{HT})$ tienden a ser inestables. [2; 321]

Cabe destacar que existen otros estimadores, como los desarrollados por Brewer (1963), Durbin (1967), Sampford (1967), Murthy (1957), Rao-Hartley-Cochran (1962), entre otros. [2;323-330]

11.- MUESTREO MONOETÁPICO SISTEMÁTICO DE CONGLOMERADOS

Este diseño se diferencia del anterior sólo en que la selección de los conglomerados se hace en forma sistemática, y no aleatoria simple. Entonces, los *N* conglomerados se dividen en n intervalos, cada uno compuesto por *k* conglomerados, donde k=N/n. Luego se escoge un número *"r"* entre *1* y *k*, y se seleccionan los conglomerados *r* , *r+k* , *r+2k* , ... , *r+(n-1)k*, donde *r+(n-1)k < N*, y *r+(n-1)k = N* si y sólo si *r=k*, para *k* entero.

Sean

y_{tij} = observación *j-ésima*, del conglomerado *i-ésimo* de la muestra *t-ésima*

$Y_{ti} = \sum_{j=1}^{M_{ti}} y_{tij}$ = total del conglomerado *i-ésimo* de la muestra *t-ésima*

M_{ti} = total de elementos del conglomerado *i-ésimo* de la muestra *t-ésima*

$m_{0t} = \sum_{i=1}^{n} M_{ti}$ = total de elementos en la muestra, de la muestra *t-ésima*

$M_0 = \sum_{t=1}^{k}\sum_{i=1}^{n} M_{ti} = \sum_{t=1}^{k} m_{0t}$ = total de elementos en la población

$y_t = \sum_{i=1}^{n}\sum_{j=1}^{M_{ti}} y_{tij}$ = total de la muestra *t-ésima*

$y_t = \frac{y_t}{n}$ = promedio por conglomerado de la muestra *t-ésima* (o total promedio muestral por conglomerado de la muestra *t-ésima*)

$y_t = \frac{\sum_{i=1}^{n} y_{ti}}{\sum_{t=1}^{n} M_{ti}} = \frac{Y_t}{m_{0t}}$ = promedio por elemento, de la muestra *t-ésima*

donde $t=1,...,k$; $i=1,..., n$; $j=1,...,M_t$

Los valores poblacionales son:

$Y = \sum_{t=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{M_{ti}} y_{tij} = \sum_{t=1}^{k}\sum_{i=1}^{n} Y_{ti} = \sum_{t=1}^{k} Y_t$ = total poblacional, que coincide con $Y = \sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij} = \sum_{i=1}^{N} Y_i$

En la notación convencional, sin considerar las muestras sistemáticas. También se tiene que,

$\overline{Y} = \frac{Y}{N}$ = promedio poblacional por conglomerado

$\overline{\overline{Y}} = \frac{Y}{M_0}$ =promedio poblacional por elemento

Adaptando las fórmulas del muestreo sistemático a los conglomerados, se tiene que los estimadores del total, el promedio por conglomerado y el promedio por elemento son,

$$\hat{Y}_{csy} = \frac{N}{n}\sum_{i=1}^{n} Y_{ti} \quad , \quad \hat{\bar{Y}}_{csy} = \frac{1}{n}\sum_{i=1}^{n} Y_{ti} = \frac{\hat{Y}_{csy}}{N} \quad , \quad \hat{\bar{\bar{Y}}}_{csy} = \frac{N}{n M_0}\sum_{i=1}^{n} Y_{ti} = \frac{\hat{Y}_{csy}}{M_0} \qquad (11.1)$$

suponiendo que de las *k* posibles muestras, la seleccionada fue la *t-ésima*. Las varianzas vienen dadas por,

$$V\left(\hat{Y}_{csy}\right) = \frac{1}{k}\sum_{t=1}^{k}\left(\hat{Y}_{csyt} - Y\right)^2$$

$$V\left(\hat{\bar{Y}}_{csy}\right) = \frac{1}{k}\sum_{t=1}^{k}\left(\hat{\bar{Y}}_{csyt} - \bar{Y}\right)^2 = \frac{1}{N^2 k}\sum_{t=1}^{k}\left(\hat{Y}_{csyt} - Y\right)^2 = \frac{1}{N^2}V\left(\hat{Y}_{csy}\right)$$

$$V\left(\hat{\bar{\bar{Y}}}_{csy}\right) = \frac{1}{k}\sum_{t=1}^{k}\left(\hat{\bar{\bar{Y}}}_{csyt} - \bar{\bar{Y}}\right)^2 = \frac{1}{M_0^2 k}\sum_{t=1}^{k}\left(\hat{Y}_{csyt} - Y\right)^2 = \frac{1}{M_0^2}V\left(\hat{Y}_{csy}\right)$$

donde $\hat{Y}_{csyt}, \hat{\bar{Y}}_{csyt}, \hat{\bar{\bar{Y}}}_{csyt}$ son las estimaciones del total, el promedio por conglomerado y el promedio por elemento arrojados por la muestra *t-ésima.*

Las varianzas estimadas, según las fórmulas del capítulo 8 son,

muestreo aleatorio simple, $$\hat{V}\left(\hat{Y}_{csy}\right) = \frac{N-n}{N n}\frac{\sum_{i=1}^{n}\left(Y_{ti} - \hat{\bar{Y}}_t\right)^2}{n-1}$$

selecciones pareadas, $$\hat{V}\left(\hat{Y}_{csy}\right) = \frac{N-n}{N n^2}\sum_{i=1}^{n/2}\left(Y_{th1} - Y_{th2}\right)^2$$

donde Y_{th1}, Y_{th2} son los totales del 1er y 2do conglomerado del *h-ésimo* par de la muestra *t-ésima*

diferencias sucesivas, $$\hat{V}\left(\hat{Y}_{csy}\right) = \frac{N-n}{2N n(n-1)}\sum_{i=1}^{n}\left(Y_{ti} - Y_{t(i+1)}\right)^2$$

para determinar las fórmulas de las varianzas estimadas del promedio por conglomerado y por elemento se utilizan las equivalencias mostradas en (11.1), esto es,

$$\hat{V}\left(\hat{\bar{Y}}_{csy}\right) = \frac{1}{N^2}\hat{V}\left(\hat{Y}_{csy}\right) \quad , \quad \hat{V}\left(\hat{\bar{\bar{Y}}}_{csy}\right) = \frac{1}{M_0^2}\hat{V}\left(\hat{Y}_{csy}\right)$$

Aplicando el muestreo sistemático replicado de conglomerados, se tiene se seleccionarán *l* muestras replicadas

sistemáticas o replicaciones, cada una de tamaño n', de manera que $n = l\,n'$, y el intervalo de selección de cada replicación es $k' = \frac{N}{n'} = l\left(\frac{N}{n}\right) = l\,k$, donde $k = \frac{N}{n}$.

$$\bar{y}_{csyr_g} = \frac{\sum_{i=1}^{n'} Y_{csyr_{gi}}}{n'} \qquad ; \qquad \hat{Y}_{csyr_g} = N\,\bar{y}_{csyr_g} \qquad g=1,2,...,l$$

Y los estimadores, $\hat{\bar{Y}}_{csyr} = \frac{\sum_{g=1}^{l} \bar{y}_{csyr_g}}{l}$; $\hat{Y}_{csyr} = N\,\hat{\bar{Y}}_{csyr}$ $\hat{\bar{\bar{Y}}}_{csyr} = \frac{\hat{Y}_{csyr}}{M_0}$.

Las varianzas vienen dadas por las siguientes expresiones,

$$V\left(\hat{Y}_{csyr}\right) = \frac{N^2(1-f)}{l}\frac{\sum_{g=1}^{k'}\left(\bar{y}_{csyr_g} - \bar{\bar{Y}}\right)^2}{k'-1} \qquad ; \qquad \hat{V}\left(\hat{Y}_{csyr}\right) = \frac{N^2(1-f)}{l}\frac{\sum_{g=1}^{l}\left(\bar{y}_{csyr_g} - \hat{\bar{\bar{Y}}}_{csyr}\right)^2}{l-1}$$

$$V\left(\hat{\bar{\bar{Y}}}_{csyr}\right) = \frac{1}{N^2}V\left(\hat{Y}_{csyr}\right) \qquad ; \qquad \hat{V}\left(\hat{\bar{\bar{Y}}}_{csyr}\right) = \frac{1}{N^2}\hat{V}\left(\hat{Y}_{csyr}\right)$$

$$V\left(\hat{\bar{\bar{Y}}}_{csyr}\right) = \frac{1}{M_0^2}V\left(\hat{Y}_{csyr}\right) \qquad ; \qquad \hat{V}\left(\hat{\bar{\bar{Y}}}_{csyr}\right) = \frac{1}{M_0^2}\hat{V}\left(\hat{Y}_{csyr}\right)$$

En los casos de estratos donde k_h no es entero, se aplican los mismos tratamientos mostrados en el capítulo 8 (Muestreo Aleatorio Sistemático).

12.- SUBMUESTREO CON UNIDADES DE IGUAL TAMAÑO

Sean *N* conglomerados, cada uno con *M* elementos, se toma una muestra aleatoria simple de *n* conglomerados, y en cada uno de ellos se seleccionan *m* elementos de manera aleatoria simple.

Se tienen las siguientes medidas:

Total de Muestras Posibles;	$\binom{N}{n}\binom{M}{m}^n$
Total de Muestras posibles donde está incluido un elemento determinado;	$\binom{N-1}{n-1}\binom{M-1}{m-1}\binom{M}{m}^{n-1}$
Total de Muestras Posibles donde está incluido un par de elementos determinados, del mismo conglomerado;	$\binom{N-1}{n-1}\binom{M-2}{m-2}\binom{M}{m}^{n-1}$
Total de Muestras Posibles donde está incluido un par de elementos determinados, de conglomerados diferentes;	$\binom{N-2}{n-2}\binom{M-1}{m-1}^2\binom{M}{m}^{n-2}$

12.1.- Total y promedio poblacional

Sean los siguientes valores poblaciones,

y_{ij} = valor de la variable *"y"* en el elemento *j-ésimo* de la unidad *i-ésima*

$Y_i = \sum_{j=1}^{M} y_{ij}$ = el total de la unidad *i-ésima*

$\overline{Y_i} = \frac{Y_i}{M}$ = el promedio por elemento de la unidad *i-ésima*

$Y = \sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} = \sum_{i=1}^{N} Y_i$ = el total poblaciona l

$\overline{Y} = \frac{Y}{N}$ = el total promedio poblaciona l por unidad

$\overline{\overline{Y}} = \frac{Y}{N\,M}$ = el promedio poblaciona l por elemento

sea además, $y_i = \sum_{j=1}^{m} y_{ij}$ = el total muestral de la unidad *i-ésima*

12.2.- Estimadores

Como la muestra en cada conglomerado seleccionado es aleatoria simple de elementos, se tienen los siguientes estimadores,

$\hat{\bar{Y}}_i = \bar{y}_i = \frac{y_i}{m}$ = el promedio muestral por elemento de la unidad *i-ésima*, que es a su vez, el estimador del promedio por elemento de la unidad *i-ésima*

$\hat{Y}_i = M\,\bar{y}_i = \frac{M}{m}\sum_{j=1}^{m} y_{ij}$ = el total estimado de la unidad *i-ésima*

Aplicando el principio de asignar a los conglomerados que no están en la muestra el promedio de los que si están, se tiene que

$$\hat{Y} = \sum_{i=1}^{n} \hat{Y}_i + \sum_{j=n+1}^{N} \hat{Y}'_j$$

donde $\hat{Y}'_j$ es el total promedio por conglomerado, de los conglomerados que están en la muestra, es decir,

$$\hat{Y}'_j = \frac{\sum_{i=1}^{n} \hat{Y}_i}{n}$$

de tal manera que,

$$\hat{Y} = \sum_{i=1}^{n} \hat{Y}_i + \sum_{j=n+1}^{N} \left(\frac{\sum_{i=1}^{n} \hat{Y}_i}{n} \right) = \sum_{i=1}^{n} \hat{Y}_i + (N-n) \left(\frac{\sum_{i=1}^{n} \hat{Y}_i}{n} \right) = \left(1 + \frac{N-n}{n} \right) \sum_{i=1}^{n} \hat{Y}_i = \frac{N}{n} \sum_{i=1}^{n} \hat{Y}_i$$

entonces,

$$\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} \hat{Y}_i = \frac{N}{n} \sum_{i=1}^{n} M\,\bar{y}_i$$

y como $\bar{Y} = \frac{Y}{N}$

entonces, $\hat{\bar{Y}} = \frac{\hat{Y}}{N} = \frac{\sum_{i=1}^{n} M\,\bar{y}_i}{n} = M\,\bar{\bar{y}}$, y $\hat{Y} = N\,\hat{\bar{Y}}$

Otra manera de hallar $\hat{Y}$ es asignándole a los elementos que no están en la muestra, el promedio de los elementos que sí están,

$$\hat{Y} = \sum_{i=1}^{n}\left(\sum_{j=1}^{m} y_{ij} + \sum_{k=m+1}^{M} y'_{ij}\right) + \sum_{h=n+1}^{N}\sum_{k=1}^{M} y'_{hk} = \sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij} + \sum_{i=1}^{n}\sum_{k=m+1}^{M} y'_{ik} + \sum_{h=n+1}^{N}\sum_{k=1}^{M} y'_{hk}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij} + n(M-m)\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}}{nm}\right) + (N-n)M\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}}{nm}\right)$$

$$= \left(1 + \frac{n(M-m)}{nm} + \frac{(N-n)M}{nm}\right)\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij} = NM\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}}{nm}\right)$$

entonces

$$\hat{Y} = NM\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}}{nm}\right)$$

y como $\bar{\bar{Y}} = \frac{Y}{NM}$ se tiene que $\hat{\bar{\bar{Y}}} = \frac{\hat{Y}}{NM} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}}{nm}$ que es el promedio muestral por elemento, es decir que,

$$\hat{\bar{\bar{Y}}} = \bar{\bar{y}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}}{nm} \quad , \quad y \quad \hat{Y} = NM\,\bar{\bar{y}}$$

Si al hacer,

$$\hat{Y} = \sum_{i=1}^{n}\left(\sum_{j=1}^{m} y_{ij} + \sum_{k=m+1}^{M} y'_{ij}\right) + \sum_{h=n+1}^{N}\sum_{k=1}^{M} y'_{hk}$$

se le asigna a los conglomerados que no están en la muestra pero pertenecen a conglomerados que sí están en la muestra, el promedio de los elementos de su conglomerado, se tiene,

$$\hat{Y} = \sum_{i=1}^{n}\left(\sum_{j=1}^{m} y_{ij} + (M-m)\frac{\sum_{j=1}^{m} y_{ij}}{m}\right) + (N-n)M\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}}{nm}\right)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij} + (M-m)\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}}{m} + (N-n)M\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}}{nm}$$

$$= \left[nm + n(M-m) + (N-n)M\right]\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}}{nm}$$

$$= NM\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}}{nm} = NM\,\bar{\bar{y}}$$

que es el mismo resultado al que se llegó anteriormente. Esto sucede porque,

- los conglomerados son de igual tamaño
- se toma la misma cantidad de elementos en cada conglomerado seleccionado
- tanto en primera como en segunda etapa, la selección es aleatoria simple.

12.3.- Varianzas

Sea $V(\bar{\bar{y}}) = \dfrac{\sum_{l=1}^{L}(\bar{\bar{y}}_l - \bar{\bar{Y}})^2}{L}$ la varianza del estimador del promedio por elemento, donde

$$L = \binom{N}{n}\binom{M}{m}^n \text{ es el total de muestras posibles}$$

$$\bar{\bar{y}}_l = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{lij}}{n\,m} \text{ es el promedio muestral de la } l\text{-ésima muestra}$$

desarrollando $V(\bar{\bar{y}})$, se tiene,

$$V(\bar{\bar{y}}) = \frac{\sum_{l=1}^{L}(\bar{\bar{y}}_l - \bar{\bar{Y}})^2}{L} = \frac{\sum_{l=1}^{L}\bar{\bar{y}}_l^2 - L\bar{\bar{Y}}^2}{L} = \frac{\sum_{l=1}^{L}\bar{\bar{y}}_l^2}{L} - \bar{\bar{Y}}^2 \qquad (12.1)$$

trabajando el primer término,

$$\sum_{l=1}^{L}\bar{y}_l^2=\sum_{l=1}^{L}\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}y_{lij}}{nm}\right)^2=\frac{1}{n^2m^2}\sum_{l=1}^{L}\left(\sum_{i=1}^{n}\sum_{j=1}^{m}y_{lij}\right)^2$$

que es la suma de todos los totales muestrales, de todas las muestras posibles. Aplicando los totales de muestras en las que está incluido un elemento determinado, sólo y en pareja,

$$=\frac{1}{n^2m^2}\left[\binom{N-1}{n-1}\binom{M-1}{m-1}\binom{M}{m}^{n-1}\sum_{i=1}^{N}\sum_{j=1}^{M}y_{ij}^2+2\binom{N-1}{n-1}\binom{M-2}{m-2}\binom{M}{m}^{n-1}\sum_{i=1}^{N}\sum_{j=1}^{M-1}\sum_{k=j+1}^{M}y_{ij}y_{ik}+2\binom{N-2}{n-2}\binom{M-1}{m-1}^2\binom{M}{m}^{n-2}\sum_{i=1}^{N-1}\sum_{h=i+1}^{N}\sum_{j=1}^{M-1}\sum_{k=j+1}^{M}y_{ij}y_{hk}\right]$$

$$=\frac{1}{n^2m^2}\binom{N-1}{n-1}\binom{M-1}{m-1}\binom{M}{m}^{n-1}\left[\sum_{i=1}^{N}\sum_{j=1}^{M}y_{ij}^2+2\left(\frac{m-1}{M-1}\right)\sum_{i=1}^{N}\sum_{j=1}^{M-1}\sum_{k=j+1}^{M}y_{ij}y_{ik}+2\left(\frac{n-1}{N-1}\right)\frac{\binom{M-1}{m-1}}{\binom{M}{m}}\sum_{i=1}^{N-1}\sum_{h=i+1}^{N}\sum_{j=1}^{M-1}\sum_{k=j+1}^{M}y_{ij}y_{hk}\right]$$

sustituyendo en (12.1) y reemplazando L por $L=\binom{N}{n}\binom{M}{m}^n$,

$$V(\bar{y})=\frac{\binom{N-1}{n-1}\binom{M-1}{m-1}\binom{M}{m}^{n-1}\left[\sum_{i=1}^{N}\sum_{j=1}^{M}y_{ij}^2+2\left(\frac{m-1}{M-1}\right)\sum_{i=1}^{N}\sum_{j=1}^{M-1}\sum_{k=j+1}^{M}y_{ij}y_{ik}+2\left(\frac{n-1}{N-1}\right)\frac{\binom{M-1}{m-1}}{\binom{M}{m}}\sum_{i=1}^{N-1}\sum_{h=i+1}^{N}\sum_{j=1}^{M-1}\sum_{k=j+1}^{M}y_{ij}y_{hk}\right]}{n^2m^2\binom{N}{n}\binom{M}{m}^n}-\bar{Y}^2$$

pero como $\binom{N}{n}\binom{M}{m}^n=\frac{N}{n}\frac{M}{m}\binom{N-1}{n-1}\binom{M-1}{m-1}\binom{M}{m}^{n-1}$ entonces,

$$V(\bar{y})=\frac{\sum_{i=1}^{N}\sum_{j=1}^{M}y_{ij}^2+2\left(\frac{m-1}{M-1}\right)\sum_{i=1}^{N}\sum_{j=1}^{M-1}\sum_{k=j+1}^{M}y_{ij}y_{ik}+2\left(\frac{n-1}{N-1}\right)\frac{\binom{M-1}{m-1}}{\binom{M}{m}}\sum_{i=1}^{N-1}\sum_{h=i+1}^{N}\sum_{j=1}^{M-1}\sum_{k=j+1}^{M}y_{ij}y_{hk}}{NMnm}-\bar{Y}^2$$

$$=\frac{\sum_{i=1}^{N}\sum_{j=1}^{M}y_{ij}^2+2\left(\frac{m-1}{M-1}\right)\sum_{i=1}^{N}\sum_{j=1}^{M-1}\sum_{k=j+1}^{M}y_{ij}y_{ik}+2\frac{m(n-1)}{M(N-1)}\sum_{i=1}^{N-1}\sum_{h=i+1}^{N}\sum_{j=1}^{M-1}\sum_{k=j+1}^{M}y_{ij}y_{hk}}{NMnm}-\bar{Y}^2$$

$$= \frac{1}{N\ M\ n\ m}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}^2 + \frac{2\,(m-1)}{N\ M\ (M-1)\ n\ m}\sum_{i=1}^{N}\sum_{j=1}^{M-1}\sum_{k=j+1}^{M} y_{ij}y_{ik} + 2\frac{2\,(n-1)}{N\,(N-1)\,M^2\ n}\sum_{i=1}^{N-1}\sum_{h=i+1}^{N}\sum_{j=1}^{M-1}\sum_{k=j+1}^{M} y_{ij}y_{hk} - \bar{\bar{Y}}^2 \quad (12.2)$$

ahora se hacen las siguientes operaciones:

al primer sumando;	se multiplica y se divide por	*(N-1) (M-1)*
	se suma y se resta	*m (n-1)*
	se multiplica y se divide por	*M*
al segundo sumando;	se multiplica y se divide por	*(N-1)*
	se suma y se resta	*2 n m*
	se multiplica y se divide por	*M*

operando sobre el cociente (la primera parte de la ecuación) de (12.2) resulta,

$$= \frac{M(NM-N-M+1-nm+m)+Mm(n-1)}{N(N-1)M^2(M-1)nm}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}^2 + \frac{2M(Nm-N-m+1+nm-nm)}{N(N-1)M^2(M-1)nm}\sum_{i=1}^{N}\sum_{j=1}^{M-1}\sum_{k=j+1}^{M} y_{ij}y_{ik} +$$
$$+ \frac{2(n-1)}{N(N-1)M^2n}\sum_{i=1}^{N-1}\sum_{h=i+1}^{N}\sum_{j=1}^{M-1}\sum_{k=j+1}^{M} y_{ij}y_{hk} - \bar{\bar{Y}}^2$$

$$= \frac{M(NM-N-M+1-nm+m)+Mm(n-1)}{N(N-1)M^2(M-1)nm}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}^2 + \frac{(n-1)}{N(N-1)M^2n}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}^2 + \frac{2M(Nm-N-nm+1)+2Mm(n-1)}{N(N-1)M^2(M-1)nm}\sum_{i=1}^{N}\sum_{j=1}^{M-1}\sum_{k=j+1}^{M} y_{ij}y_{ik} +$$
$$+ \frac{2(n-1)}{N(N-1)M^2n}\sum_{i=1}^{N-1}\sum_{h=i+1}^{N}\sum_{j=1}^{M-1}\sum_{k=j+1}^{M} y_{ij}y_{hk}$$

resolviendo, agrupando términos y sustituyendo en (12.2), se tiene,

$$= \frac{M(N-1)(M-m)-(N-1)(M-m)+m(N-n)(M-1)}{N(N-1)M^2(M-1)nm}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}^2 + 2\frac{M(Nm-N-nm+1)+m(n-1)+Nm-Nm}{N(N-1)M^2(M-1)nm}\sum_{i=1}^{N}\sum_{j=1}^{M-1}\sum_{k=j+1}^{M} y_{ij}y_{ik} - \bar{\bar{Y}}^2 +$$
$$+ \frac{(n-1)}{N(N-1)M^2n}\left[\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}^2 + 2\sum_{i=1}^{N-1}\sum_{h=i+1}^{N}\sum_{j=1}^{M-1}\sum_{k=j+1}^{M} y_{ij}y_{hk}\right]$$

$$= \frac{M-m}{NM(M-1)nm}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}^2 + \frac{m(N-n)(M-1)-(N-1)(M-m)}{N\ (N-1)\ M^2\ (M-1)\ n\ m}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}^2 +$$
$$+ 2\frac{m(N-n)(M-1)-(N-1)(M-m)}{N\ (N-1)\ M^2\ (M-1)\ n\ m}\sum_{i=1}^{N}\sum_{j=1}^{M-1}\sum_{k=j+1}^{M} y_{ij}y_{ik} - \bar{\bar{Y}}^2 + \frac{(n-1)}{N(N-1)M^2n}\left[\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\right]^2$$

$$=\frac{M-m}{NM(M-1)nm}\sum_{i=1}^{N}\sum_{j=1}^{M}y_{ij}^2+\frac{m(N-n)(M-1)-(N-1)(M-m)}{N(N-1)M^2(M-1)nm}\left[\sum_{i=1}^{N}\sum_{j=1}^{M}y_{ij}^2+2\sum_{i=1}^{N}\sum_{j=1}^{M-1}\sum_{k=j+1}^{M}y_{ij}y_{ik}\right]-\bar{\bar{Y}}^2+\frac{N(n-1)}{(N-1)n}\left[\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}y_{ij}\right]$$

$$=\frac{M-m}{NM(M-1)nm}\sum_{i=1}^{N}\sum_{j=1}^{M}y_{ij}^2+\frac{m(N-n)(M-1)-(N-1)(M-m)}{N(N-1)M^2(M-1)nm}\sum_{i=1}^{N}\bar{Y}_i^2-\bar{\bar{Y}}^2+\frac{N(n-1)}{n(N-1)}\bar{\bar{Y}}^2$$

$$=\frac{M-m}{NM(M-1)nm}\sum_{i=1}^{N}\sum_{j=1}^{M}y_{ij}^2+\frac{N-n}{N(N-1)n}\sum_{i=1}^{N}\bar{Y}_i^2+\frac{M-m}{N(M-1)nm}\sum_{i=1}^{N}\bar{Y}_i^2-\frac{N-n}{n(N-1)}\bar{\bar{Y}}^2$$

$$=\frac{N-n}{N(N-1)n}\left[\sum_{i=1}^{N}\bar{Y}_i^2-N\bar{\bar{Y}}^2\right]+\frac{M-m}{NM(M-1)nm}\left[\sum_{i=1}^{N}\sum_{j=1}^{M}y_{ij}^2-M\sum_{i=1}^{N}\bar{Y}_i^2\right]$$

$$=\frac{1-f_1}{n}\frac{\sum_{i=1}^{N}\left(\bar{Y}_i-\bar{\bar{Y}}\right)^2}{N-1}+\frac{1-f_2}{nm}\frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij}-\bar{Y}_i\right)^2}{N(M-1)}=\frac{1-f_1}{n}S_1^2+\frac{1-f_2}{nm}S_2^2$$

por lo tanto, $V(\bar{\bar{y}})=\frac{1-f_1}{n}S_1^2+\frac{1-f_2}{nm}S_2^2$

Ahora se hallará la esperanza de $\bar{\bar{y}}$,

$$E(\bar{\bar{y}})=\frac{\sum_{k=1}^{L}\bar{\bar{y}}_k}{L}=\frac{\sum_{k=1}^{L}\left(\frac{\sum_{i=1}^{N}\sum_{j=1}^{M}y_{kij}}{nm}\right)}{L}=\frac{\sum_{k=1}^{L}\sum_{i=1}^{n}\sum_{j=1}^{m}y_{kij}}{Lnm}$$

donde y_{kij} es el valor del elemento *j-ésimo* del conglomerado *i-ésimo*, de la muestra *k-ésima*, y L es el total de muestras posibles. Desarrollando el numerador,

$$E(\bar{\bar{y}}) = \frac{\binom{N-1}{n-1}\binom{M-1}{m-1}\binom{M}{m}^{n-1}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}}{\binom{N}{n}\binom{M}{m}^{n} n\,m} = \frac{\binom{N-1}{n-1}\binom{M-1}{m-1}\binom{M}{m}^{n-1}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}}{\frac{N}{n}\binom{N-1}{n-1}\frac{M}{m}\binom{M-1}{m-1}\binom{M}{m}^{n} n\,m} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}}{N\,M}$$

es decir, que $\bar{\bar{y}}$ es un estimador insesgado de $\bar{\bar{Y}}$. Como es de esperarse, $\bar{y}_i$ también es un estimador insesgado de $\bar{Y}_i$, como se aprecia a continuación,

$$E(\bar{y}_i) = \frac{\sum_{l=1}^{\binom{M}{n}} \bar{y}_{il}}{\binom{M}{m}} = \frac{\sum_{l=1}^{\binom{M}{n}} \left(\frac{\sum_{j=1}^{m} y_{ilj}}{m} \right)}{\binom{M}{m}} = \frac{\binom{M-1}{m-1}\sum_{j=1}^{M} y_{ij}}{m\binom{\frac{M}{m}}{}\binom{M-1}{m-1}} = \frac{\sum_{j=1}^{M} y_{ij}}{M} = \bar{Y}_i$$

donde y_{ilj} es el valor del elemento *j-ésimo*, de la muestra *l-ésima*, del conglomerado *i-ésimo*. Nótese que si se asocia $\bar{Y}_i$ como el valor esperado del promedio del conglomerado *i-ésimo*, entonces el estimador del promedio por elemento y su varianza son,

$$\hat{\bar{\bar{Y}}} = \frac{\sum_{i=1}^{n} \bar{Y}_i}{n} = \bar{\bar{y}}$$

$$E\left(\hat{\bar{\bar{Y}}}\right) = E(\bar{\bar{y}}) = \frac{\sum_{k=1}^{\binom{N}{n}} \hat{\bar{\bar{Y}}}_k}{\binom{N}{n}} = \frac{\sum_{k=1}^{\binom{N}{n}} \left(\sum_{i=1}^{n} \frac{\bar{Y}_{ki}}{n} \right)}{\binom{N}{n}} = \frac{\binom{N-1}{n-1}\sum_{i=1}^{N} \bar{Y}_i}{n\binom{\frac{N}{n}}{}\binom{N-1}{n-1}} = \frac{\sum_{i=1}^{N} \bar{Y}_i}{N} = \bar{\bar{Y}}$$

si se denomina a

$$\frac{\sum_{i=1}^{n} \bar{Y}_i}{n}$$

como la esperanza de segunda etapa de $\hat{\bar{\bar{Y}}} = \bar{\bar{y}}$, que consiste en obtener el valor esperado de $\bar{\bar{Y}}$ una vez seleccionados los *n* conglomerados de la muestra, contemplando todas las posibles muestras de elementos o

unidades de segunda etapa, en los conglomerados seleccionados, y se denota por $E_2\left(\hat{\bar{\bar{Y}}}\right)=E_2\left(\bar{\bar{y}}\right)$, entonces,

$$E_1\left[E_2\left(\hat{\bar{\bar{Y}}}\right)\right]=E_1\left[E_2(\bar{\bar{y}})\right]=\frac{1}{\binom{N}{n}}\sum_{k=1}^{\binom{N}{n}}\hat{\bar{\bar{Y}}}_k$$

que es la esperanza de primera etapa, que consiste en contemplar todas las posibles muestras de conglomerados o unidades de primera etapa. Es decir, que $E\left(\hat{\bar{\bar{Y}}}\right)=E_1\left[E_2\left(\hat{\bar{\bar{Y}}}\right)\right]=E_1\left[E_2(\bar{\bar{y}})\right]$.

Este planteamiento servirá para hallar la varianza de $\hat{\bar{\bar{Y}}}$, $V\left(\hat{\bar{\bar{Y}}}\right)=V(\bar{\bar{y}})$, de otra manera, que será de mucha utilidad en los próximos capítulos, para hallar la varianza cuando los conglomerados son de diferente tamaño.

Este planteamiento servirá para hallar la varianza de $\hat{\bar{\bar{Y}}}$, $V\left(\hat{\bar{\bar{Y}}}\right)=V(\bar{\bar{y}})$, de otra manera, que será de mucha utilidad en los próximos capítulos, para hallar la varianza cuando los conglomerados son de diferente tamaño.

Para hallar $V(\hat{\bar{\bar{Y}}})=V(\bar{\bar{y}})$, se hace,

$$V(\bar{\bar{y}})=E\left[\left(\bar{\bar{y}}-\bar{\bar{Y}}\right)^2\right]=E_1\left[E_2\left[\left(\bar{\bar{y}}-\bar{\bar{Y}}\right)^2\right]\right]=E_1\left[E_2(\bar{\bar{y}}^2)-2\bar{\bar{Y}}E_2(\bar{\bar{y}})+\bar{\bar{Y}}^2\right]=E_1\left[E_2(\bar{\bar{y}}^2)-\bar{\bar{Y}}^2\right]$$

pero como $V_2(\bar{\bar{y}})=E_2(\bar{\bar{y}}^2)-\left[E_2(\bar{\bar{y}})\right]^2$, entonces,

$$\begin{aligned}V(\bar{\bar{y}}) &= E_1\left[V_2(\bar{\bar{y}})+\left[E_2(\bar{\bar{y}})\right]^2-2\bar{\bar{Y}}E_2(\bar{\bar{y}})+\bar{\bar{Y}}^2\right]\\ &= E_1\left[V_2(\bar{\bar{y}})\right]+E_1\left[\left[E_2(\bar{\bar{y}})\right]^2\right]-2\bar{\bar{Y}}E_1\left[E_2(\bar{\bar{y}})\right]+\bar{\bar{Y}}^2\\ &= E_1\left[V_2(\bar{\bar{y}})\right]+E_1\left[\left[E_2(\bar{\bar{y}})\right]^2\right]-\bar{\bar{Y}}^2\end{aligned}$$

sustituyendo $\bar{\bar{Y}}=E_1\left[E_2(\bar{\bar{y}})\right]$, se hace,

$$\begin{aligned}V(\bar{\bar{y}}) &= E_1\left[V_2(\bar{\bar{y}})\right]+E_1\left[\left[E_2(\bar{\bar{y}})\right]^2\right]-\left[E_1\left[E_2(\bar{\bar{y}})\right]\right]^2\\ &= E_1\left[V_2(\bar{\bar{y}})\right]+V_1\left[E_2(\bar{\bar{y}})\right]\end{aligned}$$

desarrollando cada uno de los miembros,

$$V_1\left[E_2(\bar{\bar{y}})\right] = E_1\left[\left(E_2(\bar{\bar{y}}) - \bar{\bar{Y}}\right)^2\right] = E_1\left[\left(E_2(\bar{\bar{y}})\right)^2\right] - 2\,\bar{\bar{Y}}\;E_1\left[E_2(\bar{\bar{y}})\right] + \bar{\bar{Y}}^2 = E_1\left[\left(E_2(\bar{\bar{y}})\right)^2\right] - \bar{\bar{Y}}^2 \qquad (12.3)$$

$$E_1\left[\left(E_2(\bar{\bar{y}})\right)^2\right] = \frac{\sum_{k=1}^{\binom{N}{n}}\left[\frac{\sum_{i=1}^{n}\bar{Y}_{ki}}{n}\right]^2}{\binom{N}{n}} = \frac{\sum_{k=1}^{\binom{N}{n}}\left[\sum_{i=1}^{n}\bar{Y}_{ki}\right]^2}{n^2\binom{N}{n}}$$

$$= \frac{\sum_{k=1}^{\binom{N}{n}}\left[\sum_{i=1}^{n}\bar{Y}_{ki}^2 + 2\sum_{i=1}^{n-1}\sum_{h=i+1}^{n}\bar{Y}_{ki}\;\bar{Y}_{kh}\right]}{n^2\binom{N}{n}} = \frac{\binom{N-1}{n-1}\sum_{i=1}^{n}\bar{Y}_i^2 + 2\binom{N-2}{n-2}\sum_{i=1}^{N-1}\sum_{h=1}^{N}\bar{Y}_i\;\bar{Y}_h}{n^2\binom{N}{n}}$$

$$= \frac{\sum_{i=1}^{N}\bar{Y}_i^2 + 2\frac{n-1}{N-1}\sum_{i=1}^{N-1}\sum_{h=i+1}^{N}\bar{Y}_i\bar{Y}_h}{N\,n}$$

$$= \frac{1}{Nn}\left[\sum_{i=}^{N}\bar{Y}_i^2 + 2\frac{n-1}{N-1}\sum_{i=1}^{N-1}\sum_{h=i+1}^{N}\bar{Y}_i\bar{Y}_h + \frac{n-1}{N-1}\sum_{i=1}^{N-1}\sum_{h=i+1}^{N}\bar{Y}_i\bar{Y}_h - 2\frac{n-1}{N-1}\sum_{i=1}^{N-1}\sum_{h=i+1}^{N}\bar{Y}_i\bar{Y}_h\right]$$

$$= \frac{1}{Nn} = \left[\left(\frac{n-1}{N-1}\right)\left(\sum_{i=1}^{N}\bar{Y}_i^2 + 2\sum_{i=1}^{N-1}\sum_{h=i+1}^{N}\bar{Y}_i\bar{Y}_h\right) + \left(1 + \frac{n-1}{N-1}\right)\sum_{i=1}^{N}\bar{Y}_i^2\right]$$

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

sustituyendo en (12.3) ,

$$= \frac{1}{Nn}\left[\left(\frac{n-1}{N-1}\right)N^2\bar{\bar{Y}}^2 + \left(\frac{N-n}{N-1}\right)\sum_{i=1}^{N}\bar{Y}_i^2 - N\,n\bar{\bar{Y}}^2\right] = \frac{N-n}{N\,n}\,\frac{\sum_{i=1}^{N}\bar{Y}_i^2 - N\bar{\bar{Y}}^2}{N-1} = \frac{N-n}{N\,n}\,\frac{\sum_{i=1}^{N}(\bar{Y}_i - \bar{\bar{Y}})^2}{N-1} = \frac{1-f_1}{n}\,S_1^2$$

donde

$$f_1 = \frac{n}{N} \qquad y \qquad S_1^2 = \frac{\sum_{i=1}^{N}(\bar{Y}_i - \bar{\bar{Y}})^2}{N-1}$$

Por otro lado,

$$V(\hat{\bar{Y}}_i) = V(\bar{y}_i) = \frac{M-m}{M\ m}\frac{\sum_{i=1}^{N}(y_{ij}-\bar{Y}_i)^2}{M-1}$$

ya que se toma una muestra aleatoria simple en cada conglomerado seleccionado. Entonces,

$$V_2(\bar{\bar{y}}) = V\left(\frac{\sum_{i=1}^{n}\bar{y}_i}{n}\right) = \frac{1}{n^2}\left[\sum_{i=1}^{n}V_2(\bar{y}_i) + 2\sum\sum cov_2(\bar{y}_i,\bar{y}_h)\right]$$

pero como las muestras son independientes en cada conglomerado,

$$cov_2(\bar{y}_i,\bar{y}_h) = 0 \qquad \forall\ i \neq j \qquad i = 1,2,...,n \qquad j = 2,3,...,n$$

por lo tanto,

$$V_2(\bar{\bar{y}}) = \frac{M-m}{M\,m\,n^2}\frac{\sum_{j=1}^{M}(y_{ij}-\bar{Y}_i)^2}{M-1}$$

aplicando E_1 se tiene que,

$$E_1\left[V_2(\bar{\bar{y}})\right] = \frac{M-m}{M\,m\,n^2}\frac{\sum_{l=1}^{\binom{N}{n}}\sum_{i=1}^{n}\sum_{j=1}^{M}(y_{lij}-\bar{Y}_{li})^2}{\binom{N}{n}(M-1)} = \frac{M-m}{M(M-1)mn^2}\frac{\sum_{l=1}^{\binom{N}{n}}\sum_{i=1}^{n}\left[\sum y_{lij}^2 - M\,\bar{Y}_{li}^2\right]}{\binom{N}{n}}$$

$$= \frac{M-m}{M(M-1)mn}\frac{\sum_{i=1}^{N}\left[\sum_{j=1}^{M}\left(y_{ij}^2 - M\,\bar{Y}_i^2\right)\right]}{N} = \frac{M-m}{M}\frac{1}{m\,n}\frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{Y}_i)^2}{N(M-1)} = \frac{1-f_2}{m\,n}S_2^2$$

donde

$$f_2 = \frac{m}{M} \qquad y \qquad S_2^2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{Y}_i)^2}{N(M-1)}$$

y entonces,

$$V(\bar{\bar{y}}) = V_1\left[E_2(\bar{\bar{y}})\right] + E_1\left[V_2(\bar{\bar{y}})\right] = \frac{1-f_1}{n}S_1^2 + \frac{1-f_2}{nm}S_2^2$$

que es el mismo resultado obtenido mediante la anterior metodología.

13.- SUBMUESTREO ALEATORIO CON UNIDADES DE DIFERENTE TAMAÑO

Esta clase de muestreo es también llamada muestreo bietápico o de dos etapas, porque realmente se toma una muestra en dos etapas. Se tienen N conglomerados, con $M_1, M_2, \ldots, M_N$ elementos en cada uno de ellos respectivamente, entonces se toma una muestra de $n \leq N$, y en cada uno se seleccionan m_i elementos, con $i=1,2,\ldots,n$. Sean los siguientes parámetros,

y_{ij} = valor de la variable *"y"* en el elemento $j-ésimo$ de la unidad $i-ésima$

$Y_i = \sum_{j=1}^{M_i} y_{ij}$ = total de la unidad $i-ésima$

M_i = total de elementos en la unidad $i-ésima$

$\bar{Y}_i = \frac{Y_i}{M_i}$ = promedio por elemento de la unidad $i-ésima$

$Y = \sum_{j=1}^{M_i} y_{ij}$ = total poblaciona l

$\bar{Y} = \frac{Y}{N}$ = total promedio poblaciona l por unidad

$M_0 = \sum_{i=1}^{N} M_i$ = total de elementos en la población

$\bar{\bar{Y}} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij}}{\sum_{i=1}^{N} M_i} = \frac{Y}{M_0}$ = promedio poblaciona l por elemento

13.1.- Selección de conglomerados con iguales probabilidades

Este diseño consiste en que los conglomerados se seleccionan de manera aleatoria simple, es decir, con iguales probabilidades. Una vez seleccionada la muestra, se tienen dos formas de estimar, a través de los estimadores insesgados o de los estimadores de razón al tamaño.

13.1.1.- Estimadores insesgados - selección sin reemplazamiento

Para hallar los estimadores respectivos, se aplica el principio 1, tomando en cuenta que se selecciona una muestra en dos etapas. De manera que,

$$\hat{\bar{Y}}_i = \bar{y}_i = \frac{y_i}{m_i} = \frac{\sum_{j=1}^{m_i} y_{ij}}{m_i} \qquad ; \qquad \hat{Y}_i = m_i\, y_i = \frac{M_i}{m_i} \sum_{j=1}^{m_i} y_{ij}$$

ya que, como se dijo, en cada conglomerado se toma una muestra aleatoria simple de elementos.

Aplicando el principio 1 a los conglomerados, es decir, asignándole a cada conglomerado que no está en la muestra, el promedio de los conglomerados que sí están, se tiene,

$$\hat{Y} = \sum_{i=1}^{n} \hat{Y}_i + \sum_{j=n+1}^{N} \hat{Y}'_j = \sum_{i=1}^{n} \hat{Y}_i + \sum_{j=n+1}^{N} \left(\frac{\sum_{i=1}^{n} \hat{Y}_i}{n} \right) = \sum_{i=1}^{n} \hat{Y}_i + (N-n) \left(\frac{\sum_{i=1}^{n} \hat{Y}_i}{n} \right) = \left(1 + \frac{N-n}{n} \right) \sum_{i=1}^{n} \hat{Y}_i$$

$$= \left(1 + \frac{N-n}{n} \right) \sum_{i=1}^{n} \hat{Y}_i = \frac{N}{n} \sum_{i=1}^{n} \hat{Y}_i$$

y como

$$\bar{Y} = \frac{Y}{N}$$

entonces

$$\hat{\bar{Y}} = \frac{\hat{Y}}{N} = \frac{1}{N} \frac{N}{n} \left(\sum_{i=1}^{n} \hat{Y}_i \right) = \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{n} = \frac{\sum_{i=1}^{n} \frac{M_i}{m_i} \sum_{j=1}^{m_i} y_{ij}}{n}$$

Estos son los estimadores insesgados de Y , $\bar{Y}$ respectivamente, y para identificarlos se denotarán por $\hat{Y}_u$, $\hat{\bar{Y}}_u$, es decir

$$\hat{Y}_u = \frac{N}{n} \sum_{i=1}^{n} \hat{Y}_i \qquad ; \qquad \hat{\bar{Y}}_u = \frac{\sum_{i=1}^{n} \hat{Y}_i}{n}$$

Para verificar la propiedad de insesgamiento, se procederá a hallar su esperanza.

$$E(\hat{Y}_u) = \frac{1}{\binom{N}{n}}\left[\sum_{k=1}^{\binom{N}{n}}\frac{N}{n}\sum_{i=1}^{n}\sum_{l=1}^{\binom{M_i}{m_i}}\left(\frac{\frac{M_i}{m_i}\sum_{j=1}^{m_i} y_{kilj}}{\binom{M_i}{m_i}}\right)\right] = \frac{1}{\binom{N}{n}}\left[\sum_{k=1}^{\binom{N}{n}}\frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{m_i}\left(\frac{\sum_{l=1}^{\binom{M_i}{m_i}}\sum_{j=1}^{m_i} y_{kilj}}{\binom{M_i}{m_i}}\right)\right]$$

$$= \frac{1}{\binom{N}{n}}\left[\sum_{k=1}^{\binom{N}{n}}\frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{m_i}\left(\frac{\binom{M_i-1}{m_i-1}\sum_{j=1}^{M_i} y_{kij}}{\binom{M_i}{m_i}}\right)\right] = \frac{N}{n\binom{N}{n}}\left(\sum_{k=1}^{\binom{N}{n}}\sum_{i=1}^{n}\sum_{j=1}^{M_i} y_{kilj}\right)$$

$$= \frac{N}{n\left(\frac{N}{n}\right)\binom{N-1}{n-1}}\binom{N-1}{n-1}\sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij} = \sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij} = Y$$

por lo tanto, $\hat{Y}_u$ es un estimador insesgado de *Y*. Igual ocurre con $\hat{\bar{Y}}_u$, ya que

$$E(\hat{\bar{Y}}_u) = E\left(\frac{\hat{Y}_u}{N}\right) = \frac{1}{N}E(\hat{Y}_u) = \frac{Y}{N} = \bar{Y}$$

Para hallar la varianza de $\hat{Y}_u$, se hace uso de la igualdad

$$V(\hat{Y}_u) = V_1\left[E_2(\hat{Y}_u)\right] + E_1\left[V_2(\hat{Y}_u)\right]$$

donde,

$$E_2(\hat{Y}_u) = N\frac{\sum_{i=1}^{n} Y_i}{n}$$

aplicando la fórmula $V(\beta) = E(\beta^2) - (E(\beta)^2)$, se llega a la siguiente expresión:

$$V_1\left[E_2(\hat{Y}_u)\right] = E_1\left[\left(E_2(\hat{Y}_u)\right)^2\right] - \left[E_1\left(E_2(\hat{Y}_u)\right)\right]^2 = E_1\left[\left(E_2(\hat{Y}_u)\right)^2\right] - Y^2 \qquad (13.1)$$

entonces, se desarrolla el primer término de la diferencia,

$$E_1\left[\left(E_2(\hat{Y}_u)\right)^2\right] = \frac{\sum_{k=1}^{\binom{N}{n}}\left(N\frac{\sum_{i=1}^{n}Y_i}{n}\right)^2}{\binom{N}{n}} = \frac{N^2\sum_{k=1}^{\binom{N}{n}}\left(\sum_{i=1}^{n}Y_i\right)^2}{n^2\binom{N}{n}}$$

$$= \frac{N^2\sum_{k=1}^{\binom{N}{n}}\left[\sum_{i=1}^{n}Y_i^2 + 2\sum_{i=1}^{n-1}\sum_{h=i+1}^{n}Y_iY_h\right]}{n^2\binom{N}{n}} = \frac{N^2\left[\binom{N-1}{n-1}\sum_{i=1}^{N}Y_i^2 + 2\binom{N-2}{n-2}\sum_{i=1}^{n-1}\sum_{h=i+1}^{n}Y_iY_h\right]}{n^2\left(\frac{N}{n}\right)\binom{N-1}{n-1}}$$

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

$$= \frac{N^2}{Nn}\left[\sum_{i=1}^{N}Y_i^2 + 2\left(\frac{n-1}{N-1}\right)\sum_{i=1}^{N-1}\sum_{h=i+1}^{N}Y_iY_h + \left(\frac{n-1}{N-1}\right)\sum_{i=1}^{N}Y_i^2 - \left(\frac{n-1}{N-1}\right)\sum_{i=1}^{N}Y_i^2\right]$$

$$= \frac{N^2}{Nn}\left[\left(\frac{n-1}{N-1}\right)\left(\sum_{i=1}^{N}Y_i^2 + 2\sum_{i=1}^{N-1}\sum_{h=i+1}^{N}Y_iY_h\right) + \left(1-\frac{n-1}{N-1}\right)\sum_{i=1}^{N}Y_i^2\right]$$

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

sustituyendo este resultado en (13.1),

$$V_1\left[E_2(\hat{Y}_u)\right] = \frac{N^2}{Nn}\left[\left(\frac{n-1}{N-1}\right)Y^2 + \left(\frac{N-n}{N-1}\right)\sum_{i=1}^{N}Y_i^2\right] - Y^2 = \frac{N^2}{Nn}\left[\left(\frac{n-1}{N-1}\right)Y^2 + \left(\frac{N-n}{N-1}\right)\sum_{i=1}^{N}Y_i^2 - \frac{Nn}{N^2}Y^2\right]$$

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

$$= \frac{N^2(N-n)}{Nn}\frac{\sum_{i=1}^{N}Y_i^2 - N\bar{Y}^2}{N-1} = \frac{N^2(1-f_1)}{n}\frac{\sum_{i=1}^{N}\left(Y_i-\bar{Y}\right)^2}{N-1}$$

por lo tanto,

$$V_1\left[E_2(\hat{Y}_u)\right]=\frac{N^2(1-f_1)}{n}\frac{\sum_{i=1}^{N}\left(Y_i-\bar{Y}\right)^2}{N-1}$$

ahora se hallará $E_1\left[V_2(\hat{Y}_u)\right]$,

$$V_2(\hat{Y}_u) \;=V_2\left(\frac{N\sum_{i=1}^{n}\hat{Y}_i}{n}\right)=\frac{N^2}{n^2}V_2\left(\sum_{i=1}^{n}\hat{Y}_i\right)=\frac{N^2}{n^2}\left[\sum_{i=1}^{n}V(\hat{Y}_i)+\sum_{i=1}^{n-1}\sum_{h=i+1}^{n}COV(\hat{Y}_i,\hat{Y}_h)\right]$$

como las muestras se toman independientes en cada conglomerado, $COV(Y_i,Y_h)=0$, para todo $i\neq h$, $i,h=1,2,...,N$.

Por lo tanto,

$$V_2(\hat{Y}_u) =\frac{N^2}{n^2}\sum_{i=1}^{n}V(\hat{Y}_i)=\frac{N^2}{n^2}\sum_{i=1}^{n}\frac{M_i^2(M_i-m_i)}{m_i}\frac{\sum_{j=1}^{M_i}(y_{ij}-\bar{Y}_i)^2}{M_i-1}$$

$$=\frac{N^2}{n^2}\sum_{i=1}^{n}\frac{M_i^2(1-f_{2i})}{m_i}\frac{\sum_{j=1}^{M_i}(y_{ij}-\bar{Y}_i)^2}{M_i-1}$$

ya que en cada conglomerado se está tomando una muestra aleatoria simple de m_i elementos. Tomando la esperanza se tiene,

$$E_1\left[V_2(\hat{Y}_u)\right] =\frac{1}{\binom{N}{n}}\sum_{k=1}^{\binom{N}{n}}\frac{N^2}{n^2}\sum_{i=1}^{n}\frac{M_{ki}^2(1-f_{2ki})}{m_{ki}}\frac{\sum_{j=1}^{M_{ki}}(y_{kij}-\bar{Y}_{ki})^2}{M_{ki}-1}$$

$$=\frac{N^2}{n^2\binom{N}{n}}\sum_{k=1}^{\binom{N}{n}}\sum_{i=1}^{n}\frac{M_{ki}^2(1-f_{2ki})}{m_{ki}}\frac{\sum_{j=1}^{M_{ki}}(y_{kij}-\bar{Y}_{ki})^2}{M_{ki}-1}$$

$$=\frac{N^2}{n^2\binom{N}{n}}\binom{N-1}{n-1}\sum_{i=1}^{N}\frac{M_i^2(1-f_{2i})}{m_i}\frac{\sum_{j=1}^{M_i}(y_{ij}-\bar{Y}_i)^2}{M_i-1}$$

$$= \frac{N}{n} \sum_{i=1}^{N} \frac{M_i^2 (1 - f_{2i})}{m_i} \frac{\sum_{j=1}^{M_i} (y_{ij} - \bar{Y}_i)^2}{M_i - 1}$$

$$= \frac{N}{n} \sum_{i=1}^{N} \frac{M_i^2 (1 - f_{2i})}{m_i} S_{2i}^2$$

donde,

$$S_{2i}^2 = \frac{\sum_{j=1}^{M_i} (y_{ij} - \bar{Y}_i)^2}{M_i - 1}$$

finalmente,

$$V(\hat{Y}_u) = \frac{N^2 (1 - f_1)}{n} \frac{\sum_{i=1}^{N} (Y_i - \bar{Y})^2}{N - 1} + \frac{N}{n} \sum_{i=1}^{N} \frac{M_i^2 (1 - f_{2i})}{m_i} S_{2i}^2$$

$$V(\hat{\bar{Y}}_u) = \frac{1}{N^2} V(\hat{Y}_u) = \frac{(1 - f_1)}{n} \frac{\sum_{i=1}^{N} (Y_i - \bar{Y})^2}{N - 1} + \frac{1}{nN} \sum_{i=1}^{N} \frac{M_i^2 (1 - f_{2i})}{m_i} S_{2i}^2$$

Nótese que si los conglomerados fueran de igual tamaño y se toman muestras de igual tamaño en cada conglomerado, es decir, si

$$M_1 = M_2 = \ldots = M_N = M \quad \text{y} \quad m_1 = m_2 = \ldots = m_N = m \quad \Rightarrow \quad f_{21} = f_{22} = \ldots = f_{2N} = f$$

se obtiene la misma fórmula vista en el capítulo anterior, cuando los conglomerados eran de igual tamaño. De manera que el sub-muestreo con unidades de igual tamaño es un caso particular del sub-muestreo con unidades de diferente tamaño, como era de suponer.

Las varianzas estimadas vienen dadas por,

$$\hat{V}(\hat{Y}_u) = \frac{N^2 (1 - f_1)}{n} \frac{\sum_{i=1}^{n} (\hat{Y}_i - \hat{\bar{Y}}_u)^2}{n - 1} + \frac{N}{n} \sum_{i=1}^{n} \frac{M_i^2 (1 - f_{2i})}{m_i} s_{2i}^2$$

$$\hat{V}(\hat{\bar{Y}}_u) = \frac{1}{N^2} \hat{V}(\hat{Y}_u) = \frac{(1 - f_1)}{n} \frac{\sum_{i=1}^{n} (\hat{Y}_i - \hat{\bar{Y}}_u)^2}{n - 1} + \frac{1}{nN} \sum_{i=1}^{n} \frac{M_i^2 (1 - f_{2i})}{m_i} s_{2i}^2$$

donde, $s_{2i}^2 = \dfrac{\sum_{j=1}^{m_i}(y_{ij} - \hat{\bar{Y}}_i)^2}{m_i - 1}$, y son estimadores insesgados de sus respectivas varianzas.

Cabe destacar que estos estimadores suelen usarse para estimar el total y el promedio por conglomerados, sin embargo también pueden utilizarse para estimar el promedio por elemento. En este orden de ideas, a continuación se presenta dicho estimador , su varianza y su varianza estimada.

$$\hat{\bar{\bar{Y}}}_u = \frac{\hat{Y}_u}{M_0}$$

$$V\left(\hat{\bar{\bar{Y}}}_u\right) = \frac{1}{M_0^2} V(\hat{Y}_u) = \frac{1}{M_0^2}\left[\frac{N^2(1-f_1)}{n}\frac{\sum_{i=1}^{N}(Y_i - \bar{\bar{Y}})^2}{N-1} + \frac{N}{n}\sum_{i=1}^{N}\frac{M_i^2(1-f_{2i})}{m_i} S_{2i}^2\right]$$

$$\hat{V}\left(\hat{\bar{\bar{Y}}}_u\right) = \frac{1}{M_0^2} \hat{V}(\hat{Y}_u) = \frac{1}{M_0^2}\left[\frac{N^2(1-f_1)}{n}\frac{\sum_{i=1}^{n}(\hat{Y}_i - \hat{\bar{\bar{Y}}}_u)^2}{n-1} + \frac{N}{n}\sum_{i=1}^{n}\frac{M_i^2(1-f_{2i})}{m_i} s_{2i}^2\right]$$

Por otra parte, los estimadores de razón al tamaño, que se detallarán más adelante, son mayoritariamente utilizados para estimar el total y el promedio por elemento, pero igualmente se puede estimar el promedio por conglomerados.

13.1.2.- Estimadores insesgados - selección con reemplazamiento

Este diseño difiere del anterior en que la selección de las unidades o muestra de primera etapa, se hace a través de muestreo aleatorio simple, pero con reemplazamiento, de manera que las demostraciones se se pueden tomar del caso anterior, sustituyendo en el primer sumando de la varianza el término correspondiente expresado en el capítulo 3 *(Muestreo Aleatorio Simple con Reemplazamiento)*, por tal motivo, aquí sólo se presentarán las fórmulas. De este modo, se tiene que los estimadores del total , del total promedio por conglomerado y del promedio por elemento son,

$$\hat{Y}_{ucr} = \frac{N}{n}\sum_{i=1}^{n}\hat{Y}_i \qquad ; \qquad \hat{\bar{Y}}_{ucr} = \frac{\sum_{i=1}^{n}\hat{Y}_i}{n} = \frac{\hat{Y}_u}{N}$$

que son exactamente los mismos que en el apartado anterior, tal como ocurre en el muestreo aleatorio simple con y sin reemplazamiento. Sus respectivas varianzas son,

$$V(\hat{Y}_{ucr}) = \frac{N^2}{n}\frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}{N-1} + \frac{N}{n}\sum_{i=1}^{N}\frac{M_i^2(1-f_{2i})}{m_i}S_{2i}^2$$

$$V(\hat{\bar{Y}}_{ucr}) = \frac{1}{n}\frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}{N-1} + \frac{1}{nN}\sum_{i=1}^{N}\frac{M_i^2(1-f_{2i})}{m_i}S_{2i}^2$$

y las varianzas estimadas,

$$\hat{V}(\hat{Y}_{ucr}) = \frac{N^2}{n}\frac{\sum_{i=1}^{n}(\hat{Y}_i - \hat{\bar{Y}}_u)^2}{n-1} + \frac{N}{n}\sum_{i=1}^{n}\frac{M_i^2(1-f_{2i})}{m_i}s_{2i}^2$$

$$\hat{V}(\hat{\bar{Y}}_{ucr}) = \frac{1}{n}\frac{\sum_{i=1}^{n}(\hat{Y}_i - \hat{\bar{Y}}_u)^2}{n-1} + \frac{1}{nN}\sum_{i=1}^{n}\frac{M_i^2(1-f_{2i})}{m_i}s_{2i}^2$$

donde, $s_{2i}^2 = \frac{\sum_{j=1}^{m_i}(y_{ij} - \hat{\bar{Y}}_i)^2}{m_i - 1}$, y son estimadores insesgados de sus respectivas varianzas.

Igual que en caso anterior,

$$\hat{\bar{\bar{Y}}}_{ucr} = \frac{\hat{Y}_{ucr}}{M_0} \quad , \quad V\left(\hat{\bar{\bar{Y}}}_{ucr}\right) = \frac{1}{M_0^2}V\left(\hat{Y}_{ucr}\right) \quad , \quad \hat{V}\left(\hat{\bar{\bar{Y}}}_{ucr}\right) = \frac{1}{M_0^2}\hat{V}\left(\hat{Y}_{ucr}\right)$$

13.1.3.- Estimadores de razón al tamaño

Ahora se tiene que aunque la selección de los conglomerados se hace con iguales probabilidades, se considera el tamaño de los mismos para realizar las estimaciones, es decir, se le asigna a los elementos que no están en la muestra, el promedio de los que si están, esto da origen a los "estimadores de razón al tamaño", los mismos que se trabajaron en el apartado 10.2, pero ahora se hará en submuestreo. El desarrollo se muestra a continuación,

$$\bar{\bar{Y}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{M_i} y_{ij} + \sum_{k=n+1}^{N}\sum_{j=1}^{M_k} y_{kj}}{\sum_{i=1}^{N} M_i} \qquad \text{y su estimador} \qquad \hat{\bar{\bar{Y}}}_R = \frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{m_i} y_{ij} + \sum_{j=m_i+1}^{M_i} y'_{ij}\right) + \sum_{k=n+1}^{N}\sum_{j=1}^{M_k} y''_{kj}}{\sum_{i=1}^{N} M_i}$$

Considerando que,

$$y'_{ij} = y''_{kj} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i} y_{ij}}{\sum_{i=1}^{n} m_i} = \bar{\bar{y}} \qquad \forall\ k, j \quad , \ k = n+1,\ldots,N \qquad , \ j = 1,\ldots,M_i$$

entonces,

$$\hat{\bar{\bar{Y}}}_{RT} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i} y_{ij} + \sum_{i=1}^{n}\sum_{j=m_i+1}^{M_i}\bar{\bar{y}} + \sum_{k=n+1}^{N}\sum_{j=1}^{M_k}\bar{\bar{y}}}{\sum_{i=1}^{N} M_i} = \frac{\left(\frac{\sum_{i=1}^{n} m_i}{\sum_{i=1}^{n} m_i}\right)\sum_{i=1}^{n}\sum_{j=1}^{m_i} y_{ij} + \sum_{i=1}^{n}\sum_{j=m_i+1}^{M_i}\bar{\bar{y}} + \sum_{k=n+1}^{N}\sum_{j=1}^{M_k}\bar{\bar{y}}}{\sum_{i=1}^{N} M_i}$$

$$= \frac{\sum_{i=1}^{n} m_i\,\bar{\bar{y}} + \sum_{i=1}^{n}(M_i - m_i)\bar{\bar{y}} + \sum_{k=n+1}^{N} M_i\bar{\bar{y}}}{\sum_{i=1}^{N} M_i} = \frac{\sum_{i=1}^{n} M_i\,\bar{\bar{y}} + \sum_{k=n+1}^{N} M_i\bar{\bar{y}}}{\sum_{i=1}^{N} M_i} = \left[\frac{\sum_{i=1}^{n} M_i + \sum_{k=n+1}^{N} M_i}{\sum_{i=1}^{N} M_i}\right]\bar{\bar{y}} = \bar{\bar{y}}$$

luego, $\hat{\bar{\bar{Y}}}_{RT} = \bar{\bar{y}}$, entonces,

$$\hat{Y}_{RT} = M_0\hat{\bar{\bar{Y}}}_{RT} = M_0\bar{\bar{y}} = M_0\frac{\sum_{i=1}^{n}\sum_{j=1}^{M_i} y_{ij}}{\sum_{i=1}^{n} M_i} = \sum_{i=1}^{N} M_i\frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} M_i}$$

$$\hat{\bar{\bar{Y}}}_{RT} \;=\; \frac{\hat{Y}_{RT}}{N} \;=\; \frac{1}{N}\sum_{i=1}^{N} M_i \,\frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} M_i} \;=\; \frac{M_0}{N}\,\bar{\bar{y}} \;=\; \frac{M_0}{N}\,\hat{\bar{\bar{Y}}}_{RT}$$

Al ser estimadores del tipo razón, no son insesgados, entonces se procederá a determinar su Error Cuadrático Medio. Nótese que tienen la misma estructura que los mostrados en (10. 10), (10. 14) y (10.18), por lo tanto el componente de primera etapa del ECM se mantiene igual al error cuadrático medio mencionados arriba. En cuanto a la segunda etapa, esta vez no varía respecto a los estimadores insesgados en submuestreo con unidades de diferente tamaño, por lo tanto el componente de segunda etapa de la varianza coincide con éstos. De manera que el error cuadrático medio queda como sigue,

$$ECM(\hat{Y}_{RT}) \approx N^2\left(\frac{1-f_1}{n}\right)\frac{\sum_{i=1}^{N}\left(Y_i - M_i\bar{\bar{Y}}\right)^2}{N-1} + \frac{N}{n}\sum_{i=1}^{N}\frac{M_i^2(1-f_{2i})}{m_i}\,S_{2i}^2$$

$$ECM(\hat{\bar{Y}}_{RT}) \approx \left(\frac{1-f_1}{n}\right)\frac{\sum_{i=1}^{N}\left(Y_i - M_i\bar{\bar{Y}}\right)^2}{N-1} + \frac{1}{Nn}\sum_{i=1}^{N}\frac{M_i^2(1-f_{2i})}{m_i}\,S_{2i}^2$$

$$ECM(\hat{\bar{\bar{Y}}}_{RT}) \approx \frac{N^2}{M_0^2}\left(\frac{1-f_1}{n}\right)\frac{\sum_{i=1}^{N}\left(Y_i - M_i\bar{\bar{Y}}\right)^2}{N-1} + \frac{N}{M_0^2 n}\sum_{i=1}^{N}\frac{M_i^2(1-f_{2i})}{m_i}\,S_{2i}^2$$

Los Errores Cuadrático Medio estimados vienen dados por,

$$\hat{ECM}(\hat{Y}_{RT}) \approx N^2\left(\frac{1-f_1}{n}\right)\frac{\sum_{i=1}^{n}\left(Y_i - M_i\hat{\bar{\bar{Y}}}_{RT}\right)^2}{n-1} + \frac{N}{n}\sum_{i=1}^{N}\frac{M_i^2(1-f_{2i})}{m_i}\,s_{2i}^2$$

$$\hat{ECM}(\hat{\bar{Y}}_{RT}) \approx \left(\frac{1-f_1}{n}\right)\frac{\sum_{i=1}^{n}\left(Y_i - M_i\hat{\bar{\bar{Y}}}_{RT}\right)^2}{n-1} + \frac{1}{Nn}\sum_{i=1}^{N}\frac{M_i^2(1-f_{2i})}{m_i}\,s_{2i}^2$$

$$\hat{ECM}(\hat{\bar{\bar{Y}}}_{RT}) \approx \frac{N^2}{M_0^2}\left(\frac{1-f_1}{n}\right)\frac{\sum_{i=1}^{n}\left(Y_i - M_i\hat{\bar{\bar{Y}}}_{RT}\right)^2}{n-1} + \frac{N}{M_0^2 n}\sum_{i=1}^{N}\frac{M_i^2(1-f_{2i})}{m_i}\,s_{2i}^2$$

Donde,

$$S_{2i}^2 = \frac{\sum_{j=1}^{M_i}(y_{ij} - \bar{Y}_i)^2}{M_i - 1} \qquad , \qquad s_{2i}^2 = \frac{\sum_{j=1}^{m_i}(y_{ij} - \hat{\bar{Y}}_i)^2}{m_i - 1}$$

13.2.- Selección de conglomerados con probabilidades desiguales

En este caso la selección de las unidades primarias o de primera etapa, se seleccionan con una probabilidad asignada a cada una, y que se denotará por z_i, es decir, que z_i será la probabilidad de seleccionar la *i-ésima* unidad primaria en la primera etapa. La selección de los elementos se hace de manera aleatoria simple, a efectos de las fórmulas y desarrollos que se presentarán, pero se debe hacer la salvedad que puede hacerse de cualquier manera (sistemático, estratificado u otro), pero se deben adaptar las fórmulas.

Los estimadores y el desarrollo teórico es similar al presentado en la sección 10.2.- (Selección de Conglomerados con Probabilidades Desiguales).

El estimador del total a utilizar es el siguiente,

$$\hat{Y}_{PPZ} = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{Y}_i}{z_i} = \frac{1}{n}\sum_{i=1}^{n}\frac{M_i\bar{y}_i}{z_i}$$

y su esperanza

$$E\left(\hat{Y}_{PPZ}\right) = E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{Y}_i}{z_i}\right] = \frac{1}{n}E\left[\sum_{i=1}^{n}\frac{\hat{Y}_i}{z_i}\right]$$, nótese que es muy similar a la presentada en la sección 10.2 (Selección de Conglomerados con Probabilidades Desiguales), y sólo difiere en el término $\hat{Y}_i$, que el muestreo monoetápico es Y_i; pero como $\hat{Y}_i$ es un estimador insesgado de Y_i, al tomar esperanza queda la misma expresión, ya que también se tienen N^n muestras posibles de primera etapa, es decir,

$E\left(\hat{Y}_{PPZ}\right) = \frac{1}{n}\sum_{k=1}^{N^n}\sum_{i=1}^{n}P_k\frac{Y_{ki}}{z_i} = \sum_{i=1}^{N}\hat{Y}_i = Y$, como ya se demostró en dicha oportunidad, donde P_k es la probabilidad de selección de la *k-ésima* muestra, $k=1,2,...,N^n$. Entonces, $\hat{Y}_{PPZ}$ es un estimador insesgado de *Y*.

Para hallar la varianza, se utilizará el desarrollo de esperanzas y varianzas de primera y segunda etapa utilizadas en el punto 10.1.1. (Estimadores en Selección de Conglomerados con Iguales Probabilidades, en Muestreo Monoetápico Aleatorio de Conglomerados).

Se denominará a

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{z_i}$$

como la esperanza de segunda etapa de Y_{PPZ}, que consiste en tomar la esperanza sobre los elementos (o unidades de segunda etapa) conociendo ya la muestra de primera etapa, es decir, las n unidades de primera etapa seleccionadas; de modo que,

$$E_2\left(\hat{Y}_{PPZ}\right)=\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{z_i} \qquad => \qquad E\left(\hat{Y}_{PPZ}\right)=E_1\left[E_2\left(\hat{Y}_{PPZ}\right)\right]=Y$$

y $$V\left(\hat{Y}_{PPZ}\right)=E\left[\left(\hat{Y}_{PPZ}-Y\right)^2\right]=E_1\left[E_2\left(\hat{Y}_{PPZ}^2-2Y\hat{Y}_{PPZ}+Y^2\right)\right]=E_1\left[E_2\left(\hat{Y}_{PPZ}^2\right)-2YE_2\left(\hat{Y}_{PPZ}\right)+Y^2\right]$$

pero como $V_2\left(\hat{Y}_{PPZ}\right)=E_2\left(\hat{Y}_{PPZ}^2\right)-\left[E_2\left(\hat{Y}_{PPZ}\right)\right]^2$ => $V\left(\hat{Y}_{PPZ}\right)=E_1\left[V_2\left(\hat{Y}_{PPZ}\right)+\left(E_2\left(\hat{Y}_{PPZ}\right)\right)^2-2YE_2\left(\hat{Y}_{PPZ}\right)+Y^2\right]$

$$\begin{aligned}&=E_1\left[V_2\left(\hat{Y}_{PPZ}\right)\right]+\left[E_2\left(\hat{Y}_{PPZ}\right)\right]^2-2YE_1\left[E_2\left(\hat{Y}_{PPZ}\right)\right]+Y^2\\&=E_1\left[V_2\left(\hat{Y}_{PPZ}\right)\right]+\left[E_2\left(\hat{Y}_{PPZ}\right)\right]^2-2Y^2+Y^2\\&=E_1\left[V_2\left(\hat{Y}_{PPZ}\right)\right]+\left[E_2\left(\hat{Y}_{PPZ}\right)\right]^2-Y^2\end{aligned}$$

haciendo $Y=E_1\left[E_2\left(\hat{Y}_{PPZ}\right)\right]$,

$$V\left(\hat{Y}_{PPZ}\right)=E_1\left[V_2\left(\hat{Y}_{PPZ}\right)\right]+E_1\left[\left(E_2\left(\hat{Y}_{PPZ}\right)\right)^2\right]-\left[E_1\left[E_2\left(\hat{Y}_{PPZ}\right)\right]\right]^2=E_1\left[V_2\left(\hat{Y}_{PPZ}\right)\right]+V_1\left[E_2\left(\hat{Y}_{PPZ}\right)\right] \qquad (13.2)$$

desarrollando cada uno de los términos,

$$V_1\left[E_2\left(\hat{Y}_{PPZ}\right)\right]=E_1\left[E_2\left(\hat{Y}_{PPZ}\right)\right]^2-Y^2 \qquad (13.3)$$

$$E_1\left[E_2\left(\hat{Y}_{PPZ}\right)\right]^2=\sum_{k=1}^{N^n}P_k\left[\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{z_i}\right]=\frac{1}{n^2}\sum_{k=1}^{N^n}P_k\left(\sum_{i=1}^{n}\frac{Y_i}{z_i}\right)=\frac{1}{n^2}\left[n\sum_{i=1}^{N}\frac{Y_i^2}{z_i}+n(n-1)Y^2\right]=\frac{1}{n}\sum_{i=1}^{N}\frac{Y_i^2}{z_i}+\frac{(n-1)}{n}Y^2$$

sustituyendo en (13.2)

$$V_1\left[E_2\left(\hat{Y}_{PPZ}\right)\right]=\frac{1}{n}\sum_{i=1}^{N}\frac{Y_i^2}{z_i}+\frac{n-1}{n}Y^2-Y^2=\frac{1}{n}\sum_{i=1}^{N}\frac{Y_i^2}{z_i}-\frac{1}{n}Y^2=\frac{1}{n}\left[\sum_{i=1}^{N}\frac{Y_i^2}{z_i}-Y^2\right]=\frac{1}{n}\left[\sum_{i=1}^{N}\frac{Y_i^2}{z_i}-2Y\sum_{i=1}^{N}Y_i+Y^2\right]$$

$$\begin{aligned}&=\frac{1}{n}\left[\sum_{i=1}^{N}z_i\left(\frac{Y_i}{z_i}\right)^2-2Y\sum_{i=1}^{N}z_i\left(\frac{Y_i}{z_i}\right)+Y^2\sum_{i=1}^{N}z_i\right]=\frac{1}{n}\sum_{i=1}^{N}\left[z_i\left(\frac{Y_i}{z_i}\right)^2-2Yz_i\left(\frac{Y_i}{z_i}\right)+Y^2z_i\right]\\&=\frac{1}{n}\sum_{i=1}^{N}z_i\left[\left(\frac{Y_i}{z_i}\right)^2-2Y\left(\frac{Y_i}{z_i}\right)+Y^2\right]=\frac{1}{n}\sum_{i=1}^{N}z_i\left(\frac{Y_i}{z_i}-Y\right)^2\end{aligned} \qquad (13.4)$$

Para desarrollar el otro miembro de (13.2), se debe hallar primero la varianza de segunda etapa de $\hat{Y}_{PPZ}$ y luego

hallar su esperanza de primera etapa.

$$V_2\left(\hat{\bar{Y}}_{PPZ}\right)=V_2\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{Y}_i}{z_i}\right)=\frac{1}{n^2}V_2\left(\sum_{i=1}^{n}\frac{\hat{Y}_i}{z_i}\right)=\frac{1}{n^2}\left[\sum_{i=1}^{n}V_2\left(\frac{\hat{Y}_i}{z_i}\right)+2\sum_{i=1}^{n-1}\sum_{h=i+1}^{n}COV_2\left(\frac{\hat{Y}_i}{z_i},\frac{\hat{Y}_h}{z_h}\right)\right]$$

pero como las muestras son independientes en cada conglomerado, $COV_2\left(\frac{\hat{Y}_i}{z_i},\frac{\hat{Y}_h}{z_h}\right)=0 \quad ,\forall i\neq h\ ,\ i,h=1,2,...,N$,

entonces,

$$V_2\left(\hat{\bar{Y}}_{PPZ}\right)=\frac{1}{n^2}\sum_{i=1}^{n}V_2\left(\frac{\hat{Y}_i}{z_i}\right)=\frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{z_i^2}V_2\left(\hat{Y}_i\right)=\frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{z_i^2}\frac{M_i^2(M_i-m_i)}{M_i\,m_i}\frac{\sum_{j=1}^{M_i}\left(y_{ij}-\bar{Y}_i\right)^2}{M_i-1}$$

por lo tanto,

$$E_1\left[V_2\left(\hat{\bar{Y}}_{PPZ}\right)\right]=E_1\left[\frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{z_i^2}\frac{M_i^2(M_i-m_i)}{M_i\,m_i}\frac{\sum_{j=1}^{M_i}\left(y_{ij}-\bar{Y}_i\right)^2}{M_i-1}\right]=\frac{1}{n^2}\,E_1\left[\sum_{i=1}^{n}\frac{1}{z_i^2}\frac{M_i^2(M_i-m_i)}{M_i\,m_i}\frac{\sum_{j=1}^{M_i}\left(y_{ij}-\bar{Y}_i\right)^2}{M_i-1}\right]$$

$$=\frac{1}{n^2}\,E_1\left[\sum_{i=1}^{n}\frac{M_i^2(1-f_{2i})}{m_i\,z_i^2}\frac{\sum_{j=1}^{M_i}\left(y_{ij}-\bar{Y}_i\right)^2}{M_i-1}\right]=\frac{1}{n^2}\sum_{k=1}^{N^n}P_k\sum_{i=1}^{n}\frac{M_i^2(1-f_{2i})}{m_i\,z_i^2}\frac{\sum_{j=1}^{M_i}\left(y_{ij}-\bar{Y}_i\right)^2}{M_i-1}$$

$$=\frac{1}{n^2}\sum_{i=1}^{N}n\left(\frac{M_i^2(1-f_{2i})}{m_i\,z_i^2}\frac{\sum_{j=1}^{M_i}\left(y_{ij}-\bar{Y}_i\right)^2}{M_i-1}\right)\left(z_i\sum_{j1=1}^{N}\sum_{j2=1}^{N}\cdots\sum_{j(n-1)=1}^{N}z_{j1}\,z_{j2}\cdots z_{j(n-1)}\right)$$

$$=\frac{1}{n^2}\sum_{i=1}^{N}n\left(\frac{M_i^2(1-f_{2i})}{m_i\,z_i^2}\frac{\sum_{j=1}^{M_i}\left(y_{ij}-\bar{Y}_i\right)^2}{M_i-1}\right)\left[z_i\left(\sum_{j1=1}^{N}z_{j1}\right)\left(\sum_{j1=1}^{N}z_{j2}\right)\cdots\left(\sum_{j(n-1)=1}^{N}z_{j(n-1)}\right)\right]$$

$$=\frac{1}{n}\sum_{i=1}^{N}\frac{M_i^2(1-f_{2i})}{m_i\,z_i^2}\frac{\sum_{j=1}^{M_i}\left(y_{ij}-\bar{Y}_i\right)^2}{M_i-1}z_i$$

$$=\frac{1}{n}\sum_{i=1}^{N}\frac{M_i^2(1-f_{2i})}{m_i\,z_i}\frac{\sum_{j=1}^{M_i}\left(y_{ij}-\bar{Y}_i\right)^2}{M_i-1}$$

Por lo tanto, sustituyendo este resultado y el obtenido en (13.4) en (13.2) se obtiene la varianza definitiva de $\hat{Y}_{PPZ}$, que es,

$$V\left(\hat{Y}_{PPZ}\right)=\frac{1}{n}\sum_{i=1}^{N} z_i\left(\frac{Y_i}{z_i}-Y\right)^2+\frac{1}{n}\sum_{i=1}^{N}\frac{M_i^2(1-f_{2i})}{m_i\, z_i}\frac{\sum_{j=1}^{M_i}\left(y_{ij}-\bar{Y}_i\right)^2}{M_i-1}$$

y un estimador insesgado es,

$$\hat{V}\left(\hat{Y}_{PPZ}\right)=\frac{1}{n}\frac{\sum_{i=1}^{n} z_i\left(\frac{\hat{Y}_i}{z_i}-\hat{Y}_{PPZ}\right)^2}{n-1}$$

Para el estimador del promedio por elemento,

$$\hat{\bar{\bar{Y}}}_{PPZ}=\frac{\hat{Y}_{PPZ}}{M_0}=\frac{1}{nM_0}\sum_{i=1}^{n}\frac{\hat{Y}_i}{z_i}=\frac{1}{n}\sum_{i=1}^{n}\frac{M_i\bar{y}_i}{z_i}$$

$$V\left(\hat{\bar{\bar{Y}}}_{PPZ}\right)=\frac{1}{M_0^2}V\left(\hat{Y}_{PPZ}\right)=\frac{1}{n\,M_0^2}\sum_{i=1}^{N} z_i\left(\frac{Y_i}{z_i}-Y\right)+\frac{1}{n\,M_0^2}\sum_{i=1}^{N}\frac{M_i^2(1-f_{2i})}{m_i\, z_i}\frac{\sum_{j=1}^{M_i}\left(y_{ij}-\bar{Y}_i\right)^2}{M_i-1}$$

$$\hat{V}\left(\hat{\bar{\bar{Y}}}_{PPZ}\right)=\frac{1}{M_0^2}\hat{V}\left(\hat{Y}_{PPZ}\right)=\frac{1}{nM_0^2}\frac{\sum_{i=1}^{n} z_i\left(\frac{\hat{Y}_i}{z_i}-\hat{Y}_{PPZ}\right)^2}{n-1}$$

y para el estimador del promedio por conglomerado,

$$\hat{\bar{Y}}_{PPZ}=\frac{\hat{Y}_{PPZ}}{N}=\frac{1}{nN}\sum_{i=1}^{n}\frac{\hat{Y}_i}{z_i}=\frac{1}{n}\sum_{i=1}^{n}\frac{M_i\bar{y}_i}{z_i}$$

$$V\left(\hat{\bar{Y}}_{PPZ}\right)=\frac{1}{N^2}V\left(\hat{Y}_{PPZ}\right)=\frac{1}{n\,N^2}\sum_{i=1}^{N} z_i\left(\frac{Y_i}{z_i}-Y\right)+\frac{1}{n\,N^2}\sum_{i=1}^{N}\frac{M_i^2(1-f_{2i})}{m_i\, z_i}\frac{\sum_{j=1}^{M_i}\left(y_{ij}-\bar{Y}_i\right)^2}{M_i-1}$$

$$\hat{V}\left(\hat{\bar{Y}}_{PPZ}\right)=\frac{1}{N^2}\hat{V}\left(\hat{Y}_{PPZ}\right)=\frac{1}{nN^2}\frac{\sum_{i=1}^{n} z_i\left(\frac{\hat{Y}_i}{z_i}-\hat{Y}_{PPZ}\right)^2}{n-1}$$

Igual que en el muestreo monoetápico, cuando la selección se hace con probabilidades diferentes, un caso particular, que merece especial atención es cuando la selección se hace con probabilidad proporcional al tamaño, y

resulta de sustituir z_i por $M_i \big/ M_0$. El resultado es,

$$\hat{Y}_{PPS} = \frac{M_0}{n}\sum_{i=1}^{n}\bar{y}_i$$

$$V\left(\hat{Y}_{PPS}\right) = \frac{M_0}{n}\sum_{i=1}^{N} M_i\left(\bar{y}_i - \bar{\bar{Y}}\right)^2 + \frac{M_0^2}{n}\sum_{i=1}^{N}\frac{M_i^2(1-f_{2i})}{m_i\, z_i}\frac{\sum_{j=1}^{M_i}\left(y_{ij}-\bar{Y}_i\right)^2}{M_i-1}$$

$$\hat{V}\left(\hat{Y}_{PPS}\right) = \frac{M_0}{n}\frac{\sum_{i=1}^{n} M_i\left(\bar{y}_i - \bar{\bar{Y}}_{PPS}\right)^2}{n-1}$$

para el estimador del promedio por elemento,

$$\hat{\bar{\bar{Y}}}_{PPS} = \frac{\hat{Y}_{PPZ}}{M_0} = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i$$

$$V\left(\hat{\bar{\bar{Y}}}_{PPS}\right) = \frac{1}{M_0^2}V\left(\hat{Y}_{PPS}\right) = \frac{1}{nM_0}\sum_{i=1}^{N} M_i\left(\bar{y}_i - \bar{\bar{Y}}\right)^2 + \frac{1}{n}\sum_{i=1}^{N}\frac{M_i^2(1-f_{2i})}{m_i\, z_i}\frac{\sum_{j=1}^{M_i}\left(y_{ij}-\bar{Y}_i\right)^2}{M_i-1}$$

$$\hat{V}\left(\hat{\bar{\bar{Y}}}_{PPS}\right) = \frac{1}{M_0^2}\hat{V}\left(\hat{Y}_{PPS}\right) = \frac{1}{nM_0}\frac{\sum_{i=1}^{n} M_i\left(\bar{y}_i - \bar{\bar{Y}}_{PPS}\right)^2}{n-1}$$

y para el estimador del promedio por conglomerado,

$$\hat{\bar{\bar{Y}}}_{PPS} = \frac{\hat{Y}_{PPS}}{N} = \frac{M_0}{nN}\sum_{i=1}^{n}\bar{y}_i$$

$$V\left(\hat{\bar{\bar{Y}}}_{PPS}\right) = \frac{1}{N^2}V\left(\hat{Y}_{PPS}\right) = \frac{M_0}{nN^2}\sum_{i=1}^{N} M_i\left(\bar{y}_i - \bar{\bar{Y}}\right)^2 + \frac{M_0^2}{nN^2}\sum_{i=1}^{N}\frac{M_i^2(1-f_{2i})}{m_i\, z_i}\frac{\sum_{j=1}^{M_i}\left(y_{ij}-\bar{Y}_i\right)^2}{M_i-1}$$

$$\hat{V}\left(\hat{\bar{\bar{Y}}}_{PPS}\right) = \frac{1}{N^2}\hat{V}\left(\hat{Y}_{PPS}\right) = \frac{M_0}{nN^2}\frac{\sum_{i=1}^{n} M_i\left(\bar{y}_i - \bar{\bar{Y}}_{PPS}\right)^2}{n-1}$$

13.3.- Selección de conglomerados con probabilidades desiguales sin reemplazamiento

Este diseño consiste en que la selección de unidades en todas las etapas, se hace sin reemplazamiento. Tal como se comentó en el apartado 10.3. (Selección de Conglomerados con Probabilidades Desiguales sin Reemplazamiento), en el presente trabajo no se desarrollará este caso, sólo se mostraran los estimadores y varianzas desarrollados por Horvitz y Thompson (1952).

En dicho apartado se definió, para el caso general, donde n≥2,

π_i = probabilidad de que la i-ésima unidad esté en la muestra

π_{ij} = probabilidad de que la i-ésima y j-ésima unidades estén ambas en la muestra

El estimador de Horvitz-Thompson del total poblacional es,

$\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{\hat{Y}_i}{\pi_i}$, que es un estimador insesgado de *Y*, donde $\hat{Y}_i = \frac{M_i}{m_i} \sum_{j=1}^{m_i} y_{ij} = M_i \ \bar{y}_i$

y su varianza es,

$$V(\hat{Y}_{HT}) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(\pi_i \pi_j - \pi_{ij}\right) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2 + \sum_{i=1}^{N} \frac{M_i (M_i - m_i) S_{2i}^2}{m_i \ \pi_i}$$

[2; 369]

y la respectiva varianza estimada es,

$$\hat{V}(\hat{Y}_{HT}) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\pi_i \pi_j - \pi_{ij}\right) \left(\frac{\hat{Y}_i}{\pi_i} - \frac{\hat{Y}_j}{\pi_j} \right)^2 + \sum_{i=1}^{n} \frac{M_i (M_i - m_i) s_{2i}^2}{m_i \ \pi_i}$$

Otro desarrollo de la varianza de este estimador, mostrado en [1; 204], es,

$$V(\hat{Y}_{HT}) = \sum_{i=1}^{N} \sigma_i^2 \pi_i + \sum_{i=1}^{N} \frac{Y_i^2}{\pi_i} + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(\pi_i \pi_j - \pi_{ij}\right) \left(\frac{Y_i}{\pi_i} \frac{Y_j}{\pi_j} \right) \pi_{ij} - Y^2$$

Existen muchos otros desarrollos para estos estimadores, en principio existen extensiones para todos los de una etapa. Brewer y Hanif (1983) describen 50 procedimientos, dentro de los cuales destacan Sen, Grundy y Yates (1953), Brewer (1963-75), Durbin (1967), Rao (1962) y Cassel et al (1977), Sampford (1967), Murthy (1957), Rao-Hartley-Cochran (1962), entre otros. [1; 205] y [2;323, 369-370, 377-379]

Es de resaltar que existe otro procedimiento desarrollado por Azorín [1;206] denominado “Muestreo con probabilidades gradualmente variables (pgvpt)”.

14.- MUESTREO SUCESIVO

Esta clase de muestreo se justifica cuando no se conoce la distribución de las variables a investigar, y no hay manera de determinar el tamaño ni el diseño a utilizar, y además se dispone de tiempo.

El procedimiento consiste en tomar muestras aleatorias sucesivamente, hasta encontrar que la varianza se estabiliza, es decir, que por mucho que se aumente el tamaño de la muestra, la varianza no se reduce de manera sustancial.

Supongamos que se va a tomar una muestra aleatoria simple de una población de N elementos. En principio se toma una muestra de tamaño n_0, y se hallan los estimadores y varianzas,

$$\hat{\bar{Y}}_0 = \frac{\sum_{i=1}^{n_0} y_i}{n_0} \quad ; \quad \hat{Y}_0 = N\hat{\bar{Y}}_0 \quad ; \quad \hat{V}(\hat{\bar{Y}}_0) = \frac{N-n_0}{N\,n_0}\frac{\sum_{i=1}^{n_0}(y_i-\bar{Y}_0)^2}{n_0-1} \quad ; \quad \hat{V}(\hat{Y}_0) = N^2\hat{V}(\hat{\bar{Y}})$$

Luego se toma otra muestra de tamaño n_1, y se vuelven a hacer los cálculos de estimadores y varianzas, pero esta vez el tamaño de la muestra es n_0+n_1,

$$\hat{\bar{Y}}_1 = \frac{\sum_{i=1}^{n_0+n_1} y_i}{n_0+n_1} \quad ; \quad Y_1 = N\hat{\bar{Y}}_1 \quad ; \quad \hat{V}(\hat{\bar{Y}}_1) = \frac{N-(n_0+n_1)}{N\,(n_0+n_1)}\frac{\sum_{i=1}^{n_0+n_1}(y_i-\bar{Y}_1)^2}{(n_0+n_1)-1} \quad ; \quad \hat{V}(\hat{Y}_1) = N^2\hat{V}(\hat{\bar{Y}}_1)$$

y así se toman otras muestras de tamaño n_2, n_3, n_4,..... , hasta llegar a la muestra $(k+1)$-ésima, de tamaño n_k, donde los estimadores son:

$$\hat{\bar{Y}}_k = \frac{\sum_{i=1}^{n^*} y_i}{n^*} \quad ; \quad \hat{Y}_k = N\hat{\bar{Y}}_k \quad ; \quad \hat{V}(\hat{\bar{Y}}_k) = \frac{N-n^*}{N\,n^*}\frac{\sum_{i=1}^{n^*}(y_i-\bar{Y}_k)^2}{n^*-1} \quad ; \quad \hat{V}(\hat{Y}_k) = N^2\hat{V}(\hat{\bar{Y}}_k)$$

donde $n^* = n_1+n_2+\cdots+n_k$; que se cumpla con algún criterio de parada pre-establecido, como por ejemplo:

$$\frac{\left|\hat{V}(\hat{\bar{Y}}_k) - \hat{V}(\hat{\bar{Y}}_{k-1})\right|}{\left|\hat{V}(\hat{\bar{Y}}_{k-1}) - \hat{V}(\hat{\bar{Y}}_{k-2})\right|} \le \varepsilon \quad , \quad \varepsilon \ge 0$$

En ese caso, el tamaño de la muestra n, es $n = n_1+n_2+\cdots+n_k$..

BIBLIOGRAFÍA

[1] AZORÍN POCH, F. / SÁNCHEZ-CRESPO, J.; Métodos y Aplicaciones del Muestreo; Alianza Editorial/Alianza Universida Textos; Madrid, 1986.

[2] COCHRAN, William .C.; Técnicas de Muestreo; Compañía Editorial Continental S.A.; Cuarta Impresión; México, 1984; Traducido de Sampling Techniques.

[3] KENETT, Ron S., ZACKS, Shelemyahu; Estadística Industrial Moderna: diseño y control de la calidad y la confiabilidad; International Thompson; México , 2000; Traducido de Modern Industrial Statistics: Design and Control of Quality and Reliability.

[4] KISH, Leslie; Muestreo de Encuestas; Editorial Trillas; Primera edición en español, segunda reimpresión; México, 1979; Traducido de Survey Sampling.

[5] MARTÍN-CARO M., Harold D.; El Uso de las Técnicas de Simulación como Método Alternativo en la Evaluación de Censos; Trabajo de Ascenso, UCV; Caracas, 1995.

[6] SCHEFFER, R. / MENDENHALL, W. / OTT, L.; Elementos de Muestreo; Grupo Editorial Iberoamérica; México, 1987; Traducido de Elementary Survey Sampling, PWS Publishers, USA, 1986

[7] SEIJAS Z., Félix L.; Investigación por Muestreo; Ediciones de la Biblioteca, Ediciones FACES-UCV, Universidad Central de Venezuela, Tercera Edición, Caracas, 1999.

[8] SUKHATME, Pandurang V.; Teoría de Encuestas por Muestreo con Aplicaciones; Fondo de Cultura Económica; Primera Edición; México, 1956; Traducido de Sampling Theory of Surveys with Applications.

[9] HOOG, Robert V., TANNIS, Elliot A.; Probability & Statistical Inference; Macmillan Publishing Co., Inc.; New York, U.S.A., 1977