MONOGRAFIAS
DE LA
ACADEMIA
DE
CIENCIAS
Exactas
Físicas
Químicas y
Naturales
DE
ZARAGOZA

MUESTREO BASICO PARA CIENCIAS APLICADAS

Mariano Ruiz Espejo

N.º 8

1996

Depósito legal: Z. 558—1997

Imprime:
Coop. de Artes Gráficas
LIBRERIA GENERAL
Pedro Cerbuna, 23
50009 Zaragoza

MUESTREO BÁSICO PARA CIENCIAS APLICADAS

Mariano Ruiz Espejo

Indice

§Prólogo.

El objetivo que me propongo al publicar esta monografía básica sobre muestreo de poblaciones finitas es el de proporcionar técnicas estadísticas válidas para iniciar al lector en los principios matemáticos y en los conceptos elementales usados en la práctica de encuestas o de estudios observacionales, referidos a un conjunto finito de unidades que componen la población, sobre una base muestral.

Estas técnicas constituyen una materia cuyo conocimiento es imprescindible en la formación estadística general. Por un lado cuestiona una situación clásica de la estadística tradicional: la información por sí misma carece de valor si se desconoce cómo se ha recabado. De una muestra disponible pueden hacerse conjeturas sobre su procedencia e incluso de qué modelo cabe suponer que es el origen de tales observaciones; estas son las cuestiones que se plantean en la inferencia clásica paramétrica, no paramétrica, bayesiana, e incluso econométrica. Sin embargo tales planteamientos requieren, para su utilización práctica, la incorporación de planteamientos subjetivos del estadístico profesional que intenta explicar una realidad que sobrepasa los límites de su oficina material, su aula y su propio pensamiento.

Aspectos como la economía y las finanzas, la demografía, la población y sus condiciones sociales, la industria, la energía, la alimentación, los servicios y el desarrollo, son entre otros características que deben ser conocidas con cierta precisión por las autoridades para poder corregir cualquier deficiencia o atender nuevas necesidades en su evolución. La teoría de muestras aporta sólidos conceptos y resultados matemáticos que permiten conocer estas características con unas técnicas que son relativamente rápidas y económicas si las comparamos con un estudio similar a partir de un censo que permita observar a todas las unidades de la población.

Los requerimientos para seguir el nivel matemático y estadístico de este libro son un curso al menos de Estadística (que incida en la Probabilidad así como en los operadores Esperanza Matemática y Varianza de una variable aleatoria), así como unas mínimas bases Combinatoria y Matemática (entre cuyos contenidos incluyan los desarrollos en serie de Taylor y el método de los multiplicadores de Lagrange).

El texto puede servir como guía en Facultades de Ciencias Matemáticas, y de Ciencias Económicas y Empresariales, así como en Escuelas Universitarias de Estadística. Sin duda el contenido puede completar la formación científica en Escuelas Técnicas Superiores de Ingenieros Industriales, y en Ciencias de la Salud, lo que aportará una perspectiva más moderna en sus tradicionales formaciones estadísticas.

Quiero agradecer el interés mostrado por algunos profesores universitarios en guiarme por estos contenidos en un principio, así como sus desinteresadas ayudas en muchos momentos.

Espero que el interés mostrado hasta estas líneas por el lector, siga vivo a lo largo de los capítulos en la confianza de que le proporcionarán ciencia e instrumentos útiles para conocer hechos reales.

Mariano Ruiz Espejo
Madrid, Enero de 1996

§Presentación.

El muestreo de poblaciones finitas se basa en la hipótesis fáctica y realista de que queremos realizar estimaciones sobre parámetros, o mejor funciones paramétricas, basadas en las observaciones posibles que podrán obtenerse de un número finito de unidades o elementos de la población o universo.

Para ello se proponen estimadores concretos o funciones reales de las observaciones tomadas en una muestra obtenida mediante un diseño que asigne una probabilidad determinada de cada posible muestra de población finita. La muestra siempre será finita.

Una propiedad deseable del estimador es que sea insesgado. Además una medida de dispersión del estimador viene dada por la varianza. La medida de dispersión usada para un estimador sesgado es su "error cuadrático medio".

Es muy deseable antes de poner en práctica una estrategia de muestreo, compuesta por un diseño muestral y un estimador, estudiar su conveniencia comparándola con otras también candidatas a ser usadas en un estudio concreto. Para ello conviene analizar en qué estrategia se minimiza la dispersión del estimador, y por tanto será más preciso que los restantes, o al menos, investigar en qué condiciones será más deseable si se dispone de información auxiliar adicional a las propias observaciones, u otras características incorporables al diseño muestral o al propio estimador, etc.

Por último, es deseable dar una estimación de la dispersión de la estrategia usada, haciendo uso de la información muestral y de informaciones disponibles por el investigador. De este modo el estudio terminará con una estimación concreta del parámetro o función paramétrica, acompañada por una estimación de su dispersión.

Capítulo 1

Introducción.

El muestreo de poblaciones finitas consiste en seleccionar una parte de una "colección finita" de unidades (llamada "población") y seguidamente hacer inferencias sobre la colección entera basándose en las observaciones tomadas en la "parte seleccionada" (llamada "muestra").

La aleatorización surgida de modo natural es la debida al método de selección de unidades en la muestra.

En nuestro modelo de muestreo suponemos que a cada unidad de la población se le asocia un número real "y", desconocido y fijo, que es el valor de la "variable en estudio" o "variable de interés". Por ejemplo, la variable de interés puede ser la "renta familiar anual" en la población finita compuesta por "los hogares de una determinada ciudad".

Además se exige al modelo la "identificabilidad" de unidades, que permitirá al estadístico profesional observar cualquier muestra seleccionada de modo aleatorio. Así, la distribución de un estimador dado será algo que el estadístico crea y controla.

Una "población finita" es una colección de N unidades, donde el número entero N cumple que $0 < N < \infty$ y se le llama "tamaño de la población".

La "identificabilidad" de unidades consiste en la posibilidad de acceso a cualquier unidad si es seleccionada en la muestra aleatoria. En un caso concreto podría ser la lista de nombres y direcciones o teléfonos si las unidades fueran personas. Las unidades de una población finita se dicen identificables si pueden ser numeradas unívocamente de 1 a N, y el número de cada unidad es conocido.

De este modo la "población finita", o "universo" U, de unidades identificables puede representarse como

$$U = \{1, 2, \ldots, k, \ldots, N\}.$$

Cada unidad numerada k tiene asociado un número y_k cuando la característica en estudio es "y", resultado de la medida sin error de la variable "y" en la unidad "k". Así la "observación numerada" será el par (k, y_k).

El vector $\mathbf{y} = (y_1, \ldots, y_N)$ es el "vector paramétrico" de la población finita, donde las unidades k de 1 a N están localizadas por su posición en el parámetro o vector $\mathbf{y}$.

El "espacio paramétrico", donde puede variar el vector $\mathbf{y} = (y_1, y_2, \ldots, y_N)$, puede ser $\mathbb{R}^N$ (si $y_1 \in \mathbb{R}$, $y_2 \in \mathbb{R}$, $\ldots$, $y_N \in \mathbb{R}$), $\mathbb{R}_+^N$ (si $y_1 \in \mathbb{R}_+$, $y_2 \in \mathbb{R}_+$, $\ldots$, $y_N \in \mathbb{R}_+$, ó $\{0,1\}^N$

(si y_k puede tomar el valor 0, si la unidad k no posee cierta cualidad, ó 1, si la unidad k-ésima posee la cualidad; siendo $k = 1, 2, \ldots,$ ó N).

Una función real definida en el espacio paramétrico se llama "función paramétrica". La inferencia en poblaciones finitas se centra en una función paramétrica especificada, y a veces sobre el propio parámetro $\mathbf{y}$.

Dos funciones paramétricas importantes son la "media poblacional"

$$\overline{y} = \frac{\sum_{k=1}^{N} y_k}{N}$$

y la "varianza poblacional"

$$\sigma^2 = \frac{\sum_{k=1}^{N} (y_k - \overline{y})^2}{N}$$

donde aparece $\overline{y}$ definida previamente como media poblacional.

Llamamos "muestra ordenada" a la secuencia $\mathbf{s} = \left(k_1, \ldots, k_{n(\mathbf{s})}\right)$ tal que $k_i \in U$ es la unidad situada en el i-ésimo lugar de la secuencia. El "tamaño muestral" es $n(\mathbf{s})$, que puede ser mayor que N si aparecen elementos o unidades repetidos. Se llama muestra ordenada porque la secuencia conserva el orden de selección de unidades por un procedimiento determinado. Notaremos

$$\mathbf{S} = \{\mathbf{s} : \mathbf{s} \text{ es muestra ordenada}\}.$$

Así, por ejemplo, si $U = \{1, 2, 3\}$, una muestra ordenada puede ser $\mathbf{s}_1 = (1, 2)$ ó $\mathbf{s}_2 = (3, 1)$ ó $\mathbf{s}_3 = (2, 2)$ ó $\mathbf{s}_4 = (1, 3, 2, 1)$.

Se llama "tamaño muestral efectivo" de una secuencia $\mathbf{s}$ al número de componentes distintas que tiene, y se denota $\nu(\mathbf{s})$. Así, en el ejemplo $\nu(\mathbf{s}_1) = \nu(\mathbf{s}_2) = 2$, $\nu(\mathbf{s}_3) = 1$ y $\nu(\mathbf{s}_4) = 3$.

Dada la secuencia $\mathbf{s}$, podemos construir el conjunto

$$s = \{k : k \text{ es componente de } \mathbf{s}\}$$

y entonces $\nu(\mathbf{s}) = \text{card}(s)$, donde $\text{card}(A)$ es el número de elementos del conjunto A; como s es un conjunto contenido en U, podemos calcular su número de elementos en s, que será $\text{card}(s)$.

Llamamos "muestra no ordenada" a todo conjunto s no vacío, con $s \subset U$. Sea $\mathcal{S}$ el conjunto de muestras no ordenadas (no vacías)

$$\mathcal{S} = \{s : \emptyset \neq s \subset U\} = \wp(U) - \{\emptyset\}$$

con $\text{card}(\mathcal{S}) = 2^N - 1$, pues si $\text{card}(A) = N$ se demuestra matemáticamente que $card(\wp(A)) = 2^N$, siendo $\wp(A)$ el conjunto cuyos elementos son todos los posibles subconjuntos de A incluyendo el conjunto $\emptyset$. Así, por ejemplo, si $U = \{1, 2\}$, $\mathcal{S} = \{\{1\}, \{2\}, \{1, 2\}\}$ y

$\text{card}(\mathcal{S}) = 2^2 - 1 = 3$ es el número de muestras no ordenadas no vacías. Se llama "muestra no ordenada" porque no influye el orden de selección de las componentes en el conjunto s, así como la multiplicidad de unidades en la muestra.

Ahora el "tamaño muestral efectivo" será $\nu(s) = \text{card}(s)$.

Hemos denotado a la muestra $\mathbf{s}$ ó s por la inicial de "sample" que significa muestra en inglés.

Se llama "función de reducción" a la aplicación $r : \mathbf{S} \to \mathcal{S}$ tal que

$$r(\mathbf{s}) = \{k \in U : k \text{ es componente de } \mathbf{s}\} = s$$

es decir r elimina el orden y la multiplicidad de unidades en la muestra ordenada $\mathbf{s}$, transformándola en una muestra no ordenada s.

Por ejemplo, si $\mathbf{s} = (1,1,2)$ entonces $r(\mathbf{s}) = s = \{1,2\}$.

Un "diseño muestral" es una función de probabilidad definida sobre $\mathbf{S}$ ó $\mathcal{S}$. Un "diseño muestral ordenado" es una función $p : \mathbf{S} \to [0,1]$ tal que $p(\mathbf{s}) \geq 0$ para toda $\mathbf{s} \in \mathbf{S}$, y

$$\sum_{\mathbf{s} \in \mathbf{S}} p(\mathbf{s}) = 1.$$

Un "diseño muestral no ordenado" es una función $p : \mathcal{S} \to [0,1]$ tal que $p(s) \geq 0$ para toda $s \in \mathcal{S}$, y

$$\sum_{s \in \mathcal{S}} p(s) = 1$$

El diseño muestral puede introducirse a partir de un diseño ordenado $p(\mathbf{s})$, correspondiendo

$$p(s) = \sum_{\mathbf{s} \in r^{-1}(s)} p(\mathbf{s})$$

siendo $r^{-1}(s) = \{\mathbf{s} \in \mathbf{S} : r(\mathbf{s}) = s\} \subset \mathbf{S}$ el conjunto de muestras ordenadas, $\mathbf{s}$, tales que $r(\mathbf{s}) = s$. También $p(s)$ se puede postular como punto de partida. Por ejemplo, si $s = \{1,2\}$, $r^{-1}(s)$ contiene las siguientes muestras ordenadas: (1,2), (1,1,2), (1,2,1), (2,1,1), (1,2,2), (2,1,2), (2,2,1), etc.

Se define "probabilidad de inclusión" π_k de la unidad $k \in U$ en la muestra aleatoria $\mathbf{s}$ ó s, a

$$\pi_k = \sum_{\mathbf{s} \in \mathbf{S}_k} p(\mathbf{s}) = \sum_{s \in \mathcal{S}_k} p(s)$$

donde $\mathbf{S}_k = \{\mathbf{s} : k \in \mathbf{s}\}$ y $\mathcal{S}_k = \{s : k \in s\}$, es decir π_k es la suma de las probabilidades de las muestras, ordenadas o no, que tengan como componente la unidad $k \in U$.

La "probabilidad de inclusión de 2º orden" π_{km} de las unidades k y m en la muestra es

$$\pi_{km} = \sum_{\mathbf{s} \in \mathbf{S}_{km}} p(\mathbf{s}) = \sum_{s \in \mathcal{S}_{km}} p(s)$$

donde $\mathbf{S}_{km} = \{\mathbf{s} : k, m \in \mathbf{s}\}$ y $\mathcal{S}_{km} = \{s : k, m \in s\}$. En este caso se suman las probabilidades de las muestras que tengan como componentes las unidades $k \in U$ y $m \in U$.

Del mismo modo se obtienen las probabilidades de inclusión de órdenes superiores.

Un diseño ordenado p se llama "diseño de tamaño fijo" igual a n si el número de componentes de $\mathbf{s}$, $n(\mathbf{s})$, es constante para toda $\mathbf{s} \in \mathbf{S}$ tal que $p(\mathbf{s}) > 0$, y lo denotaremos TF(n). Un diseño ordenado (o no ordenado) se llama "diseño de tamaño efectivo fijo" e igual a ν, si el tamaño muestral efectivo $\nu(\mathbf{s})$ (ó $\nu(s)$) es constante para toda $\mathbf{s} \in \mathbf{S}$ ($s \in \mathcal{S}$) tal que $p(\mathbf{s}) > 0$ ($p(s) > 0$) y lo denotaremos TEF(ν) donde

$$\overline{\nu} = \sum_{\mathbf{s} \in \mathbf{S}} \nu(\mathbf{s}) p(\mathbf{s}) = \sum_{s \in \mathcal{S}} \nu(s) p(s)$$

es el tamaño muestral efectivo esperado.

Ejemplo 1.1 *Sea* $U = \{1, 2, 3, 4, 5\}$ *y tenemos por diseño no ordenado el siguiente*

$$p(\{1, 2\}) = \frac{1}{3}$$

$$p(\{3, 4, 5\}) = \frac{1}{3}$$

$$p(\{3, 4\}) = \frac{1}{3}.$$

Así $\nu(\{1, 2\}) = 2$, $\nu(\{3, 4, 5\}) = 3$, $\nu(\{3, 4\}) = 2$ mientras que el tamaño muestral efectivo esperado es

$$\overline{\nu} = 2\frac{1}{3} + 3\frac{1}{3} + 2\frac{1}{3} = \frac{7}{3}.$$

En este caso,

$$\pi_1 = p(\{1, 2\}) = \frac{1}{3}$$

$$\pi_2 = p(\{1, 2\}) + p(\{2, 4\}) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$$

También

$$\pi_3 = \pi_5 = \frac{1}{3}$$

y

$$\pi_4 = \frac{2}{3}$$

además tenemos por ejemplo que

$$\pi_{1,3} = 0, \ \pi_{2,4} = \frac{1}{3}, \text{etc.}$$

Ejemplo 1.2 *Si tenemos la población* $U = \{1,2,3,4,5,6,7\}$ *y el diseño ordenado*

$$p(1,1,2) = \frac{1}{5}$$

$$p(3,2,5) = \frac{1}{5}$$

$$p(4,6,7) = \frac{1}{5}$$

$$p(6,2,5) = \frac{1}{5}$$

$$p(7,1,7) = \frac{1}{5}.$$

Ahora $\nu(1,1,2) = 2 = \nu(7,1,7)$, $\nu(3,2,5) = \nu(4,6,7) = \nu(6,2,5) = 3$.
También

$$\overline{\nu} = 2 \cdot \frac{1}{5} + 3 \cdot \frac{1}{5} + 3 \cdot \frac{1}{5} + 3 \cdot \frac{1}{5} + 2 \cdot \frac{1}{5} = \frac{13}{5}$$

$$\pi_7 = p(4,6,7) + p(7,1,7) = \frac{2}{5}$$

$$\pi_2 = p(1,1,2) + p(3,2,5) + p(6,2,5) = \frac{3}{5}$$

$$\pi_{2,7} = 0$$

$$\pi_{4,7} = \frac{1}{5}$$

$$\pi_{2,5} = \frac{2}{5}$$

etc.

§1.1 Algunos resultados básicos.

Como medida de dispersión de los estimadores en el muestreo, se suele utilizar la varianza del estimador, que no es sino una cantidad positiva.

Veamos alguna propiedad que justifica su uso:

Lema 1.1 *Si* ξ *y* η *son variables aleatorias cualesquiera tal que* $E(\xi^2)$ *y* $E(\eta^2)$ *existen, entonces*

$$|E(\xi\eta)| \leq \sqrt{E(\xi^2)E(\eta^2)}$$

(Desigualdad de Schwarz)

Demostración.- Sea $\tau_\lambda = (\xi - \lambda\eta)^2$ una variable aleatoria con $\lambda \in \mathbb{R}$.

$$E(\tau_\lambda) = E\left[(\xi - \lambda\eta)^2\right] = E\left(\xi^2\right) - 2\lambda E(\xi\eta) + \lambda^2 E\left(\eta^2\right).$$

Como $\tau_\lambda \geq 0$ implica que $E(\tau_\lambda) \geq 0$ para todo $\lambda \in \mathbb{R}$, y esto a su vez implica que $E(\tau_\lambda)$ tiene a lo sumo una raíz en λ, por lo que el discriminante es negativo o cero:

$$4[E(\xi\eta)]^2 - 4E\left(\xi^2\right)E\left(\eta^2\right) \leq 0$$

por lo que podemos concluir que

$$|E(\xi\eta)| \leq \sqrt{E(\xi^2)E(\eta^2)}.$$

Teorema 1.1 *Sea ξ una variable aleatoria cualquiera, entonces*

$$E[|\xi - E(\xi)|] \leq \sqrt{V(\xi)}.$$

Demostración.-

$$\{E[|\xi - E(\xi)|]\}^2 = E^2(|\xi - E(\xi)| \cdot 1) \leq E\left[(|\xi - E(\xi)|)^2\right] \cdot E(1^2) = V(\xi)$$

haciendo uso del Lema previo ó de la desigualdad de Jensen (ver Mood et al, 1974).

Así la desviación absoluta media está acotada por la desviación típica o raíz cuadrada de la varianza de la variable aleatoria. Aunque la desviación absoluta media de un estimador es su medida de dispersión más natural y deseable, respecto a su esperanza matemática, tiene la desventaja de que no es fácilmente utilizable en el desarrollo matemático. Esto no ocurre con la varianza del estimador, que una vez calculada o estimada sabemos que su desviación típica acota superiormente a la desviación absoluta media.

Una medida del error de un estimador, t, respecto a su parámetro o función paramétrica, θ, a estimar, es su error cuadrático medio, que coincide con su varianza mas su sesgo al cuadrado.

$$ECM(t) = E\left[(t-\theta)^2\right] = E\left\{[t - E(t) + E(t) - \theta]^2\right\} =$$

$$= E\left\{[t - E(t)]^2 + [E(t) - \theta]^2 + 2[t - E(t)][E(t) - \theta]\right\} =$$

$$= E\left\{[t - E(t)]^2\right\} + E\left\{[E(t) - \theta]^2\right\} + 0 = V(t) + B^2(t),$$

siendo $B(t) = E(t) - \theta$ el sesgo (bias en inglés) de t. Basta indicar que

$$E\{2[t - E(t)][E(t) - \theta]\} = 2[E(t) - \theta] \cdot E[t - E(t)] = 2[E(t) - \theta] \cdot 0 = 0.$$

Capítulo 2

Muestreo aleatorio simple con reemplazamiento (masr)

§2.1 Diseño masr.

Es aquel diseño ordenado p sobre $\mathbf{S}$ tal que asigna una probabilidad $p(\mathbf{s}) = \frac{1}{N^n}$ para cada secuencia de tamaño muestral $n(\mathbf{s}) = n$, y $p(\mathbf{s}) = 0$ para las restantes secuencias.

Otra caracterización de éste mismo diseño sería la selección de una bola de una urna conteniendo N bolas, numeradas del 1 al N. Una vez extraída una bola, se anota en la primera componente de la secuencia y seguidamente se reincorpora la bola extraída a la urna, de modo que en la segunda extracción pueda seleccionarse con igual probabilidad cualquier unidad de la 1 a la N, independientemente de la primera extracción.

Repitiendo este proceso se selecciona una secuencia de tamaño muestral n, con $0 < n$.

Con este diseño ordenado, las distribuciones marginales de la secuencia son iguales e independientes entre sí.

Además es un diseño, pues

$$p(\mathbf{s}) = \frac{1}{N^n} > 0$$

para toda $\mathbf{s} \in \mathbf{S}$ tal que $n(\mathbf{s}) = n$, y

$$p(\mathbf{s}) = 0$$

para toda muestra ordenada $\mathbf{s}$ restante, y

$$\sum_{\mathbf{s}\in\mathbf{S}} \frac{1}{N^n} = N^n \frac{1}{N^n} = 1,$$

pues existen N^n muestras ordenadas $\mathbf{s}$ en $\mathbf{S}$ de tamaño fijo n (TF(n)).

Las probabilidades de inclusión serán

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

para toda $k \in U$, y

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

para todas $k \neq m \in U$, ya que

$$p(k \notin \mathrm{s}) = \left(1 - \frac{1}{N}\right)^n$$

implica que

$$\pi_k = p(k \in \mathrm{s}) = 1 - p(k \notin \mathrm{s}) = 1 - \left(1 - \frac{1}{N}\right)^n$$

utilizando que si A es un suceso $p(A) = 1 - p(\overline{A})$, siendo $\overline{A}$ el suceso complementario de A. También si $k \notin \mathrm{s}$ equivale a que $\mathrm{s} = (k_1, k_2, \ldots, k_n)$ verifica que $k_i \neq k$ para todo $i = 1, 2, \ldots, n$, cuya probabilidad es

$$p(k \notin \mathrm{s}) = \prod_{i=1}^{n} p(k_i \neq k) = \prod_{i=1}^{n} \frac{N-1}{N} = \left(1 - \frac{1}{N}\right)^n.$$

Por otro lado, si $k \neq m \in U$ tenemos por la propiedad $p(A \cup B) = p(A) + p(B) - p(A \cap B)$, siendo A y B dos sucesos cualesquiera, ahora

$$p(k \text{ ó } m \notin \mathrm{s}) = p(k \notin \mathrm{s}) + p(m \notin \mathrm{s}) - p(k \text{ y } m \notin \mathrm{s}) =$$
$$= \left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{1}{N}\right)^n - \left(1 - \frac{2}{N}\right)^n = 2\left(1 - \frac{1}{N}\right)^n - \left(1 - \frac{2}{N}\right)^n,$$

luego

$$\pi_{km} = p(k \text{ y } m \in \mathrm{s}) = 1 - p(k \text{ ó } m \notin \mathrm{s}) = 1 - 2\left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{2}{N}\right)^n.$$

§2.2 Estimación de la media poblacional.

La media muestral (siendo j_i la unidad de la población U seleccionada en i–ésimo lugar, $i = 1, 2, \ldots, n$) es

$$\overline{y}_{\mathrm{s}} = \frac{1}{n}\sum_{i=1}^{n} y_{j_i} = \frac{1}{n}\sum_{i \in \mathrm{s}} y_i$$

(tomándose n sumandos, $n = n(\mathrm{s})$) y es insesgada para estimar la media poblacional

$$\overline{y} = \frac{1}{N}\sum_{i \in U} y_i$$

siendo $\mathrm{s} = \left(y_{j_1}, y_{j_2}, \ldots, y_{j_n}\right)$ la secuencia obtenida por diseño masr, es decir el subíndice j_i indica la unidad seleccionada en i–ésimo lugar, y por ello no tiene por qué coincidir con j cada una. En efecto, la esperanza matemática de $\overline{y}_{\mathrm{s}}$ es

$$E(\overline{y}_S) = E\left(\frac{1}{n}\sum_{i\in S} y_i\right) = \frac{1}{n}E\left(\sum_{i\in S} y_i\right) =$$

$$= \frac{1}{n}E\left(\sum_{i=1}^{n} y_{j_i}\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(y_{j_i}\right) = \frac{1}{n}n\overline{y} = \overline{y}$$

por distribuírse idénticamente y_{j_i} a la variable "y" en la población finita.

La varianza será

$$V(\overline{y}_S) = V\left(\frac{1}{n}\sum_{i=1}^{n} y_{j_i}\right) = \frac{1}{n^2}V\left(\sum_{i=1}^{n} y_{j_i}\right)$$

y por ser independientes los sumandos en la última expresión queda

$$V(\overline{y}_S) = \frac{1}{n^2}\sum_{i=1}^{n} V\left(y_{j_i}\right) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.$$

Además con diseño masr, la cuasivarianza muestral

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_{j_i} - \overline{y}_S\right)^2$$

es un estimador insesgado de la varianza poblacional σ^2, pues

$$E\left(s^2\right) = \frac{1}{n-1}E\left[\sum_{i=1}^{n}\left(y_{j_i} - \overline{y}_S\right)^2\right]$$

y sumando y restando $\overline{y}$ dentro del paréntesis,

$$E\left(s^2\right) = \frac{1}{n-1}E\left[\sum_{i=1}^{n}\left(y_{j_i} - \overline{y} + \overline{y} - \overline{y}_S\right)^2\right]$$

que desarrollando el cuadrado de dos sumandos $\left(y_{j_i} - \overline{y}\right)$ e $(\overline{y} - \overline{y}_S)$,

$$E\left(s^2\right) = \frac{1}{n-1}E\left[\sum_{i=1}^{n}\left(y_{j_i} - \overline{y}\right)^2 + n(\overline{y} - \overline{y}_S)^2 + 2\sum_{i=1}^{n}\left(y_{j_i} - \overline{y}\right)(\overline{y} - \overline{y}_S)\right]$$

y por las propiedades de la esperanza matemática E,

$$E\left(s^2\right) = \frac{1}{n-1}\left\{n\sigma^2 + n\frac{\sigma^2}{n} + 2E\left[(\overline{y} - \overline{y}_S)\sum_{i=1}^{n}\left(y_{j_i} - \overline{y}\right)\right]\right\},$$

pero como

$$2\sum_{i=1}^{n}\left(y_{j_i} - \overline{y}\right) = 2n(\overline{y}_S - \overline{y})$$

tenemos

$$E\left(s^2\right) = \frac{1}{n-1}\left\{n\sigma^2 + \sigma^2 - 2nE\left[(\overline{y} - \overline{y}_S)^2\right]\right\} =$$

$$= \frac{1}{n-1}\left[(n+1)\sigma^2 - 2n\frac{\sigma^2}{n}\right] = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2$$

pues

$$E\left[(\overline{y} - \overline{y}_S)^2\right] = V(\overline{y}_S) = \frac{\sigma^2}{n}.$$

Luego, un estimador insesgado de

$$V(\overline{y}_S) = \frac{\sigma^2}{n}$$

será

$$\widehat{V}(\overline{y}_S) = \frac{s^2}{n}$$

puesto que $E(s^2) = \sigma^2$, deduciéndose que

$$E\left[\widehat{V}(\overline{y}_S)\right] = E\left(\frac{s^2}{n}\right) = \frac{1}{n}E\left(s^2\right) = \frac{1}{n}\sigma^2 = V(\overline{y}_S).$$

§2.3 Estimación del total poblacional.

La función paramétrica "total poblacional" es

$$T = N\overline{y} = \sum_{i=1}^{N} y_i$$

Un estimador insesgado del total poblacional es $N\overline{y}_S$. En efecto,

$$E(N\overline{y}_S) = NE(\overline{y}_S) = N\overline{y} = T$$

y su varianza es

$$V(N\overline{y}_S) = N^2 V(\overline{y}_S) = N^2\frac{\sigma^2}{n},$$

por lo que un estimador insesgado de $V(N\overline{y}_S)$ es

$$\widehat{V}(N\overline{y}_S) = \frac{N^2}{n}s^2$$

puesto que $E(s^2) = \sigma^2$, y de aquí

$$E\left[\widehat{V}(N\overline{y}_S)\right] = \frac{N^2}{n}E\left(s^2\right) = \frac{N^2}{n}\sigma^2 = V(N\overline{y}_S).$$

§2.4 Estimación de la proporción poblacional.

La proporción poblacional es un caso particular de la media poblacional, cuando la variable de interés toma valor 1 ó 0 según posea o no una cualidad en una unidad concreta; por ejemplo, tener sexo varón o no, si la población es de personas. La proporción poblacional será

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

con

$$y_i = \begin{cases} 1 \text{ si la unidad } i \text{ posee cierta cualidad} \\ 0 \text{ si la unidad } i \text{ no posee cierta cualidad;} \end{cases}$$

el estimador insesgado propuesto es

$$\overline{y}_S = \frac{1}{n}\sum_{i\in S} y_i = \widehat{p}$$

llamado "proporción muestral", que verifica

$$E(\widehat{p}) = E(\overline{y}_S) = \overline{y} = P$$

y

$$V(\widehat{p}) = \frac{\sigma^2}{n} = \frac{PQ}{n}$$

siendo $Q = 1 - P$, ya que como $\sigma^2 = \alpha_2 - \alpha_1^2$ y si y_i toma sólo los valores 1 ó 0, $y_i^2 = y_i$ por lo que

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})^2 = \frac{1}{N}\sum_{i=1}^{N} y_i^2 - \overline{y}^2 = P - P^2 = P(1-P) = PQ.$$

El estimador insesgado de $PQ = \sigma^2$ será ahora la cuasivarianza muestral

$$s^2 = \frac{n}{n-1}\widehat{p}\widehat{q}$$

siendo $\widehat{q} = 1 - \widehat{p}$, y por tanto un estimador de $V(\widehat{p})$ insesgado será

$$\widehat{V}(\widehat{p}) = \frac{s^2}{n} = \frac{\widehat{p}\widehat{q}}{n-1}.$$

Basta observar que $\widehat{p}\widehat{q}$ es la "varianza muestral".

§2.5 Tamaño de la muestra.

¿Qué tamaño muestral n es necesario para alcanzar un error máximo de muestreo "e" con la probabilidad "$1-\alpha$" (llamado nivel de confianza, y siendo α el nivel de significación)?

Por la desigualdad de Chebycheff

$$P\left[|\overline{y}_S - \overline{y}| < e\right] \geq 1 - \frac{V(\overline{y}_S)}{e^2} = 1 - \alpha.$$

Luego

$$\alpha = \frac{V(\overline{y}_S)}{e^2} = \frac{\sigma^2}{ne^2}$$

que implica

$$n = \frac{\sigma^2}{\alpha e^2}.$$

Como σ^2 (varianza poblacional) es desconocida, puede ser estimada por s^2 (cuasi-varianza muestral) mediante una muestra piloto o previa al estudio. Así, $n = \frac{s^2}{\alpha e^2}$ es el tamaño muestral aproximado para obtener un error de muestreo "e" con una probabilidad superior a "$1-\alpha$", mediante la estimación anticipada de la varianza poblacional σ^2 por una muestra piloto.

Sin embargo, al tratarse este diseño "masr" de un muestreo cuyas observaciones son idénticas a la población e independientes entre sí, podemos utilizar el Teorema Central del Límite y obtener otros tamaños muestrales siendo éstos suficientemente grandes.

Así tenemos que

$$n\overline{y}_S = \sum_{i=1}^{n} y_{j_i}$$

con

$$E(n\overline{y}_S) = n\overline{y}$$

y

$$V(n\overline{y}_S) = n^2\frac{\sigma^2}{n} = n\sigma^2.$$

Por tanto, tipificando la variable aleatoria $n\overline{y}_S$,

$$\frac{n\overline{y}_S - n\overline{y}}{\sqrt{n\sigma^2}} \doteq N(0,1)$$

es la distribución aproximada cuando n es grande (mayor que 35 según algunos autores). Luego al nivel de confianza $1-\alpha$, tenemos λ_α tomada de las tablas de la distribución normal (0,1)

$$-\lambda_\alpha < \frac{n(\overline{y}_S - \overline{y})}{\sqrt{n}\sigma} < \lambda_\alpha$$

que implica

$$\left|\frac{\sqrt{n}\,(\overline{y}_{\mathbf{S}}-\overline{y})}{\sigma}\right| < \lambda_\alpha$$

es decir

$$|\overline{y}_{\mathbf{S}}-\overline{y}| < \frac{\lambda_\alpha \sigma}{\sqrt{n}} = e,$$

por lo que en función del error máximo de muestreo "e", tenemos

$$n = \frac{\lambda_\alpha^2 \sigma^2}{e^2}$$

Como σ^2 es desconocida podemos estimarla por s^2, y obtendremos una estimación de n, que será

$$\widehat{n} = \frac{\lambda_\alpha^2 s^2}{e^2}$$

verificando $E(\widehat{n}) = n$, por ser $E(s^2) = \sigma^2$, teniendo así un estimador insesgado del tamaño muestral.

Ejercicio 2.1 *Disponemos de una población finita de tamaño $N = 5$ y queremos estimar la media poblacional con diseño masr de tamaño $n = 3$. Proponer un estimador insesgado de la media poblacional, calcular la estimación y dar una estimación insesgada de su varianza, en estos casos:*

a) Suponiendo que la muestra es $\mathbf{s} = (1,2,2)$ *e* $y_1 = 4$ *e* $y_2 = 8$.

b) Suponiendo que la muestra es $\mathbf{s}' = (1,3,2)$ *con* $y_3 = 6$.

Resolución.

a) El estimador será $\overline{y}_{\mathbf{S}}$. La estimación de la media poblacional será

$$\overline{y}_{\mathbf{S}} = \frac{y_1+y_2+y_3}{3} = \frac{4+8+8}{3} = \frac{20}{3}.$$

La estimación insesgada de la varianza es

$$\widehat{V}\,(\overline{y}_{\mathbf{S}}) = \frac{s^2}{n}$$

donde

$$s^2 = \frac{1}{n-1}\sum_{k\in\mathbf{S}}(y_k-\overline{y}_{\mathbf{S}})^2 = \frac{1}{2}\left[\left(\frac{8}{3}\right)^2+\left(\frac{4}{3}\right)^2+\left(\frac{4}{3}\right)^2\right] = \frac{58}{9}.$$

Luego

$$\widehat{V}(\overline{y}_S) = \frac{\frac{58}{9}}{3} = \frac{58}{27}.$$

b)

$$\overline{y}_{S'} = \frac{y_1 + y_2 + y_3}{3} = \frac{4+6+8}{3} = 6$$

y

$$s^2 = \frac{1}{2}\left[2^2 + 0 + 2^2\right] = 4$$

por lo que

$$\widehat{V}(\overline{y}_{S'}) = \frac{4}{3}.$$

Ejercicio 2.2 *A partir del problema anterior, ¿cuál muestra es más precisa para estimar la media poblacional?*

Respuesta.

Ambas muestras son resultados o concreciones del mismo diseño masr. No puede afirmarse que una muestra sea más precisa que otra, ya que la precisión es la inversa de la varianza del estimador, y esta varianza es común a ambas muestras que son casos particulares del mismo diseño.

La pregunta admitiría respuesta si conociéramos la media poblacional, $\overline{y}$, comparando las desviaciones $|\overline{y}_S - \overline{y}|$ e $|\overline{y}_{S'} - \overline{y}|$.

A la menor desviación le correspondería la mayor precisión. Pero como la media poblacional es desconocida, no podemos compararlas.

Ejercicio 2.3 *Suponemos que tenemos una población de tamaño $N = 1000$ y queremos estimar la media poblacional con un error máximo de muestreo $e = 2$ y con la probabilidad $1 - \alpha = 0,95$. ¿Qué tamaño muestral debe tener la muestra para que el diseño masr verifique tales condiciones? (Suponemos que hemos estimado la varianza poblacional por 7 a partir de una muestra piloto).*

Resolución.

Aplicando la desigualdad de Chebycheff,

$$n = \frac{\widehat{\sigma}^2}{\alpha \cdot e^2} \doteq \frac{7}{0.05 \cdot 4} = 35.$$

Aceptando la hipótesis de normalidad en la distribución de $\overline{y}_S$, $\alpha = 0,05$ implica $\lambda_\alpha = 1,96$ (mirando en las tablas de la distribución normal de parámetros 0 y 1). Así,

$$n \doteq \frac{\lambda_\alpha^2 \cdot \widehat{\sigma}^2}{e^2} = \frac{1.96^2 \cdot 7}{4} = 6.7.$$

Luego $n = 7$.

Nota aclaratoria. Recordar que el número de permutaciones con repetición de N elementos tomados de n en n, es N^n.

Si $U = \{1, 2, 3\}$, por tanto $N = 3$, y el tamaño muestral es $n = 2$, tendremos las siguientes muestras ordenadas o secuencias con diseño masr:

$$\begin{array}{ccc} (1,1) & (2,1) & (3,1) \\ (1,2) & (2,2) & (3,2) \\ (1,3) & (2,3) & (3,3) \end{array}$$

Número de secuencias $= N^n = 3^2 = 9$. La distribución conjunta será:

$1^{\underline{a}}$ Componente \ \ $2^{\underline{a}}$ Componente	1	2	3	Marginal de la $1^{\underline{a}}$ Componente
1	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{3}$
2	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{3}$
3	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{3}$
Marginal de la $2^{\underline{a}}$ Componente	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	

Observar que las funciones de cuantía marginales de las dos componentes son independientes e idénticamente distribuídas.

Ejercicio 2.4 *Se desea estimar la renta total mensual de un colectivo de 186 trabajadores de una planta industrial. A este efecto se selecciona una muestra aleatoria simple con reemplazamiento de tamaño 20, resultando una media muestral de 158.000 pts. y una cuasivarianza de* $4 \cdot 10^8$ *pts. al cuadrado. Proponer un estimador insesgado de la renta total así como un estimador insesgado de su varianza.*

Resolución.

En este problema el tamaño de la población es el número de trabajadores de la planta industrial, es decir $N = 186$. El tamaño muestral es el número de trabajadores seleccionados en la muestra, es decir $n = 20$. Además la renta media de los trabajadores de la muestra es $\overline{y}_S = 158.000$ pts. y la cuasivarianza muestral es $s^2 = 4 \cdot 10^8$ pts. al cuadrado.

El estimador insesgado de la renta total de los 186 trabajadores es

$$N\overline{y}_S = 186 \cdot 158.000 = 29.388.000 \text{ pts.}$$

Por otro lado, un estimador un estimador insesgado de la varianza del estimador $N\overline{y}_S$, $V(N\overline{y}_S)$, es

$$\widehat{V}(N\overline{y}_S) = \frac{N^2}{n} s^2 = \frac{186^2}{20} \cdot 4 \cdot 10^8 = 6.919,2 \cdot 10^8 (\text{pts})^2.$$

Ejercicio 2.5 *Se quiere conocer una medida de la proporción de productos que superan la calidad necesaria para cierto mercado, de entre los 3000 productos terminados. Con diseño masr de tamaño 100 se controla la calidad de los productos identificados por su número de control, resultando una proporción del 4% de defectuosos, es decir* $\widehat{p} = 0,04$. *Estimar sin sesgo la varianza del estimador* $\widehat{p}$.

Resolución.

El tamaño poblacional es 3000, el tamaño muestral es $n = 100$. La proporción de defectuosos en la muestra, es $\widehat{p} = 0,04 = \frac{4}{100}$, por lo que la proporción de no defectuosos es $\widehat{q} = 1 - \widehat{p} = 0,96$. El estimador pedido resulta ser:

$$\widehat{V}(\widehat{p}) = \frac{\widehat{p}\widehat{q}}{n-1} = \frac{0,04 \cdot 0,96}{99} \simeq 0,003\,\widehat{87}$$

Observar que el dato $N = 3000$ no ha sido necesario utilizarlo para la cuestión planteada. Este dato sería necesario, por ejemplo, para estimar el total de productos que superan la calidad necesaria:

$$N\widehat{p} = 3000 \cdot 0,04 = 120$$

pero este estimador insesgado no se ha pedido.

Capítulo 3

Muestreo aleatorio simple sin reemplazamiento (mas)

§3.1 Diseño mas

Es también conocido como "muestreo irrestrictamente aleatorio" ó "muestreo aleatorio sin reemplazamiento con probabilidades iguales".

Suponemos que disponemos de una población U de tamaño N, $U = \{1,2, \ldots ,N\}$ donde tenemos identificados, por su número de orden desde 1 hasta N en la población, a las unidades que la componen.

El "mas" es un diseño o distribución de probabilidad definida sobre las posibles muestras o subconjuntos no vacíos de la población U, denotadas por s (del inglés "sample"=muestra), de cierto tamaño fijado n, con $0 < n \leq N$.

El número de muestras–conjunto de tamaño n que podemos disponer es

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

para el diseño mas.

Este diseño básico entre las técnicas de muestreo recibe el nombre de "sin reemplazamiento con probabilidades iguales" pues si tuvieramos una urna que contuviera N bolas numeradas del 1 al N, la selección de una muestra s entre las $\binom{N}{n}$ posibles, se podrá realizar extrayendo una primera bola de la urna y anotaríamos su número como componente de la muestra no ordenada o conjunto no vacío s; seguidamente no reincorporaríamos a la urna la bola ya seleccionada con lo cual la urna preparada para la segunda extracción tendría $N - 1$ bolas del 1 al N faltando la bola ya extraída y por tanto su número identificativo no se podrá seleccionar en adelante. La segunda extracción selecciona una segunda bola que tampoco se reintegrará a la urna y por tanto ya no se repetirá en las siguientes extracciones de la misma urna. Así actuaríamos hasta seleccionar n unidades $(0 < n \leq N)$ ordenadamente.

De este modo obtendríamos una secuencia o vector n–dimensional $\mathbf{s}$ que tendrá como probabilidad de selección (usando el teorema de producto con sucesos dependientes)

$$p(\mathrm{s}) = \frac{1}{N} \cdot \frac{1}{N-1} \cdot \cdots \cdot \frac{1}{N-n+1} = \frac{1}{\frac{N!}{(N-n)!}}.$$

Ahora bien, como las muestras–conjunto s de tamaño n está compuesta por tantas muestras–vector s como permutaciones de n elementos o unidades distintas, obtendríamos que la probabilidad de seleccionar al azar una muestra–conjunto s será

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

es decir para cada muestra–conjunto de tamaño n, su probabilidad de ser obtenida es la misma e igual a $\frac{1}{\binom{N}{n}}$. Evidentemente la probabilidad total será 1, es decir

$$\sum_{s \in \mathcal{S}} p(s) = \binom{N}{n} \frac{1}{\binom{N}{n}} = 1$$

donde $\mathrm{S} = \{s \subset U : \mathrm{card}(s) = n\}$ es el conjunto de muestras no ordenadas con probabilidad positiva.

La probabilidad de inclusión de la unidad i $(1 \leq i \leq N)$ en la muestra–conjunto s es (utilizando la regla de Laplace al ser equiprobables las muestras de S)

$$\pi_i = p(i \in s) = \frac{\text{casos favorables}}{\text{casos posibles}} = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{\frac{(N-1)!}{(n-1)!(N-n)!}}{\frac{N!}{n!(N-n)!}} = \frac{n}{N}$$

para el diseño mas. La probabilidad de inclusión de las unidades i y j $(i \neq j)$ en la muestra es

$$\pi_{ij} = p(i, j \in s) = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{\frac{(N-2)!}{(n-2)!(N-n)!}}{\frac{N!}{n!(N-n)!}} = \frac{n(n-1)}{N(N-1)}$$

Análogamente se obtienen las probabilidades de inclusión de órdenes superiores.

§3.2 Estimación de la media poblacional.

La media muestral es

$$\overline{y}_s = \frac{1}{n} \sum_{i \in s} y_i$$

siendo s la muestra–conjunto de tamaño n seleccionada, e y_i la variable de interés en la unidad $i \in s$; este estimador $\overline{y}_s$ con diseño mas es insesgado para estimar la media poblacional

$$\overline{y} = \frac{1}{N} \sum_{i \in U} y_i$$

La demostración es la siguiente.

$$E(\overline{y}) = \sum_{s \in \mathcal{S}} \overline{y}_s p(s) = \sum_{s \in \mathcal{S}} \frac{1}{n} (y_{i_1} + y_{i_2} + \cdots + y_{i_n}) \frac{1}{\binom{N}{n}}$$

y sumando tantas veces y_i como muestras s contengan la unidad i

$$E(\overline{y}_s) = \sum_{i \in U} \frac{1}{n} y_i \operatorname{card}(\{s : i \in s\}) \frac{1}{\binom{N}{n}}$$

siendo

$$\operatorname{card}(\{s : i \in s\}) = \binom{N-1}{n-1}$$

el número de muestras aludido

$$E(\overline{y}_s) = \sum_{i \in U} y_i \frac{1}{n} \binom{N-1}{n-1} \frac{1}{\binom{N}{n}} =$$

$$= \sum_{i \in U} y_i \frac{1}{n} \cdot \frac{\dfrac{(N-1)!}{(n-1)!(N-n)!}}{\dfrac{N!}{n!(N-n)!}} = \frac{1}{N} \sum_{i \in U} y_i = \overline{y}.$$

Otra demostración es la siguiente. Llamando j_i a la unidad de U que es seleccionada en la muestra s con el orden i $(1 \leq i \leq n)$

$$E(\overline{y}_s) = E\left(\frac{1}{n} \sum_{i=1}^{n} y_{j_i}\right) = \frac{1}{n} \sum_{i=1}^{n} E\left(y_{j_i}\right) = \frac{1}{n} n E\left(\overline{y}_{j_i}\right) = \overline{y}$$

con $1 \leq i \leq n$. Para justificar que $E\left(y_{j_i}\right) = \overline{y}$, tenemos

$$E\left(y_{j_i}\right) = \sum_{\mathrm{s} \in \mathrm{S}} y_{j_i} p(\mathrm{s}) =$$

$$= \sum_{i \in U} y_i \frac{1}{\dfrac{N!}{(N-n)!}} \operatorname{card}(\{\mathrm{s} : i \text{ sea la } k - \text{ésima componente de s}\}) =$$

$$= \sum_{i \in U} y_i \frac{1}{\frac{N!}{(N-n)!}} \operatorname{card}(\{\mathrm{s} : \mathrm{s} = (j_1, j_2, \ldots, i = j_k, \ldots, j_n)\})$$

siendo $i = j_k$ una unidad fijada y las restantes no fijadas, y como

$$\operatorname{card}(\{\mathrm{s} : \mathrm{s} = (j_1, j_2, \ldots, i = j_k, \ldots, j_n)\}) =$$

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

pues $(N-1)$ es el número de unidades diferentes de $i = j_k$ que pueden ocupar el primer lugar de la secuencia s, $(N-2)$ es el número de unidades diferentes de j_1 y j_k que pueden ocupar el segundo lugar en la secuencia s, etc. siendo 1 el factor k–ésimo por ser $i = j_k$ la única unidad que ocupa el lugar k–ésimo de s. Por todo ello,

$$E\left(y_{j_i}\right) = \sum_{i \in U} y_i \frac{1}{\frac{N!}{(N-n)!}} \frac{(N-1)!}{(N-n)!} = \frac{1}{N} \sum_{i \in U} y_i = \overline{y}.$$

La varianza de la media muestral $\overline{y}_s$ bajo diseño mas es

$$V(\overline{y}_s) = \frac{N-n}{N} \frac{S^2}{n} \tag{3.1}$$

donde $S^2 = \dfrac{N\sigma^2}{(N-1)}$ es la cuasivarianza poblacional. La demostración es la siguiente.

$$V(\overline{y}_s) = E\left[(\overline{y}_s - \overline{y})^2\right] = E\left[\left(\frac{1}{n}\sum_{i \in s} y_i - \overline{y}\right)^2\right] = E\left\{\left[\frac{1}{n}\sum_{i \in s}(y_i - \overline{y})\right]^2\right\} =$$

$$= E\left\{\frac{1}{n^2}\left[\sum_{i \in s}(y_i - \overline{y})^2 + \sum_{i \neq j \in s}(y_i - \overline{y})(y_j - \overline{y})\right]\right\} =$$

$$= \frac{1}{n^2} E\left[\sum_{i \in s}(y_i - \overline{y})^2\right] + \frac{1}{n^2} E\left[\sum_{i \neq j \in s}(y_i - \overline{y})(y_j - \overline{y})\right] = \tag{3.2}$$

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

$$= \frac{N-n}{N-1}\frac{\sigma^2}{n} = \frac{N-n}{N}\frac{S^2}{n}.$$

Veamos ahora el paso 3.2,

$$E\left[\sum_{i \in s}(y_i - \overline{y})^2\right] = \sum_{s \in \mathcal{S}} \sum_{i \in s} (y_i - \overline{y})^2 p(s) =$$

$$= \sum_{i \in U} (y_i - \overline{y})^2 \operatorname{card}(\{s : i \in s\})\, p(s) = \sum_{i \in U} (y_i - \overline{y})^2 \binom{N-1}{n-1} \frac{1}{\binom{N}{n}} =$$

$$= \sum_{i \in U} (y_i - \overline{y})^2 \frac{\frac{(N-1)!}{(n-1)!(N-n)!}}{\frac{N!}{n!(N-n)!}} = n \frac{1}{N} \sum_{i \in U} (y_i - \overline{y})^2 = n\sigma^2,$$

y también

$$E\left[\sum_{i \neq j \in s} (y_i - \overline{y})(y_j - \overline{y})\right] = \sum_{s \in \mathcal{S}} \sum_{i \neq j \in s} (y_i - \overline{y})(y_j - \overline{y}) p(s) =$$

$$= \sum_{i \neq j \in U} (y_i - \overline{y})(y_j - \overline{y}) \operatorname{card}(\{s : i, j \in s\})\, p(s) =$$

$$\sum_{i \neq j \in U} (y_i - \overline{y})(y_j - \overline{y}) \binom{N-2}{n-2} \frac{1}{\binom{N}{n}} = \tag{3.3}$$

$$= -N\sigma^2 \frac{\frac{(N-2)!}{(n-2)!(N-n)!}}{\frac{N!}{n!(N-n)!}} = -\frac{n(n-1)}{N(N-1)} N\sigma^2 = -\frac{n(n-1)}{(N-1)} \sigma^2.$$

La igualdad 3.3 se obtiene así. Como

$$\sum_{i \in U} (y_i - \overline{y}) = 0$$

esto implica que

$$\left[\sum_{i \in U} (y_i - \overline{y})\right]^2 = 0$$

de aquí obtenemos

$$\sum_{i \in U} (y_i - \overline{y})^2 + \sum_{i \neq j \in U} (y_i - \overline{y})(y_j - \overline{y}) = 0$$

de donde

$$\sum_{i \neq j \in U} (y_i - \overline{y})(y_j - \overline{y}) = -(N-1)S^2 = -N\sigma^2.$$

§3.3 Estimación del total poblacional.

El total poblacional de una variable de interés "y" es

$$T = N\overline{y} = \sum_{i \in U} y_i.$$

Un estimador insesgado del total es $\widehat{T} = N\overline{y}_s$ pues

$$E\left(\widehat{T}\right) = E\left(N\overline{y}_s\right) = NE\left(\overline{y}_s\right) = N\overline{y} = T$$

y su varianza es

$$V\left(\widehat{T}\right) = V\left(N\overline{y}_s\right) = N^2 V\left(\overline{y}_s\right) = N^2 \frac{N-n}{N} \frac{S^2}{n} = N(N-n)\frac{S^2}{n}.$$

§3.4 Estimación de la proporción poblacional.

Basta ver que la proporción poblacional P es la media poblacional de una variable de interés que puede tomar valor 0 ó 1 según no tenga o sí una cualidad,

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

donde

$$y_i = \begin{cases} 0 \text{ cuando no tiene la cualidad la unidad } i \\ 1 \text{ cuando sí tiene la cualidad la unidad } i. \end{cases}$$

El estimador proporción muestral $\widehat{p}$ es un caso particular de media muestral,

$$\widehat{p} = \frac{\sum_{i \in s} y_i}{n}$$

y por tanto $E(\widehat{p}) = P$, y además su varianza se puede obtener teniendo en cuenta ahora de $\sigma^2 = P(1-P) = PQ$ por lo que siendo $Q = 1 - P$, de la fórmula 3.1

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

§3.5 Estimación de la varianza de la media muestral.

La varianza de la media muestral con diseño mas, hemos visto que es

$$V\left(\overline{y}_s\right) = \frac{N-n}{N} \frac{S^2}{n}$$

donde n y N son conocidos y S^2 es la cuasivarianza poblacional que no es conocida y por tanto, al ser una función paramétrica, sería posible ser estimada insesgadamente. En efecto, la cuasivarianza muestral

$$s^2 = \frac{1}{n-1}\sum_{i\in s}(y_i - \overline{y}_s)^2$$

es un estimador insesgado de S^2 bajo diseño mas.

$$E\left(s^2\right) = \frac{1}{n-1}E\left[\sum_{i\in s}(y_i - \overline{y}_s)^2\right] = \frac{1}{n-1}\sum_{s\in\mathcal{S}}\sum_{i\in s}(y_i - \overline{y}_s)^2 p(s)$$

y sumando y restando $\overline{y}$

$$E\left(s^2\right) = \frac{1}{n-1}\sum_{s\in\mathcal{S}}\sum_{i\in s}\left[(y_i - \overline{y}) + (\overline{y} - \overline{y}_s)\right]^2 p(s) =$$

$$= \frac{1}{n-1}\left[\sum_{s\in\mathcal{S}}\sum_{i\in s}(y_i - \overline{y})^2 + \sum_{s\in\mathcal{S}}\sum_{i\in s}(\overline{y} - \overline{y}_s)^2 + 2\sum_{s\in\mathcal{S}}\sum_{i\in s}(y_i - \overline{y})(\overline{y} - \overline{y}_s)\right] p(s) =$$

$$= \frac{1}{n-1}\left[\sum_{i\in U}(y_i - \overline{y})^2 \left(\text{card}\{s : i \in s\}\right) p(s) + V(\overline{y}_s)\,\text{card}\left(\{i : i \in s\}\right) + \right.$$

$$\left. +2\sum_{s\in\mathcal{S}}(\overline{y} - \overline{y}_s)\sum_{i\in s}(y_i - \overline{y})\,p(s)\right],$$

siendo $\text{card}\left(\{i : i \in s\}\right) = n$, pero como

$$\sum_{i\in s}(y_i - \overline{y}) = n(\overline{y}_s - \overline{y}),$$

tenemos

$$E\left(s^2\right) = \frac{1}{n-1}\left[\sum_{i\in U}(y_i - \overline{y})^2 \binom{N-1}{n-1}\frac{1}{\binom{N}{n}} + nV(\overline{y}_s) - \right.$$

$$\left. -2n\sum_{s\in\mathcal{S}}(\overline{y}_s - \overline{y})^2 p(s)\right] =$$

$$= \frac{1}{n-1}\left[N\sigma^2\frac{\frac{(N-1)!}{(n-1)!(N-n)!}}{\frac{N!}{n!(N-n)!}} + nV(\overline{y}_s) - 2nV(\overline{y}_s)\right] =$$

$$= \frac{1}{n-1}\left[n\sigma^2 - nV(\overline{y}_s)\right] = \frac{n}{n-1}\left[\sigma^2 - \frac{N-n}{N-1}\frac{\sigma^2}{n}\right] =$$

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

$$= \frac{n\sigma^2}{n-1}\,\frac{Nn-n-N+n}{Nn-n} = \frac{n\sigma^2}{n-1}\,\frac{N}{n}\,\frac{n-1}{N-1} = \frac{N\sigma^2}{N-1} = S^2.$$

Por tanto, un estimador insesgado de la varianza de la media muestral con diseño mas será

$$\widehat{V}(\overline{y}_s) = \frac{N-n}{N}\,\frac{s^2}{n}.$$

En el caso del estimador del total $\widehat{T} = N\overline{y}_s$, la varianza estimada insesgadamente es

$$\widehat{V}\left(\widehat{T}\right) = \widehat{V}(N\overline{y}_s) = N(N-n)\frac{s^2}{n}.$$

Al estimar la proporción poblacional, como $s^2 = \dfrac{n\sigma_s^2}{(n-1)}$ siendo s^2 la cuasivarianza muestral y σ_s^2 la varianza muestral, $s^2 = \dfrac{n\widehat{p}\widehat{q}}{(n-1)}$ donde $\widehat{p}$ es la proporción muestral y $\widehat{q} = 1 - \widehat{p}$. Entonces el estimador insesgado de la varianza de la proporción muestral es, bajo diseño mas de TEF(n),

$$\widehat{V}(\widehat{p}) = \frac{N-n}{N-1}\,\frac{\widehat{\sigma}^2}{n} = \frac{N-n}{N}\,\frac{\widehat{S}^2}{n} = \frac{N-n}{N}\,\frac{s^2}{n} = \frac{N-n}{N}\,\frac{\dfrac{n\widehat{p}\widehat{q}}{n-1}}{n} = \frac{N-n}{N(n-1)}\widehat{p}\widehat{q}.$$

§3.6 Tamaño de la muestra.

¿Cuál es el tamaño muestral "n" necesario para alcanzar un error máximo de muestreo "e" con la probabilidad "$1-\alpha$"?

Por la desigualdad de Chebycheff,

$$P[|\overline{y}_s - \overline{y}| < e] \geq 1 - \frac{V(\overline{y}_s)}{e^2} = 1 - \alpha$$

ya que no es conocida la distribución del estadístico $\overline{y}_s$. Entonces

$$\alpha = \frac{V(\overline{y}_s)}{e^2} = \frac{(N-n)\sigma^2}{e^2(N-1)n} = \frac{N\sigma^2}{e^2(N-1)n} - \frac{\sigma^2}{e^2(N-1)}$$

de donde

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

que implica, despejando n,

$$n = \frac{\dfrac{N\sigma^2}{e^2(N-1)}}{\alpha + \dfrac{\sigma^2}{e^2(N-1)}} = \frac{1}{\dfrac{\alpha e^2(N-1)}{N\sigma^2} + \dfrac{1}{N}} = \frac{1}{\dfrac{\alpha e^2}{S^2} + \dfrac{1}{N}} = \frac{S^2}{\alpha e^2 + \dfrac{S^2}{N}} \tag{3.4}$$

donde N es conocido, S^2 puede ser estimado insesgadamente por s^2, y "α" y "e" vienen dados por el "nivel de confianza" y la "precisión" respectivamente, solicitados.

Si N es suficientemente grande, $n_\infty = \dfrac{S^2}{\alpha e^2}$. Entonces, dividiendo numerador y denominador por (αe^2) en la fórmula 3.4,

$$n = \frac{n_\infty}{1 + \dfrac{n_\infty}{N}} < n_\infty$$

Obligando a que $n_\infty - n < 1$, para obtener el valor de n a partir del cual no se debe seguir obteniendo unidades muestrales, tenemos

$$n_\infty - n = n_\infty - \frac{n_\infty}{1 + \dfrac{n_\infty}{N}} = n_\infty \left(1 - \frac{1}{1 + \dfrac{n_\infty}{N}}\right) =$$

$$= n_\infty \left(1 - \frac{N}{N + n_\infty}\right) = n_\infty \frac{N + n_\infty - N}{N + n_\infty} = \frac{n_\infty^2}{N + n_\infty} < 1$$

verificándose si y sólo si

$$n_\infty^2 < N + n_\infty$$

que implica

$$n_\infty^2 - n_\infty = n_\infty (n_\infty - 1) < N$$

es decir, si $n_\infty = \dfrac{S^2}{\alpha e}$ verifica que $n_\infty(n_\infty - 1) < N$, tomaremos como tamaño muestral $n = n_\infty$, y en otro caso, de la fórmula 3.4

$$n = \frac{S^2}{\alpha e^2 + \dfrac{S^2}{N}}.$$

En el caso de estimación del total poblacional $N\overline{y}$, el tamaño de la muestra n para un error máximo "e" y un nivel de confianza "$1 - \alpha$", tenemos que por la desigualdad de Chebycheff,

$$P\left[|N\overline{y}_s - N\overline{y}| < e\right] \geq 1 - \frac{V\left(N\overline{y}_s\right)}{e^2} = 1 - \alpha$$

de donde

$$\alpha = \frac{V(N\overline{y}_s)}{e^2} = \frac{N(N-n)\dfrac{S^2}{n}}{e^2} = \frac{\dfrac{N^2S^2}{n} - NS^2}{e^2}$$

luego

$$\alpha e^2 = \frac{N^2S^2}{n} - NS^2$$

que implica

$$\alpha e^2 + NS^2 = \frac{N^2S^2}{n}$$

es decir, despejando n,

$$n = \frac{N^2S^2}{\alpha e^2 + NS^2} = \frac{S^2}{\dfrac{\alpha e^2}{N^2} + \dfrac{S^2}{N}}.$$

En el caso particular de la proporción muestral $\hat{p}$ como estimador insesgado de la proporción poblacional P, el tamaño muestral n para un error máximo "e" y un nivel de confianza "$1-\alpha$", tendremos al aplicar en este caso la desigualdad de Chebycheff

$$P[|\hat{p} - P| < e] \geq 1 - \frac{V(\hat{p})}{e^2} = 1 - \alpha$$

por lo que

$$\alpha = \frac{V(\hat{p})}{e^2} = \frac{\dfrac{N-n}{N-1}\dfrac{PQ}{n}}{e^2} = \frac{PQ\dfrac{N}{(N-1)n} - \dfrac{PQ}{N-1}}{e^2}$$

que implica

$$\alpha e^2 = PQ\frac{N}{(N-1)n} - \frac{PQ}{N-1}$$

o bien

$$\alpha e^2 + \frac{PQ}{N-1} = \frac{PQN}{(N-1)n}$$

es decir, despejando n,

$$n = \frac{PQ\dfrac{N}{N-1}}{\alpha e^2 + \dfrac{PQ}{N-1}} \tag{3.5}$$

Cuando N es suficientemente grande, $n_\infty = \dfrac{PQ}{\alpha e^2}$; de donde dividiendo numerador y denominador por αe^2,

$$n = \frac{n_\infty \frac{N}{N-1}}{1 + \frac{n_\infty}{N-1}} = \frac{n_\infty \left(1 + \frac{1}{N-1}\right)}{1 + \frac{n_\infty}{N-1}}.$$

y entonces $n_\infty - n < 1$ si y sólo si

$$n_\infty - n = n_\infty \left(1 - \frac{1 + \frac{1}{N-1}}{\frac{N-1+n_\infty}{N-1}}\right) =$$

$$= n_\infty \left(1 - \frac{\frac{N}{N-1}}{\frac{N-1+n_\infty}{N-1}}\right) = n_\infty \left(1 - \frac{N}{N-1+n_\infty}\right) < 1$$

o bien,

$$n_\infty \left(\frac{N-1+n_\infty - N}{N-1+n_\infty}\right) = n_\infty \left(\frac{n_\infty - 1}{N-1+n_\infty}\right) < 1$$

que equivale a que

$$n_\infty (n_\infty - 1) < N - 1 + n_\infty$$

o también, si

$$n_\infty (n_\infty - 2) < N - 1$$

tomamos n_∞ como tamaño muestral, mientras que si

$$n_\infty (n_\infty - 2) \geq N - 1$$

tomaremos n, dado en la fórmula 3.5, como tamaño muestral.

3.6.1 Tamaño muestral con hipótesis de normalidad.

Al estimar una proporción P con diseño mas, es usual en la práctica suponer que el estimador $\widehat{p}$ se distribuye de modo normal con parámetros P como media y $\sqrt{V(\widehat{p})}$ como desviación típica.

En este caso si N es inferior o igual a 100.000 unidades, con muestreo aleatorio simple sin reemplazamiento, el tamaño muestral n para los niveles de confianza $1 - \alpha_1 = 0.955$ ($\lambda_{\alpha_1} = 2$) ó $1 - \alpha_2 = 0.997$ ($\lambda_{\alpha_2} = 3$), se recoge en la fórmula general para un error máximo de muestreo e,

$$n = \frac{\lambda_\alpha PQN}{e^2(N-1) + \lambda_\alpha^2 PQ}$$

de modo que si $P = Q = \frac{1}{2}$ se hace máximo PQ, y en cualquier caso podrá acotarse superiormente PQ por $\frac{1}{4}$.

Si N es superior a 100.000 los diseños mas y masr son muy aproximados, y el tamaño muestral n con hipótesis de normalidad para $\widehat{p}$ como estimador de P, para un error de muestreo "e" (y $\lambda_{\alpha_1} = 2$ ó $\lambda_{\alpha_2} = 3$) vendrá dado por la fórmula

$$n = \frac{\lambda_\alpha^2 PQ}{e^2},$$

y acotando $PQ \leq \frac{1}{4}$ tenemos

$$n \leq \frac{\lambda_\alpha^2}{4e^2}.$$

Sin embargo, aunque sea muy habitual hacer esta hipótesis de normalidad existen críticas sobre su uso con diseño mas (ver Plane y Gordon, 1982).

§3.7 Comparación de precisiones dadas por masr y mas.

Usualmente se define precisión de un estimador como el inverso de la varianza del estimador. Como tenemos

$$V(\text{masr}, \overline{y}_{\mathbf{S}}) = \frac{\sigma^2}{n} \quad \text{y} \quad V(\text{mas}, \overline{y}_s) = \frac{N-n}{N-1}\,\frac{\sigma^2}{n},$$

si $n > 1$, tenemos a su vez que $V(\text{masr}, \overline{y}_{\mathbf{S}}) > V(\text{mas}, \overline{y}_s)$ para el mismo tamaño muestral n, por lo que es más precisa la estrategia (mas, $\overline{y}_s$) que (masr, $\overline{y}_{\mathbf{S}}$).

Sin embargo, al poder tener unidades repetidas en la muestra ordenada **s** obtenida por el diseño masr, puede ser más económica la obtención de datos pues pueden imputarse directamente, a partir de la primera observación de una unidad, en la segunda o sucesivas apariciones en la muestra ordenada o secuencia.

Ejercicio 3.1 *Dada una población finita de tamaño $N = 2000$, se toma una muestra de tamaño $n = 20$ por diseño mas, de modo que la media muestral es $\overline{y}_s = 537$ y la cuasi-varianza muestral es $s^2 = 100$. Se pide una estimación de la media poblacional, así como de la varianza del estimador de la media poblacional propuesto, utilizando estimadores insesgados.*

Solución.
$\widehat{\overline{y}} = \overline{y}_s = 537$ y

$$\widehat{V}(\overline{y}_s) = \frac{N-n}{N}\,\frac{s^2}{n} = \frac{2000-20}{2000}\,\frac{100}{20} = 4.95.$$

Ejercicio 3.2 *Acotar la varianza de una proporción muestral con diseño mas en cualquier caso, independientemente de los posibles valores que pueda tomar la proporción poblacional.*

Solución.
La varianza de la proporción muestral es

$$V(\hat{p}) = \frac{N-n}{N-1}\,\frac{PQ}{n} \leq \frac{N-n}{4(N-1)n}$$

pues $f(P) = P(1-P) = PQ$ se minimiza en $P = \frac{1}{2}$, es decir, $f'(P) = 1 - 2P = 0$, luego el punto crítico es $P = \frac{1}{2}$. Al ser $f''(P) = -2 < 0$ el punto crítico es un máximo. Así $P(1-P) \leq \frac{1}{4} = f(\frac{1}{2})$.

Ejercicio 3.3 *Calcular el tamaño muestral necesario para obtener un error máximo de muestreo* $e = 10^5$ *al nivel de confianza* $0,90 = 1-\alpha$ *para estimar el total de una población finita de tamaño* $N = 30.000$ *(De una muestra piloto, se estima* S^2 *por 50).*

Solución.

$$n = \frac{N^2 S^2}{\alpha e^2 + N S^2} \simeq 45$$

Ejercicio 3.4 *Una muestra aleatoria simple sin reemplazamiento ha sido seleccionada de la población compuesta por las familias residentes en cierta provincia, con objeto de estimar el número medio de hijos varones por familia. Se han observado* $n = 10$ *familias del total de las mismas* $N = 39000$. *Los datos vienen recogidos por la siguiente tabla*

Familia i	*nº hijos varones,* y_i	y_i^2
1	0	0
2	2	4
3	1	1
4	3	9
5	1	1
6	0	0
7	6	36
8	0	0
9	4	16
10	2	4

Estimar insesgadamente la media provincial de hijos varones por familia y estimar insesgadamente la varianza del estimador insesgado.

Solución.
Como estimador de la media tomamos la media muestral

$$\overline{y}_s = \frac{\sum_{i=1}^{10} y_i}{10} = 1.9$$

y como estimador insesgado de la varianza de $\overline{y}_s$, tenemos que

$$\hat{V}(\overline{y}_s) = \frac{N-n}{N}\,\frac{s^2}{n} = \frac{38990}{39000}\,\frac{7.8487778}{10} = 0.7846765$$

con $N = 39000$, $n = 10$ y,

$$s^2 = \frac{\sum_{i=1}^{10}(y_i - \overline{y}_s)^2}{n-1} = \frac{\sum_{i=1}^{10} y_i^2 - \frac{\left(\sum_{i=1}^{10} y_i\right)^2}{10}}{9} = \frac{71 - \frac{1.9^2}{10}}{9} = 7.8487778.$$

Ejercicio 3.5 *Una industria tiene interés en conocer el tiempo semanal que los empleados gastan en ciertas actividades no productivas. Las fichas control del tiempo de una muestra aleatoria simple sin reemplazamiento de $n = 70$ empleados muestran que el tiempo promedio dedicado a esas actividades es de 16.45 horas, con una cuasivarianza muestral de $s^2 = 3.01$. La empresa emplea $N = 1250$ empleados. Estimar el número total de horas–hombre que se pierden por semana en tareas no productivas y dar una estimación de la varianza de tal estimación inicial.*

Solución.

La población consiste en $N = 1250$ empleados, de los que se extrae una muestra aleatoria simple sin reemplazamiento de $n = 70$ empleados. La cantidad promedio de tiempo que se pierde por los 70 empleados es de $\overline{y}_s = 16.45$ horas por semana.

Luego la estimación del total es

$$\widehat{N\overline{y}} = N\overline{y}_s = 1250 \cdot 16.45 = 20562.5 \text{ horas}$$

Para estimar insesgadamente la varianza de $\widehat{N\overline{y}}$, tenemos

$$\hat{V}\left(\widehat{N\overline{y}}\right) = N^2 \frac{N-n}{N} \frac{s^2}{n} = N(N-n)\frac{s^2}{n} = 1250 \cdot 1180 \frac{3.01}{70} = 63425$$

horas al cuadrado.

Ejercicio 3.6 *Para estimar la renta familiar disponible al año, en promedio, de una población, se sabe que existen en total $N = 200000$ familias y que tras una encuesta piloto se ha estimado que la cuasivarianza de la renta familiar es $S^2 = 2000000$.*

Determinar el tamaño muestral n para estimar la media poblacional o renta media $\overline{y}$, mediante la media muestral $\overline{y}_s$ obtenida por muestreo aleatorio simple sin reemplazamiento, para alcanzar un error máximo de muestreo de $e = 200000$ pts, con una probabilidad de $1 - \alpha = 0.95$.

Solución.

Directamente obtenemos que este tamaño muestral debe ser

$$n = \frac{S^2}{\alpha e^2 + \frac{S^2}{N}} = \frac{2000000}{0.05 \cdot 200000^2 + \frac{2000000}{200000}} = \frac{2 \cdot 10^6}{2 \cdot 10^9 + 10} = 1$$

luego con el tamaño muestral $n = 1$ se obtienen dichas características, si la muestra piloto ha sido bien realizada para calcular $S^2 = 2000000$.

Ejercicio 3.7 *Una empresa productora de aves para el consumo alimenticio está interesada en estimar la ganancia total de peso de $N = 2000$ aves a lo largo de un mes mediante la alimentación de las aves con una ración. Frente a la alternativa de tener que pesar las 2000 aves un mes después, se diseña un método de estimación del total $N\overline{y}$, siendo y_i el peso del ave i ($i = 1, 2, \ldots, 2000$), por el que se pesarán n aves de modo que el error máximo de muestreo, e, sea igual a $e = 3000$ gramos=3Kg. al nivel de confianza de $1 - \alpha = 0.90$ (90%). Usando datos de anteriores estudios sobre nutrición se estima que la cuasivarianza es $S^2 = 40$ gramos al cuadrado. Determinar el tamaño muestral n.*

Solución.
Ahora tendremos que al estimar un total poblacional,

$$n = \frac{S^2}{\frac{\alpha e^2}{N^2} + \frac{S^2}{N}} = \frac{40}{\frac{0.10 \cdot 3000^2}{2000^2} + \frac{40}{2000}} = 163.26531$$

así que tomando $n = 164$ aves podemos estimar el peso total con dichos requerimientos.

Ejercicio 3.8 *Una muestra aleatoria simple sin reemplazamiento de tamaño $n = 100$ se ha seleccionado para estimar:*

a) la fracción de $N = 300$ estudiantes de C.O.U. que asistirán a la Universidad y

b) la fracción de estudiantes que han trabajado a tiempo parcial durante su estancia en el instituto.

Sean y_i y x_i ($i = 1, 2, \ldots, 100$) las respuestas del i–ésimo estudiante seleccionado: $y_i = 0$ si el i–ésimo estudiante no planea ir a la Universidad, e $y_i = 1$ si lo planea. Del mismo modo, $x_i = 0$ si él no ha tenido alguna vez trabajo a tiempo parcial durante su estancia en el instituto, y $x_i = 1$ si lo ha tenido.
Sea:

$$\sum_{i=1}^{100} y_i = 25$$

y

$$\sum_{i=1}^{100} x_i = 30.$$

Usando estos datos muestrales estimar p_1 (proporción de estudiantes del último año planean asistir a la Universidad) y p_2 (proporción de estudiantes del último año que ha trabajado a tiempo parcial durante sus cursos en el instituto, incluyendo veranos).

Solución.

$$\widehat{p}_1 = \frac{\sum_{i=1}^{100} y_i}{100} = \frac{25}{100} = 0.25$$

y

$$\widehat{p}_2 = \frac{\sum_{i=1}^{100} x_i}{100} = \frac{30}{100} = 0.30.$$

Además las varianzas estimadas para $V(\widehat{p}_1)$ y $V(\widehat{p}_2)$ son:

$$\widehat{V}(\widehat{p}_1) = \frac{N-n}{N}\,\frac{\widehat{p}_1\widehat{q}_1}{n-1} = \frac{300-100}{300}\,\frac{0.25\cdot 0.75}{99} = 0.00126$$

y

$$\widehat{V}(\widehat{p}_2) = \frac{N-n}{N}\,\frac{\widehat{p}_2\widehat{q}_2}{n-1} = \frac{300-100}{300}\,\frac{0.30\cdot 0.70}{99} = 0.00141.$$

Ejercicio 3.9 *Una empresa tiene a su cargo un total de $N = 2000$ obreros y el jefe de personal quiere estimar la proporción de obreros que llevan trabajando en la empresa más de 10 años. A tal efecto decide realizar un sondeo entre los obreros, ya que realizar un censo sería inapropiado debido a la rapidez con que debe disponer de los datos. Si se selecciona una muestra aleatoria simple sin reemplazamiento para estimar tal proporción P, determinar el tamaño muestral n, aceptando que $\widehat{p}$ estima suficientemente bien la proporción poblacional P (y $\widehat{q}$ a Q), cuando el error máximo admisible es $e = 0.1$ al nivel de confianza $1-\alpha = 0.95$. (Suponer $\widehat{p} = 50\%$).*

Solución.
Si $\widehat{p} = P$ y $\widehat{q} = Q$ entonces

$$n = \frac{\widehat{p}\widehat{q}\dfrac{N}{N-1}}{\alpha e^2 + \dfrac{\widehat{p}\widehat{q}}{N-1}} = \frac{0.5\cdot 0.5\dfrac{2000}{1999}}{0.05\cdot 0.1^2 + \dfrac{0.5\cdot 0.5}{1999}} = 400.2 \text{ obreros}$$

Bastará observar 401 obreros.

Capítulo 4

Muestreo estratificado.

§4.1 Introducción.

Este tipo de muestreo se produce cuando la población U de tamaño N se clasifica en L estratos o clases de modo que si el tamaño en el estrato h ($h = 1, 2, \ldots,$ ó L) es N_h tendremos

$$\sum_{h=1}^{L} N_h = N.$$

El tamaño relativo del estrato h es $\frac{N_h}{N} = P_h$, de modo que

$$\sum_{h=1}^{L} P_h = 1.$$

§4.2 Diseño estratificado.

Una muestra estratificada se obtiene al seleccionar aleatoriamente n_h elementos del estrato h, con $0 < n_h \leq N_h$ ($h = 1, 2, \ldots, L$). Además la selección dentro de cada estrato es independiente del resto de estratos, es decir no hay ninguna dependencia entre las unidades seleccionadas en uno y otro estratos cualesquiera. El tamaño de la muestra estratificada es

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

Si "y" fuera la variable de interés o de estudio, notaremos y_{hi} a la característica "y" medida en la i–ésima unidad del estrato h.

Tendremos

$$\overline{y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} y_{hi}$$

(Media del estrato h),

$$N_h \overline{y}_h = \sum_{i=1}^{N_h} y_{hi}$$

(Total del estrato h),

$$\sigma_h^2 = \frac{1}{N_h} \sum_{i=1}^{N_h} (y_{hi} - \overline{y}_h)^2$$

(Varianza del estrato h), siendo $\sigma_h^2 = \dfrac{(N_h - 1)S_h^2}{N_h}$, donde S_h^2 es la cuasivarianza en el estrato h.

La media poblacional es

$$\overline{y} = \frac{\sum_{h=1}^{L} \sum_{i=1}^{N_h} y_{hi}}{N} = \frac{\sum_{h=1}^{L} N_h \overline{y}_h}{N} = \sum_{h=1}^{L} P_h \overline{y}_h.$$

El total de la población es

$$N\overline{y} = \sum_{h=1}^{L} N_h \overline{y}_h.$$

La varianza de la población es

$$\sigma^2 = \frac{1}{N} \sum_{h=1}^{L} \sum_{i=1}^{N_h} (y_{hi} - \overline{y})^2 = \frac{N-1}{N} S^2$$

siendo S^2 la cuasivarianza poblacional.

§4.3 Análisis de la varianza de una población estratificada.

$$\sigma^2 = \frac{1}{N} \sum_{h=1}^{L} \sum_{i=1}^{N_h} (y_{hi} - \overline{y})^2$$

y sumando y restando $\overline{y}_h$, queda

$$\sigma^2 = \frac{1}{N} \sum_{h=1}^{L} \sum_{i=1}^{N_h} (y_{hi} - \overline{y}_h + \overline{y}_h - \overline{y})^2 =$$

$$= \frac{1}{N} \left[\sum_{h=1}^{L} \sum_{i=1}^{N_h} (y_{hi} - \overline{y}_h)^2 + \sum_{h=1}^{L} N_h (\overline{y}_h - \overline{y})^2 + 2 \sum_{h=1}^{L} \sum_{i=1}^{N_h} (y_{hi} - \overline{y}_h)(\overline{y}_h - \overline{y}) \right] =$$

$$= \frac{1}{N} \left[\sum_{h=1}^{L} N_h \sigma_h^2 + \sum_{h=1}^{L} N_h (\overline{y}_h - \overline{y})^2 + 0 \right] = \sum_{h=1}^{L} P_h \sigma_h^2 + \sum_{h=1}^{L} P_h (\overline{y}_h - \overline{y})^2 \qquad (4.1)$$

El paso 4.1 se debe a que

$$2\sum_{h=1}^{L}\sum_{i=1}^{N_h}(y_{hi}-\overline{y}_h)(\overline{y}_h-\overline{y})=2\sum_{h=1}^{L}(\overline{y}_h-\overline{y})\sum_{i=1}^{N_h}(y_{hi}-\overline{y}_h)=0$$

pues

$$\sum_{i=1}^{N_h}(y_{hi}-\overline{y}_h)=N_h\overline{y}_h-N_h\overline{y}_h=0.$$

Es decir, la conclusión es que la variabilidad total ó varianza poblacional σ^2 puede descomponerse en dos sumandos, uno que representa la variación dentro de estratos y otro que indica la variación entre estratos.

§4.4 Estimación de la media poblacional.

Si $\overline{y}_{s(h)}$ es la media muestral observada en el estrato h–ésimo a partir de la muestra $s(h)$, un estimador insesgado de la media poblacional $\overline{y}$ es

$$\overline{y}_{st}=\sum_{h=1}^{L}P_h\overline{y}_{s(h)}.$$

En efecto,

$$E(\overline{y}_{st})=E\left(\sum_{h=1}^{L}P_h\overline{y}_{s(h)}\right)=\sum_{h=1}^{L}P_hE\left(\overline{y}_{s(h)}\right)=\sum_{h=1}^{L}P_h\overline{y}_h=\overline{y}.$$

Además la varianza del estimador $\overline{y}_{st}$, suponiendo que dentro de cada estrato se selecciona la muestra $s(h)$ por muestreo aleatorio simple sin reemplazamiento de tamaño n_h, es

$$V(\overline{y}_{st})=V\left(\sum_{h=1}^{L}P_h\overline{y}_{s(h)}\right)=\sum_{h=1}^{L}V\left(P_h\overline{y}_{s(h)}\right)$$

por ser independientes las selecciones de unidades en los estratos, y por ello, siendo $P_h=\dfrac{N_h}{N}$,

$$V(\overline{y}_{st})=\sum_{h=1}^{L}P_h^2V\left(\overline{y}_{s(h)}\right)=\sum_{h=1}^{L}P_h^2\frac{N_h-n_h}{N_h}\frac{S_h^2}{n_h}=\sum_{h=1}^{L}\frac{N_h(N_h-n_h)}{N^2}\frac{S_h^2}{n_h}.$$

Un estimador insesgado de $V(\overline{y}_{st})$ será

$$\widehat{V}(\overline{y}_{st})=\sum_{h=1}^{L}\frac{N_h(N_h-n_h)}{N^2}\frac{s_h^2}{n_h}$$

pues $E(s_h^2)=S_h^2$, o sea, la cuasivarianza muestral s_h^2 es insesgada para estimar la cuasivarianza poblacional en el estrato h, S_h^2, con diseño mas en cada estrato.

§4.5 Estimación del total poblacional.

El estimador usual, en muestreo estratificado, del total poblacional $T = N\overline{y}$ es $\widehat{T} = N\overline{y}_{st}$. Este estimador es insesgado,

$$E\left(N\overline{y}_{st}\right) = NE\left(\overline{y}_{st}\right) = N\overline{y}.$$

La varianza de $N\overline{y}_{st}$ es

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

y un estimador insesgado de ésta última varianza es

$$\widehat{V}\left(N\overline{y}_{st}\right) = \sum_{h=1}^{L}N_h\left(N_h - n_h\right)\frac{s_h^2}{n_h}$$

pues $E(s_h^2) = S_h^2$ $(h = 1, 2, \ldots, L)$ con diseño mas de tamaño n_h.

§4.6 Estimación de la proporción poblacional.

Como hemos visto, la proporción (ya sea poblacional o muestral) es una media aritmética donde la variable "y" toma valores de cero o uno, y por ello la proporción muestral $\widehat{P}$ es un estimador insesgado de la proporción poblacional P, que es

$$\overline{y}_{st} = \widehat{P} = \sum_{h=1}^{L}P_h\widehat{p}_h = \sum_{h=1}^{L}P_h\overline{y}_{s(h)}$$

donde $\widehat{p}_h = \overline{y}_{s(h)}$ es la proporción muestral en el estrato h. Así,

$$E\left(\widehat{P}\right) = E\left(\sum_{h=1}^{L}P_h\widehat{p}_h\right) = \sum_{h=1}^{L}P_hE\left(\widehat{p}_h\right) = \sum_{h=1}^{L}P_hp_h = P.$$

La varianza de este estimador es (siendo p_h la proporción del estrato h y $q_h = 1 - p_h$),

$$V\left(\widehat{P}\right) = V\left(\sum_{h=1}^{L}P_h\widehat{p}_h\right) = \sum_{h=1}^{L}P_h^2V\left(\widehat{p}_h\right) = \sum_{h=1}^{L}P_h^2\frac{N_h - n_h}{N_h - 1}\frac{p_hq_h}{n_h}.$$

pues la varianza del estrato h es $\sigma_h^2 = p_hq_h$.

Un estimador insesgado de $V(\widehat{P})$ es

$$\widehat{V}\left(\widehat{P}\right) = \sum_{h=1}^{L}P_h^2\frac{N_h - n_h}{N_h}\frac{\widehat{p}_h\widehat{q}_h}{n_h - 1}$$

porque un estimador insesgado de $S_h^2 = N_h\dfrac{\sigma_h^2}{(N_h - 1)}$ es la cuasivarianza muestral $s_h^2 = n_h\dfrac{\widehat{p}_h\widehat{q}_h}{(n_h - 1)}$, pues

$$E\left[\hat{V}\left(\hat{P}\right)\right] = \sum_{h=1}^{L} P_h^2 \frac{N_h - n_h}{N_h} \frac{E\left(\hat{p}_h \hat{q}_h\right)}{n_h - 1} = V\left(\hat{P}\right)$$

ya que

$$\frac{n_h}{n_h - 1} E\left(\hat{p}_h \hat{q}_h\right) = p_h q_h \frac{N_h}{N_h - 1}$$

como ya conocíamos.

§4.7 Afijación muestral.

Dado el tamaño muestral n, se denomina afijación muestral al reparto de n en L números n_h $(h = 1, 2, \ldots, L)$ de modo que

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

y n_h sea el tamaño muestral en el estrato h $(h = 1, 2, \ldots, L)$. Algunos tipos de afijación son:

4.7.1 Afijación igual (ig.).

Consiste en asignar el mismo tamaño muestral en cada estrato, es decir $n_1 = n_2 = \cdots = n_h = \cdots = n_L$. Como

$$n = \sum_{h=1}^{L} n_h = L n_h \quad \Rightarrow \quad n_h = \frac{n}{L} \quad (h = 1, 2, \ldots, L).$$

4.7.2 Afijación proporcional (prop.).

Consiste en asignar a cada estrato h, un tamaño muestral n_h proporcional al tamaño de dicho estrato N_h. Si n_h es proporcional a N_h, $n_h \alpha N_h$

$$n = \sum_{h=1}^{L} n_h = \sum_{h=1}^{L} k N_h = k \sum_{h=1}^{L} N_h = kN \quad \Rightarrow \quad k = \frac{n}{N}$$

donde k es la constante de proporcionalidad. Luego

$$n_h = k N_h = \frac{n}{N} N_h = n P_h \quad (h = 1, 2, \ldots, L).$$

4.7.3 Afijación mínima (mín.).

Fijado el tamaño muestral n, consiste en asignar a cada estrato un tamaño muestral n_h de modo que la varianza $V(\overline{y}_{st})$ sea mínima.

Para ello utilizamos el método de los multiplicadores de Lagrange con la restricción

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

El lagrangiano es L^*,

$$L^* = V(\overline{y}_{st}) + \lambda\left(\sum_{h=1}^{L} n_h - n\right) = \sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{N^2}\,\frac{S_h^2}{n_h} + \lambda\left(\sum_{h=1}^{L} n_h - n\right)$$

donde λ es el multiplicador de Lagrange. Resolviendo,

$$\frac{\partial L}{\partial n_h} = -\frac{N_h^2 S_h^2}{N^2 n_h^2} + \lambda = 0 \quad (h = 1, 2, \ldots, L)$$

de donde

$$\lambda = \frac{N_h^2 S_h^2}{N^2 n_h^2} \quad (h = 1, 2, \ldots, L)$$

luego

$$\sqrt{\lambda} = \frac{N_1 S_1}{N n_1} = \cdots = \frac{N_h S_h}{N n_h} = \cdots = \frac{N_L S_L}{N n_L} =$$

$$= \frac{\sum_{h=1}^{L} N_h S_h}{N \sum_{h=1}^{L} n_h} = \frac{\sum_{h=1}^{L} N_h S_h}{Nn}$$

y por esto,

$$n_h = n\frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} \quad (h = 1, 2, \ldots, L)$$

es decir la afijación mínima consiste en asignar al estrato h un tamaño muestral n_h directamente proporcional al producto $N_h S_h$.

Este mismo resultado de afijación se daría si se fija la varianza $V = V(\overline{y}_{st})$, de modo que se minimice el tamaño muestral

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

4.7.4 Afijación óptima con costes variables (ópt.)

Si suponemos ahora que el coste de observación de una unidad muestral del estrato h es C_h, y el coste total de la muestra es C, podemos minimizar la varianza $V(\overline{y}_{st})$ sujeta a que

$$C = \sum_{h=1}^{L} C_h n_h.$$

El mismo resultado se dará cuando se minimiza el coste C sujeto a una precisión o varianza prefijada. Recordemos que la precisión de un estimador insesgado es el inverso de su varianza.

El lagrangiano será ahora

$$L^* = \sum_{h=1}^{L} \frac{N_h (N_h - n_h)}{N^2} \frac{S_h^2}{n_h} + \lambda \left(\sum_{h=1}^{L} C_h n_h - C \right)$$

y derivando parcialmente L^* respecto a n_h e igualando a cero,

$$\frac{\partial L}{\partial n_h} = -\frac{N_h^2 S_h^2}{N^2 n_h^2} + \lambda C_h = 0 \quad (h = 1, 2, \ldots, L)$$

de donde

$$\sqrt{\lambda} = \frac{\frac{N_h S_h}{\sqrt{C_h}}}{N n_h} = \frac{\sum_{h=1}^{L} \frac{N_h S_h}{\sqrt{C_h}}}{N \sum_{h=1}^{L} n_h} = \frac{\sum_{h=1}^{L} \frac{N_h S_h}{\sqrt{C_h}}}{Nn}$$

y por tanto

$$n_h = n \frac{\frac{N_h S_h}{\sqrt{C_h}}}{\sum_{h=1}^{L} \frac{N_h S_h}{\sqrt{C_h}}} \quad (h = 1, 2, \ldots, L)$$

es decir n_h es proporcional a $\frac{N_h S_h}{\sqrt{C_h}}$. Esta afijación encuentra una solución de compromiso entre el coste y la precisión. El tamaño muestral en el estrato h, n_h, puede expresarse en función del coste prefijado C. En efecto,

$$C = \sum_{h=1}^{L} C_h n_h = n \frac{\sum_{h=1}^{L} N_h S_h \sqrt{C_h}}{\sum_{h=1}^{L} \frac{N_h S_h}{\sqrt{C_h}}}$$

luego

$$n = C \frac{\sum_{h=1}^{L} \frac{N_h S_h}{\sqrt{C_h}}}{\sum_{h=1}^{L} N_h S_h \sqrt{C_h}}$$

de donde sustituyendo, tenemos finalmente

$$n_h = n\frac{\frac{N_h S_h}{\sqrt{C_h}}}{\sum_{h=1}^{L}\frac{N_h S_h}{\sqrt{C_h}}} = C\frac{\frac{N_h S_h}{\sqrt{C_h}}}{\sum_{h=1}^{L} N_h S_h\sqrt{C_h}} \quad (h = 1, 2, \ldots, L).$$

4.7.5 Afijación fijada (fij.)

Si n_h $(h = 1, 2, \ldots, L)$ están dados previamente, el tamaño muestral n será también prefijado,

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

4.7.6 Afijación valoral (val.)

Esta afijación consiste en que dado el tamaño muestral n, distribuir el tamaño muestral n_h en el estrato h de forma que n_h sea proporcional al total del estrato h, $N_h\overline{y}_h$. Es decir,

$$\frac{n_1}{N_1\overline{y}_1} = \cdots = \frac{n_h}{N_h\overline{y}_h} = \cdots = \frac{n_L}{N_L\overline{y}_L} =$$

$$= \frac{\sum_{h=1}^{L} n_h}{\sum_{h=1}^{L} N_h\overline{y}_h} = \frac{n}{N\overline{y}}$$

de donde

$$n_h = n\frac{N_h\overline{y}_h}{N\overline{y}} = n\frac{P_h\overline{y}_h}{\overline{y}} \quad (h = 1, 2, \ldots, L).$$

4.7.7 Afijación especial (spc.)

Consiste en asignar un tamaño muestral en el estrato L (usualmente denominado de unidades grandes), n_L, de modo que $n_L = N_L$, es decir se selecciona todo el último estrato en la muestra. Así, para los restantes estratos h $(h = 1, 2, \ldots, L-1)$ puede asignarse cualquier tipo de afijación ya visto para el tamaño muestral restante $n - N_L$.

§4.8 Comparaciones.

Vamos a comparar la estrategia (mas, $\overline{y}_s$) con las estrategias estratificadas (prop, $\overline{y}_{st}$) y (mín, $\overline{y}_{st}$) para un tamaño muestral n común, en el caso en que N y N_h sean suficientemente grandes. En tal caso, S^2 y S_h^2 son muy próximos a σ^2 y σ_h^2, por lo que

$$V\,(\text{mas}, \overline{y}_s) \doteq \frac{\sigma^2}{n}$$

$$V\,(\text{prop}, \overline{y}_{st}) \doteq \frac{1}{n} \sum_{h=1}^{L} P_h \sigma_h^2$$

sustituyendo en $V(\overline{y}_{st})$ el valor n_h por nP_h, y

$$V\,(\text{mín}, \overline{y}_{st}) \doteq \frac{1}{n} \left(\sum_{h=1}^{L} P_h \sigma_h \right)^2$$

sustituyendo en $V(\overline{y}_{st})$ el valor n_h por

$$\frac{n N_h S_h}{\sum_{h=1}^{L} N_h S_h}$$

Ahora, como

$$\sigma^2 = \sum_{h=1}^{L} P_h \sigma_h^2 + \sum_{h=1}^{L} P_h \left(\overline{y}_h - \overline{y} \right)^2 \geq \sum_{h=1}^{L} P_h \sigma_h^2$$

por la descomposición tradicional del análisis de la varianza, tenemos

$$V\,(\text{mas}, \overline{y}_s) \geq V\,(\text{prop}, \overline{y}_{st})$$

al dividir

$$\sigma^2 \quad \text{y} \quad \sum_{h=1}^{L} P_h \sigma_h^2 \quad \text{por} \quad n.$$

También como

$$0 \leq \sum_{h=1}^{L} P_h \left(\sigma_h - \overline{\sigma} \right)^2 = \sum_{h=1}^{L} P_h \sigma_h^2 - \overline{\sigma}^2 = \sum_{h=1}^{L} P_h \sigma_h^2 - \left(\sum_{h=1}^{L} P_h \sigma_h \right)^2$$

siendo σ_h la desviación típica en el estrato h y

$$\overline{\sigma} = \sum_{h=1}^{L} P_h \sigma_h$$

la desviación típica esperada.

Deducimos que

$$\sum_{h=1}^{L} P_h \sigma_h^2 \geq \left(\sum_{h=1}^{L} P_h \sigma_h \right)^2$$

o bien, dividiendo por n dichas expresiones tenemos

$$V(\text{prop}, \overline{y}_{st}) \geq V(\text{mín}, \overline{y}_{st}).$$

En conclusión las varianzas de las estrategias comparadas verifican

$$V(\text{mas}, \overline{y}_s) \geq V(\text{prop}, \overline{y}_{st}) \geq V(\text{mín}, \overline{y}_{st})$$

si N_h y N son suficientemente grandes.

Con lo visto en el diseño mas, es ahora fácil (como ejercicio) proponer estimadores insesgados de las varianzas obtenidas por cada afijación muestral.

§4.9 Estimador insesgado de la varianza poblacional.

Partimos de que el diseño empleado dentro de cada estrato es masr.

Sabemos que la varianza poblacional puede descomponerse de la forma siguiente:

$$\sigma^2 = \sum_{h=1}^{L} P_h \sigma_h^2 + \sum_{h=1}^{L} P_h (\overline{y}_h - \overline{y})^2.$$

Para estimar σ^2, el primer sumando no presenta problema con diseño masr puesto que la cuasivarianza muestral s_h^2 es un estimador insesgado de σ_h^2. En cuanto al segundo sumando, sustituyamos $\overline{y}_h$ e $\overline{y}$ por sus estimadores $\overline{y}_{s(h)}$ e $\overline{y}_{st}$, y calculemos la esperanza matemática:

$$E\left[\sum_{h=1}^{L} P_h \left(\overline{y}_{s(h)} - \overline{y}_{st}\right)^2\right] =$$

(sumando y restando $\overline{y}_h - \overline{y}$)

$$= E\left\{\sum_{h=1}^{L} P_h \left[(\overline{y}_h - \overline{y}) + \left(\overline{y}_{s(h)} - \overline{y}_h\right) - (\overline{y}_{st} - \overline{y})\right]^2\right\} =$$

$$= \underbrace{E\left[\sum_{h=1}^{L} P_h (\overline{y}_h - \overline{y})^2\right]}_{(1)} + \underbrace{E\left[\sum_{h=1}^{L} P_h \left(\overline{y}_{s(h)} - \overline{y}_h\right)^2\right]}_{(2)} +$$

$$+ \underbrace{E\left[\sum_{h=1}^{L} P_h (\overline{y}_{st} - \overline{y})^2\right]}_{(3)} \underbrace{-2\, E\left[\sum_{h=1}^{L} P_h \left(\overline{y}_{s(h)} - \overline{y}_h\right)(\overline{y}_{st} - \overline{y})\right]}_{(4)} +$$

$$\underbrace{+2\, E\left[\sum_{h=1}^{L} P_h (\overline{y}_h - \overline{y})\left(\overline{y}_{s(h)} - \overline{y}_h\right)\right]}_{(5)} \underbrace{-2\, E\left[\sum_{h=1}^{L} P_h (\overline{y}_h - \overline{y})(\overline{y}_{st} - \overline{y})\right]}_{(6)}$$

donde

$$(1) = \sum_{h=1}^{L} P_h \left(\overline{y}_h - \overline{y}\right)^2.$$

$$(2) = \sum_{h=1}^{L} P_h V\left(\overline{y}_{s(h)}\right) = \sum_{h=1}^{L} P_h \frac{\sigma_h^2}{n_h}.$$

$$(3) = \sum_{h=1}^{L} P_h V\left(\overline{y}_{st}\right) = \sum_{h=1}^{L} P_h \frac{\sigma_h^2}{n_h}.$$

$$(4) = E\left[\left(\overline{y}_{st} - \overline{y}\right)^2\right] = V\left(\overline{y}_{st}\right).$$

$$(5) = E\left[\sum_{h=1}^{L} P_h \overline{y}_h \left(\overline{y}_{s(h)} - \overline{y}_h\right) - \overline{y} \sum_{h=1}^{L} P_h \left(\overline{y}_{s(h)} - \overline{y}_h\right)\right] =$$

$$= \sum_{h=1}^{L} P_h \overline{y}_h E\left(\overline{y}_{s(h)} - \overline{y}_h\right) - \overline{y} E\left[\sum_{h=1}^{L} P_h \left(\overline{y}_{s(h)} - \overline{y}_h\right)\right] = 0 - \overline{y} E\left(\overline{y}_{st} - \overline{y}\right) = 0.$$

$$(6) = E\left[\left(\overline{y}_{st} - \overline{y}\right) \sum_{h=1}^{L} P_h \left(\overline{y}_h - \overline{y}\right)\right] = E\left[\left(\overline{y}_{st} - \overline{y}\right)\left(\overline{y} - \overline{y}\right)\right] = 0.$$

Luego

$$E\left[\sum_{h=1}^{L} P_h \left(\overline{y}_{s(h)} - \overline{y}_{st}\right)^2\right] = (1) + (2) + (3) - 2 \cdot (4) =$$

$$= \sum_{h=1}^{L} P_h \left(\overline{y}_h - \overline{y}\right)^2 + \sum_{h=1}^{L} P_h \frac{\sigma_h^2}{n_h} + V\left(\overline{y}_{st}\right) - 2V\left(\overline{y}_{st}\right) =$$

$$= \sum_{h=1}^{L} P_h \left(\overline{y}_h - \overline{y}\right)^2 + \sum_{h=1}^{L} P_h \frac{\sigma_h^2}{n_h} - \sum_{h=1}^{L} P_h^2 \frac{\sigma_h^2}{n_h} =$$

$$= \sum_{h=1}^{L} P_h \left(\overline{y}_h - \overline{y}\right)^2 + \sum_{h=1}^{L} P_h \left(1 - P_h\right) \frac{\sigma_h^2}{n_h},$$

por lo que el segundo sumando de ésta última fórmula es el sesgo de

$$\sum_{h=1}^{L} P_h \left(\overline{y}_{s(h)} - \overline{y}_{st}\right)^2$$

como estimador de

$$\sum_{h=1}^{L} P_h \left(\overline{y}_h - \overline{y}\right)^2$$

y en consecuencia el estimador insesgado de σ^2 con diseño estratificado y muestreo masr dentro de cada estrato es

$$\widehat{\sigma}^2 = \sum_{h=1}^{L} P_h^2 s_h^2 + \sum_{h=1}^{L} P_h \left(\overline{y}_{s(h)} - \overline{y}_{st}\right)^2 - \sum_{h=1}^{L} P_h \left(1 - P_h\right) \frac{s_h^2}{n_h}$$

Una fórmula algo más compleja se puede dar para estimar σ^2 con diseño estratificado y diseño mas dentro de cada estrato (Mirás, 1985).

§4.10 Postestratificación.

A veces se utiliza un diseño no estratificado para seleccionar la muestra, pero una vez seleccionada se decide estratificarla y estimar la media poblacional, $\overline{y}$, por una media postestratificada. De este modo el tamaño muestral en el estrato h, n_h, es aleatorio antes de seleccionar la muestra, y fijo una vez seleccionada. Sea $\overline{y}_{s(h)}$ la media muestral en el estrato h, y N_h el tamaño del estrato h, conocido para poder construir el estimador

$$\overline{y}_{poststst} = \sum_{h=1}^{L} \frac{N_h}{N} \overline{y}_{s(h)} = \sum_{h=1}^{L} P_h \overline{y}_{s(h)}$$

con

$$\overline{y}_{s(h)} = \frac{1}{n} \sum_{i \in s(h)} y_{hi}$$

siendo y_{hi} el i-ésimo valor del estrato h, que aparece en la muestra no estratificada. Ahora los valores n_h no son fijados previamente con el diseño estratificado, sino que son consecuencias o resultados constatados con la muestra obtenida, es decir, son aleatorios. Para calcular la esperanza y la varianza de $\overline{y}_{poststst}$, se considerará previamente fijados los n_h y despues se introduce la aleatoriedad de los n_h. Así haciendo uso de resultados que se demostrarán en el capítulo 10,

$$E\left(\overline{y}_{poststst}\right) = E\left[E\left(\overline{y}_{poststst} \mid n_h\right)\right] =$$

$$= E\left[\sum_{h=1}^{L} P_h E\left(\overline{y}_{s(h)} \mid n_h\right)\right] = E\left(\sum_{h=1}^{L} P_h \overline{y}_h\right) = E\left(\overline{y}\right) = \overline{y}$$

por lo que $\overline{y}_{poststst}$ es insesgado. Su varianza se calcula así,

$$V\left(\overline{y}_{poststst}\right) = V\left[E\left(\overline{y}_{poststst} \mid n_h\right)\right] + E\left[V\left(\overline{y}_{poststst} \mid n_h\right)\right]$$

pero como

$$E\left(\overline{y}_{poststst} \mid n_h\right) = \overline{y} \quad \Rightarrow \quad V\left(\overline{y}\right) = 0 \quad \Rightarrow \quad V\left[E\left(\overline{y}_{poststst} \mid n_h\right)\right] = 0$$

y

$$V\left(\overline{y}_{poststst} \mid n_h\right) = \sum_{h=1}^{L} P_h^2 V\left(\overline{y}_{s(h)} \mid n_h\right) =$$

$$= \sum_{h=1}^{L} P_h^2 \frac{N_h - n_h}{N_h n_h} S_h^2 = \sum_{h=1}^{L} P_h^2 \frac{S_h^2}{n_h} - \frac{1}{N} \sum_{h=1}^{L} P_h S_h^2$$

de donde

$$E\left[V\left(\overline{y}_{postst} \mid n_h\right)\right] = E\left[\sum_{h=1}^{L} P_h^2 S_h^2 \frac{1}{n_h} - \frac{1}{N} \sum_{h=1}^{L} P_h S_h^2\right] =$$

$$= \sum_{h=1}^{L} P_h^2 S_h^2 E\left(\frac{1}{n_h}\right) - \frac{1}{N} \sum_{h=1}^{L} P_h S_h^2.$$

Sea ahora $p_h = \dfrac{n_h}{n}$ el tamaño relativo de la muestra en el estrato h, mientras que $P_h = \dfrac{N_h}{N}$ es el tamaño relativo del estrato h en la población. Ahora p_h estima sin sesgo a P_h. Por tanto, escribiendo

$$n_h = n\frac{n_h}{n} = np_h = n\left(p_h - P_h + P_h\right) = nP_h\left(1 + \frac{p_h - P_h}{P_h}\right)$$

luego

$$\frac{1}{n_h} = \frac{1}{nP_h} \frac{1}{1 + \dfrac{p_h - P_h}{P_h}}$$

pero si

$$\left|\frac{p_h - P_h}{P_h}\right| < 1$$

cosa razonable pues p_h estima sin sesgo a P_h, y además converge en probabilidad, nos permite expresar

$$\frac{1}{n_h} = \frac{1}{nP_h}\left[1 - \frac{p_h - P_h}{P_h} + \frac{(p_h - P_h)^2}{P_h^2} - \cdots\right] \simeq$$

$$\simeq \frac{1}{nP_h}\left[1 - \frac{p_h - P_h}{P_h} + \frac{(p_h - P_h)^2}{P_h^2}\right]$$

aproximando $\dfrac{1}{n_h}$ por los tres primeros términos del desarrollo en serie. Tomando esperanzas,

$$E\left(\frac{1}{n_h}\right) \simeq \frac{1}{nP_h}\left[1 - 0 + \frac{V(p_h)}{P_h^2}\right] = \frac{1}{nP_h}\left[1 + \frac{N-n}{(N-1)n}\frac{P_h Q_h}{P_h^2}\right]$$

pues $E(p_h - P_h) = 0$; y siendo $Q_h = 1 - P_h$. Luego

$$V\left(\overline{y}_{postst}\right) \simeq \sum_{h=1}^{L} P_h^2 S_h^2 \frac{1}{nP_h}\left[1 + \frac{N-n}{(N-1)n}\frac{Q_h}{P_h}\right] - \frac{1}{N}\sum_{h=1}^{L} P_h S_h^2 =$$

$$= \sum_{h=1}^{L} P_h S_h^2 \left[\frac{1}{n} \left(1 + \frac{N-n}{(N-1)n} \frac{Q_h}{P_h} \right) - \frac{1}{N} \right].$$

Por último vamos a ver que éste método de estimación se basa en que los tamaños relativos P_h son conocidos. Si calculáramos

$$\overline{y}_{poststo} = \sum_{h=1}^{L} P_{h_0} \overline{y}_{s(h)}$$

con P_{h_0} cualquiera, tendremos que

$$\overline{y}_{postst} = \overline{y}_{poststo} + \sum_{h=1}^{L} (P_h - P_{h_0}) \overline{y}_{s(h)}$$

por lo que

$$E\left(\overline{y}_{poststo}\right) = E\left(\overline{y}_{postst}\right) - \sum_{h=1}^{L} (P_h - P_{h_0}) E\left[\overline{y}_{s(h)}\right] = \overline{y} - \sum_{h=1}^{L} (P_h - P_{h_0}) \overline{y}_h$$

luego el sesgo de $\overline{y}_{poststo}$ será

$$B\left(\overline{y}_{poststo}\right) = E\left(\overline{y}_{poststo}\right) - \overline{y} = -\sum_{h=1}^{L} (P_h - P_{h_0}) \overline{y}_h = B$$

por lo que su error cuadrático medio será

$$\mathrm{ECM}\left(\overline{y}_{poststo}\right) = V\left(\overline{y}_{poststo}\right) + B^2$$

donde B es el sesgo de $\overline{y}_{poststo}$ que es independiente del tamaño muestral n. Por lo tanto, no es recomendable usar la estratificación a posteriori o postestratificación si los tamaños relativos de los estratos, P_h $(h = 1, 2, \ldots, L)$, no son conocidos.

Ejercicio 4.1 *En un estudio por muestreo estratificado se decide utilizar afijación especial para el tercer estrato (de unidades grandes) y utilizar afijación igual en los dos primeros estratos donde se emplean las estrategias (masr, $\overline{y}_{S(1)}$) y (mas, $\overline{y}_{s(2)}$) respectivamente. Proponer un estimador insesgado de la media poblacional para este diseño y calcular su varianza.*

Solución.
Un estimador insesgado de $\overline{y}$ será

$$\overline{y}_{st} = P_1 \overline{y}_{S(1)} + P_2 \overline{y}_{s(2)} + P_3 \overline{y}_3$$

siendo $P_h = \dfrac{N_h}{N}$ como hemos denotado en teoría. Su varianza será:

$$V(\overline{y}_{st}) = P_1^2 \frac{\sigma_1^2}{n_1} + P_2^2 \frac{N_2 - n_2}{N_2 - 1} \frac{\sigma_2^2}{n_2}$$

con $n_1 = n_2 = \dfrac{(n - n_3)}{2}$, siendo n el tamaño muestral total y n_h el tamaño muestral en el estrato h(=1,2,3), con $n_3 = N_3$.

Ejercicio 4.2 *En las condiciones del ejercicio anterior, proponer un estimador insesgado de la varianza del estimador de* $\overline{y}$.

Solución.
Será

$$\widehat{V}(\overline{y}_{st}) = P_1^2 \frac{s_1^2}{n_1} + P_2^2 \frac{N_2 - n_2}{N_2} \frac{s_2^2}{n_2}$$

siendo s_1^2 la cuasivarianza muestral obtenida con diseño masr de tamaño fijo n_1, en el primer estrato, y s_2^2 la cuasivarianza muestral obtenida con diseño mas de tamaño efectivo fijo n_2 en el segundo estrato.

Ejercicio 4.3 *En una población estratificada en 2 estratos, se ha obtenido que* $P_1 = 0.4$ *y* $P_2 = 0.6$*; y por una muestra piloto se sabe que aproximadamente* $\sigma_1^2 = 100$, $\sigma_2^2 = 81$ *y* $\sigma^2 = 225$.

Suponiendo que N *(tamaño poblacional) y los* N_h *(tamaños de los estratos,* $h = 1, 2$*) son suficientemente grandes, calcular el tamaño muestral* n *para que una muestra con afijación mínima proporcione la misma varianza que un diseño mas sobre toda la población de tamaño muestral efectivo* $n^* = 150$, *para estimar la media poblacional* $\overline{y}$.

Solución.

$$V(\text{mas}, \overline{y}_s) = \frac{\sigma^2}{n^*} = \frac{225}{150} = 1.5,$$

$$V(\text{mín}, \overline{y}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{2} P_h \sigma_h\right)^2 = \frac{1}{n}(0.4 \cdot 10 + 0.6 \cdot 9)^2 = \frac{88.36}{n},$$

luego si

$$1.5 = V(\text{mas}, \overline{y}_s) = V(\text{mín}, \overline{y}_{st}) = \frac{88.36}{n}$$

se deduce que

$$n = \frac{88.36}{1.5} = 59.$$

Ejercicio 4.4 *Determinar la afijación proporcional en cada estrato, si el tamaño muestral total es* $n = 1000$, *y hay* 5 *estratos de tamaños relativos* $P_1 = 0.2$, $P_2 = 0.3$, $P_3 = 0.1$, $P_4 = 0.25$ *y* $P_5 = 0.15$. *¿Cuál es la mayor diferencia absoluta de tamaños muestrales con respecto a la afijación igual?.*

Solución.

Afijación proporcional:

$$n_h = nP_h = \begin{cases} 200 & \text{si} \quad \text{h=1} \\ 300 & \text{si} \quad \text{h=2} \\ 100 & \text{si} \quad \text{h=3} \\ 250 & \text{si} \quad \text{h=4} \\ 150 & \text{si} \quad \text{h=5} \end{cases}$$

Afijación igual:

$$n_h = \frac{1000}{5} = 200 \ (h = 1, 2, 3, 4 \text{ y } 5)$$

En los estratos 2 y 3 se dan las mayores diferencias absolutas entre ambas afijaciones, pues $\mid 300 - 200 \mid = \mid 100 - 200 \mid = 100$.

Ejercicio 4.5 *Para estimar la proporción poblacional P de inclinación de voto a cierto partido en el conjunto de españoles con derecho a voto, se ha dividido geográficamente España en dos estratos: litoral y centro, de modo que el tamaño relativo de ambos es $P_1 = P_2 = \frac{1}{2}$, o bien $N_1 = N_2 = 10000000$ de votantes. Se decide usar afijación igual $n_1 = n_2 = 5000$ y resultan, con diseño mas en cada estrato, las proporciones $\widehat{p}_1 = 0.35$ y $\widehat{p}_2 = 0.28$. Estimar P por muestreo estratificado aleatorio con afijación igual y estimar insesgadamente la varianza de tal estimador $\widehat{P}$.*

Solución.

$$\widehat{P} = \sum_{h=1}^{2} P_h \widehat{p}_h = \frac{1}{2}(0.35 + 0.28) = 0.315.$$

luego la estimación del porcentaje de voto favorable es del 31.5%.

Un estimador de $V(\widehat{P})$, insesgado es:

$$\widehat{V}\left(\widehat{P}\right) = \sum_{h=1}^{2} P_h^2 \frac{N_h - n_h}{N_h} \frac{\widehat{p}_h \widehat{q}_h}{n_h - 1} =$$

$$= \frac{1}{4} \frac{9995000}{10000000} \frac{1}{4999} (0.35 \cdot 0.65 + 0.28 \cdot 0.72) \simeq 0.0000214$$

que representa un error de muestreo muy pequeño, por lo que el estimador $\widehat{P}$ es muy preciso.

Ejercicio 4.6 *En una comarca compuesta por tres pueblos A, B y C, se desea conocer la edad media de sus habitantes. Para ello se dispone de un presupuesto de 10000 ptas, y se supone que el costo por observación es de $c_A = c_B = 8$ y $c_C = 12$ ptas (por unidad). Determinar el tamaño muestral n_A , n_B y n_C en cada pueblo, y el tamaño total muestral n, si de una encuesta previa se ha estimado que las cuasivarianzas son $S_A = 30$, $S_B = 32$ y $S_C = 40$, y que se dispone de la información del total de habitantes en cada pueblo $N_A = 25000$, $N_B = 12000$ y $N_C = 2000$. El objetivo es obtener la máxima precisión a coste fijo.*

Solución.

n_h debe ser proporcional a $\frac{N_h S_h}{\sqrt{C_h}}$ (h=1,2,3 ó A,B,C)

$$\begin{aligned} n_A \alpha \frac{N_A S_A}{\sqrt{C_A}} &= 25000 \cdot \frac{30}{\sqrt{8}} = 265165.04 \quad n_A = t \cdot 265165.04 \\ n_B \alpha \frac{N_B S_B}{\sqrt{C_B}} &= 12000 \cdot \frac{32}{\sqrt{8}} = 135764.5 \quad n_B = t \cdot 135764.5 \\ n_C \alpha \frac{N_C S_C}{\sqrt{C_C}} &= 2000 \cdot \frac{40}{\sqrt{12}} = 23094.011 \quad n_C = t \cdot 23094.011 \end{aligned} \tag{4.2}$$

Por otro lado tenemos que el coste total es (siendo t la constante de proporcionalidad)

$$10000 = C = \sum_{h=1}^{3} n_h C_h = 8n_A + 8n_B + 12n_C = t(3484564.5)$$

de donde

$$t = \frac{10000}{3484564.5} = 0.0028697991$$

Luego de 4.2,

n_A	= 760.97	$\simeq$	761	habitantes del pueblo A
n_B	= 389.60	$\simeq$	390	habitantes del pueblo B
n_C	= 66.27	$\simeq$	67	habitantes del pueblo C
n		=	1.218	habitantes en total

Ejercicio 4.7 *Una empresa de publicidad quiere estimar la proporción $\widehat{p}_{st}$ de hogares en un municipio donde se ve cierto programa televisivo. El municipio es dividido en tres estratos 1, 2 y 3. Los tamaños de los estratos son $N_1 = 155$, $N_2 = 62$ y $N_3 = 93$ hogares respectivamente. Una muestra estratificada de $n = 40$ hogares se selecciona con afijación proporcional. Estimar p_{st} y dar una estimación insesgada de su varianza. Datos:*

Estrato	*Tamaño muestra*	*Número de hogares donde se ve el programa*	$\widehat{p}_{s(h)}$
1	$n_1 = 20$	*16*	*0.80*
2	$n_2 = 8$	*2*	*0.25*
3	$n_3 = 12$	*6*	*0.50*

Solución.

$N = N_1 + N_2 + N_3 = 155 + 62 + 93 = 310$ hogares en total en el municipio.

$$\widehat{p}_{st} = \sum_{h=1}^{3} \frac{N_h}{N} \widehat{p}_{s(h)} = \frac{155}{310} 0.80 + \frac{62}{310} 0.25 + \frac{93}{310} 0.50 = 0.60$$

es la proporción pedida; y una estimación insesgada de la varianza de $\widehat{p}_{st}$ es

$$\widehat{V}\left(\widehat{p}_{st}\right) = \sum_{h=1}^{3} \frac{N_h^2}{N^2} \widehat{V}\left(\widehat{p}_{s(h)}\right)$$

donde

$$\widehat{V}\left(\widehat{p}_{s(h)}\right) = \frac{N_h - n_h}{N_h} \frac{\widehat{p}_h \widehat{q}_h}{n_h - 1} = \begin{cases} 0.007 & \text{si} \quad h = 1 \\ 0.024 & \text{si} \quad h = 2 \\ 0.020 & \text{si} \quad h = 3 \end{cases}$$

luego sustituyendo

$$\widehat{V}\left(\widehat{p}_{st}\right) = 0.0045.$$

Capítulo 5

Estimador de la razón.

§5.1 Introducción.

Si además de la variable de interés "y", que puede observarse por muestreo aleatorio simple sin reemplazamiento, disponemos de una variable auxiliar "x" conocida para todas las unidades de la población finita, es posible estimar la "razón poblacional" R,

$$R = \frac{N\overline{y}}{N\overline{x}} = \frac{\overline{y}}{\overline{x}}$$

por medio de la "razón muestral" $\widehat{R}$,

$$\widehat{R} = \frac{\overline{y}_s}{\overline{x}_s} = \frac{n\overline{y}_s}{n\overline{x}_s}.$$

§5.2 Sesgo del estimador de la razón.

$$\mathrm{Cov}\left(\widehat{R}, \overline{x}_s\right) = E\left(\widehat{R}\overline{x}_s\right) - E\left(\widehat{R}\right) E\left(\overline{x}_s\right) =$$

$$= E\left(\overline{y}_s\right) - E\left(\widehat{R}\right)\overline{x}_s = \overline{y} - E\left(\widehat{R}\right)\overline{x}$$

y dividiendo ambas igualdades extremas por $\overline{x}$, tenemos

$$\frac{\mathrm{Cov}\left(\widehat{R}, \overline{x}_s\right)}{\overline{x}} = R - E\left(\widehat{R}\right)$$

es decir, el sesgo (o "bias" en inglés) exacto es la esperanza del estimador menos la función paramétrica a estimar

$$B\left(\widehat{R}\right) = E\left(\widehat{R}\right) - R = -\frac{\mathrm{Cov}\left(\widehat{R}, \overline{x}_s\right)}{\overline{x}}.$$

§5.3 Sesgo aproximado.

$$\hat{R} - R = \frac{\overline{y}_s}{\overline{x}_s} - R = \frac{\overline{y}_s - R\overline{x}_s}{\overline{x}_s} = \frac{\overline{y}_s - R\overline{x}_s}{\overline{x}} \frac{\overline{x}}{\overline{x}_s}$$

suponiendo que la variable de interés "y" es siempre positiva, así como la variable auxiliar "x". Ahora,

$$\frac{\overline{x}}{\overline{x}_s} = \frac{\overline{x}}{\overline{x} + \overline{x}_s - \overline{x}} = \frac{1}{1 + \dfrac{\overline{x}_s - \overline{x}}{\overline{x}}} =$$

$$= 1 - \frac{\overline{x}_s - \overline{x}}{\overline{x}} + \frac{(\overline{x}_s - \overline{x})^2}{\overline{x}^2} - \cdots$$

siempre y cuando

$$\left| \frac{\overline{x}_s - \overline{x}}{\overline{x}} \right| < 1$$

con lo que disponemos del desarrollo en serie

$$\hat{R} - R = \frac{\overline{y}_s - R\overline{x}_s}{\overline{x}} \left[1 - \frac{\overline{x}_s - \overline{x}}{\overline{x}} + \frac{(\overline{x}_s - \overline{x})^2}{\overline{x}^2} - \cdots \right]$$

de donde el sesgo expresado asintóticamente es

$$B\left(\hat{R}\right) = E\left(\hat{R} - R\right) =$$

$$= -\frac{E\left[(\overline{y}_s - R\overline{x}_s)(\overline{x}_s - \overline{x})\right]}{\overline{x}^2} + \frac{E\left[(\overline{y}_s - R\overline{x}_s)(\overline{x}_s - \overline{x})^2\right]}{\overline{x}^3} - \cdots$$

pues el primer sumando del desarrollo verifica

$$E\left(\frac{\overline{y}_s - R\overline{x}_s}{\overline{x}}\right) = \frac{1}{\overline{x}}(\overline{y} - \overline{y}) = 0.$$

En concreto, si aproximamos el sesgo por los dos primeros términos del desarrollo en serie dado en $\hat{R} - R$, tenemos que como el primer sumando es 0,

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

$$= -\frac{E\left[\overline{y}_s(\overline{x}_s - \overline{x})\right] + RE\left[\overline{x}_s(\overline{x}_s - \overline{x})\right]}{\overline{x}^2} =$$

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

$$= \frac{-\operatorname{Cov}(\overline{y}_s, \overline{x}_s) + RV(\overline{x}_s)}{\overline{x}^2}.$$

§5.4 Varianza aproximada.

Considerando el primer término del desarrollo en serie de $\widehat{R} - R$ obtenemos que

$$\widehat{R} - R \doteq \frac{\overline{y}_s - R\overline{x}_s}{\overline{x}}$$

Con esta aproximación, $E(\widehat{R}) \doteq R$ y

$$V\left(\widehat{R}\right) \doteq \frac{1}{\overline{x}^2} V\left(\overline{y}_s - R\overline{x}_s\right) = \frac{V\left(\overline{y}_s\right) + R^2 V\left(\overline{x}_s\right) - 2R\,\mathrm{Cov}\left(\overline{y}_s, \overline{x}_s\right)}{\overline{x}^2}.$$

Ahora bien, se demuestra que (de Hansen, Hurwitz y Madow, 1953, p. 97)

$$\mathrm{Cov}\left(\overline{y}_s, \overline{x}_s\right) = \frac{N-n}{Nn} S_{yx}$$

siendo S_{yx} la cuasicovarianza poblacional

$$S_{yx} = \frac{1}{N-1} \sum_{i=1}^{N} \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right) = S_{xy}.$$

Por ello,

$$V\left(\widehat{R}\right) = \frac{N-n}{Nn\overline{x}^2}\left(S_y^2 + R^2 S_x^2 - 2RS_{yx}\right).$$

§5.5 Tamaño muestral del estimador de razón.

Si queremos calcular el tamaño muestral n para que el "estimador de razón", $t_R = \widehat{R}\overline{x}$, de $\overline{y}$, difiera de la media poblacional menos que el error máximo absoluto admisible, e, con cierto nivel de confianza $1 - \alpha$, recurrimos a la desigualdad de Chebycheff, pues hemos visto que $E(t_R) \doteq \overline{y}$.

$$P\left[|t_R - \overline{y}| < e\right] \geq 1 - \frac{V\left(t_R\right)}{e^2} = 1 - \alpha$$

de donde

$$\alpha e^2 = V\left(t_R\right) = V\left(\widehat{R}\overline{x}\right) = \overline{x}^2 V\left(\widehat{R}\right) =$$

$$= \overline{x}^2 \frac{N-n}{Nn\overline{x}^2}\left(S_y^2 + R^2 S_x^2 - 2RS_{xy}\right) =$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right)\left(S_y^2 + R^2 S_x^2 - 2RS_{xy}\right)$$

luego

$$\frac{1}{n} = \frac{1}{N} + \frac{\alpha e^2}{S_y^2 + R^2 S_x^2 - 2RS_{xy}}$$

y por ello

$$n = \frac{1}{\frac{1}{N} + \frac{\alpha e^2}{S_y^2 + R^2 S_x^2 - 2RS_{xy}}}$$

siendo N, α y e valores conocidos; por lo que para estimar n debemos calcular S_y^2, R^2, S_x^2, R y S_{xy} de algún modo, por ejemplo proponiendo en lo posible estimadores insesgados de estas funciones paramétricas.

§5.6 Ganancia en precisión.

Para comparar las estrategias (mas, $\overline{y}_s$) y (mas, t_R), escribimos sus varianzas

$$V(\text{mas}, \overline{y}_s) = \frac{N-n}{Nn} S_y^2$$

y

$$V(\text{mas}, t_R) = \frac{N-n}{Nn}\left(S_y^2 + R^2 S_x^2 - 2RS_{xy}\right)$$

Entonces, $V(\text{mas}, t_R) \leq V(\text{mas}, \overline{y}_s)$ si y sólo si

$$R^2 S_x^2 - 2RS_{xy} \leq 0$$

o bien, como $S_{xy} = \rho S_x S_y$, queda equivalente a que

$$R^2 S_x^2 - 2R\rho S_x S_y \leq 0$$

es decir si y sólo si

$$\rho \geq \frac{R^2 S_x^2}{2RS_x S_y} = \frac{RS_x}{2S_y}$$

siendo ρ el coeficiente de correlación entre las variables "y" y "x".

§5.7 Estimador de la razón en el muestreo estratificado.

Existen dos tipos principales de estimadores de la razón combinado con el muestreo estratificado: el estimador separado de la razón y el estimador combinado de la razón.

5.7.1 Estimador separado de la razón.

Si dentro de cada estrato empleamos el estimador de razón, el estimador separado será

$$t_S = \sum_{h=1}^{L} P_h \widehat{R}_h \overline{x}_h = \frac{1}{N} \sum_{h=1}^{L} N_h \widehat{R}_h \overline{x}_h$$

y su varianza aproximada es

$$V(t_S) = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 \overline{x}_h^2 V\left(\widehat{R}_h\right) \doteq \sum_{h=1}^{L} P_h^2 \overline{x}_h^2 \frac{N_h - n_h}{N_h n_h}\left(S_{hy}^2 + R_h^2 S_{hx}^2 - 2R_h S_{hxy}\right)$$

donde

$$S_{hy}^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}(y_{hi} - \overline{y}_h)^2$$

$$R_h = \frac{\overline{y}_h}{\overline{x}_h}$$

$$S_{hxy} = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}(x_{hi} - \overline{x}_h)(y_{hi} - \overline{y}_h)$$

etc.

5.7.2 Estimador combinado de la razón.

Será

$$t_C = \widehat{R}_C \overline{x} = \frac{\sum_{h=1}^{L} P_h \overline{y}_{s(h)}}{\sum_{h=1}^{L} P_h \overline{x}_{s(h)}}\overline{x} = \frac{\overline{y}_{st}}{\overline{x}_{st}}\overline{x}$$

y su varianza aproximada (debido a que

$$\widehat{R}_C - R \doteq \frac{\overline{y}_{st} - R\overline{x}_{st}}{\overline{x}},$$

y por ello $E\left(\widehat{R}_C\right) \doteq R$)

$$V(t_C) \doteq V(\overline{y}_{st}) + R^2 V(\overline{x}_{st}) - 2R\,\mathrm{Cov}(\overline{y}_{st}, \overline{x}_{st}) =$$

$$= \sum_{h=1}^{L} P_h^2 \frac{N_h - n_h}{N_h n_n}\left(S_{hy}^2 + R^2 S_{hx}^2 - 2RS_{hxy}\right)$$

donde

$$S_{hy}^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}(y_{hi} - \overline{y}_h)^2,$$

$$R = \frac{\overline{y}}{\overline{x}},$$

etc.

§5.8 Estimador producto.

Supongamos que entre la variable auxiliar "x" y la de interés "y" existe una relación de proporcionalidad inversa, o bien una estabilidad en los productos $y_k x_k$ $(k = 1, 2, \ldots, N)$. Se propone en estos casos el estimador producto t_P, bajo diseño mas,

$$t_P = \overline{y}_s \frac{\overline{x}_s}{\overline{x}}$$

Para conocer sus características,

$$t_P = \frac{1}{\overline{x}} (\overline{y}_s - \overline{y} + \overline{y}) (\overline{x}_s - \overline{x} + \overline{x}) = \overline{y} \left[1 + \frac{\overline{y}_s - \overline{y}}{\overline{y}}\right] \left[1 + \frac{\overline{x}_s - \overline{x}}{\overline{x}}\right] =$$

$$= \overline{y} \left[1 + \frac{\overline{y}_s - \overline{y}}{\overline{y}} + \frac{\overline{x}_s - \overline{x}}{\overline{x}} + \frac{\overline{y}_s - \overline{y}}{\overline{y}} \frac{\overline{x}_s - \overline{x}}{\overline{x}}\right] \tag{5.1}$$

de donde

$$E(t_P) = \overline{y} + \overline{y} \frac{\mathrm{Cov}(\overline{x}_s, \overline{y}_s)}{\overline{y}\ \overline{x}} = \overline{y} + \frac{N-n}{Nn} \frac{S_{xy}}{\overline{x}}$$

luego el sesgo de t_P, $B(t_P)$, es

$$B(t_P) = E(t_P) - \overline{y} = \frac{N-n}{Nn} \frac{S_{xy}}{\overline{x}}$$

asintóticamente (cuando n crece) nulo.

Para calcular su error cuadrático medio aproximado, de 5.1

$$t_P - \overline{y} = \overline{y} \left[\frac{\overline{y}_s - \overline{y}}{\overline{y}} + \frac{\overline{x}_s - \overline{x}}{\overline{x}} + \frac{\overline{y}_s - \overline{y}}{\overline{y}} \frac{\overline{x}_s - \overline{x}}{\overline{x}}\right],$$

entonces

$$(t_P - \overline{y})^2 \simeq (\overline{y}_s - \overline{y})^2 + 2\frac{\overline{y}}{\overline{x}} (\overline{y}_s - \overline{y}) (\overline{x}_s - \overline{x}) + \frac{\overline{y}^2}{\overline{x}^2} (\overline{x}_s - \overline{x})^2$$

de donde

$$ECM(t_P) = E\left[(t_P - \overline{y})^2\right] \simeq \frac{N-n}{Nn} \left[S_y^2 + 2RS_{xy} + R^2 S_x^2\right]$$

siendo $R = \frac{\overline{y}}{\overline{x}}$. Por tanto t_P mejora a la media muestral $\overline{y}_s$ bajo diseño mas si y sólo si

$$2S_{xy} + RS_x^2 < 0$$

o bien

$$2\rho S_x S_y + RS_x^2 < 0$$

es decir si y sólo si

$$\rho < -\frac{RS_x}{2S_y}$$

es decir cuando el coeficiente de correlación entre las variables "y" y "x" verifica la última condición.

Es inmediato proponer estimadores producto simple y combinado en muestreo estratificado.

Ejercicio 5.1 *Se desea estimar la producción de trigo total en cada cierta comarca. Para ello se toma como unidad de muestreo la parcela dedicada a dicho cultivo, y se conoce como variable auxiliar la superficie de terreno de las parcelas individualmente. Si se supone que la producción de trigo es proporcional a la superficie sembrada en cada parcela o unidad, justificar que el estimador de razón para la producción total es un apropiado estimador.*

Solución.

Si $y_i = cx_i$ $(i = 1, 2, \ldots, N)$ aproximadamente, siendo N el total de parcelas sembradas en la comarca, y_i la producción de trigo en la unidad i, x_i la superficie sembrada en la unidad i, y "c" la constante de proporcionalidad; tenemos que el estimador del total producido de razón es

$$Nt_R = N\overline{y}_s \frac{\overline{x}}{\overline{x}_s}$$

pero como $\overline{y}_s = c\overline{x}_s$ aproximadamente

$$Nt_R = Nc\overline{x}_s \frac{\overline{x}}{\overline{x}_s} = Nc\overline{x} = N\overline{y}$$

con lo que queda demostrada su adecuación pues

$$c\overline{x} = c\frac{\sum_{i=1}^{N} x_i}{N} = \frac{\sum_{i=1}^{N} cx_i}{N} = \frac{\sum_{i=1}^{N} y_i}{N} = \overline{y},$$

siendo $N\overline{y}$ el total producido.

Ejercicio 5.2 *Determinar el tamaño muestral n para que el estimador de razón $t_R = \overline{y}_s \frac{\overline{x}}{\overline{x}_s}$ de la media poblacional $\overline{y}$ de cierta variable de interés "y", difiera de tal función paramétrica menos que $e = 5$ al nivel de confianza del 95% $(1 - \alpha = 0.95)$. Además $N = 1000$, y de una muestra piloto se estima que $S_y^2 = 30$, $R = 2$, $S_x^2 = 15$ y $S_{xy} = 3$.*

Solución.

$$n = \frac{1}{\frac{1}{N} + \frac{\alpha e^2}{S_y^2 + R^2 S_x^2 - 2RS_{xy}}} \doteq 59.$$

Ejercicio 5.3 *Para estimar el consumo medio de las familias de un país se ha utilizado el estimador de razón con la variable auxiliar "renta familiar". Indicar la conveniencia o no de tal estimador para tal objetivo.*

Respuesta.

Razonando como en el ejercicio 5.1, el estimador de razón será apropiado cuando exista una proporcionalidad entre consumo y renta familiar en tal país. Es decir el estimador de razón es deseable cuando la dependencia entre el consumo y la renta familiar sea aproximadamente lineal (una recta) que además pase por el origen. Si la dependencia es lineal pero no pasa por el origen (tal recta), puede utilizarse el estimador de regresión razonando de modo análogo a como se hará en el ejercicio 6.1.

Ejercicio 5.4 *La experiencia de ciertos directivos de unos grandes almacenes les hace admitir que las ventas de cierto producto en un día es inversamente proporcional a su precio de venta al público. En esta situación qué estimador propondría, como asesor de la empresa, para la venta media mensual pudiendo conocer las ventas en 5 días diferentes seleccionados con diseño mas, y sabiendo los precios de venta de los 25 días que abre al público dichos almacenes en ese mes.*

Respuesta.

Llamamos y_i a las ventas del día i (=1, 2, ..., 25), y x_i al precio del producto en ese mismo día; admitimos que aproximadamente $y_i x_i = c$, según nos informan los directivos. Nos interesa estimar la venta media mensual a lo largo de los 25 días laborables del mes. Tal media es denotada por $\overline{y}$, y la media muestral de ventas es $\overline{y}_s$. El precio medio mensual es $\overline{x}$, y la media muestral de precios es $\overline{x}_s$. En estas condiciones y ya que se da una relación aproximada entre y_i y x_i de proporcionalidad inversa, un estimador deseable de $\overline{y}$ es el estimador producto

$$t_P = \overline{y}_s \frac{\overline{x}_s}{\overline{x}}.$$

Capítulo 6

Estimador de regresión.

§6.1 Introducción.

Cuando los puntos (y_i, x_i) con $i = 1, 2, \ldots, N$, donde y_i es la variable de interés y x_i es la variable auxiliar, están situados sobre una línea recta que pasa por el orígen $y_i = ax_i$ el estimador de la razón es el más indicado. Si la relación es lineal del tipo

$$y_i = a + bx_i$$

(línea recta que no pasa por el orígen), entonces es más indicado el estimador de regresión lineal para la media poblacional $\overline{y}$, con diseño mas, pues se dará que

$$\overline{y} = a + b\overline{x}$$

e

$$\overline{y}_s = a + b\overline{x}_s$$

y restando la segunda de la primera igualdad,

$$\overline{y} - \overline{y}_s = b\left(\overline{x} - \overline{x}_s\right)$$

por lo que se propone como estimador

$$\overline{y}_{rg} = \overline{y}_s + b\left(\overline{x} - \overline{x}_s\right),$$

siendo b una variable aleatoria.

Este estimador es sesgado, pues

$$E\left(\overline{y}_{rg}\right) = E\left(\overline{y}_s\right) + \overline{x}E\left(b\right) - E\left(b\overline{x}_s\right) = \overline{y} - \mathrm{Cov}\left(b, \overline{x}_s\right).$$

Luego el sesgo $B(\overline{y}_{rg})$ será

$$B\left(\overline{y}_{rg}\right) = E\left(\overline{y}_{rg}\right) - \overline{y} = -\mathrm{Cov}\left(b, \overline{x}_s\right).$$

§6.2 Varianza mínima del estimador de regresión.

Si b es constante

$$V\left(\overline{y}_{rg}\right) = V\left(\overline{y}_s\right) + b^2 V\left(\overline{x}_s\right) - 2b\,\mathrm{Cov}\left(\overline{y}_s, \overline{x}_s\right) = \frac{N-n}{Nn}\left(S_y^2 + b^2 S_x^2 - 2bS_{yx}\right)$$

pues

$$\mathrm{Cov}\left(\overline{y}_s, \overline{x}_s\right) = \frac{N-n}{Nn} S_{yx}.$$

Llamando $f(b) = V(\overline{y}_{rg})$, b alcanza su mínimo cuando $f'(b) = 0$, o bien

$$\frac{N-n}{Nn}\left(2bS_x^2 - 2S_{yx}\right) = 0$$

de donde despejando b, tenemos

$$b = \frac{S_{yx}}{S_x^2}$$

que es mínimo pues si $n < N$ y $S_x^2 > 0$

$$f''(b) = \frac{N-n}{Nn} 2S_x^2 > 0.$$

Para este valor mínimo de b la varianza toma el valor

$$V_{mín}\left(\overline{y}_{rg}\right) = \frac{N-n}{Nn}\left[S_y^2 + \frac{S_{yx}^2}{S_x^2} - 2\frac{S_{yx}^2}{S_x^2}\right] =$$

$$= \frac{N-n}{Nn}\left[S_y^2 - \frac{S_{yx}^2}{S_x^2}\right] = \frac{N-n}{Nn} S_y^2\left(1-\rho^2\right)$$

siendo $S_{yx}^2 = \rho^2 S_y^2 S_x^2$ (ρ es el coeficiente de correlación de "y" y "x").

En realidad el valor mínimo de b así obtenido es una función paramétrica, por lo que para aplicar estos resultados aproximadamente tendremos que estimar b por

$$\widehat{b} = \frac{\sum_{i\in s}\left(y_i - \overline{y}_s\right)\left(x_i - \overline{x}_s\right)}{\sum_{i\in s}\left(x_i - \overline{x}_s\right)^2}$$

que es el estimador mínimo-cuadrático de $b = \dfrac{S_{yx}}{S_x^2}$ (Cochran, 1977).

§6.3 Comparación de varianzas.

Ya sabemos que en el diseño mas,

$$V(\overline{y}_s) = \frac{N-n}{Nn} S_y^2$$

$$V(t_R) \doteq \frac{N-n}{Nn} \left(S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x \right)$$

$$V\left(\overline{y}_{rg}\right) \doteq \frac{N-n}{Nn} S_y^2 \left(1-\rho^2\right)$$

Por tanto

$$V(\overline{y}_s) \geq V\left(\overline{y}_{rg}\right)$$

correspondiendo el signo igual si y sólo si

$$\rho = \rho_{yx}^2 = 0$$

Además $V(t_R) \geq V(\overline{y}_{rg})$ si y sólo si

$$R^2 S_x^2 - 2R\rho S_y S_x + \rho^2 S_y^2 \geq 0$$

o bien cuando (cosa que se verifica siempre),

$$(RS_x - \rho S_y)^2 \geq 0$$

correspondiendo el signo igual cuando

$$R = \rho \frac{S_y}{S_x}.$$

§6.4 El estimador de regresión en el muestreo estratificado.

El estimador de regresión simple en el muestreo estratificado es

$$\overline{y}_{rgst} = \sum_{h=1}^{L} P_h \overline{y}_{rgh}$$

siendo $\overline{y}_{rgh}$ el estimador de regresión lineal en el estrato h. Su varianza es

$$V\left(\overline{y}_{rgst}\right) = \sum_{h=1}^{L} P_h^2 V\left(\overline{y}_{rgh}\right) = \sum_{h=1}^{L} P_h^2 \frac{N_h - n_h}{N_h n_h} \left(S_{hy}^2 + b_h^2 S_{hx}^2 - 2b_h S_{hyx} \right)$$

siendo b_h es estimador constante de b para el estrato h.

El estimador de regresión lineal combinado es

$$\overline{y}_{rgc} = \overline{y}_{st} + b\left(\overline{x} - \overline{x}_{st}\right)$$

y para b fijo, su varianza es

$$V\left(\overline{y}_{rgc}\right) = V\left(\overline{y}_{st}\right) + b^2 V\left(\overline{x}_{st}\right) - 2b\,\mathrm{Cov}\left(\overline{y}_{st}, \overline{x}_{st}\right) =$$

$$= \sum_{h=1}^{L} \frac{P_h^2\left(N_h - n_h\right)}{N_h n_h}\left(S_{hy}^2 + b^2 S_{hx}^2 - 2bS_{hyx}\right).$$

Ejercicio 6.1 *En una urbanización de N viviendas se dispone de la información auxiliar (x) número de residentes por vivienda. Se sabe además que se verifica que la superficie en metros cuadrados de las viviendas o variable de interés (y) mantiene la relación aproximada: $y_i = a + bx_i$ $(i = 1, 2, \ldots, N)$. Estudiar la conveniencia del estimador de regresión lineal para estimar la superficie media de las viviendas de dicha urbanización.*

Solución.

Como $y_i = a + bx_i$, entonces $\overline{y}_s = a + b\overline{x}_s$ y podemos escribir

$$\overline{y}_{rg} = \overline{y}_s + \frac{\sum_{i\in s}\left(y_i - \overline{y}_s\right)\left(x_i - \overline{x}_s\right)}{\sum_{i\in s}\left(x_i - \overline{x}_s\right)^2}\left(\overline{x} - \overline{x}_s\right) =$$

$$= a + b\overline{x}_s + \frac{\sum_{i\in s}\left(a + bx_i - a - b\overline{x}_s\right)\left(x_i - \overline{x}_s\right)}{\sum_{i\in s}\left(x_i - \overline{x}_s\right)^2}\left(\overline{x} - \overline{x}_s\right) =$$

$$= a + b\overline{x}_s + b\left(\overline{x} - \overline{x}_s\right) = a + b\overline{x} = \overline{y}.$$

Luego es un estimador apropiado si se da la relación aproximada $y_i = a + bx_i$ $(i = 1, 2, \ldots, N)$.

Ejercicio 6.2 *Si se sabe que el coeficiente de correlación $\rho_{yx} = 0.4$, determinar la ganancia en precisión del estimador $\overline{y}_{rg}$ de regresión lineal con respecto a la estrategia $\overline{y}_s$ media muestral, ambas con diseño mas.*

Solución.

$$V\left(\overline{y}_s\right) = \frac{N - n}{Nn} S_y^2$$

$$V\left(\overline{y}_{rg}\right) = \frac{N - n}{Nn} S_y^2\left(1 - \rho^2\right) = 0.84 V\left(\overline{y}_s\right)$$

la ganancia en "precisión" (inversa de la varianza) será

$$\frac{1}{V\left(\overline{y}_{rg}\right)} - \frac{1}{V\left(\overline{y}_s\right)} = \frac{1}{V\left(\overline{y}_s\right)}\left(\frac{1}{0.84} - 1\right) = 0.19\frac{1}{V\left(\overline{y}_s\right)} > 0$$

(ganancia positiva).

Ejercicio 6.3 *Estimar la media poblacional $\overline{y}$ por el método de regresión sabiendo que se dispone de los datos siguientes:*
Medias muestrales $\overline{y}_s = 5$ y $\overline{x}_s = 3$.
Media poblacional de la variable auxiliar: $\overline{x} = 4$
Estimador mínimo cuadrático de b, $\widehat{b} = 2$.

Solución.
El estimador de regresión es

$$\overline{y}_{rg} = \overline{y}_s + \widehat{b}\,(\overline{x} - \overline{x}_s) = 5 + 2(4 - 3) = 7.$$

Capítulo 7

Muestreo sistemático.

§7.1 Conceptos básicos.

En el caso en que las N unidades de la población, constituyan un número divisible por el tamaño muestral n, $\frac{N}{n} = k$. Entonces existirán k muestras no ordenadas o conjuntos de tamaño n, que se seleccionan del siguiente modo:

a) Se selecciona una unidad entre las k primeras unidades de la población, cada unidad con probabilidad $\frac{1}{k}$.

b) Las restantes $n-1$ unidades de la muestra son las que ocupan los lugares relativos idénticos en los $n-1$ restantes grupos de k unidades de la población de tamaño N.

Habrá entonces k muestras posibles:

$$\begin{aligned}(s1) &= \{1, k+1, 2k+1, \ldots, N-k+1\}\\ (s2) &= \{2, k+2, 2k+2, \ldots, N-k+2\}\\ &\vdots\\ (sk) &= \{k, 2k, 3k, \ldots, N\}\end{aligned}$$

cada una con probabilidad de selección igual a $\frac{1}{k} = \frac{n}{N}$.

Ventajas de este método de muestreo:

a) La muestra se extiende a toda la población.

b) Puede recoger el efecto de estratificación debido al orden en que se numeran las unidades de la población.

c) Es de aplicación y comprobación sencillas.

Inconvenientes del muestreo sistemático:

a) En caso de periodicidad de la variable de interés, podría aumentar la varianza de los estimadores medias muestrales.

b) El problema teórico que se presenta en la estimación de varianzas pues no existen estimadores insesgados de las varianzas de las medias muestrales.

§7.2 Características de la distribución en el muestreo.

Si se selecciona la muestra (si) con probabilidad $\frac{1}{k} = \frac{n}{N}$, tendremos como estimador media muestral

$$\overline{y}_{si} = \frac{1}{n}\sum_{j=1}^{n} y_{i+k(j-1)} \quad (i = 1, 2, \ldots, k)$$

que es insesgado para $\overline{y}$, pues

$$E(\overline{y}_{si}) = \frac{1}{k}\sum_{i=1}^{k}\overline{y}_{si} = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n}\sum_{j=1}^{n} y_{i+k(j-1)} = \frac{N}{N}\overline{y} = \overline{y}$$

de donde como k es el número de muestras posibles, su varianza será

$$V(\overline{y}_{si}) = \frac{1}{k}\sum_{i=1}^{k}(\overline{y}_{si} - \overline{y})^2$$

o bien empleando algunos artificios,

$$\sum_{j=1}^{n} V(\overline{y}_{si}) = nV(\overline{y}_{si}) = \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{n}(\overline{y}_{si} - \overline{y})^2 \tag{7.1}$$

de donde, al ser $N = nk$,

$$NV(\overline{y}_{si}) = nkV(\overline{y}_{si}) \overset{(7.1)}{=} \sum_{i=1}^{k}\sum_{j=1}^{n}(\overline{y}_{si} - \overline{y})^2 \overset{(*)}{=}$$

$$\overset{(*)}{=} N\sigma^2 - \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{i+k(j-1)} - \overline{y}_{si}\right)^2$$

usando en $(*)$ el análisis de la varianza. Por tanto,

$$V(\overline{y}_{si}) = \sigma^2 - \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{i+k(j-1)} - \overline{y}_{si}\right)^2 = \sigma^2 - \frac{n-1}{n}S_{ws}^2 \tag{7.2}$$

donde

$$S_{ws}^2 = \frac{n}{n-1}\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{i+k(j-1)} - \overline{y}_{si}\right)^2$$

es una suma de desviaciones dentro, con otra notación. La fórmula (7.2) nos da la varianza de la media muestral en muestreo sistemático.

También podemos considerar que el muestreo sistemático es un muestreo por conglomerados de igual tamaño unictápico o sin submuestreo de un solo conglomerado, y entonces aplicando los resultados del muestreo por conglomerado (ver capítulo 10), tenemos que $k = L$, $n = \overline{N}$, $1 = n$ donde a la izquierda de cada igualdad situamos los

parámetros del muestreo sistemático y a la derecha los correspondientes del muestreo por conglomerados unitápico. Luego,

$$V(\overline{y}_{si}) \doteq \frac{N-n}{N-1}\frac{\sigma^2}{n} + \frac{N-n}{N-1}\frac{\sigma^2}{n}(n-1)\delta = \frac{N-n}{N-1}\frac{\sigma^2}{n}[1+(n-1)\delta],$$

siendo δ el coeficiente de correlación intraconglomerados que se verá en el capítulo 10. De la teoría del muestreo por conglomerados unietápico, el estimador $\overline{y}_{si}$ es insesgado, como ya hemos visto en el presente capítulo.

§7.3 Comparación con el diseño mas.

Si

$$S_{ws}^2 > \frac{N}{N-1}\sigma^2 = S^2,$$

el muestreo sistemático es más preciso que la estrategia (mas, $\overline{y}_s$), puesto que

$$V(\overline{y}_{si}) = \frac{N-1}{N}S^2 - \frac{n-1}{n}S_{ws}^2 <$$

$$< \frac{N-1}{N}S^2 - \frac{n-1}{n}S^2 = \frac{N-n}{Nn}S^2 = V(\text{mas}, \overline{y}_s).$$

Si $S_{ws}^2 < S^2$, entonces $V(\overline{y}_{si}) > V(\text{mas}, \overline{y}_s)$.
Si $S_{ws}^2 = S^2$, entonces $V(\overline{y}_{si}) = V(\text{mas}, \overline{y}_s)$.

También pueden hacerse estas comparaciones en función del valor de δ (si $\delta > 0$, $V(\overline{y}_{si}) > V(\text{mas}, \overline{y}_s)$; si $\delta < 0$, $V(\overline{y}_{si}) < V(\text{mas}, \overline{y}_s)$; si $\delta = 0$, $V(\overline{y}_{si}) = V(\text{mas}, \overline{y}_s)$).

Capítulo 8

Muestreo en ocasiones sucesivas.

§8.1 Conceptos básicos.

En algunos casos es necesario estimar el cambio que experimenta la media poblacional entre dos ocasiones sucesivas, designadas por los instantes t_1 y t_2, con una muestra de tamaño n. A este cambio entre dos ocasiones le denotamos $\delta = \mu_2 - \mu_1$, siendo μ_i la media poblacional en el instante t_i $(i = 1, 2)$.

Utilizando la media muestral en ambas ocasiones, el estimador simple del cambio será el estimador insesgado para δ,

$$\widehat{\delta} = \overline{y}_2 - \overline{y}_1 = \frac{1}{n}\sum_{i=1}^{n}(y_{2i} - y_{1i}).$$

En estas condiciones, se puede optar por una alternativa entre:

a) Utilizar la misma muestra, denominada "panel", en ambas ocasiones.

b) Conservar en la segunda ocasión c unidades de la primera muestra, eliminar $n - c$ y añadir $n - c$ nuevas unidades.

c) Utilizar en la segunda ocasión una muestra independiente de la primera.

La alternativa a) nos permitiría conocer los cambios individuales entre las dos ocasiones. Será una vía imposible si las mediciones fueran destructivas. Pero en los casos en que fuese posible, no sería deseable por los sesgos que una exposición continuada a los métodos de encuesta pueden originar en la conducta de los entrevistados. Se suele decir en estos casos que la muestra se "contamina" con el tiempo.

§8.2 Muestreo en dos ocasiones.

Tendremos que

$$\overline{y}_1 = \frac{n-c}{n}\overline{y}_{1\bar{c}} + \frac{c}{n}\overline{y}_{1c}$$

e

$$\overline{y}_2 = \frac{c}{n}\overline{y}_{2c} + \frac{n-c}{n}\overline{y}_{2\overline{c}},$$

siendo $\overline{c} = n - c$. Así, supuesto que de una ocasión a otra no varíe la cuasivarianza poblacional S^2,

$$V(\overline{y}_1) = \frac{N-n}{N}\frac{S^2}{n},$$

$$V(\overline{y}_2) = \frac{N-n}{N}\frac{S^2}{n},$$

y

$$\mathrm{Cov}(\overline{y}_1, \overline{y}_2) = \frac{c^2}{n^2}\mathrm{Cov}(\overline{y}_{1c}, \overline{y}_{2c}) = \frac{c^2}{n^2}\rho_{12}\sqrt{\frac{N-c}{N}}\frac{S}{\sqrt{c}}\sqrt{\frac{N-c}{N}}\frac{S}{\sqrt{c}} =$$

$$= \frac{N-c}{N}\rho_{12}\frac{c}{n^2}S^2 = \frac{N-c}{N}\rho_{12}\frac{S^2}{n}\pi_c,$$

y sustituyendo estos valores en la varianza de δ tenemos

$$V(\hat{\delta}) = V(\overline{y}_1) + V(\overline{y}_2) - 2\,\mathrm{Cov}(\overline{y}_1, \overline{y}_2) =$$

$$= \frac{N-n}{N}\left[\frac{S^2}{n} + \frac{S^2}{n}\right] - 2\frac{N-c}{N}\frac{S^2}{n}\rho_{12}\pi_c =$$

$$= \frac{2S^2}{n}\left[\frac{N-n}{N} - \frac{N-c}{N}\rho_{12}\pi_c\right]$$

donde ρ_{12} es el coeficiente de correlación entre los valores comunes a ambas ocasiones y π_c la proporción de unidades comunes en las dos muestras.

Si lo que queremos estimar no es el cambio entre dos ocasiones, δ, sino la media sobre dos ocasiones, consideremos el estimador de la media, $\mu = \frac{1}{2}(\mu_1 + \mu_2)$, a

$$\overline{y} = \frac{1}{2}(\overline{y}_1 + \overline{y}_2)$$

media de las medias en ambas ocasiones; su varianza es

$$V(\overline{y}) = \frac{1}{4}[V(\overline{y}_1) + V(\overline{y}_2) - 2\,\mathrm{Cov}(\overline{y}_1, \overline{y}_2)] =$$

$$= \frac{1}{4}\left\{\left[\frac{N-n}{N}2\frac{S^2}{n}\right] + \frac{N-c}{N}\frac{2S^2}{n}\rho_{12}\pi_c\right\} =$$

$$= \frac{S^2}{2n}\left[\frac{N-n}{N} + \frac{N-c}{N}\rho_{12}\pi_c\right]$$

y éste valor es mínimo cuando $\pi_c = 0$, si $\rho_{12} > 0$ como es habitual en la práctica.

Ejercicio 8.1 *Estimar el "cambio" de las ventas por empresa expendedora de cierto producto de primera necesidad en dos instantes de tiempo* 0 *y* 1*, así como la media sobre las dos ocasiones, si* $\overline{x}_1 = 50$ *y* $\overline{x}_2 = 65$*, dando la varianza de ambos estimadores. Explicar cuales pueden ser las unidades de la población.*

Solución.

a) Estimación del "cambio": $\delta = \overline{x}_2 - \overline{x}_1 = 15$

$$V\left(\widehat{\delta}\right) = \frac{2S^2}{n}\left[\frac{N-n}{N} - \frac{N-c}{N}\rho_{12}\pi_c\right].$$

b) Estimación de la "media sobre dos ocasiones": $\overline{x} = \frac{1}{2}\left(\overline{x}_1 + \overline{x}_2\right) = 57.5$

$$V\left(\overline{x}\right) = \frac{S^2}{2n}\left[\frac{N-n}{N} + \frac{N-c}{N}\rho_{12}\pi_c\right].$$

c) Las unidades de la población o de muestreo serán las empresas expendedoras.

Ejercicio 8.2 *Admitiendo que en dos instantes de tiempo consecutivos se presentan dos cuasivarianzas diferentes* S_1^2 *y* S_2^2*, calcular las varianzas* $V\left(\overline{y}_1\right)$ *y* $V\left(\overline{y}_2\right)$*, y la covarianza* $\mathrm{Cov}\left(\overline{y}_1, \overline{y}_2\right)$.

Solución.

$$V\left(\overline{y}_1\right) = \frac{N-n}{N}\frac{S_1^2}{n},$$

$$V\left(\overline{y}_2\right) = \frac{N-n}{N}\frac{S_2^2}{n}$$

y

$$\mathrm{Cov}\left(\overline{y}_1, \overline{y}_2\right) = \frac{c^2}{n^2}\rho_{12}\frac{N-c}{N}\frac{\sqrt{S_1^2 S_2^2}}{c}$$

Capítulo 9

Muestreo con probabilidades desiguales.

§9.1 Muestreo con probabilidades proporcionales al tamaño con reemplazamiento (pptr).

Es aquel diseño ordenado cuya función $p(\mathbf{s})$ asigna la probabilidad p_k de seleccionar la unidad k en cada una de las n extracciones independientes, donde el resultado de la i–ésima extracción es el i–ésimo componente de la muestra o vector $\mathbf{s}$. Los valores p_k son números positivos conocidos tales que

$$\sum_{k\in U} p_k = 1$$

Así el diseño pptr es un diseño de tamaño fijo n pero no de tamaño efectivo fijo. O sea $n(\mathbf{s}) = n$ pero $\nu(\mathbf{s})$ no es constante, ya que se realizan n extracciones con reemplazamiento en una urna cuya composición es de n_k proporcional a p_k ($n_k \alpha p_k$, siendo n_k el número de bolas con la anotación k). De este modo,

$$M = \sum_{k\in U} n_k \alpha \sum_{k\in U} p_k = 1$$

siendo M la constante de proporcionalidad ($n_k = Mp_k$, $k = 1, 2, \ldots, N$). Observar que en el caso particular en que $p_k = \dfrac{1}{N}$ $(k = 1, 2, \ldots, N)$ tenemos el diseño masr. Otro caso particular de importancia de éste diseño es aquel en que $p_k \alpha x_k$, es decir la probabilidad de selección de la unidad k es proporcional al valor positivo ($x_k > 0$) de la variable auxiliar en la unidad k, x_k. Si y_k es la producción de trigo en la parcela k, x_k podría ser la superficie de la unidad k que es conocida antes de observar y_k. En este caso,

$$p_k = \frac{x_k}{\sum_{i\in U} x_i} = \frac{x_k}{N\overline{x}} \quad (k = 1, 2, \ldots, N)$$

donde

$$\overline{x} = \sum_{i \in U} \frac{x_i}{N}.$$

También, si x_k es un número entero positivo (para todo $k \in U$), $n_k = x_k$ puede ser la composición de la urna para seleccionar una muestra s bajo diseño pptr.

Las probabilidades de inclusión son:

$$\pi_k = 1 - (1 - p_k)^n$$

y

$$\pi_{km} = 1 - (1 - p_k)^n - (1 - p_m)^n + (1 - p_k - p_m)^n$$

etc. generalizando el caso de diseño masr.

9.1.1 Estimador insesgado de $\overline{y}$ bajo diseño pptr.

El estimador más importante asociado con éste diseño es el estimador de Hansen-Hurwitz (1943) definido así

$$t_{HH} = \sum_{k \in \mathbf{S}} \frac{y_k}{Nnp_k} = \sum_{k \in U} \frac{y_k e_k}{Nnp_k}$$

que es insesgado para estimar la media poblacional

$$\overline{y} = \frac{1}{N} \sum_{k \in U} y_k$$

En efecto, en una muestra ordenada s la unidad k puede pertenecer a s un número de veces e_k $(= 0, 1, 2, \ldots, n)$ si p_k es la constante probabilidad de selección de la unidad k en cada selección o componente de la muestra ordenada s. El modelo que se ha creado, al ser independientes las diversas selecciones que obtienen las unidades ordenadas de la muestra es el de una distribución multinomial, y por ello tenemos que (siendo $(e_1, e_2, \ldots, e_k, \ldots, e_N)$ una variable aleatoria multinomial),

$$E(e_k) = np_k \quad ; \quad V(e_k) = np_k(1 - p_k)$$

y si $k \neq m$,

$$\mathrm{Cov}(e_k, e_m) = -np_k p_m$$

y por tanto, si $k \neq m$

$$E(e_k e_m) = \mathrm{Cov}(e_k, e_m) + E(e_k) E(e_m) =$$

$$= -np_k p_m + n^2 p_k p_m = \left(n^2 - n\right) p_k p_m.$$

Así,

$$E\left(t_{HH}\right)=E\left(\sum_{k\in U}\frac{y_k e_k}{Nnp_k}\right)=\sum_{k\in U}\frac{y_k}{Nnp_k}E\left(e_k\right)=\sum_{k\in U}\frac{y_k}{Nnp_k}np_k=\overline{y}$$

luego t_{HH} es insesgado para $\overline{y}$.

9.1.2 Varianza del estimador Hansen–Hurwitz.

$$V\left(t_{HH}\right)=E\left[\left(t_{HH}-\overline{y}\right)^2\right]=$$

$$=E\left[\left(\sum_{k\in U}\frac{y_k e_k}{Nnp_k}-\overline{y}\right)^2\right]=E\left(\sum_{k\in U}\frac{y_k e_k}{Nnp_k}\right)^2-\overline{y}^2=$$

$$=E\left(\sum_{k\in U}\frac{y_k^2 e_k^2}{N^2n^2p_k^2}+\sum_{k\neq m\in U}\frac{y_k y_m e_k e_m}{N^2n^2p_k p_m}\right)-\overline{y}^2=$$

$$=\frac{1}{N^2n^2}\left[\sum_{k\in U}\frac{y_k^2}{p_k^2}E\left(e_k^2\right)+\sum_{k\neq m\in U}\frac{y_k y_m}{p_k p_m}E\left(e_k e_m\right)\right]-\overline{y}^2, \tag{9.1}$$

pero como

$$E\left(e_k^2\right)=V\left(e_k\right)+\left[E\left(e_k\right)\right]^2=np_k\left(1-p_k\right)+n^2p_k^2=np_k-np_k^2+n^2p_k^2$$

de aquí, la fórmula (9.1) se obtiene sustituyendo

$$V\left(t_{HH}\right)=\frac{1}{N^2n^2}\left[\sum_{k\in U}\frac{y_k^2}{p_k^2}\left(np_k-np_k^2+n^2p_k^2\right)+\right.$$

$$\left.+\sum_{k\neq m\in U}\frac{y_k y_m}{p_k p_m}\left(n^2p_k p_m-np_k p_m\right)\right]-\overline{y}^2=$$

$$=\frac{1}{N^2n^2}\left[\sum_{k\in U}\frac{y_k^2}{p_k^2}np_k-\sum_{k\in U}\frac{y_k^2}{p_k^2}np_k^2+\sum_{k\in U}\frac{y_k^2}{p_k^2}n^2p_k^2+\right.$$

$$\left.+\sum_{k\neq m\in U}\frac{y_k y_m}{p_k p_m}n^2p_k p_m-\sum_{k\neq m\in U}\frac{y_k y_m}{p_k p_m}np_k p_m\right]-\overline{y}^2=$$

$$=\frac{1}{N^2}\left[\underbrace{\frac{1}{n}\sum_{k\in U}\frac{y_k^2}{p_k^2}p_k}_{(2)}-\underbrace{\frac{1}{n}\sum_{k\in U}y_k^2}_{(3)}+\underbrace{\sum_{k\in U}y_k^2}_{(1)}+\underbrace{\sum_{k\neq m\in U}y_k y_m}_{(1)}-\underbrace{\frac{1}{n}\sum_{k\neq m\in U}y_k y_m}_{(3)}\right]-\overline{y}^2=$$

$$=\frac{1}{N^2}\left[N^2\overline{y}^2+\frac{1}{n}\sum_{k\in U}\frac{y_k^2}{p_k^2}p_k-\frac{N^2\overline{y}^2}{n}\right]-\overline{y}^2=$$

$$= \frac{1}{N^2}\left[\frac{1}{n}\sum_{k\in U}\frac{y_k^2}{p_k^2}p_k - \frac{N^2\overline{y}^2}{n}\right] = \frac{1}{N^2 n}\left[\sum_{k\in U}\frac{y_k^2}{p_k^2}p_k - N^2\overline{y}^2\right], \tag{9.2}$$

donde usando la relación $\sigma^2 = \alpha_2 - \alpha_1^2$ tenemos continuando (9.2) que

$$V(t_{HH}) = \frac{1}{N^2 n}\sum_{k\in U}\left(\frac{y_k}{p_k} - N\overline{y}\right)^2 p_k \tag{9.3}$$

que en el caso particular de diseño masr ($p_k = \frac{1}{N}$) obtenemos

$$V(t_{HH}) = V(\overline{y}_S) = \frac{\sigma^2}{n}.$$

9.1.3 Estimador insesgado de $V(t_{HH})$.

Es el siguiente

$$\widehat{V}(t_{HH}) = \frac{\sum_{i\in S}\left(\frac{y_i}{p_i} - Nt_{HH}\right)^2}{N^2 n(n-1)},$$

o bien

$$\widehat{V}(t_{HH}) = \frac{\sum_{i\in S}\frac{y_i^2}{p_i^2} - nN^2 t_{HH}^2}{N^2 n(n-1)}$$

puesto que (al ser $\sum_{i\in S}\frac{y_i}{p_i} = nNt_{HH}$),

$$\sum_{i\in S}\left(\frac{y_i}{p_i} - Nt_{HH}\right)^2 = \sum_{i\in S}\frac{y_i^2}{p_i^2} - 2\sum_{i\in S}\frac{y_i}{p_i}Nt_{HH} + nN^2 t_{HH}^2 =$$

$$= \sum_{i\in S}\frac{y_i^2}{p_i^2} - 2nN^2 t_{HH}^2 + nN^2 t_{HH}^2 = \sum_{i\in S}\frac{y_i^2}{p_i^2} - nN^2 t_{HH}^2,$$

y como esta varianza es invariante por cambio de orígen,

$$\widehat{V}(t_{HH}) = \frac{\sum_{i\in S}\left(\frac{y_i}{p_i} - \overline{y}\right)^2 - n(Nt_{HH} - N\overline{y})^2}{N^2 n(n-1)}$$

o también desarrollando el numerador tenemos

$$\sum_{i\in S}\left(\frac{y_i}{p_i} - \overline{y}\right)^2 - n(Nt_{HH} - N\overline{y})^2 =$$

$$= \sum_{i\in S}\frac{y_i^2}{p_i^2} - 2N\overline{y}\sum_{i\in S}\frac{y_i}{p_i} + nN^2\overline{y}^2 - nN^2 t_{HH}^2 - nN^2\overline{y}^2 + 2nN^2\overline{y}t_{HH} =$$

$$= \sum_{i \in \mathbf{S}} \frac{y_i^2}{p_i^2} - nN^2 t_{HH}^2.$$

De este modo queda,

$$\widehat{V}(t_{HH}) = \frac{1}{N^2 n(n-1)} \left[\sum_{i \in U} \left(\frac{y_i}{p_i} - N\overline{y} \right)^2 e_i - n \left(N t_{HH} - N\overline{y} \right)^2 \right].$$

Luego

$$E\left[\widehat{V}(t_{HH})\right] = \frac{1}{N^2 n(n-1)} \left[\sum_{i \in U} \left(\frac{y_i}{p_i} - N\overline{y} \right)^2 E(e_i) - nN^2 E\left(t_{HH} - \overline{y}\right)^2 \right] =$$

$$= \frac{1}{N^2 n(n-1)} \left[\sum_{i \in U} \left(\frac{y_i}{p_i} - N\overline{y} \right)^2 np_i - nN^2 V(t_{HH}) \right] = \text{ (por (9.3))}$$

$$= \frac{1}{N^2(n-1)} \left[N^2 n V(t_{HH}) - N^2 V(t_{HH}) \right] = \frac{1}{n-1}(n-1)V(t_{HH}) = V(t_{HH}).$$

Entonces, $\widehat{V}(t_{HH})$ es un estimador insesgado de $V(t_{HH})$.

§9.2 Muestreo con probabilidades proporcionales al tamaño sin reemplazamiento (ppt).

Denota al diseño $p(\mathbf{s})$ que en n extracciones sucesivas sin reemplazamiento, la j–ésima extracción se realiza con probabilidades p_k entre las restantes unidades no extraídas previamente, para $j = 1, 2, \ldots, n \leq N$. Esto es, en la primera extracción se selecciona la unidad k_1 con probabilidad p_{k_1}. Y en la j–ésima extracción ($j > 1$) seleccionamos la unidad k_j con probabilidad

$$\frac{p_{k_j}}{1 - \sum_{i=1}^{j-1} p_{k_i}}$$

supuesto conocidas las unidades $k_1, k_2, \ldots, k_{j-1}$ obtenidas en las $j-1$ primeras extracciones y $k_j \neq k_i$ ($i = 1, 2, \ldots, j-1$). El diseño ppt es de tamaño fijo n y de tamaño efectivo fijo n.

Un caso particular de diseño ppt en el que $p_k \alpha x_k$ (es decir, p_k es proporcional a la variable auxiliar positiva x en la unidad k), verifica que en la primera extracción $p_k = \dfrac{x_k}{N\overline{x}}$ para todo $k \in U$, y en la j–ésima extracción se selecciona la unidad $k_j \neq k_1, \ldots, k_{j-1}$ (distinta de las ya obtenidas en las primeras $j-1$ etapas) con probabilidad

$$\frac{x_{k_j}}{\sum_{i=j}^{N} x_{k_i}}.$$

Un estimador propuesto para este diseño, que es insesgado para $\overline{y}$, es el estimador de Raj (1956) definido así

$$t_R = \frac{\sum_{i=1}^{n} t_i}{nN}$$

donde

$$t_1 = \frac{y_{k_1}}{p_{k_1}}$$

y para todo $i = 2, 3, \ldots, n$,

$$t_i = \sum_{j=1}^{i-1} y_{k_j} + \frac{\left(1 - \sum_{j=1}^{i-1} p_{k_j}\right) y_{k_i}}{p_{k_i}}$$

Este estimador fue mejorado en precisión por Murthy (1967) modificando el estimador y conservando el mismo diseño.

§9.3 Muestreo con probabilidades de inclusión proporcionales al tamaño (pipt).

Este diseño denota al muestreo de tamaño efectivo fijo n, $p(s)$, tal que las probabilidades de inclusión son $\pi_k = np_k$ $(k = 1, 2, \ldots, N)$ donde

$$p_k = \frac{x_k}{N\overline{x}} \quad (k = 1, 2, \ldots, N)$$

siendo x una variable auxiliar positiva, con la condición $nx_k \leq N\overline{x}$ $(k = 1, 2, \ldots, N)$. Dados los valores x_1, x_2, ..., x_N de la variable auxiliar, puede existir más de un diseño pipt.

9.3.1 Estimador insesgado.

En estas condiciones, un estimador insesgado de la media poblacional $\overline{y}$, viene dado por el estimador Horvitz-Thompson (1952) que se define así:

$$t_{HT} = \sum_{k \in s} \frac{y_k}{N\pi_k} = \sum_{k \in s} \frac{y_k e_k}{N\pi_k}$$

denotando e_k a la variable aleatoria auxiliar tal que

$$e_k = \begin{cases} 1 & \text{si} \quad k \in s \\ 0 & \text{si} \quad k \notin s \end{cases}$$

es decir e_k es una variable aleatoria distribuída como una binomial de parámetros $n = 1$ (prueba) y π_k como probabilidad de éxito, $e_k \equiv \mathrm{B}(1, \pi_k)$,

$$E(e_k) = 1 \cdot p(k \in s) + 0 \cdot p(k \notin s) = 1 \cdot \pi_k = \pi_k$$

$$V(e_k) = E\left(e_k^2\right) - [E(e_k)]^2 = E(e_k) - [E(e_k)]^2 = \pi_k - \pi_k^2 = \pi_k(1-\pi_k).$$

También $e_i e_j$ se distribuye como una variable aleatoria binomial de parámetros 1 y π_{ij}, $e_i e_j \equiv \mathrm{B}(1, \pi_{ij})$. De este modo,

$$e_i e_j = \begin{cases} 1 & \text{si } i,j \in s \\ 0 & \text{en caso contrario,} \end{cases}$$

$E\left(e_i e_j\right) = 1 \cdot p(i,j \in s) + 0 \cdot p(i \text{ ó } j \notin s) = 1 \cdot \pi_{ij} = \pi_{ij}$
$\mathrm{Cov}\left(e_i, e_j\right) = E\left(e_i e_j\right) - E(e_i)\,E\left(e_j\right) = \pi_{ij} - \pi_i \pi_j,$
con lo que

$$E(t_{HT}) = E\left(\sum_{i\in U} \frac{y_i e_i}{N\pi_i}\right) = \sum_{i\in U} \frac{y_i}{N\pi_i} E(e_i) = \sum_{i\in U} \frac{y_i}{N\pi_i}\pi_i = \frac{1}{N}\sum_{i\in U} y_i = \overline{y}.$$

9.3.2 Varianza del estimador Horvitz–Thompson.

$$V(t_{HT}) = \frac{1}{N^2} V\left(\sum_{i\in U} \frac{y_i e_i}{\pi_i}\right) =$$

$$= \frac{1}{N^2}\left[\sum_{i\in U} V\left(\frac{y_i e_i}{\pi_i}\right) + \sum_{i\neq j\in U} \mathrm{Cov}\left(\frac{y_i e_i}{\pi_i}, \frac{y_j e_j}{\pi_j}\right)\right] =$$

$$= \frac{1}{N^2}\left[\sum_{i\in U} \frac{y_i^2}{\pi_i^2} V(e_i) + \sum_{i\neq j\in U} \frac{y_i y_j}{\pi_i \pi_j} \mathrm{Cov}\left(e_i, e_j\right)\right] =$$

$$= \frac{1}{N^2}\left[\sum_{i\in U} \frac{y_i^2}{\pi_i^2}\pi_i(1-\pi_i) + \sum_{i\neq j\in U} \frac{y_i y_j}{\pi_i \pi_j}\left(\pi_{ij} - \pi_i \pi_j\right)\right].$$

9.3.3 Estimador insesgado de la varianza.

Si $\pi_{ij} > 0$ para todos los pares $i \neq j \in U$,

$$\widehat{V}(t_{HT}) = \frac{1}{N^2}\left[\sum_{i\in s} \frac{y_i^2}{\pi_i^2}(1-\pi_i) + \sum_{i\neq j\in s} \frac{y_i y_j}{\pi_i \pi_j}\frac{\left(\pi_{ij} - \pi_i \pi_j\right)}{\pi_{ij}}\right] =$$

$$= \frac{1}{N^2}\left[\sum_{i\in U} \frac{y_i^2}{\pi_i^2}(1-\pi_i)\, e_i + \sum_{i\neq j\in U} \frac{y_i y_j}{\pi_i \pi_j}\frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} e_i e_j\right].$$

Como $E(e_i) = \pi_i$, y si $i \neq j$ $E(e_i e_j) = \pi_{ij}$, tenemos que

$$E\left[\widehat{V}(t_{HT})\right] = \frac{1}{N^2}\left[\sum_{i\in U} \frac{y_i^2}{\pi_i^2}(1-\pi_i)\, E(e_i) +\right.$$

$$+ \sum_{i \neq j \in U} \frac{y_i y_j}{\pi_i \pi_j} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} E(e_i e_j) \Bigg] = V(t_{HT}).$$

Otro estimador de la varianza $V(t_{HT})$ es el estimador de Yates y Grundy (1953),

$$\widehat{V}_{YG}(t_{HT}) = \frac{1}{N^2} \left[\sum_{i<j \in s} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \right].$$

Ambos estimadores propuestos de la varianza del estimador Horvitz–Thompson son insesgados, pero pueden tomar valores negativos.

Ejercicio 9.1 *Se selecciona una muestra ordenada de tamaño fijo igual a 4 con diseño pptr. La muestra es* $\mathbf{s} = (4, 3, 5, 7)$ *y los valores de la variable de interés observada son respectivamente* $(33, 21, 15, 9)$. *Estimar la media poblacional* $\overline{y}$ *con el estimador de Hansen-Hurwitz para* $N = 20$, $p_4 = p_3 = \frac{1}{20}$ *y* $p_5 = p_7 = \frac{1}{40}$, *así como estimar la varianza de dicho estimador (de modo insesgado) con la misma muestra seleccionada* $\mathbf{s}$.

Solución.

$$t_{HH} = \sum_{k \in \mathbf{S}} \frac{y_k}{N p_k} = 25.5$$

$$\widehat{V}(t_{HH}) = \frac{\sum_{i \in \mathbf{S}} \left(\frac{y_i}{p_i} - N t_{HH} \right)^2}{N^2 n(n-1)} = 12.75.$$

Ejercicio 9.2 *Se obtiene una muestra con diseño pipt de tamaño fijo* $n = 2$. *Esta muestra resulta ser* $s = \{2, 1\}$ *y los valores observados son* $y_2 = 4$ *e* $y_1 = 8$. *Si* $\pi_1 = \frac{1}{3}$, $\pi_2 = \frac{2}{3}$ *y* $\pi_{12} = \frac{2}{9}$, *estimar la media poblacional* $\overline{y}$ *y dar una estimación de la varianza del estimador de* $\overline{y}$ *(ambas insesgadas), si* $N = 4$.

Solución.

$$t_{HT} = \sum_{k \in s} \frac{y_k}{N \pi_k} = 7.5$$

$$\widehat{V}(t_{HT}) = \frac{1}{N^2} \left[\sum_{i \in s} \frac{y_i^2}{\pi_i^2} (1 - \pi_i) + \sum_{i \neq j \in s} \frac{y_i y_j}{\pi_i \pi_j} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \right] = 24.75$$

Capítulo 10

Muestreo por conglomerados.

§10.1 Introducción.

Un conglomerado es una clase o parte de una clasificación de la población finita. Así, si tenemos L conglomerados, cada uno de ellos contiene varias unidades elementales de la población. Si el conglomerado i $(1 \leq i \leq L)$ contiene N_i unidades llamadas secundarias, el número total de unidades secundarias o elementos de la población será

$$N = \sum_{i=1}^{L} N_i.$$

En el muestreo unietápico por conglomerados (o muestreo por conglomerados sin submuestreo), se seleccionan n conglomerados de entre los L que constituyen el colectivo, y dentro de cada uno de estos n conglomerados se observan todas las unidades secundarias que contienen. Así los conglomerados son las unidades de muestreo y las unidades secundarias son las unidades de observación ya que es de las unidades secundarias de donde se obtiene la información.

La variable de interés u observable es y_{ij} para $i = 1, 2, \ldots, L$; $j = 1, 2, \ldots, N_i$.

La media del conglomerado i será, $(1 \leq i \leq L)$

$$\overline{y}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij} = \mu_i;$$

el total del conglomerado i es

$$N_i \overline{y}_i = \sum_{j=1}^{N_i} y_{ij} = N_i \mu_i.$$

La media poblacional es

$$\overline{y} = \frac{1}{N} \sum_{i=1}^{L} \sum_{j=1}^{N_i} y_{ij} = \frac{1}{N} \sum_{i=1}^{L} N_i \overline{y}_i = \frac{1}{N} \sum_{i=1}^{L} N_i \mu_i = \mu.$$

La cuasivarianza del conglomerado i es

$$S_i^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} \left(y_{ij} - \overline{y}_i\right)^2 = \frac{N_i}{N_i - 1}\sigma_i^2$$

y la cuasivarianza poblacional es

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{L} \sum_{j=1}^{N_i} \left(y_{ij} - \overline{y}\right)^2 = \frac{N}{N-1}\sigma^2.$$

En estas condiciones el análisis de la varianza o variación total se puede descomponer en una variación dentro de conglomerados y una variación entre conglomerados de modo análogo al muestreo estratificado,

$$\sigma^2 = \sum_{i=1}^{L} \frac{N_i}{N}\sigma_i^2 + \sum_{i=1}^{L} \frac{N_i}{N} \left(\mu_i - \mu\right)^2.$$

El coeficiente de correlación intraconglomerados es

$$\delta = \frac{\displaystyle\sum_{i=1}^{L} \frac{\displaystyle\sum_{j \neq k}^{N_i} \left(y_{ij} - \overline{y}\right)\left(y_{ik} - \overline{y}\right)}{LN_i(N_i - 1)}}{\dfrac{\displaystyle\sum_{i=1}^{L}\sum_{j=1}^{N_i} \left(y_{ij} - \overline{y}\right)^2}{N}}$$

que es un indicador del grado de homogeneidad de los conglomerados, donde en el denominador aparece la varianza poblacional σ^2.

10.1.1 Estimador de la media con conglomerados del mismo tamaño.

En el caso en que los conglomerados sean del mismo tamaño ($N_i = \overline{N}$, $i = 1, 2, \ldots, L$) la media muestral sería

$$\overline{y}_{cs} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\overline{N}} \sum_{j=1}^{\overline{N}} y_{ij} = \frac{1}{n} \sum_{i=1}^{n} \overline{y}_i = \frac{1}{n} \sum_{i \in s} \overline{y}_i \quad (1 \leq n \leq L)$$

siendo s una muestra de n unidades primarias o conglomerados. Este estimador es insesgado para estimar $\overline{y}$ en muestreo aleatorio simple sin reemplazamiento de conglomerados. En efecto,

$$E\left(\overline{y}_{cs}\right) = E\left(\frac{1}{n}\sum_{i \in s} \overline{y}_i\right) = \frac{1}{n} E\left(\sum_{i \in s} \overline{y}_i\right) = \frac{1}{n} \sum_{s \in \mathcal{S}} \sum_{i \in s} \overline{y}_i p(s) =$$

$$= \frac{1}{n} \sum_{i=1}^{L} \overline{y}_i \operatorname{card}\left(\{s : i \in s\}\right) \frac{1}{\binom{L}{n}} = \frac{1}{n} \frac{1}{\overline{N}} \sum_{i=1}^{L} \overline{N}\, \overline{y}_i \frac{\binom{L-1}{n-1}}{\binom{L}{n}} =$$

$$= \frac{1}{n\,\overline{N}} N\overline{y} \frac{\frac{(L-1)!}{(n-1)!(L-n)!}}{\frac{L!}{n!(L-n)!}} = \frac{N\overline{y}}{\overline{N}L} = \frac{N\overline{y}}{N} = \overline{y}.$$

10.1.2 Varianza de la estrategia (mas, $\overline{y}_{cs}$).

$$V(\overline{y}_{cs}) = \frac{L-n}{L} \frac{S^2_{\overline{y}_i}}{n}$$

por analogía con $V(mas, \overline{y}_s)$. Además

$$S^2_{\overline{y}_i} = \frac{1}{L-1} \sum_{i=1}^{L} (\overline{y}_i - \overline{y})^2$$

pues

$$\frac{1}{L} \sum_{i=1}^{L} \overline{y}_i = \frac{1}{L} \sum_{i=1}^{L} \frac{1}{\overline{N}} \sum_{j=1}^{\overline{N}} y_{ij} = \overline{y}$$

ya que $L\overline{N} = N$. Así tenemos,

$$V(\overline{y}_{cs}) = \frac{L-n}{L(L-1)n} \sum_{i=1}^{L} (\overline{y}_i - \overline{y})^2 = \frac{L-n}{L(L-1)n} \sum_{i=1}^{L} \left(\frac{1}{\overline{N}} \sum_{j=1}^{\overline{N}} y_{ij} - \overline{y} \right)^2 =$$

$$= \frac{L-n}{L(L-1)n\overline{N}^2} \sum_{i=1}^{L} \left[\sum_{j=1}^{\overline{N}} \left(y_{ij} - \overline{y} \right) \right]^2 =$$

$$= \frac{L-n}{L(L-1)n\overline{N}^2} \sum_{i=1}^{L} \left[\sum_{j=1}^{\overline{N}} (y_{ij} - \overline{y})^2 + \sum_{j \neq k}^{\overline{N}} (y_{ij} - \overline{y})(y_{ik} - \overline{y}) \right] =$$

$$= \frac{L-n}{L(L-1)n\overline{N}^2} \left[N\sigma^2 + N(\overline{N}-1)\sigma^2\delta \right] =$$

$$= \frac{(L-n)N\sigma^2}{L(L-1)n\overline{N}^2} \left[1 + (\overline{N}-1)\delta \right].$$

Ahora, si N y L son suficientemente grandes, multiplicando y dividiendo por $\overline{N}$ y haciendo uso de la aproximación $(L-1)\overline{N} \doteq N-1$, concluímos que

$$V(\overline{y}_{cs}) \doteq \frac{N - n\overline{N}}{N-1} \frac{\sigma^2}{n\overline{N}} \left[1 + (\overline{N}-1)\delta \right]$$

donde el primer factor es la varianza de la media muestral bajo diseño mas y el tamaño muestral es $n\overline{N}$ unidades simples, y el segundo factor representa el incremento o disminución de la varianza de $\overline{y}_{cs}$, frente a la de $\overline{y}_s$, cuando el muestreo es por L conglomerados de tamaño $\overline{N}$.

Si $\delta = \dfrac{-1}{N-1}$ (valor mínimo) entonces $V(\overline{y}_{cs}) \doteq 0$.

Si $\delta = 0$ entonces $V(\overline{y}_{cs}) \doteq V(\text{mas}, \overline{y}_s)$ de igual tamaño muestral.

Si $\delta > 0$ entonces $V(\overline{y}_{cs}) > V(\text{mas}, \overline{y}_s)$ para el mismo tamaño muestral.

La consecuencia es que lo ideal será agrupar unidades en conglomerados de forma que las unidades secundarias sean diferentes o heterogéneas entre sí según la variable de interés.

10.1.3 Estimación del total con conglomerados de igual tamaño.

El estimador será

$$N\overline{y}_{cs} = \frac{N}{n}\sum_{i=1}^{n}\overline{y}_i$$

y su varianza aproximada será

$$V(N\overline{y}_{cs}) = N^2 V(\overline{y}_{cs}) \doteq N^2 \frac{N-n\overline{N}}{N-1}\,\frac{\sigma^2}{n\overline{N}}\left[1+(\overline{N}-1)\delta\right].$$

10.1.4 Estimación de la proporción con conglomerados de igual tamaño.

Si $\widehat{p}$ es la proporción muestral (media muestral de valores 0 ó 1) será insesgado para la proporción poblacional P, y su varianza es

$$V(\widehat{p}) = \frac{N-n\overline{N}}{N-1}\,\frac{PQ}{n\overline{N}}\left[1+(\overline{N}-1)\delta\right]$$

ya que $\sigma^2 = PQ$, con $Q = 1-P$.

10.1.5 Tamaño de la muestra con conglomerados de igual tamaño.

Si fijamos el error absoluto máximo e, para un coeficiente de confianza $1-\alpha$ en la estimación de la media poblacional $\overline{y}$, por la desigualdad de Chebycheff tendremos

$$P\left[|\overline{y}_{cs}-\overline{y}| \le e\right] \ge 1 - \frac{\sigma^2_{\overline{y}_{cs}}}{e^2} = 1-\alpha$$

de donde

$$\alpha = \frac{\sigma^2_{\overline{y}_{cs}}}{e^2} = \frac{1}{e^2}\left[\frac{N\sigma^2}{(N-1)n\overline{N}} - \frac{\sigma^2}{N-1}\right]\left[1+(\overline{N}-1)\delta\right]$$

por lo que el número de conglomerados a tomar será n,

$$n = \frac{N\sigma^2}{(N-1)\overline{N}\left[\dfrac{\alpha e^2}{1+(\overline{N}-1)\delta} + \dfrac{\sigma^2}{N-1}\right]}$$

que depende de σ^2 y δ que en principio son desconocidos.

Del mismo modo se obtendrían los tamaños muestrales para la estimación del total o la proporción cuando los conglomerados son del mismo tamaño $\overline{N}$.

10.1.6 Tamaño óptimo de un conglomerado.

El tamaño del conglomerado $\overline{N}$ óptimo es un problema no resuelto hasta ahora, si bien se han dado soluciones empíricas aproximadas debidas a Smith, Jessen ó Hansen, Hurwitz y Madow según se recoge en el libro de Azorín y Sánchez-Crespo (1986).

§10.2 Muestreo por conglomerados de tamaño desigual.

Sea ahora

$$y_i = \sum_{j=1}^{N_i} y_{ij} = N_i \overline{y}_i$$

el total del conglomerado i $(1 \leq i \leq L)$, de tamaño N_i y de media $\overline{y}_i$.

Dada una muestra aleatoria simple de n unidades primarias o conglomerados de los L que componen la población, una estimación insesgada del total poblacional, $N\overline{y}$, de la variable de interés "y" es

$$t = \widehat{N\overline{y}} = \frac{L}{n} \sum_{i=1}^{n} y_i = \frac{L}{n} \sum_{i \in s} y_i$$

siendo

$$N\overline{y} = \sum_{i=1}^{L} y_i = \sum_{i=1}^{L} \sum_{j=1}^{N_i} y_{ij},$$

y su varianza es

$$V(t) = V\left(\widehat{N\overline{y}}\right) = L^2 V\left(\frac{1}{n} \sum_{i=1}^{n} y_i\right) = L^2 \frac{L-n}{L-1} \frac{\frac{1}{L} \sum_{i=1}^{L} (y_i - \overline{y}_t)^2}{n}$$

donde

$$\overline{y}_t = \frac{1}{L} \sum_{i=1}^{L} y_i.$$

10.2.1 Caso de probabilidades proporcionales al tamaño del conglomerado.

10.2.1.1. Con reposición.

Si de la población de L unidades primarias o conglomerados, se extrae una muestra de tamaño n, con reposición de modo que en todas y cada una de las extracciones la unidad i tiene una probabilidad $p_i = \frac{N_i}{N}$, siendo

$$0 \leq p_i \leq 1 \quad \text{y} \quad \sum_{i=1}^{L} p_i = 1$$

este diseño corresponde al estudiado pptr. Consideramos estimadores del total del tipo

$$t = \widehat{N\overline{y}} = \sum_{i \in \mathbf{S}} c_i y_i = \sum_{i \in U} c_i y_i e_i$$

siendo s una muestra obtenida por diseño pptr de conglomerados, c_i son constantes, $y_i = N_i \overline{y}_i$ totales de conglomerados, U la población de L conglomerados y e_i la variable aleatoria auxiliar que mide el número de veces que aparece la unidad primaria i en la muestra s.

$$E(t) = \sum_{i \in U} c_i y_i E(e_i) = \sum_{i \in U} c_i y_i n p_i = L\overline{y}_t = \sum_{i \in U} y_i$$

cuando

$$c_i = \frac{1}{np_i} \quad (i = 1, 2, \ldots, N).$$

Y su varianza será

$$V(t) = V\left(\sum_{i \in U} c_i y_i e_i\right) = \sum_{i \in U} c_i^2 y_i^2 V(e_i) + \sum_{i \neq j \in U} c_i c_j y_i y_j \operatorname{Cov}\left(e_i, e_j\right) =$$

$$= \sum_{i \in U} c_i^2 y_i^2 n p_i (1 - p_i) - \sum_{i \neq j \in U} c_i c_j y_i y_j n p_i p_j =$$

$$= n\left(\sum_{i \in U} c_i^2 y_i^2 p_i - \sum_{i \in U} c_i^2 y_i^2 p_i^2 - \sum_{i \neq j \in U} c_i c_j y_i y_j p_i p_j\right) =$$

$$= n\left[\sum_{i \in U} c_i^2 y_i^2 p_i - \left(\sum_{i \in U} c_i y_i p_i\right)^2\right]$$

y sustituyendo c_i por $\dfrac{1}{np_i}$, la varianza del estimador

$$t = \sum_{i \in \mathbf{S}} \frac{y_i}{np_i}$$

quedará

$$V(t) = \frac{1}{n}\left[\sum_{i \in U} \frac{y_i^2}{p_i} - (L\overline{y}_t)^2\right] = \frac{1}{n} \sum_{i \in U} p_i \left(\frac{y_i}{p_i} - L\overline{y}_t\right)^2$$

pues

$$E\left[\left(\frac{y_i}{p_i}\right)^2\right] - \left[E\left(\frac{y_i}{p_i}\right)\right]^2 = V\left(\frac{y_i}{p_i}\right).$$

El estimador insesgado de $V(t)$ es

$$\widehat{V}(t) = \frac{1}{n(n-1)} \sum_{i \in S} \left(\frac{y_i}{p_i} - t \right)^2 .$$

En efecto,

$$\sum_{i \in S} \left(\frac{y_i}{p_i} - t \right)^2 = \sum_{i \in S} \left(\frac{y_i^2}{p_i^2} + t^2 - 2t \frac{y_i}{p_i} \right) =$$

$$= \sum_{i \in S} \frac{y_i^2}{p_i^2} - nt^2 = \sum_{i \in U} \frac{y_i^2}{p_i^2} e_i - nt^2$$

de donde

$$E\left[\widehat{V}(t)\right] = E\left[\frac{1}{n(n-1)} \left(\sum_{i \in U} \frac{y_i^2}{p_i^2} e_i - nt^2 \right) \right] =$$

$$= \frac{1}{n(n-1)} \left\{ \sum_{i \in U} \frac{y_i^2}{p_i^2} E(e_i) - n \left[V(t) + (L\overline{y}_t)^2 \right] \right\},$$

y usando (9.1)

$$E\left[\widehat{V}(t)\right] = \frac{1}{n-1} \sum_{i \in U} \frac{y_i^2}{p_i} - \frac{1}{n(n-1)} \sum_{i \in U} \frac{y_i^2}{p_i} + \frac{1}{n(n-1)} (L\overline{y}_t)^2 - \frac{1}{n-1} (L\overline{y}_t)^2 =$$

$$= \frac{1}{n} \left[\sum_{i \in U} \frac{y_i^2}{p_i} - (L\overline{y}_t)^2 \right] = V(t)$$

pues

$$\frac{1}{n(n-1)} - \frac{1}{n-1} = -\frac{n-1}{n(n-1)} = -\frac{1}{n}.$$

10.2.1.2. Sin reposición.

Si la muestra de n conglomerados se selecciona mediante algún procedimiento aleatorio sin reposición, tenemos como estimador del total $L\overline{y}_t$ a

$$t = \widehat{L\overline{y}_t} = \sum_{i \in s} c_i y_i = \sum_{i \in U} c_i y_i e_i$$

donde la variable aleatoria auxiliar e_i puede tomar ahora los valores 0 y 1 (para todo $i \in U$).

$$E(t) = E\left(\sum_{i \in U} c_i y_i e_i \right) = \sum_{i \in U} c_i y_i E(e_i) = \sum_{i \in U} c_i y_i \pi_i$$

y para que t sea insesgado para el total $L\overline{y}_t$ basta tomar $c_i = \frac{1}{\pi_i}$.

Luego la expresión explícita del estimador t es

$$t = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in U} \frac{y_i}{\pi_i} e_i$$

$$V(t) = V\left(\sum_{i \in U} \frac{y_i}{\pi_i} e_i\right) = \sum_{i \in U} \frac{y_i^2}{\pi_i^2} V(e_i) + \sum_{i \neq j \in U} \frac{y_i y_j}{\pi_i \pi_j} \operatorname{Cov}\left(e_i, e_j\right) =$$

$$= \sum_{i \in U} \frac{y_i^2}{\pi_i^2} \pi_i (1 - \pi_i) + \sum_{i \neq j \in U} \frac{y_i y_j}{\pi_i \pi_j} \left(\pi_{ij} - \pi_i \pi_j\right),$$

y un estimador insesgado de esta varianza es

$$\widehat{V}(t) = \sum_{i \in s} \frac{y_i^2}{\pi_i^2} (1 - \pi_i) + \sum_{i \neq j \in s} \frac{y_i y_j}{\pi_i \pi_j} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}}.$$

§10.3 Submuestreo.

En el muestreo en dos etapas o bietápico, se seleccionan primero n conglomerados o unidades primarias. Después se seleccionan un número especificado de subunidades o unidades secundarias o finales de cada uno de los n conglomerados extraídos.

10.3.1 Teorema de Madow.

Para calcular el promedio de un estimador en muestreo bietápico, primero se promedia la estimación sobre todas las selecciones de la segunda etapa que pueden extraerse de un conjunto fijo de n unidades elegido por el diseño muestral. Posteriormente, se promedian sobre todas las posibles selecciones de n unidades por el diseño. Para una estimación $\widehat{\theta}$, este método se expresa como

$$E\left(\widehat{\theta}\right) = E_1\left[E_2\left(\widehat{\theta}\right)\right]$$

donde E denota el valor esperado o promedio sobre todas las muestras, E_2 el promedio sobre las posibles selecciones (en la segunda etapa), de un conjunto fijo de unidades primarias, y E_1 denota el promedio sobre todas las selecciones de la primera etapa. La demostración es sencilla,

$$E(u) = E[E(u \mid v)]$$

siendo u y v dos variables aleatorias discretas que pueden tomar N y M valores distintos respectivamente,

$$E[E(u \mid v)] = \sum_{j=1}^{M} p_j E\left(u \mid v_j\right) = \sum_{j=1}^{M} p_j \sum_{i=1}^{N} u_i p_{i|j}\left(u_i \mid v_j\right) =$$

$$= \sum_{j=1}^{M} \sum_{i=1}^{N} u_i p\left(u = u_i, v = v_j\right) = \sum_{j=1}^{M} \sum_{i=1}^{N} u_i p_{ij} =$$

$$= \sum_{i=1}^{N} u_i \sum_{j=1}^{M} p_{ij} = \sum_{i=1}^{N} u_i p_i = E(u),$$

donde hemos usado que $p_{i|j} = \frac{p_{ij}}{p_j}$. Este razonamiento es válido también para variables aleatorias contínuas.

Para la varianza de $\widehat{\theta}$ tenemos (teorema de Madow)

$$V\left(\widehat{\theta}\right) = V_1\left[E_2\left(\widehat{\theta}\right)\right] + E_1\left[V_2\left(\widehat{\theta}\right)\right]$$

donde $V_2\left(\widehat{\theta}\right)$ es la varianza sobre la submuestra, para unas unidades primarias determinadas. $V_1[\cdot]$ es la varianza sobre la muestra de unidades primarias seleccionadas. La prueba es la siguiente:

Si $\theta = E(\widehat{\theta})$,

$$V\left(\widehat{\theta}\right) = E\left[\left(\widehat{\theta} - \theta\right)^2\right] = E_1 E_2\left[\left(\widehat{\theta} - \theta\right)^2\right].$$

Pero

$$E_2\left(\widehat{\theta} - \theta\right)^2 = E_2\left(\widehat{\theta}^2\right) - 2\theta E_2\left(\widehat{\theta}\right) + \theta^2 =$$

$$= \left[E_2\left(\widehat{\theta}\right)\right]^2 + V_2(\theta) - 2\theta E_2\left(\widehat{\theta}\right) + \theta^2.$$

Ahora se promedia sobre las selecciones de la primera etapa. Como $\theta = E_1 E_2(\widehat{\theta})$,

$$V\left(\widehat{\theta}\right) = E_1\left[E_2\left(\widehat{\theta}\right)\right]^2 - \theta^2 + E_1\left[V_2\left(\widehat{\theta}\right)\right] = V_1\left[E_2\left(\widehat{\theta}\right)\right] + E_1\left[V_2\left(\widehat{\theta}\right)\right].$$

La fórmula de Madow puede generalizarse a tres o más etapas, del modo

$$V\left(\widehat{\theta}\right) = V_1\left\{E_2\left[E_3\left(\widehat{\theta}\right)\right]\right\} + E_1\left\{V_2\left[E_3\left(\widehat{\theta}\right)\right]\right\} + E_1\left\{E_2\left[V_3\left(\widehat{\theta}\right)\right]\right\}$$

donde

$$E\left(\widehat{\theta}\right) = E_1 E_2 E_3\left(\widehat{\theta}\right).$$

10.3.2 Estudio de una muestra bietápica con unidades de primera etapa iguales.

Denotamos

y_{ij} = valor de la j-ésima subunidad en la i-ésima unidad primaria.

$$\overline{y}_{s(i)} = \frac{1}{m}\sum_{j=1}^{m} y_{ij} =$$

= media muestral por subunidad en la i-ésima unidad primaria.

$$\overline{\overline{y}} = \frac{1}{n}\sum_{i=1}^{n}\overline{y}_{s(i)} =$$

= media global de muestra por subunidades.

$$S_1^2 = \frac{\sum_{i=1}^{L}(\overline{y}_i - \overline{\overline{y}})^2}{L-1} =$$

= cuasivarianza entre medias de unidades primarias.

$$S_2^2 = \frac{\sum_{i=1}^{L}\sum_{j=1}^{\overline{N}}\left(y_{ij} - \overline{y}_i\right)^2}{L(\overline{N}-1)} =$$

= media de cuasivarianzas de unidades secundarias dentro de unidades primarias.

10.3.3 Estimador de la media.

Teorema 10.1 *Si n= número de unidades primarias seleccionadas, m =número de subunidades, extraídas ambas por diseño mas,* $\overline{\overline{y}}$ *es una estimación insesgada de* $\overline{y}$ *con varianza*

$$V\left(\overline{\overline{y}}\right) = \frac{L-n}{L}\,\frac{S_1^2}{n} + \frac{\overline{N}-m}{\overline{N}}\,\frac{S_2^2}{mn}$$

Demostración.

$$E\left(\overline{\overline{y}}\right) = E_1\left[E_2\left(\overline{\overline{y}}\right)\right] = E_1\left(\frac{1}{n}\sum_{i=1}^{n}\overline{y}_i\right) = \frac{1}{L}\sum_{i=1}^{L}\overline{y}_i = \overline{y}.$$

La varianza se calcula así,

$$V\left(\overline{\overline{y}}\right) = V_1E_2\left(\overline{\overline{y}}\right) + E_1V_2\left(\overline{\overline{y}}\right).$$

$$E_2\left(\overline{\overline{y}}\right) = \frac{1}{n}\sum_{i=1}^{n}\overline{y}_i \;\Rightarrow\; V_1\left[E_2\left(\overline{\overline{y}}\right)\right] = \frac{L-n}{L}\,\frac{S_1^2}{n} \tag{10.1}$$

$$V_2\left(\overline{\overline{y}}\right) = V_2\left(\frac{1}{n}\sum_{i=1}^{n}\overline{y}_{s(i)}\right) = \frac{1}{n^2}\sum_{i=1}^{n}V_2\left(\overline{y}_{s(i)}\right) = \tag{10.2}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\frac{\overline{N}-m}{\overline{N}}\,\frac{S_{2i}^2}{m} = \frac{\overline{N}-m}{\overline{N}n^2}\,\frac{\sum_{i=1}^{n}S_{2i}^2}{m}$$

siendo

$$S_{2i}^2 = \frac{1}{\overline{N}-1}\sum_{j=1}^{\overline{N}}\left(y_{ij}-\overline{y}_i\right)^2$$

la cuasivarianza de subunidades para la i–ésima unidad primaria.

Como

$$E_1\left(\frac{1}{n}\sum_{i=1}^{n}S_{2i}^2\right) = \frac{\sum_{i=1}^{L}S_{2i}^2}{L} = \frac{\sum_{i=1}^{L}\sum_{j=1}^{\overline{N}}\left(y_{ij}-\overline{y}_i\right)^2}{L(\overline{N}-1)} = S_2^2$$

luego tendremos

$$E_1\left[V_2\left(\overline{\overline{y}}\right)\right] = \frac{\overline{N}-m}{\overline{N}}\,\frac{S_2^2}{mn}.$$

Por todo ello

$$V\left(\overline{\overline{y}}\right) = \frac{L-n}{L}\,\frac{S_1^2}{n} + \frac{\overline{N}-m}{\overline{N}}\,\frac{S_2^2}{mn}$$

10.3.4 Estimador de la varianza de la media muestral.

Un estimador insesgado de $V\left(\overline{\overline{y}}\right)$ es

$$\widehat{V}\left(\overline{\overline{y}}\right) = \frac{L-n}{L}\,\frac{s_1^2}{n} + \frac{\overline{N}-m}{\overline{N}mL}s_2^2$$

con

$$s_1^2 = \frac{\sum_{i=1}^{n}\left(\overline{y}_{s(i)}-\overline{\overline{y}}\right)^2}{n-1}$$

y

$$s_2^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(y_{ij}-\overline{y}_{s(i)}\right)^2}{n(m-1)}.$$

Veamos la prueba:

$$(n-1)s_1^2 = \sum_{i=1}^{n}\left(\overline{y}_{s(i)}-\overline{\overline{y}}\right)^2 = \sum_{i=1}^{n}\overline{y}_{s(i)}^2 - n\overline{\overline{y}}^2,$$

por tanto

$$(n-1)E_2\left(s_1^2\right) = \sum_{i=1}^{n}\overline{y}_i^2 + \sum_{i=1}^{n}\frac{\overline{N}-m}{\overline{N}m}S_{2i}^2 - n\left(\frac{1}{n}\sum_{i=1}^{n}\overline{y}_i\right)^2 - \frac{1}{n}\sum_{i=1}^{n}\frac{\overline{N}-m}{\overline{N}m}S_{2i}^2$$

pues

$$E_2\left(\overline{y}^2_{s(i)}\right) = \left[E_2\left(\overline{y}_{s(i)}\right)\right]^2 + V_2\left(\overline{y}_{s(i)}\right)$$

y

$$E_2\left(\overline{\overline{y}}^2\right) = \left[E_2\left(\overline{\overline{y}}\right)\right]^2 + V_2\left(\overline{\overline{y}}\right),$$

por (10.1) y (10.2).

Luego

$$(n-1)E_2\left(s_1^2\right) = \sum_{i=1}^{n}\left(\overline{y}_i - \frac{1}{n}\sum_{i=1}^{n}\overline{y}_i\right)^2 + \frac{(n-1)(\overline{N}-m)}{\overline{N}mn}\sum_{i=1}^{n}S_{2i}^2$$

que multiplicando por

$$\frac{L-n}{Ln(n-1)}$$

queda,

$$\frac{(L-n)}{Ln}E_2\left(s_1^2\right) = \frac{L-n}{Ln(n-1)}\sum_{i=1}^{n}\left(\overline{y}_i - \frac{1}{n}\sum_{i=1}^{n}\overline{y}_i\right)^2 + \frac{(L-n)(\overline{N}-m)}{Ln^2m\overline{N}}\sum_{i=1}^{n}S_{2i}^2$$

y promediando sobre la primera etapa de diseño mas,

$$E_1\left[\frac{(L-n)}{Ln}E_2\left(s_1^2\right)\right] = E\left(\frac{L-n}{Ln}s_1^2\right) =$$

$$= \frac{L-n}{Ln}E_1\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(\overline{y}_i - \frac{1}{n}\sum_{i=1}^{n}\overline{y}_i\right)^2\right] + \frac{(L-n)(\overline{N}-m)}{Ln^2m\overline{N}}E_1\left(\sum_{i=1}^{n}S_{2i}^2\right) =$$

$$= \frac{L-n}{Ln}\underbrace{\frac{1}{L-1}\sum_{i=1}^{L}(\overline{y}_i - \overline{\overline{y}})^2}_{S_1^2} + \frac{(L-n)(\overline{N}-m)}{Ln^2m\overline{N}}nS_2^2.$$

Ahora , como $E_1E_2(s_2^2) = S_2^2$, y el término anterior de S_2^2 (mírese $V(\overline{\overline{y}})$) tiene un coeficiente $\dfrac{(L-n)}{L}$ previo, bastará multiplicar el término en S_2^2 por $\dfrac{n}{L}$ y sumar a la expresión anterior, quedando

$$\hat{V}\left(\overline{\overline{y}}\right) = \frac{L-n}{Ln}s_1^2 + \frac{n}{L}\frac{(\overline{N}-m)}{\overline{N}mn}s_2^2$$

10.3.5 Distribución de la muestra en dos etapas.

Si admitimos la función de coste del tipo

$$C = c_1 n + c_2 nm,$$

es decir, siendo c_1 el costo de seleccionar una unidad primaria y c_2 el coste de observación por unidad secundaria. Como tenemos

$$V(\overline{\overline{y}}) = \frac{L-n}{L}\frac{S_1^2}{n} + \frac{\overline{N}-m}{\overline{N}}\frac{S_2^2}{nm} = \frac{1}{n}\left(S_1^2 - \frac{S_2^2}{\overline{N}}\right) + \frac{1}{nm}S_2^2 - \frac{1}{L}S_1^2,$$

al minimizar $V(\overline{\overline{y}})$ para C fijo, tenemos el langrangiano

$$\phi = \frac{1}{n}\left(S_1^2 - \frac{S_2^2}{\overline{N}}\right) + \frac{1}{nm}S_2^2 - \frac{1}{L}S_1^2 + \lambda\left(C - c_1 n - c_2 nm\right)$$

que se resuelve así,

$$\frac{\partial \phi}{\partial n} = -\frac{1}{n^2}\left(S_1^2 - \frac{S_2^2}{\overline{N}}\right) - \frac{1}{n^2 m}S_2^2 - \lambda\left(c_1 + c_2 m\right) = 0$$

$$\frac{\partial \phi}{\partial m} = -\frac{1}{nm^2}S_2^2 - \lambda c_2 n = 0.$$

Luego

$$\lambda = -\frac{1}{n^2\left(c_1 + c_2 m\right)}\left(S_1^2 - \frac{S_2^2}{\overline{N}}\right) - \frac{1}{n^2 m\left(c_1 + c_2 m\right)}S_2^2 = -\frac{1}{c_2 n^2 m^2}S_2^2$$

o bien, multiplicando por $n^2 m^2 c_2\left(c_1 + c_2 m\right)$

$$m^2 c_2\left(S_1^2 - \frac{S_2^2}{\overline{N}}\right) + m c_2 S_2^2 = \left(c_1 + c_2 m\right)S_2^2$$

que reordenando y simplificando resulta la ecuación en m

$$m^2 c_2\left(S_1^2 - \frac{S_2^2}{\overline{N}}\right) - c_1 S_2^2 = 0$$

que resolviendo queda el valor óptimo,

$$m = \pm\sqrt{\frac{c_1 S_2^2}{c_2\left(S_1^2 - \frac{S_2^2}{\overline{N}}\right)}}$$

y nos quedamos con el signo adecuado ya que $m > 0$. Finalmente

$$n = \frac{C}{\left(c_1 + c_2 m\right)}.$$

10.3.6 Muestra bietápica con unidades de primera etapa desiguales.

Las unidades primarias se seleccionan con probabilidades proporcionales a p_i

$$\left(p_i > 0,\ \sum_{i=1}^{L} p_i = 1\right),$$

con diseño pptr (con reemplazamiento). La submuestra es de tamaño m_i subunidades de la unidad primaria i, con diseño mas. Si la unidad primaria i se selecciona más de una vez, se restituye la totalidad de la submuestra seleccionando independientemente m_i unidades secundarias con diseño mas (sin reemplazamiento). Un estimador insesgado del total poblacional $N\overline{y}$ es

$$\widehat{N\overline{y}} = \frac{1}{n}\sum_{i=1}^{n} \frac{N_i \overline{y}_{s(i)}}{p_i}$$

en efecto,

$$E\left(\widehat{N\overline{y}}\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} \frac{N_i \overline{y}_{s(i)}}{p_i}\right) = \frac{1}{n}E\left(\sum_{i=1}^{L} \frac{N_i \overline{y}_{s(i)}}{p_i} e_i\right) =$$

$$= \frac{1}{n}E_1\left(\sum_{i=1}^{L} \frac{N_i E_2\left(\overline{y}_{s(i)}\right)}{p_i} e_i\right) = \frac{1}{n}\sum_{i=1}^{L} \frac{N_i \overline{y}_i}{p_i} E_1(e_i) = \sum_{i=1}^{L} N_i \overline{y}_i = N\overline{y}$$

siendo $e_i = k$, si el conglomerado i se selecciona k veces, con $E_1(e_i) = np_i$ $(i = 1, 2, \ldots, L)$.

La varianza de $\widehat{N\overline{y}}$ se obtiene así, haciendo uso del teorema de Madow,

$$V\left(\widehat{N\overline{y}}\right) = E_1 V_2\left(\widehat{N\overline{y}}\right) + V_1 E_2\left(\widehat{N\overline{y}}\right).$$

$$V_2\left(\widehat{N\overline{y}}\right) = V_2\left(\frac{1}{n}\sum_{i=1}^{L} \frac{N_i \overline{y}_{s(i)}}{p_i} e_i\right) = \frac{1}{n^2}\sum_{i=1}^{L} e_i^2 \frac{N_i^2}{p_i^2} V_2\left(\overline{y}_{s(i)}\right) =$$

$$= \frac{1}{n^2}\sum_{i=1}^{L} e_i^2 \frac{N_i^2}{p_i^2} \frac{N_i - m_i}{N_i m_i} S_i^2,$$

siendo S_i^2 la cuasivarianza dentro del conglomerado i $(1 \leq i \leq L)$.

$$E_1 V_2\left(\widehat{N\overline{y}}\right) = \frac{1}{n^2}\sum_{i=1}^{L} E_1\left(e_i^2\right) \frac{N_i^2}{p_i^2} \frac{N_i - m_i}{N_i m_i} S_i^2 =$$

$$= \sum_{i=1}^{L} \frac{1 - p_i + np_i}{np_i} N_i^2 \frac{N_i - m_i}{N_i m_i} S_i^2,$$

pues

$$E_1\left(e_i^2\right) = V_1(e_i) + [E_1(e_i)]^2 = np_i(1 - p_i) + n^2 p_i^2 = np_i(1 - p_i + np_i).$$

$$E_2\left(\widehat{N\overline{y}}\right) = E_2\left(\frac{1}{n}\sum_{i=1}^{L}\frac{N_i\overline{y}_{s(i)}}{p_i}e_i\right) = \frac{1}{n}\sum_{i=1}^{L}\frac{N_i\overline{y}_i}{p_i}e_i,$$

pues $E_2(\overline{y}_{s(i)}) = \overline{y}_i$.

$$V_1E_2\left(\widehat{N\overline{y}}\right) = V_1\left(\sum_{i=1}^{n}\frac{N\overline{y}_i}{np_i}e_i\right) = \frac{1}{n^2}\sum_{i=1}^{L}\frac{N_i^2\overline{y}_i^2}{p_i^2}V_1(e_i) =$$

$$= \frac{1}{n^2}\sum_{i=1}^{L}\frac{N_i^2\overline{y}_i^2}{p_i^2}np_i(1-p_i) = \sum_{i=1}^{L}\frac{N_i^2\overline{y}_i^2}{np_i}(1-p_i).$$

Luego,

$$V\left(\widehat{N\overline{y}}\right) = \sum_{i=1}^{L}\frac{1-p_i+np_i}{np_i}N_i^2\frac{N_i-m_i}{N_im_i}S_i^2 + \sum_{i=1}^{L}\frac{N_i^2\overline{y}_i^2}{np_i}(1-p_i).$$

El caso de estimación de proporciones es idéntico al de estimación de medias y éste es prácticamente igual a la estimación de totales $N\overline{y}$.

Ejercicio 10.1 *Un establecimiento comercial dispone de* 1.500 *facturas que recogen los ingresos durante un mes de trabajo; se desea estimar el total facturado mediante muestreo sistemático, tomando un arranque aleatorio entre las* 15 *primeras facturas. Por la experiencia pasada se sabe que el coeficiente de correlación intraconglomerados es aproximadamente* $\delta = 0.5$. *Se desea saber si la varianza del estimador usual en muestreo sistemático es más del doble que la varianza de la estrategia (mas,* $\overline{y}_s$*) con idéntico tamaño muestral.*

Solución.
El tamaño muestral es

$$n = \frac{N}{K} = \frac{1500}{15} = 100$$

$$V(\overline{y}_{cs}) \doteq \frac{N-n}{N-1}\frac{\sigma^2}{n}[1+(n-1)\delta] = V(\overline{y}_s)[1+(n-1)\delta] > 2V(\overline{y}_s),$$

cierto, pues

$$[1+(n-1)\delta] = 1+99\cdot 0.5 = 50.5 > 2.$$

Ejercicio 10.2 *Con el fin de estimar la calidad de cierta marca de cerillas se examina la producción que está empaquetada en cajas de* $\overline{N} = 50$ *fósforos. El número de cajas producido es de* $N = 300$. *El objetivo es estimar la producción de cerillas defectuosas para lo cual se prueban* $n = 5$ *cajas de modo destructivo. La proporción estimada de unidades defectuosas por caja en las* 5 *muestreadas es de* $\widehat{p} = 0.3$. *¿Cuál será la varianza de este estimador si suponemos que* $\widehat{p}$ *estima bien la proporción poblacional* P *y el coeficiente de correlación intraconglomerados es* $\delta = 0$*?*

Solución.

$$V(\widehat{p}) \doteq \frac{N-n\overline{N}}{N-1}\,\frac{PQ}{n\overline{N}}\left[1+(\overline{N}-1)\delta\right] \doteq 0.0001405$$

Ejercicio 10.3 *En el problema anterior, ¿Cuál será el tamaño muestral n de cajas o unidades primarias para que el error absoluto máximo de muestreo, e, sea igual a* 0.001 *para un coeficiente de confianza* $1-\alpha = 0.90$*?*

Solución.

$$n = \frac{NPQ}{(N-1)\overline{N}\left[\dfrac{\alpha e^2}{1+(\overline{N}-1)\delta}+\dfrac{PQ}{N-1}\right]} \doteq 6.$$

Ejercicio 10.4 *Averiguar el tamaño muestral n necesario para asegurar que el estimador* $\overline{y}_{cs}$*, de la media* $\overline{y}$*, con conglomerados del mismo tamaño* $\overline{N} = 30$*, fijado el error absoluto máximo* $e = 0.05$ *para un coeficiente de confianza* $1-\alpha = 0.95$*. El tamaño poblacional es* $N = 30000$ *unidades. Además la experiencia en estudios anteriores nos da una estimación de la varianza poblacional* $\sigma^2 = 0.15$ *y del coeficiente de correlación intraconglomerados* $\delta = 0.1$*.*

Solución.

$$n = \frac{N\sigma^2}{(N-1)\overline{N}\left[\dfrac{\alpha e^2}{1+(\overline{N}-1)\delta}+\dfrac{\sigma^2}{N-1}\right]} =$$

$$= \frac{30000\cdot 0.15}{29999\cdot 30\left[\dfrac{0.05\cdot 0.05^2}{1+29\cdot 0.1}+\dfrac{0.15}{29999}\right]} \doteq 136.$$

Ejercicio 10.5 *En una empresa industrial se empaquetan los productos en lotes de* $\overline{N} = 10$ *unidades, produciéndose diariamente* $L = 2000$ *lotes. Con el fin de estimar la calidad del producto se procede a la estimación de la media poblacional* $\overline{y}$ *de cierta característica "y", en muestreo bietápico seleccionando* $n = 20$ *lotes por muestreo aleatorio simple sin reemplazamiento (mas), y dentro de cada lote extraído se examinan* $m = 3$ *unidades por mas. Estimar de modo insesgado la varianza del estimador usual*

$$\overline{\overline{y}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m} y_{ij} = \frac{1}{n}\sum_{i=1}^{n}\overline{y}_{s(i)}$$

sabiendo que de la muestra obtenemos que

$$s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\overline{y}_{s(i)}-\overline{\overline{y}}\right)^2 = 0.3$$

y

$$s_2^2 = \frac{1}{n(m-1)} \sum_{i=1}^{n} \sum_{j=1}^{m} \left(y_{ij} - \overline{y}_{s(i)}\right)^2 = 3.$$

Solución.

$$\widehat{V}\left(\overline{\overline{y}}\right) = \frac{L-n}{L} \frac{s_1^2}{n} + \frac{\overline{N}-m}{\overline{N}mL} s_2^2 = \frac{2000-20}{2000} \frac{0.3}{20} + \frac{10-3}{10 \cdot 3 \cdot 2000} 3 = 0.0152.$$

Ejercicio 10.6 *En una primera fase se selecciona, con diseño masr de tamaño $n = 5$, una muestra. En segunda fase, de la muestra anterior se selecciona una submuestra con diseño masr de tamaño $m = 2$. Obtener la varianza de la media muestral de los datos observados en la segunda fase, en función de la varianza poblacional.*

Solución.

En la primera fase seleccionamos una muestra de tamaño $n = 5$, y denotamos por $\overline{x}'$, $\overline{S}^2$ y $\overline{\sigma}^2$ a la media muestral, cuasivarianza muestral y varianza muestral respectivamente en la primera fase. En la segunda fase se submuestrea la muestra de la primera fase, con el nuevo tamaño muestral $m = 2$, cuya media muestral, $\overline{x}$, es de la que se pide su varianza. Usando el teorema de Madow en dos fases,

$$V_2\left(\overline{x}\right) = \frac{\overline{\sigma}^2}{m} = \frac{n-1}{mn} \overline{S}^2$$

de donde

$$E_1\left[V_2\left(\overline{x}\right)\right] = E_1\left(\frac{n-1}{mn} \overline{S}^2\right) = \frac{n-1}{mn} E_1\left(\overline{S}^2\right) = \frac{n-1}{mn} \sigma^2 = \frac{2}{5}\sigma^2.$$

$$E_2\left(\overline{x}\right) = \overline{x}',$$

y entonces

$$V_1\left[E_2\left(\overline{x}\right)\right] = V_1\left(\overline{x}'\right) = \frac{\sigma^2}{n} = \frac{\sigma^2}{5}$$

Luego:

$$V\left(\overline{x}\right) = E_1\left[V_2\left(\overline{x}\right)\right] + V_1\left[E_2\left(\overline{x}\right)\right] = \frac{2}{5}\sigma^2 + \frac{\sigma^2}{5} = \frac{3}{5}\sigma^2,$$

siendo σ^2 la varianza poblacional.

Ejercicio 10.7 *¿Cuándo se podría llamar una partición en conglomerados con el nombre de estratificación?*

Solución.

Cuando se seleccionan todos los conglomerados para ser observados total o parcialmente.

Ejercicio 10.8 *Si en una tercera etapa el estimador $\widehat{\theta}$ es constante, obtener la varianza incondicional de $\widehat{\theta}$.*

Solución.
En cualquier caso si $\widehat{\theta} = c$ (constante)

$$\begin{aligned} V\left(\widehat{\theta}\right) &= E_1E_2V_3\left(\widehat{\theta}\right) + E_1V_2E_3\left(\widehat{\theta}\right) + V_1E_2E_3\left(\widehat{\theta}\right) = \\ &= E_1E_2\left(0\right) + E_1V_2\left(c\right) + V_1E_2\left(c\right) = \\ &= E_1\left(0\right) + E_1\left(0\right) + V_1\left(c\right) = \\ &= 0 + 0 + 0 = 0. \end{aligned}$$

Directamente también puede calcularse:

$$V\left(\widehat{\theta}\right) = V\left(c\right) = 0.$$

Capítulo 11

Diseño de encuestas.

§11.1 Población y marco. Tipos de unidades.

Llamamos "población objetivo" a un conjunto de unidades del que se requiere información. Para obtenerla, es necesario medir u observar en cada unidad uno o varios caracteres.

En realidad la población disponible no suele tenerse depurada, es decir no aparecen todas las unidades (omisiones), en otros casos existen duplicaciones (repeticiones) y unidades extrañas (o vacías). Por otro lado la información no podrá ser observada por su inaccesibilidad para unos medios dados, por su negación a ser medida o por su ausencia. Esto hace que el conjunto que es realmente objeto de investigación, llamado "población investigada" difiera de la población objetivo (la "población investigada" que contiene las unidades observadas y aquellas que podrían haberlo sido si se hubiera propuesto, debe estar contenida en la " población marco", aunque podría no ser así en algún caso de muestreo no probabilístico). Llamamos " población marco" al conjunto de unidades a partir del cual se selecciona la muestra.

§11.2 Recogida de datos.

Si disponemos de una población marco, dispondremos de los medios que nos permitan identificar cada unidad de la muestra seleccionada y localizarla. Si las unidades fueran personas dispondríamos de sus direcciones y/o teléfono para su accesibilidad por parte del encuestador, quien será el encargado de medir la o las características en estudio para cada unidad seleccionada en la muestra.

§11.3 Cuestionarios. Trabajo de campo.

La tarea de redactar el cuestionario es muy delicada. Las preguntas realizadas deben exponerse con la mayor sencillez y claridad para no dar lugar a errores de interpretación. A este objetivo contribuye el realizar estudios pilotos o previos, que indicarán las posibles malinterpretaciones y se pueda modificar el cuestionario a fin de que el definitivo sea claro en estudios de poblaciones humanas.

En el trabajo de campo, el entrevistador debe estar adecuadamente advertido de los posibles errores o malinterpretaciones por parte del encuestado, con el objeto de eliminar este tipo de errores, que se traducen en sesgos en la posterior estimación. En algunos casos una conglomeración adecuada puede ahorrar mucho coste económico del presupuesto; así como estratificaciones o muestreos sistemáticos, etc.

§11.4 Tabulación de resultados.

A la hora de presentar los resultados investigados se ha de tener precaución de no dar el orígen de la información y conservar el adecuado secreto estadístico y por tanto no deben darse datos individualizados sino conjuntos, con lo que se evitan posibles filtraciones de información y se salvaguarda el anonimato de los encuestados.

La presentación de datos se hará con completa claridad y evitando en lo posible ambigüedades de cualquier tipo. Pueden resultar útiles diversos métodos de representación gráfica de los resultados (gráficos de sectores, pirámides, diagramas, etc.).

Capítulo 12

Fuentes de error en las encuestas.

§12.1 Calidad de los datos censales.

Además del "costo superior" de un censo frente a una muestra, la "mayor lentitud" en la recolección y presentación de datos, y "menos posibilidades" ya que el personal utilizado es menos cualificado que el que podría ser empleado para una encuesta; en una encuesta por muestreo al reducir el volúmen de trabajo se puede emplear personal más capacitado y someterlo a un entrenamiento intensivo, pudiéndose dar una supervisión cuidadosa del trabajo de campo y del procesamiento de resultados, así como que puede producir resultados más exactos que la enumeración completa o censal.

§12.2 La no respuesta.

Usualmente en teoría de muestras se suponen que

a) la población marco coincide con la población objetivo,

b) que todas las unidades de la muestra son observadas y

c) que la información obtenida es correcta.

Si falla la suposición a) da lugar a los errores de cobertura (ya sea por defecto -omisiones- o por exceso -duplicaciones y unidades extrañas-).

Si falla la suposición c) da lugar a los errores de medida.

Si falla la suposición b), no se obtiene información en todas las unidades de la muestra y diremos que existe falta de respuesta o no respuesta. La no respuesta puede deberse a: la ausencia temporal del respondiente durante las horas de entrevista, negativa absoluta a colaborar, falta de conocimientos o capacidad por parte del informante, método de recogida de datos, condiciones personales y grado de adiestramiento de los entrevistadores, motivación de los informantes, etc.

Si la característica que tratamos de estimar es la media poblacional $\overline{y}$, y utilizamos solo unidades del estrato (1) de unidades que contestan, se produce un sesgo al usar la media muestral $\overline{y}_{s(1)}$, dado por

$$B\left(\overline{y}_{s(1)}\right) = E\left(\overline{y}_{s(1)}\right) - \overline{y} = \overline{y}_1 - \left(P_1\overline{y}_1 + P_2\overline{y}_2\right) =$$

$$= \left(1 - P_1\right)\overline{y}_1 + P_2\overline{y}_2 = P_2\left(\overline{y}_1 - \overline{y}_2\right)$$

que resulta proporcional al peso del estrato (2) de unidades que no contestan, P_2, y a la diferencia entre las medias de ambos estratos, $(\overline{y}_1 - \overline{y}_2)$. Existen algunos métodos de tratamiento de la falta de respuesta que permiten corregir este sesgo.

Uno de ellos es el método de Hansen y Hurwitz (1946); en una primera etapa se selecciona una muestra aleatoria simple sin reemplazamiento de n unidades. Habrá n_1 que no contestan y n_2 que no responden, $n_1 + n_2 = n$. De entre las n_2 que no contestan se disuade, a una muestra aleatoria simple sin reemplazamiento de tamaño $n_{21} \leq n_2$, de responder a la pregunta requerida.

El estimador será para el total $N\overline{y} = Y$

$$\widehat{Y} = \frac{N}{n}\left(n_1\overline{y}_{s(1)} + n_2\overline{y}_{s(21)}\right)$$

donde $\overline{y}_{s(1)}$ es la media muestral de las observaciones tomadas en la primera fase, y donde $\overline{y}_{s(21)}$ es la media muestral obtenida en la segunda fase.

$$E\left(\widehat{Y}\right) = \frac{N}{n}E_1\left[n_1\overline{y}_{s(1)} + n_2\,\mathrm{E}_2\left(\overline{y}_{s(21)}\right)\right] = \frac{N}{n}E_1\left(n\overline{y}_s\right) = N\overline{y}$$

resulta ser insesgado, y su varianza es

$$V\left(\widehat{Y}\right) = V_1E_2\left(\widehat{Y}\right) + E_1V_2\left(\widehat{Y}\right)$$

donde

$$V_1E_2\left(\widehat{Y}\right) = V_1\left(N\overline{y}_s\right) = N^2\frac{N-n}{N}\,\frac{S^2}{n}$$

$$V_2\left(\widehat{Y}\right) = V_2\left(\frac{N}{n}n_2\overline{y}_{s(21)}\right) = \frac{N^2}{n^2}n_2^2\frac{n_2 - n_{21}}{n_2}\,\frac{S_{21}^2}{n_{21}}$$

donde S_{21}^2 es la cuasivarianza de la muestra $s(21)$,

$$E_1V_2\left(\widehat{Y}\right) = \frac{N^2}{n^2}n_2\frac{n_2 - n_{21}}{n_{21}}E_1\left(S_{21}^2\right) = \frac{N^2}{n^2}n_2\frac{n_2 - n_{21}}{n_{21}}S_2^2$$

Luego:

$$V\left(\widehat{Y}\right) = N\left(N - n\right)\frac{S^2}{n} + \frac{N^2n_2\left(n_2 - n_{21}\right)}{n^2n_{21}}S_2^2$$

siendo S_2^2 la cuasivarianza del estrato de no respuesta.

Capítulo 13

Otras técnicas.

§13.1 Métodos de muestreo no aleatorios.

Corresponden a técnicas de muestreo no basadas en un diseño muestral, sino por ejemplo:

a) Una parte fácilmente accesible de la población (ejemplo: muestra de carbón a 15 ó 20 cm. de la superficie de un vagón).

b) Se selecciona a la ventura (ejemplo: se sacan 20 conejos de experimentación de una jaula según los alcance con la mano el investigador).

c) Cuando la población es pequeña y heterogénea, el investigador inspecciona la totalidad de ésta y selecciona una pequeña muestra de unidades típicas, unidades que a su parecer estén cercanas al promedio de la población ("selección intencional" o "de juicio").

d) La muestra consta de entrevistados voluntarios (en estudios donde el proceso de medición es desagradable o penoso para el que está siendo investigado).

e) El muestreo por cuotas es un método muy utilizado en la práctica, si bien tampoco tiene base probabilística. El procesamiento consiste en hacer una muestra representativa de la población cuando ésta está estratificada según una serie de variables auxiliares (ejemplo, si se trata de personas se puede conocer la proporción de varones y de hembras, en el caso de una variable auxiliar "sexo"; también podría subdividirse cada estrato por la "edad" si el individuo tiene mayoría o no de edad, etc).

   Estas informaciones auxiliares permiten clasificar la población y dar la frecuencia relativa de cada clase. Obligando a que la muestra guarde esas mismas frecuencias relativas que existen en la población para las submuestras de cada clase, imponemos un criterio de representatividad. Los entrevistadores tienen asignados un número o cuota del tamaño muestral de cada clase y deben de hacer entrevistas hasta cumplimentar el tamaño previsto, eligiendo las unidades de un modo similar de selección a la ventura, lo que conlleva sus riesgos debidos a los posibles sesgos introducibles por el encuestador, como su horario o lugar de trabajo, etc.

Ninguno de éstos métodos tiene posibilidad de calcular el error de muestreo al no basarse en diseños aleatorios o probabilísticos.

§13.2 La investigación de mercados.

Generalmente darán lugar a series cronológicas y constituyen uno de los elementos básicos del análisis de la demanda, a dos niveles (macroeconómico –de la economía nacional– y el microeconómico –de empresas, comercial o de negocios–).

Tres campos principales de aplicaciones del muestreo a los problemas de mercado son:

a) Estructura del mercado (composición de la población por sexo, edades, clases socioeconómicas, etc.).

b) Preferencias del consumidor (conocimiento de hábitos, necesidades, gustos, experiencias, etc. sobre productos considerados por consumidores efectivos o potenciales).

c) Técnicas de venta (y trata de aprovechar conocimientos adquiridos en experiencias anteriores para lanzar al mercado ciertos productos en condiciones óptimas).

Capítulo 14

Tablas de números aleatorios.

§14.1 Presentación.

Las tablas de números aleatorios recogen dígitos del 0 al 9 obtenidos con equiprobabilidad cada uno, e independientemente cada uno de los restantes. Hemos recogido las tablas de los autores Royo y Ferrer (1955) obtenidas a partir de selecciones de la Lotería Nacional.

Ejercicio 14.1 *Por simplicidad, supongamos que existen* $N = 10.000$ *unidades en la población finita y debemos seleccionar una muestra de tamaño* $n = 20$ *por mas. Las unidades se consideran numeradas de 0001 a 9999 y 0000 (como unidad 10000). Determinar las unidades seleccionadas.*

Solución.

Como disponemos de las tablas de números aleatorios de Royo y Ferrer, cada dígito está seleccionado por un sorteo independiente entre los números 0,1, 2,... al 9. Tomando de 4 en 4 dígitos tendremos las unidades selecionadas, con la condición de que si se repiten dos grupos de cuatro números consecutivos debemos rechazar el último grupo ya que el muestreo es sin reemplazamiento, y continuar obteniendo grupos que no han aparecido antes.

Operando así, las unidades selecionadas, haciendo uso de las tablas de que disponemos, son

4266	5984	7689	7808	9416	3821	5103	9964	2004	3708
1309	4371	5017	5120	6376	5333	5717	1895	8872	1851

entre las que no han aparecido ningún grupo repetido.

Comentarios.

Si se nos hubiera pedido una muestra con diseño masr el procedimiento será análogo, sin rechazar grupos repetidos en el caso en que surgieran 2 ó mas iguales.

Si $N = 5000$ podríamos seleccionarlos así: los tres últimos dígitos serían como aparecen en grupos de 4 dígitos o cifras, y el primero será 0 si aparece 0 ó 5 en dicho lugar, 1 si 1 ó 6, 2 si 2 ó 7, 3 si 3 ó 8, 4 si 4 ó 9 y nunca aparecerán números en primer lugar superiores o iguales a 5 (salvo 5000 que es también el 0000).

Ejercicio 14.2 *Explicar cómo selecionar una muestra con diseño pptr de una población finita de tamaño N, haciendo uso de las tablas de números aleatorios.*

Solución.

Formaríamos la tabla siguiente, cuando el entero $x_i \alpha p_i$, siendo p_i la probabilidad de selección de la unidad i en cada selección con reemplazamiento.

Unidad	x_i		Intervalo acumulado		
1	Np_1	De	1	hasta	x_1
2	Np_2	"	x_1+1	"	x_1+x_2
3	Np_3	"	x_1+x_2+1	"	$x_1+x_2+x_3$
⋮	⋮		⋮		⋮
N	Np_N	"	$\sum_{i=1}^{N-1} x_i + 1$	"	$\sum_{i=1}^{N} x_i$

Se seleccionan al azar basándonos en tablas de números aleatorios las n unidades muestrales comprendidas entre 1 y $\sum_{i=1}^{N} x_i$ siendo los números x_i enteros positivos.

Si por ejemplo, el número aleatorio está comprendido entre $x_1 + x_2 + 1$ y $x_1 + x_2 + x_3$, tendríamos seleccionada en la muestra la unidad 3.

Del mismo modo se seleccionan sucesivamente las $n - 1$ restantes unidades, entre las cuales podría haber unidades repetidas ya que el diseño es con reemplazamiento.

Comentario.

Con diseño ppt habría sido análogo al diseño pptr, omitiendo unidades repetidas.

§14.2 Tablas.

Tablas Auxiliares de Estadística. (J. Royo y S. Ferrer, 1955). C.S.I.C.–Instituto de Investigaciones Estadísticas, Madrid.

4266 5984 7689 7808 9416 3821 5103 9964 2004 3708
1309 4371 5017 5120 6376 5333 5717 1895 8872 1851
3648 2534 9294 6591 3342 7830 3508 3922 4252 8671
7534 3116 6821 1719 5193 5793 0978 8909 1774 1198
6087 7806 5941 5365 2645 2542 4186 1228 8089 4913

5484 2607 0459 7898 4171 0501 6654 0616 0407 7483
9687 0543 8025 0546 7419 3311 4173 7511 1094 6382
4788 4882 9833 0127 1330 0248 3949 0921 1969 3749
1010 5374 7004 8872 6589 7121 6479 3969 5540 9720
9929 7293 3803 7098 1966 1513 3585 0767 5147 6985

0799 0712 2573 5850 4777 6153 0121 5981 5696 9890
9009 5594 5679 0592 7914 6248 2663 6192 2510 5942
0428 5595 1353 4217 4224 3856 7954 4673 8436 9047
5698 0637 6758 7151 9247 3544 8183 4994 2098 0886
3215 1232 4170 2250 5770 1794 9845 4191 8966 1344

0972 0099 9895 2543 8798 8504 3951 4964 8764 8671
5447 9226 9386 8680 6399 6851 6472 8926 2531 0694
8003 9292 5326 3624 7385 0934 7561 3027 5119 9116
1810 1710 1644 1072 0480 0305 4030 2966 0281 7608
5246 0754 4485 3242 2146 0225 7421 7452 3245 4655

1398 2715 1217 7028 9092 0822 4588 7869 4702 0912
7120 8333 2882 3424 3193 3489 9598 6899 1941 9192
5893 3493 5380 7803 2479 9297 1447 3718 3526 8181
9811 8271 5471 5791 8364 7815 7069 3500 8447 7502
0106 6478 5441 3153 5963 2516 5123 1027 2061 4919

4266 5984 7689 7808 9416 3821 5103 9964 2004 3708
0014 2400 8807 8234 7704 7447 8848 9253 1389 9730
4476 8685 8991 2053 0757 2088 0078 0816 2774 4423
1499 8808 6658 4776 2209 8664 0808 3907 5708 3583
3642 4322 0421 2567 4666 9243 0260 3819 8703 4820

2604 0666 4434 9694 8126 9095 4742 1667 2951 5524
8226 2441 6407 9259 6941 4872 9321 2933 3617 2744
7912 8940 7212 5749 3706 6140 1669 4840 1520 9489
3545 3899 0999 1167 6240 2367 9463 6382 8786 4393
6761 2116 2183 0876 9998 2811 9188 4871 9402 9893

1572 4289 9480 9946 8744 7448 3462 9797 2422 5354
9552 4028 0501 1307 7586 1182 2636 5555 2987 3974
8818 3900 8666 8174 2884 6548 4652 5549 4745 8701
7627 8247 3003 6805 9439 2860 6639 7124 8703 8899
0593 1953 3713 7968 1764 3920 1060 4058 3343 8767

6740 0197 7073 8388 8819 9576 3737 4873 4607 6238
2415 2455 5979 9641 8681 9683 9953 3313 3822 4158
7260 5564 8850 7685 8803 9103 1654 3388 1729 0699
6115 1183 4548 4242 1909 0036 7947 5330 7512 7499
1863 4545 3164 8991 8051 4004 7601 8211 1638 6019

1230 6697 5260 8017 4855 8952 2896 9537 6081 6531
3081 0860 4387 4660 1336 6250 3015 2829 6381 3600
2691 3500 4646 9446 0027 0956 0000 9743 0567 2569
1278 3730 6672 5396 6052 9000 8785 5920 4385 4879
1720 0553 5174 0812 5556 9589 8498 8074 0902 2536

Referencias

[1] Azorín, F. y Sánchez-Crespo, J.L. (1986). *Métodos y Aplicaciones del Muestreo.* Alianza, Madrid.

[2] Cochran, W.G. (1977). *Sampling Techniques* (3ª edición). Wiley, New York.

[3] Hansen, M.H. y Hurwitz, W.N. (1943). "On the theory of sampling from finite populations". *Ann. Math. Statist.* **14**, 333-362.

[4] Hansen, M.H. y Hurwitz, M.N. (1946). The problem of the non-response in sample surveys. *J. Amer. Statist. Assoc.* **41**,517-529.

[5] Hansen, M.H., Hurwitz, W.N., y Madow, W.G. (1953). *Sample Survey Methods and Theory* (Vol.II). Wiley, New York.

[6] Horvitz, D.G. y Thompson, D.J. (1952). "A generalisation of sampling without replacement from a finite universe". *J. Amer. Statist. Assoc.* **47**, 663-685.

[7] Mirás, J. (1985). *Elementos de Muestreo para Poblaciones Finitas.* I.N.E., Madrid.

[8] Mood, A.M., Graybill, F.A. y Boes, D.C. (1974). *Introduction to the Theory of Statistics.* (3ª edición). McGraw-Hill, Tokyo.

[9] Murthy, M.N. (1967). *Sampling Theory and Methods.* Statistical Publishing Society, Calcutta.

[10] Plane, D.R. y Gordon, K.R. (1982). "A simple proof of the nonaplicability of the Central Limit theorem to finite populations" *Amer. Statistician* **36**, 175-176.

[11] Raj, D. (1956). "Some estimators in sampling with varying probabilities without replacement". *J. Amer. Statist. Assoc.* **51**, 269-284.

[12] Royo, J. y Ferrer, S. (1955). *Tablas Auxiliares de Estadística.* C.S.I.C.–Instituto de Investigaciones Estadísticas, Madrid.

[13] Yates, F. y Grundy, P.M. (1953). "Selection without replacement from within strata with probability proportional to size" *J. Roy. Statist. Soc. B* **15**, 235-261.