Tema 7

Muestreo por Conglomerados

Tema 7. Introducción. Extracción de los conglomerados con probabilidades iguales y sin reemplazo. Extracción de los conglomerados con probabilidades desiguales. Muestreo por conglomerados combinado con estratificación. Muestreo bietápico. Teorema de Madow. Selección con probabilidades iguales en cada etapa. Selección con probabilidades desiguales en la primera etapa y probabilidades iguales en la segunda. Generalización a tres etapas. Método de los conglomerados últimos en muestreo polietápico

7.1. Introducción.

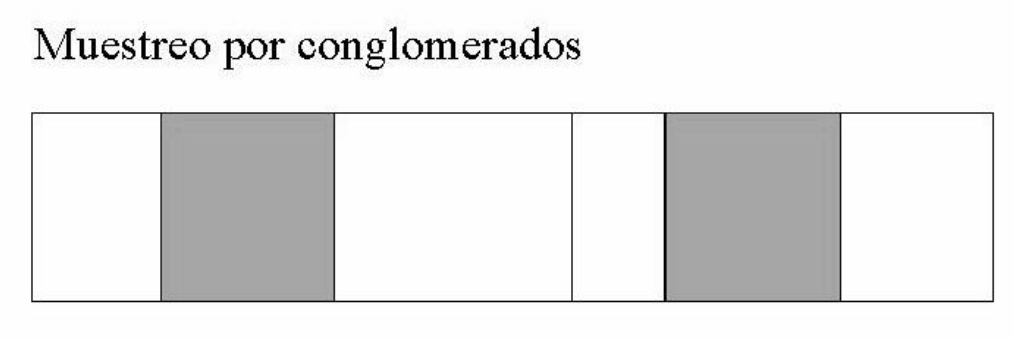
En los tipos de muestreo anteriormente considerados hemos supuesto que las unidades de muestreo eran las mismas que constituían el objeto de nuestro estudio. Vamos a ocuparnos ahora del caso más general, en que las unidades de muestreo comprenden dos o más unidades últimas. En este caso se dice que cada unidad de muestreo constituye un conglomerado de unidades últimas, y que el muestreo es por conglomerados.

Existen diversas razones para el empleo de conglomerados. La razón fundamental de su uso es que para efectuar un muestreo aleatorio simple o un muestreo estratificado, hace falta disponer de una lista de todos los elementos de la población, y en la práctica no suele disponerse de tales listas, salvo en casos particulares. Es preferible la división previa de la población en conglomerados de los cuales se selecciona cierto número, para lo cual sólo se necesita disponer de la lista de los conglomerados.

Otra razón importante para el uso de conglomerados es que este tipo de muestreo es menos costoso que el muestreo aleatorio simple, si el costo por obtener observaciones se incrementa con la distancia que separa los elementos. Así por

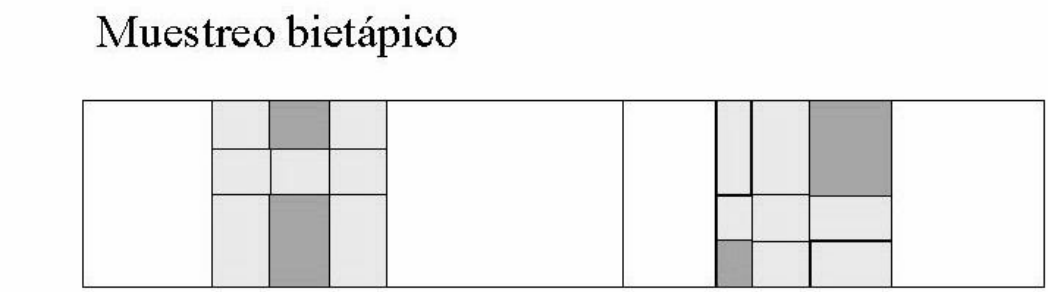
ejemplo si se selecciona una m.a.s. de hogares en una ciudad, el coste de realizar entrevistas en los hogares dispersos es muy grande debido al tiempo de transporte de los entrevistadores y otros gastos relacionados. El uso del muestreo por conglomerados es un método efectivo para reducir los gastos: los elementos dentro de un conglomerado deben estar geográficamente cerca uno de otro, y entonces los gastos se reducen.

Las secciones censales y las manzanas de una ciudad se usan frecuentemente como conglomerados de hogares en los estudios de mercado y en las propios estudios realizados por las oficinas de estadística de los distintos países. Los mismos hogares son usados como conglomerados de personas, los hospitales son conglomerados de pacientes,...

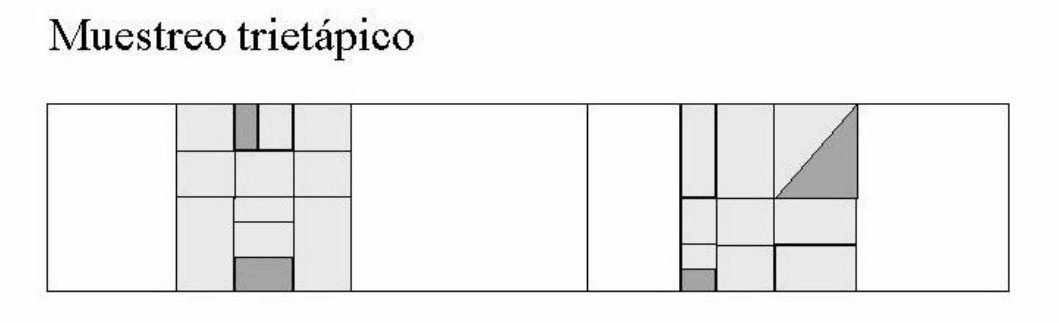


En el muestreo por conglomerados, una vez seleccionados algunos conglomerados, la muestra está formada por todas las unidades que componen el conglomerado. Por tanto para que la muestra sea representativa, los elementos que componen el conglomerado deben ser muy heterogéneos entre sí, frente al muestreo estratificado que busca la homogeneidad de las unidades dentro del estrato.

Si las unidades de un conglomerado seleccionado proporcionan resultados similares, parece antieconómico medirlos todos. Una práctica común es seleccionar y medir una muestra en cada conglomerado seleccionado. Esta técnica se conoce con el nombre de muestreo bietápico o submuestreo.



Esta situación puede extenderse a más etapas dando lugar al muestreo polietápico.



A continuación vamos a estudiar el muestreo por conglomerados monoetápico, llamado también simplemente muestreo por conglomerados. El estudio de este modelo proporcionará orientaciones útiles para la puesta en práctica de los muestreos polietápicos.

Existe una gran diversidad de situaciones en el muestreo por conglomerados: en ciertos casos los conglomerados son conocidos y de tamaños iguales, en otros de tamaños desiguales, y frecuentemente son desconocidos los tamaños.

Dada la dificultad de exponer una teoría general que sirva para todos los casos, nos vamos a restringir al estudio de aquellos más importantes. En adelante vamos a utilizar la siguiente notación:

N = número de conglomerados de la población.

n = número de conglomerados seleccionados en la muestra.

M_i = número de elementos en el conglomerado i, $i = 1, \ldots, N$.

$\overline{m} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} M_i$ = tamaño promedio del conglomerado en la muestra.

$M = \displaystyle\sum_{i=1}^{N} M_i$ = número de elementos de la población.

$\overline{M} = \dfrac{M}{N}$ = tamaño promedio del conglomerado en la población.

x_{ij} = valor de la variable en la unidad j del i-ésimo conglomerado.

$X_i = \displaystyle\sum_{j=1}^{M_i} X_{ij}$ = total de la variable en el i-ésimo conglomerado.

$X = \displaystyle\sum_{i=1}^{N} X_i$ = total poblacional de la variable .

$\overline{X}_i = \displaystyle\sum_{j=1}^{M_i} \frac{X_{ij}}{M_i}$ = media de la variable en el conglomerado i-ésimo.

$\overline{\overline{X}} = \dfrac{1}{M} X$ = media poblacional de la variable.

$$\delta = \frac{\dfrac{\displaystyle\sum_{i=1}^{N}\sum_{j\neq k}^{M_i} (x_{ij} - \overline{X})(x_{ik} - \overline{X})}{NM_i(M_i - 1)}}{\dfrac{\displaystyle\sum_{i=1}^{N}\sum_{j=1}^{M_i} (x_{ij} - \overline{X})^2}{M}} = \text{coeficiente de correlación intraconglomerados.}$$

$S_i^2 = \dfrac{1}{M_i - 1} \displaystyle\sum_{j=1}^{M_i} (x_{ij} - \overline{X}_i)^2$ = cuasivarianza en el conglomerado i-ésimo.

$S^2 = \dfrac{1}{M - 1} \displaystyle\sum_{i=1}^{N}\sum_{j=1}^{M_i} (x_{ij} - \overline{X})^2$ = cuasivarianza poblacional.

$S_b^2 = \dfrac{1}{N - 1} \displaystyle\sum_{i=1}^{N}\sum_{j=1}^{M_i} (\overline{X}_i - \overline{\overline{X}})^2$ = cuasivarianza entre las medias de los conglomerados.

Además para el estudio de proporciones utilizaremos:

A_{ij} = variable que toma el valor 1 ó 0 si la unidad j del conglomerado i-ésimo tiene o no una característica deseada.

$A_i = \displaystyle\sum_{j=1}^{M_i} A_{ij}$ = número de individuos en el conglomerado i-ésimo que tienen la característica deseada.

$P_i = \dfrac{A_i}{M_i}$ = proporción de individuos en el conglomerado i-ésimo que tienen la característica deseada.

$P = \dfrac{1}{M} \displaystyle\sum_{i=1}^{N} A_i$ = proporción poblacional de individuos con la característica

deseada.

7.2. Extracción de los conglomerados con probabilidades iguales y sin reemplazo.

7.2.1. Conglomerados del mismo tamaño

En este caso $M_i = \overline{M}\ \forall i$ y $M = N\overline{M}$.
Los estimadores del total, media y proporción poblacional son, respectivamente:

$$\widehat{X} = N\frac{1}{n}\sum_{i=1}^{n} X_i = N\overline{M}\sum_{i=1}^{n}\overline{X}_i$$

$$\overline{x} = \frac{\widehat{X}}{N\overline{M}} = \frac{1}{n}\sum_{i=1}^{n}\overline{X}_i$$

$$\widehat{P} = \frac{1}{n}\sum_{i=1}^{n} P_i = \frac{1}{n\overline{M}}\sum_{i=1}^{n} A_i$$

Veamos que son insesgados:

$$E(\overline{x}) = E(\frac{1}{n}\sum_{i\in s}\overline{X}_i) = E(\frac{1}{n}\sum_{i=1}^{N}\overline{X}_i e_i)$$

donde

$$e_i = \begin{cases} 1 & \text{si el conglomerado } i \in s \\ 0 & \text{si el conglomerado } i \notin s \end{cases}$$

Entonces

$$E(\overline{x}) = \frac{1}{n}\sum_{i=1}^{N}\overline{X}_i E(e_i) = \frac{1}{n}\sum_{i=1}^{N}\overline{X}_i\frac{n}{N} = \frac{1}{N}\sum_{i=1}^{N}\overline{X}_i = \overline{\overline{X}}$$

y por tanto el estimador $\overline{x}$ es insesgado al igual que los estimadores del total y proporción poblacional.
Las varianzas de los estimadores anteriores son:

$$V(\widehat{X}) = N^2\overline{M}^2 V(\overline{x})$$

$$V(\overline{x}) = \left(1 - \frac{n}{N}\right)\frac{S_b^2}{n\overline{M}}$$

$$V(\widehat{P}) = \left(1 - \frac{n}{N}\right) \frac{\sum_{i=1}^{N}\sum_{i=1}^{\overline{M}} (P_i - P)^2}{N-1}$$

Demostración.

$$V(\overline{x}) = V(\frac{1}{n}\sum_{i\in s}\overline{X}_i) = \left(1 - \frac{n}{N}\right) \frac{\sum_{i=1}^{N} (\overline{X}_i - \overline{\overline{X}})^2}{N-1} = \left(1 - \frac{n}{N}\right) \frac{S_b^2}{n\overline{M}}$$

puesto que

$$S_b^2 = \frac{\sum_{i=1}^{N}\sum_{i=1}^{\overline{M}} (X_i - \overline{\overline{X}})^2}{N-1}$$

Observamos que la expresión de la $V(\overline{x})$ es análoga a la del m.a.s. sustituyendo S^2 por $\frac{S_b^2}{\overline{\overline{M}}}$

Estas expresiones teóricas de las varianzas dependen de los valores desconocidos y se utilizan para comparar las precisiones de los estimadores respecto a otros esquemas de muestreo.

Estimaciones de las varianzas.

Puesto que:

$$\widehat{S}_b^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{\overline{M}} (\overline{X}_i - \overline{x})^2}{n-1}$$

es un estimador de S_b^2, se obtiene que los estimadores de las varianzas son:

$$\widehat{V}(\widehat{X}) = N^2\overline{M}^2\widehat{V}(\overline{x})$$

$$\widehat{V}(\overline{x}) = (1-f)\frac{\widehat{S}_b^2}{n\overline{M}} = \frac{1-f}{n}\frac{\sum_{i=1}^{n} (\overline{X}_i - \overline{x})^2}{n-1} =$$

$$= \frac{1-f}{n\overline{M}^2}\frac{\sum_{i=1}^{n} X_i^2 - n\overline{M}^2\overline{x}^2}{n-1}$$

$$\widehat{V}(\widehat{P}) = \frac{1-f}{n}\frac{\sum_{i=1}^{n}(P_i-\widehat{P})^2}{n-1} = \frac{1-f}{n}\frac{\sum_{i=1}^{n}P_i^2 - n\widehat{P}^2}{n-1}$$

Ejemplo 1
Un empresario quiere estimar el número de tubos de dentrífico usados por una familia al trimestre en una comunidad de 4000 hogares divididos en 400 bloques geográficos de 10 hogares cada uno. Se selecciona una muestra aleatoria simple de 4 bloques que proporciona los siguientes resultados:

Bloque	Número de tubos	Total
1	1,2,1,3,3,2,1,4,1,1	19
2	1,3,2,2,3,1,4,1,1,2	20
3	2,1,1,1,1,3,2,1,3,1	16
4	1,1,3,2,1,5,1,2,3,1	20

Estima el promedio de tubos gastados por familia y la varianza del estimador usado.
Resolución.-
Tomando cada zona geográfica como un conglomerado, se trataría de estimar una media en un muestreo por conglomerados con probabilidades iguales y sin reemplazo, con conglomerados del mismo tamaño, $M_i = \overline{M} = 10$. Así:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n}\overline{X}_i = \frac{1}{n\overline{M}}\sum_{i=1}^{n}\sum_{j=1}^{10}x_{ij} = \frac{19+20+16+20}{4\cdot 10} = 1{,}875,$$

$$\widehat{V}(\overline{x}) = \frac{N-n}{Nn\overline{M}^2}\frac{\sum_{1\le i\le 4}\left(X_i-\overline{x}\overline{M}\right)^2}{n-1} = \frac{N-n}{Nn\overline{M}^2}\frac{1}{n-1}\left(\sum_{1\le i\le n}X_i^2 - n\overline{M}^2\overline{x}^2\right) =$$

$$= \frac{(400-4)}{40\cdot 4\cdot 10^2}\frac{1}{3}10{,}75 = 0{,}0089.$$

7.2.2. Conglomerados de distinto tamaño

Supongamos que en general los tamaños de cada conglomerado M_i son distintos.
Estimadores.
Los estimadores del total, media y proporción poblacional son, respectivamente:

$$\widehat{X} = N\overline{x}_t = N\frac{\sum_{i=1}^{n} X_i}{n}$$

$$\overline{x} = \frac{\widehat{X}}{M}$$

$$\widehat{P} = N\frac{\sum_{i=1}^{n} A_i}{n}$$

con varianza:

$$V(\widehat{X}) = N^2\frac{N-n}{N-1}\frac{\frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X}_t)^2}{n}$$

donde:

$$\overline{X}_t = \frac{1}{N}\sum_{i=1}^{N} X_i$$

y con una estimación de la varianza dada por:

$$\widehat{V}(\widehat{X}_t) = N^2\frac{N-n}{Nn}\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{x}_t)^2 =$$

$$= N^2\frac{N-n}{Nn}\frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2\right)$$

Se pueden construir otros estimadores alternativos basados en la razón:

$$\widehat{X}_r = M\overline{x}_r = M\frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} M_i}$$

$$\overline{x}_r = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} M_i} \;;\; \widehat{P}_r = \frac{\sum_{i=1}^{n} A_i}{\sum_{i=1}^{n} M_i}$$

Las varianzas de los estimadores anteriores son:

$$V(\widehat{X}_r) = M^2 V(\overline{x}_r)$$

$$V(\overline{x}_r) = \frac{1-f}{n\overline{M}^2}\frac{1}{N-1}\sum_{i=1}^{N}(\overline{X}_i - \overline{\overline{X}})^2$$

$$V(\widehat{P}_r) = \frac{1-f}{n\overline{M}^2}\frac{1}{N-1}\sum_{i=1}^{N} M_i^2(P_i - P)^2$$

Las estimaciones de las varianzas son:

$$\widehat{V}(\overline{x}_r) = \frac{N-n}{Nn\overline{M}^2}\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{x}M_i)^2 =$$

$$= \frac{N-n}{Nn\overline{M}^2}\frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - 2\cdot\overline{x}\sum_{i=1}^{n} X_i M_i + \overline{x}^2\sum_{i=1}^{n} M_i^2\right)$$

$$\widehat{V}(\widehat{P}_r) = \frac{N-n}{Nn\overline{M}^2}\frac{1}{n-1}\left(\sum_{i=1}^{n} A_i^2 - 2\cdot\widehat{P}\sum_{i=1}^{n} A_i M_i + \widehat{P}^2\sum_{i=1}^{n} M_i^2\right)$$

$$\widehat{V}(\widehat{X}_r) = M^2\widehat{V}(\overline{x}_r)$$

Ejemplo 2
Un sociólogo quiere estimar el ingreso medio por persona en una ciudad en la que no hay una lista disponible de residentes. Para ello divide la ciudad en 415 bloques rectangulares. El sociólogo dispone de tiempo y dinero para muestrear 25 bloques y entrevistar todos los hogares del mismo, obteniendo los siguientes resultados:

Bloque (i)	1	2	3	4	5	6	7	8	9	10
Número de residentes (M_i)	8	12	4	5	6	6	7	5	8	3
Ingreso total (X_i)	96	121	42	65	52	40	75	65	45	50

11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	Total
2	6	5	10	9	3	6	5	5	4	6	8	7	3	8	151
85	43	54	49	53	50	32	22	45	37	51	30	39	47	41	1329

Estima el ingreso medio por persona y la varianza del estimador usado.
Resolución.-
Como en el anterior problema, se trataría de estimar la media en muestreo por conglomerados con la extracción de estos con probabilidades iguales y

sin reemplazo, siendo ahora los conglomerados de tamaños desiguales. Así, el estimador de la media será:

$$\overline{x}_r = \frac{\sum\limits_i X_i}{\sum\limits_i M_i} = \frac{1329}{151} = 8{,}80132,$$

y su error cuadrático medio estimado:

$$\widehat{ECM}(\overline{x}_r) = \frac{N-n}{Nn\overline{M}^2} \sum_{1\leq i\leq n} \frac{(X_i - \overline{x}M_i)^2}{n-1}.$$

$\overline{M}$ es desconocido por lo que debe ser estimado por:

$$\frac{\sum\limits_{1\leq i\leq n} M_i}{n} = \frac{151}{25} = 6{,}04$$

y

$$\sum_{1\leq i\leq n} (X_i - \overline{x}M_i)^2 = \sum_{1\leq i\leq n} X_i^2 - 2\overline{x} \sum_{1\leq i\leq n} X_i M_i + \overline{x}^2 \sum_{1\leq i\leq n} M_i^2 =$$

$$= 82039 - 2 \cdot 8{,}80132 \cdot 8403 + 8{,}80132^2 \cdot 1047 = 15228,$$

por lo que:

$$\widehat{ECM}(\overline{x}_r) = \frac{415-25}{415 \cdot 25 \cdot 6{,}04^2} \frac{15228}{24} = 0{,}653784.$$

7.3. Extracción de los conglomerados con probabilidades desiguales.

Consideramos que extraemos una muestra de n conglomerados con probabilidades desiguales con reemplazo. En cada extracción el conglomerado i se extrae con probabilidad p_i, siendo:

$$\sum_{i=1}^{N} p_i = 1$$

Según los resultados del muestreo con probabilidades desiguales antes estudiado se tienen los estimadores del total, media y proporción, respectivamente:
Estimadores.

$$\widehat{X} = \frac{1}{n} \sum_{i=1}^{n} \frac{X_i}{p_i}$$

$$\overline{x} = \frac{1}{M} \sum_{i=1}^{n} \frac{X_i}{np_i}$$

$$\widehat{P} = \frac{1}{M} \sum_{i=1}^{n} \frac{A_i}{np_i}$$

Varianzas.
Las varianzas de los estimadores anteriores son, respectivamente:

$$V(\widehat{X}) = \frac{1}{n} \sum_{i=1}^{N} p_i \left(\frac{X_i}{p_i} - X \right)^2$$

$$V(\overline{x}) = \frac{1}{M^2} V(\widehat{X})$$

$$V(\widehat{P}) = \frac{1}{M^2} \sum_{i=1}^{N} p_i \left(\frac{A_i}{p_i} - MP \right)^2$$

Estimaciones de las varianzas.

$$\widehat{V}(\widehat{X}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{X_i}{p_i} - \widehat{X} \right)^2$$

$$\widehat{V}(\overline{x}) = \frac{1}{M^2} \widehat{V}(\widehat{X})$$

$$\widehat{V}(\widehat{P}) = \frac{1}{M^2} \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{A_i}{p_i} - M\widehat{P} \right)^2$$

El caso más importante del muestreo con probabilidades distintas es el caso en que la probabilidad de cada extracción es elegida proporcionalmente al tamaño. En este caso, las fórmulas de estimadores y varianzas se deducen inmediatamente sustituyendo p_i por $\frac{M_i}{M}$.

7.4. Muestreo por conglomerados combinado con estratificación.

Para todos los casos antes considerados se concluye que son deseables los conglomerados de tamaño reducido y, sobre todo, de medias semejantes.
Esta última condición es difícil de respetar en la práctica. Para resolver este problema se procede a estratificar los conglomerados, metiendo en un mismo

estrato aquellos que sean más homogéneos. Dentro de cada estrato el muestreo por conglomerados producirá buenas estimaciones.
Denotaremos por $k = 1, \ldots, L$ el estrato considerado.
Estimadores.
Los estimadores de la media, total y proporción poblacional son, respectivamente:

$$\overline{x}_{RC} = \frac{\sum_{k=1}^{L} N_k \overline{x}_{kt}}{\sum_{k=1}^{L} N_k \overline{m}_k}$$

donde

$$\overline{x}_{kt} = \frac{\sum_{i=1}^{n_k} X_{ki}}{n_k}$$

es la media de los totales de los conglomerados en la muestra para el estrato k y

$$\overline{m}_k = \frac{\sum_{i=1}^{n_k} M_{ki}}{n_k}$$

$$\widehat{X}_{RC} = M\overline{x}_{RC}$$

$$\widehat{P}_{RC} = \frac{\sum_{k=1}^{L} N_k \overline{A}_{kt}}{\sum_{k=1}^{L} N_k \overline{m}_k}$$

con

$$\overline{A}_{kt} = \frac{\sum_{i=1}^{n_k} A_{ki}}{n_k}$$

Estimaciones de las varianzas.
Las estimaciones de las varianzas de los estimadores anteriores vienen dadas por las expresiones:

$$\widehat{V}(\overline{x}_{RC}) = \frac{1}{M^2} \sum_{k=1}^{L} \left[\frac{N_k(N_k - n_k)}{n_k(n_k - 1)} \sum_{i=1}^{n_k} [(X_{ki} - \overline{x}_{kt}) - \overline{x}_{RC}(M_{ki} - \overline{m}_k)]^2 \right]$$

$$\widehat{V}(\widehat{X}_{RC}) = M^2 \widehat{V}(\overline{x}_{RC})$$

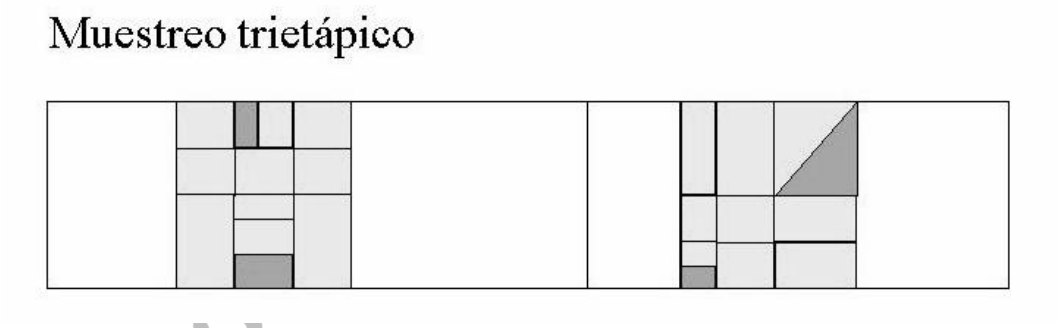
$$\widehat{V}(\widehat{P}_{RC}) = \frac{1}{M^2} \sum_{k=1}^{L} \left[\frac{N_k(N_k - n_k)}{n_k(n_k - 1)} \sum_{i=1}^{n_k} \left[(A_{ki} - \overline{A}_{kt}) - \widehat{P}_{RC}(M_{ki} - \overline{m}_k) \right]^2 \right]$$

7.5. Muestreo bietápico. Introducción

El muestreo por conglomerados en dos etapas es una extensión del concepto de muestreo por conglomerados. Un conglomerado puede tener demasiados elementos para poder ser investigados todos, o bien éstos son tan parecidos que la medición de unos pocos de ellos nos pueden dar una información bastante precisa del conglomerado completo. En esta situación es útil seleccionar una muestra de conglomerados y después seleccionar otra muestra de elementos dentro de los conglomerados de la primera muestra. Este procedimiento se conoce con el nombre de muestreo bietápico o muestreo con submuestreo.

file=cong2.eps, width=12cm

Cada conglomerado puede dividirse a su vez en conglomerados más pequeños formados por varias unidades últimas, dando lugar a un muestreo trietápico. Esta situación puede generalizarse a más etapas dando lugar al muestreo polietápico.



El muestreo bietápico se usa generalmente en las encuestas grandes que muestrean unidades poblacionales. La encuesta Gallup en EEUU selecciona distritos

electorales en una primera etapa, para luego seleccionar en una segunda etapa algunos hogares de estos distritos. La Encuesta de la Población Activa se realiza también mediante un muestreo polietápico en cada ciudad (estrato).
Vamos a comenzar con el estudio del caso más sencillo: el muestreo bietápico. Supongamos que la población está dividida en N conglomerados, llamados unidades primarias. Un plan de muestreo bietápico define:

- el modo de selección de una muestra de unidades primarias
- en cada unidad primaria extraída, el modo de selección de una muestra de unidades de la población llamadas unidades secundarias.

En vista de esto, existen muchos tipos de muestreos bietápicos, dependiendo de los métodos de selección que se utilicen en cada etapa: m.a.s., muestreo con probabilidades desiguales, muestreo estratificado, etc.
A continuación vamos a considerar los modelos teóricos más sencillos que servirán como base para la construcción de modelos más complejos. Previo a este estudio hemos de presentar un resultado importante que nos servirá para el cálculo de los momentos de los estimadores.

7.6. Teorema de Madow

Supongamos la población dividida en N unidades primarias de las cuales extraemos una muestra de tamaño n. Dentro de cada unidad primaria i extraemos una muestra de tamaño m_i. Tenemos pues dos conjuntos de unidades de muestreo cuya selección a su vez origina dos tipos de variación: el debido al submuestreo dentro de un conjunto fijo de unidades primarias, que representamos por el subíndice 2, y el correspondiente al muestreo de unidades primarias que denotaremos por el subíndice 1.
Con esta notación la esperanza de un estimador será igual a:

$$E(\widehat{\theta}) = E_1 E_2(\widehat{\theta}) = E_1(E_2(\widehat{\theta}/n))$$

que es la esperanza, sobre todas la muestras posibles de n unidades primarias, de la esperanza, condicionada a un conjunto fijo de n unidades primarias, sobre todas las submuestras posibles dentro de dicho conjunto. Es inmediato pues que:

$$V(\widehat{\theta}) = E_1 E_2(\widehat{\theta} - \theta)^2$$

Para su cálculo en la práctica se utiliza el siguiente resultado:
Teorema de Madow.

La varianza incondicional de un estimador es igual a la esperanza de la varianza condicional más la varianza de la esperanza condicional:

$$V(\widehat{\theta}) = E_1(V_2(\widehat{\theta})) + V_1(E_2(\widehat{\theta}))$$

Demostración

$$V(\widehat{\theta}) = E_1E_2(\widehat{\theta} - \theta)^2 = E_1(E_2(\widehat{\theta^2} + \theta^2 - 2\theta\widehat{\theta})) =$$

$$= E_1V_2(\widehat{\theta}) + E_1E_2^2(\widehat{\theta}) + \theta^2 - 2\theta E_1E_2(\widehat{\theta}) = E_1V_2(\widehat{\theta}) + E_1E_2^2(\widehat{\theta}) + \theta^2 - 2\theta^2 =$$

$$= E_1V_2(\widehat{\theta}) + E_1E_2^2(\widehat{\theta}) - \theta^2 = E_1V_2(\widehat{\theta}) + E_1E_2^2(\widehat{\theta}) - (E_1E_2(\widehat{\theta}))^2 = E_1V_2(\widehat{\theta}) + V_1E_2(\widehat{\theta})$$

7.7. Selección con probabilidades iguales en cada etapa

Seleccionamos en primer lugar n conglomerados, y dentro de cada conglomerado (de tamaño M_i) seleccionamos una muestra de tamaño m_i, siendo todas las extracciones mediante m.a.s.
Denotamos:

$$\overline{x}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} x_{ij}$$

7.7.1. Tamaños iguales de los conglomerados

En este caso $M_i = \overline{M}$ y $m_i = \overline{m}$.
Estimadores
Los estimadores de la media, total y proporción poblacional son, respectivamente:

$$\overline{\overline{x}} = \frac{1}{n}\sum_{i=1}^{n} \overline{x}_i,$$

$$\widehat{X} = M\overline{\overline{x}}, \text{ y}$$

$$\widehat{P} = \frac{1}{n}\sum_{i=1}^{n} p_i,$$

donde:

$$p_i = \frac{1}{m_i}\sum_{j=1}^{m_i} A_{ij}$$

Comprobemos que estos estimadores son insesgados:

$$E(\overline{\overline{x}}) = E_1E_2(\overline{\overline{x}}) = E_1(E_2(\frac{1}{n}\sum_{1\leq i\leq n}\overline{x}_i) =$$

$$= E_1(\frac{1}{n}\sum_{1\leq i\leq n}E_2(\overline{x}_i)) = E_1(\frac{1}{n}\sum_{1\leq i\leq n}\overline{X}_i) = \frac{1}{n}E_1(\sum_{1\leq i\leq N}\overline{X}_ie_i) =$$

$$= \frac{1}{n}(\sum_{1\leq i\leq N}\overline{X}_iE(e_i)) = \frac{1}{n}\sum_{1\leq i\leq N}\overline{X}_i\frac{n}{N} = \overline{\overline{X}}$$

Análogamente se comprobaría para los otros dos estimadores.

Varianzas de los estimadores

Para el cálculo de las varianzas utilizamos el teorema de Madow. Calculamos los dos términos de la varianza:

$$E_1V_2(\overline{\overline{x}}) = E_1V_2(\frac{1}{n}\sum_{1\leq i\leq n}\overline{x}_i) = E_1(\frac{1}{n^2}\sum_{1\leq i\leq n}V_2(\overline{x}_i)) =$$

$$= E_1(\frac{1}{n^2}\sum_{1\leq i\leq n}\frac{1-f_2}{\overline{m}}\frac{\sum_{1\leq j\leq\overline{M}}(x_{ij}-\overline{X}_i)^2}{\overline{M}-1}) =$$

$$= \sum_{1\leq i\leq N}\frac{1-f_2}{\overline{m}n^2(\overline{M}-1)}(\sum_{1\leq j\leq\overline{M}}(x_{ij}-\overline{X}_i)^2)\frac{n}{N} = (1-f_2)\frac{S_w^2}{n\overline{m}}$$

donde $f_2 = \frac{\overline{m}}{\overline{M}}$ es la fracción de muestreo en la segunda etapa.

Por otra parte:

$$V_1E_2(\overline{\overline{x}}) = V_1(\frac{1}{n}\sum_{1\leq i\leq n}\overline{X}_i) = (1-f_1)\frac{\sum_{1\leq i\leq N}(\overline{X}_i-\overline{\overline{X}})^2}{n(N-1)} =$$

$$= (1-f_1)\frac{S_b^2}{n\overline{M}}$$

donde

$$S_b^2 = \frac{\sum_{i=1}^{N}\sum_{i=1}^{\overline{M}}(X_i-\overline{\overline{X}})^2}{N-1}$$

$$S_w^2 = \frac{\sum_i^N \sum_j^{\overline{M}} (x_{ij} - \overline{X}_i)^2}{N(\overline{M} - 1)} = \frac{1}{N} \sum S_i^2$$

siendo S_i^2 la cuasivarianza del conglomerado i.
Sumando ambas componentes se llega a la expresión

$$V(\overline{\overline{x}}) = (1 - f_1)\frac{S_b^2}{n\overline{M}} + (1 - f_2)\frac{S_w^2}{n\overline{m}}$$

Estimaciones de las varianzas

$$\widehat{\widehat{V}}\left(\overline{\overline{x}}\right) = (1 - f_1)\frac{\widehat{S}_1^2}{n} + f_1(1 - f_2)\frac{\widehat{S}_2^2}{n\overline{m}},$$

$$\widehat{\widehat{V}}(\widehat{X}) = M^2 \widehat{\widehat{V}}\left(\overline{\overline{x}}\right), \text{ y}$$

$$\widehat{\widehat{V}}(\widehat{P}) = \frac{1 - f_1}{n} \sum_{i=1}^{n} \frac{(p_i - \widehat{P})^2}{n - 1} + \frac{f_1(1 - f_2)}{n^2(\overline{m} - 1)} \sum_{i=1}^{n} p_i q_i,$$

donde:

$$f_1 = \frac{N - n}{N}, \quad f_2 = \frac{\overline{M} - \overline{m}}{\overline{M}},$$

$$\widehat{S}_1^2 = \frac{1}{n - 1} \sum_{i=1}^{n} (\overline{x}_i - \overline{\overline{x}})^2 \quad \text{y} \quad \widehat{S}_2^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{\overline{m}} \frac{(x_{ij} - \overline{x}_i)^2}{\overline{m} - 1}.$$

En ocasiones es deseable que todas las unidades últimas tengan la misma probabilidad de pertenecer a la muestra. Para ello se toma el tamaño de cada submuestra proporcional al tamaño del conglomerado del que se selecciona. Así esta probabilidad viene dada por $\pi_i = \frac{n}{N}\frac{m_i}{M_i} = cte$. En este caso diremos que el muestreo es autoponderado.

7.7.2. Tamaños desiguales de los conglomerados

En el apartado anterior hemos considerado el caso de que los conglomerados eran de igual tamaño. Sin embargo es muy corriente encontrar situaciones en las que las unidades de muestreo de una cierta etapa no contienen un número fijo de unidades de la etapa siguiente. Así ocurriría, por ejemplo, si los conglomerados fueran edificios de casa, en las cuales el número de casas o el número de familias que viven en los edificios son en general distintos. Vamos pues a estudiar este caso general. Cada unidad primaria supondremos tiene tamaño M_i y vamos a extraer una m.a.s. de tamaño m_i en cada una.

Estimadores

Los estimadores vienen dados por las expresiones:

$$\widehat{\overline{\overline{X}}} = N \sum_{i=1}^{n} \frac{M_i \overline{x}_i}{n}.$$

$$\overline{\overline{x}} = \frac{N}{M} \sum_{i=1}^{n} \frac{M_i \overline{x}_i}{n}, \text{ y}$$

$$\widehat{\overline{\overline{P}}} = \frac{N}{M} \sum_{i=1}^{n} \frac{M_i p_i}{n}.$$

También se pueden proponer los siguientes estimadores basados en la razón:

$$\overline{\overline{x}}_r = \frac{\sum_{i=1}^{n} M_i \overline{x}_i}{\sum_{i=1}^{n} M_i} \quad \text{y} \quad \widehat{P}_r = \frac{\sum_{i=1}^{n} M_i p_i}{\sum_{i=1}^{n} M_i}.$$

$$\widehat{X}_r = M \overline{\overline{x}}_r$$

Estimaciones de las varianzas o de los errores cuadráticos medios

Para el total:

$$\widehat{V}\left(\widehat{\overline{\overline{X}}}\right) = \frac{N-n}{N} \frac{N^2}{n} \widehat{S}_b^2 + \frac{N}{n} \sum_{i=1}^{n} M_i^2 \frac{M_i - m_i}{M_i} \frac{\widehat{S}_i^2}{m_i},$$

donde:

$$\widehat{S}_b^2 = \frac{1}{n-1} \sum_{i=1}^{n} (M_i \overline{x}_i - \overline{M}\overline{\overline{x}})^2, \text{ y}$$

$$\widehat{S}_i^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} (x_{ij} - \overline{x}_i)^2.$$

$$\widehat{V}\left(\overline{\overline{x}}\right) = \frac{N-n}{N} \frac{1}{n\overline{M}^2} \widehat{S}_b^2 + \frac{1}{nN\overline{M}^2} \sum_{i=1}^{n} M_i^2 \frac{M_i - m_i}{M_i} \frac{\widehat{S}_i^2}{m_i}, \text{ y}$$

$$\widehat{V}\left(\widehat{\overline{\overline{P}}}\right) = \left(\frac{N-n}{N}\right) \frac{1}{n\overline{M}^2} \frac{1}{n-1} \sum_{i=1}^{n} \left(M_i p_i - \overline{M}\widehat{\overline{\overline{P}}}\right)^2 +$$

$$+ \frac{1}{nN\overline{M}^2} \sum_{i=1}^{n} \frac{M_i - m_i}{M_i} \frac{p_i q_i}{m_i - 1}.$$

y

$$\widehat{ECM}\left(\overline{\overline{x}}_r\right) = \frac{N-n}{N}\frac{1}{n\overline{M}^2}\widehat{S}_t^2 + \frac{1}{nN\overline{M}^2}\sum_{i=1}^{n} M_i^2\frac{M_i-m_i}{M_i}\frac{\widehat{S}_i^2}{m_i},$$

donde:

$$\widehat{S}_t^2 = \sum_{i=1}^{n}\frac{M_i^2(\overline{x}_i-\overline{\overline{x}}_t)^2}{n-1}, \text{ y}$$

$$\widehat{ECM}(\widehat{P}_r) = \frac{N-n}{N}\frac{1}{n\overline{M}^2}\widehat{S}_{pt}^2 + \frac{1}{nN\overline{M}^2}\sum_{i=1}^{n} M_i^2\frac{M_i-m_i}{M_i}\frac{p_iq_i}{m_i-1},$$

donde:

$$\widehat{S}_{pt}^2 = \frac{1}{n-1}\sum_{i=1}^{n} M_i^2(p_i-\widehat{p})^2.$$

$$\widehat{ECM}(\widehat{X}_r) = \widehat{ECM}\left(\overline{\overline{x}}_r\right)/M^2$$

Ejemplo 3
Se quiere estimar el número total de árboles afectados por una enfermedad en una provincia de 10 municipios. Se utiliza un muestreo de 4 municipios con probabilidades iguales y, dentro de cada municipio, se seleccionan sólo seis zonas con árboles, aleatoriamente.
Los resultados fueron:

Municipio	Zonas	Zonas muestreadas	Árboles enfermos
1	12	6	15, 14, 21, 13, 9, 10
2	15	6	4, 6, 10, 9, 8, 5
3	14	6	10, 11, 14, 10, 9, 5
4	21	6	8, 3, 4, 1, 2, 5

Estima el número total de árboles enfermos y establece el error de estimación.
Resolución.-
Se trata de un muestreo por conglomerados con submuestreo y probabilidades iguales en cada etapa, con $N = 10$, $n = 4$ y $m_i = 6$. Un estimador para el total X viene dado por la expresión:

$$\widehat{\widehat{X}} = N\sum_{i=1}^{n}\frac{M_i\overline{x}_i}{n}.$$

Calculamos en primer lugar las medias muestrales en cada conglomerado:

$\overline{x}_1$	$\overline{x}_2$	$\overline{x}_2$	$\overline{x}_4$
13,666	7	9,83	3,833

Entonces

$$\widehat{\widehat{X}} = \frac{10}{4}\left(12 \cdot 13{,}666 + 15 \cdot 7 + 14 \cdot 9{,}83 + 21 \cdot 3{,}833\right) = 1217{,}76,$$

con lo que el valor estimado sería de 1218 árboles enfermos.
La precisión del estimador se calcularía determinando su varianza estimada:

$$\widehat{V}(\widehat{\widehat{X}}) = \frac{N-n}{N}\frac{N^2}{n}\widehat{S}_b^2 + \frac{N}{n}\sum_{i=1}^{n} M_i^2 \frac{M_i - m_i}{M_i}\frac{\widehat{S}_i^2}{m_i},$$

siendo

$$\widehat{S}_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}(M_i\overline{x}_i - \overline{M}\overline{\overline{x}})^2, \text{ y}$$

$$\widehat{S}_i^2 = \frac{1}{m_i - 1}\sum_{i=1}^{m_i}(x_{ij} - \overline{x}_i)^2.$$

Las cuasivarianzas en cada una de las muestras obtenidas de los conglomerados son:

$\widehat{S}_1^2$	$\widehat{S}_2^2$	$\widehat{S}_3^2$	$\widehat{S}_4^2$
18,26	5,6	8,56	6,16

,

y así:

$$\widehat{S}_b^2 = \frac{1}{3}\left(\sum_{i=1}^{n} M_i^2\overline{x}_i^2 + \overline{M}^2\overline{\overline{x}}^2 - 2\overline{M}\overline{\overline{x}}\sum_{i=1}^{n} M_i\overline{x}_i\right) = \frac{1}{3}\left(63336{,}8 - \frac{9279{,}28}{4}\right) = 20339.$$

Por tanto

$$\widehat{V}(\widehat{\widehat{X}}) = \frac{6}{10}\frac{10^2}{4}20339 + \frac{10}{4}828{,}307 = 307156,$$

y el error de muestreo estimado

$$\widehat{\text{em}}(\widehat{\widehat{X}}) = \sqrt{307156} = 554{,}217.$$

7.8. Selección de las unidades con probabilidades desiguales en la primera etapa y con probabilidades iguales en la segunda

Supongamos que se seleccionan las unidades primarias con probabilidades desiguales y con reemplazo. La submuestra de m_i unidades se selecciona en cada conglomerado con un muestreo aleatorio simple.
Sea $z_i,\ i = 1, \ldots, N$ la probabilidad de seleccionar el i-ésimo conglomerado:

$$z_i > 0 \qquad \sum_{i=1}^{N} z_i = 1.$$

Estimadores
Los estimadores insesgados son:

$$\widehat{X} = \frac{1}{n} \sum_{i=1}^{n} \frac{M_i \overline{x}_i}{z_i},$$

$$\overline{\overline{x}} = \frac{1}{nM} \sum_{i=1}^{n} \frac{M_i \overline{x}_i}{z_i}, \text{ y}$$

$$\widehat{P} = \frac{1}{nM} \sum_{i=1}^{n} M_i \frac{p_i}{z_i}.$$

Estimaciones de las varianzas

$$\widehat{V}(\widehat{X}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{M_i \overline{x}_i}{z_i} - \widehat{X} \right)^2,$$

$$\widehat{V}\left(\overline{\overline{x}}\right) = \frac{1}{n(n-1)M^2} \sum_{i=1}^{n} \left(\frac{M_i \overline{x}_i}{z_i} - \widehat{X} \right)^2, \text{ y}$$

$$\widehat{V}(\widehat{P}) = \frac{1}{n(n-1)M^2} \sum_{i=1}^{n} \left(\frac{M_i p_i}{z_i} - M\widehat{P} \right)^2.$$

Para el caso particular de que las probabilidades sean proporcionales al tamaño del conglomerado $z_i = \dfrac{M_i}{M}$, se obtienen las fórmulas:
Estimadores

$$\widehat{X}_{ppt} = \frac{M}{n} \sum_{i=1}^{n} \overline{x}_i,$$

$$\overline{\overline{x}}_{ppt} = \frac{1}{n}\sum_{i=1}^{n} \overline{x}_i, \text{ y}$$

$$\widehat{P}_{ppt} = \frac{1}{n}\sum_{i=1}^{n} p_i.$$

Estimaciones de las varianzas

$$\widehat{V}(\widehat{X}_{ppt}) = M^2 \widehat{V}(\overline{\overline{x}}_{ppt}),$$

$$\widehat{V}\left(\overline{\overline{x}}_{ppt}\right) = \frac{1}{n(n-1)}\sum_{i=1}^{n} (\overline{x}_i - \overline{\overline{x}}_{ppt})^2, \text{ y}$$

$$\widehat{V}(\widehat{P}_{ppt}) = \frac{1}{n(n-1)}\sum_{i=1}^{n} (p_i - \widehat{P}_{ppt})^2.$$

Ejemplo 4
El gerente de una empresa quiere estimar el número medio de ausencias por empleado en el trimestre pasado. La empresa tiene 8 divisiones, con diferente número de empleados por división. El gerente decide muestrear 3 divisiones con probabilidades proporcionales al tamaño. Las divisiones seleccionadas tienen tamaños 2100, 1910 y 3200, respectivamente, y el número total de dias de ausencia en cada división fue 4320, 4160 y 5790, respectivamente. Estima el nmero promedio de das de ausencia por persona de toda la empresa.

Resolución.-
Es un muestreo con probabilidades desiguales donde $p_i = \dfrac{M_i}{M}$, para el que se obtiene:

$$\overline{x}_{ppt} = \frac{1}{n}\sum_{1\le i\le n} \frac{X_i}{M_i} =$$

$$= \frac{1}{3}\left[\frac{4320}{2100} + \frac{4160}{1910} + \frac{5709}{3200}\right] = 2{,}02$$

$$\widehat{V}\left(\overline{x}_{ppt}\right) = \frac{1}{n\,(n-1)}\sum_{1\le i\le n}\left(\frac{X_i}{M_i} - \overline{x}_{ppt}\right)^2 =$$

$$= \frac{1}{3\cdot 2}\left[(2{,}06-2{,}02)^2 + (2{,}18-2{,}02)^2 + (1{,}81-2{,}02)^2\right] = 0{,}0119$$

7.9. Generalización a tres etapas

Como ya hemos dicho anteriormente, un conglomerado puede subdividirse a su vez en otros conglomerados más pequeños, los cuales están constituidos por varias unidades últimas. Tenemos pues tres unidades distintas: las unidades primarias (que son los conglomerados), las unidades secundarias (que son los subconglomerados) y las unidades terciarias que componen las unidades poblacionales. Si seleccionamos en primer lugar algunos conglomerados, y dentro de éstos se selecciona una muestra de subconglomerados, para terminar seleccionando algunas unidades elementales de éstos, tenemos un muestreo trietápico. El paso del muestreo bietápico al trietápico no lleva más complejidad que la debida a la mayor complejidad de la notación. Hay que tener en cuenta que hay en este caso tres niveles de aleatorización debidos a la selección aleatoria de cada tipo de unidades.

Puesto que en cada etapa de selección se puede elegir un tipo de muestreo distinto, y además puede haber distintas situaciones (que se conozcan o no los tamaños de las unidades, que tengan o no igual tamaño, etc.) la gama de muestreos trietápicos es muy amplia. A continuación vamos a considerar el caso más simple en que la selección se realice mediante m.a.s. en cada etapa y los conglomerados sean todos del mismo tamaño.

Estimadores y varianzas

Supongamos la población dividida en n unidades primarias todas de igual tamaño, de las que extraemos una m.a.s. de n unidades. Dentro de cada unidad primaria extraemos una m.a.s. de m unidades. Por último seleccionamos una m.a.s. de t unidades terciarias entre las T que componen cada unidad secundaria.

Entonces si x_{ijk} es el valor de la variable para la unidad k del j-ésimo subconglomerado del conglomerado i, tenemos los siguientes estimadores y varianzas:

$$\overline{\overline{\overline{x}}} = \frac{1}{nmt} \sum_{1\le i\le n} \sum_{1\le j\le m} \sum_{1\le k\le t} x_{ijk}$$

$$V(\overline{\overline{\overline{x}}}) = (1-f_1)\frac{S_1^2}{n} + (1-f_2)\frac{S_w^2}{nm} + (1-f_3)\frac{S_{ww}^2}{nmt}$$

donde

$$S_{ww}^2 = \sum_{1\le i\le N} \sum_{1\le j\le M} \sum_{1\le k\le T} \frac{(x_{ijk} - \overline{X}_{ij})^2}{NM(T-1)}$$

y un estimador insesgado de la varianza viene dado por

$$\widehat{V}(\overline{\overline{\overline{x}}}) = (1-f_1)\frac{\widehat{S}_1^2}{n} + f_1(1-f_2)\frac{\widehat{S}_w^2}{nm} + f_1 f_2(1-f_3)\frac{\widehat{S}_{ww}^2}{nmt}$$

con

$$\widehat{S}_1^2 = \frac{1}{n-1} \sum_{1 \le i \le n} (\overline{\overline{x}}_i - \overline{\overline{\overline{x}}})^2$$

$$\widehat{S}_w^2 = \frac{1}{(m-1)n} \sum_{1 \le i \le n} \sum_{1 \le j \le m} (\overline{x}_{ij} - \overline{\overline{x}}_i)^2$$

$$\widehat{S}_{ww}^2 = \sum_{1 \le i \le n} \sum_{1 \le j \le m} \sum_{1 \le k \le t} \frac{(x_{ijk} - \overline{x}_{ij})^2}{nm(t-1)}$$

7.10. Método de los conglomerados últimos en muestreo polietápico

Este método ofrece una forma alternativa y sencilla para estimar la varianza de los estimadores usados en el muestreo por conglomerados en dos o más etapas, cuando las unidades primarias se extraigan con probabilidades iguales.
Al conjunto de unidades finales (secundarias en el caso de muestreo por conglomerados con submuestreo) que pertenecen a una unidad primaria (conglomerado) se le llama *conglomerado último*.
Dadas n unidades primarias y n estimadores insesgados, $\theta_1, \ldots, \theta_n$, del parámetro de interés, θ, se puede conseguir un estimador insesgado de θ mediante

$$\widehat{\widehat{\theta}} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\theta}_i$$

con varianza

$$V(\widehat{\widehat{\theta}}) = \frac{1}{n} \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{\theta}_i - \theta \right)^2$$

y una estimación insesgada de esta varianza (aproximadamente si no hay reemplazo) dada por

$$\widehat{V}(\widehat{\widehat{\theta}}) = \frac{1}{n} \frac{1}{n-1} \sum_{i=1}^{n} \left(\widehat{\theta}_i - \widehat{\widehat{\theta}} \right)^2 .$$

Así, por ejemplo, cuando se quiere estimar una media, el estimador en muestreo bietápico con selección con probabilidades iguales en cada etapa es

$$\overline{\overline{x}} = \frac{1}{n} \sum_{i=1}^{n} \overline{x}_i \quad \text{con} \quad \overline{x}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} x_{ij},$$

con lo que tomando $\widehat{\theta}_i = \overline{x}_i$ obtendríamos la estimación

$$\widehat{V}(\overline{\overline{x}}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\overline{x}_i - \overline{\overline{x}}\right)^2,$$

para el caso de ser los tamaños de los conglomerados iguales.
En el caso de ser desiguales (y en el supuesto de que M sea conocido),

$$\overline{\overline{x}} = \frac{1}{n} \sum_{i=1}^{n} \frac{N}{M} M_i \overline{x}_i,$$

con lo que llamando $\theta_i = \frac{N}{M} M_i \overline{x}_i$ obtendríamos la estimación

$$\widehat{V}(\overline{\overline{x}}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{N}{M} M_i \overline{x}_i - \overline{\overline{x}}\right)^2.$$

De forma análoga se hace para el total, $\theta = X$, tomando $\widehat{\theta}_i = N M_i \overline{x}_i$.

Ejemplo 5
En una población de $N = 10$ conglomerados se efectúa un muestreo en dos etapas, obteniéndose los siguientes resultados:

Unidades primarias de la muestra	tamao M_i	valores observados $x_{ij}, \quad (m_i) = 5$
1	50	8, 6, 12, 14, 10
2	60	8, 10, 14, 14, 16
3	80	8, 10, 10, 16, 12

Sabiendo que la suma de los tamaños es $M = 600$, construye un estimador del total X en el caso en que la selección se realice mediante un muestreo con probabilidades iguales en cada etapa. Estima el error de muestreo por el mtodo de los conglomerados últimos.
Resolución.
Si el muestreo es con probabilidades iguales, un estimador del total viene dado por

$$\widehat{\widehat{X}} = N \sum_{i=1}^{n} \frac{M_i \overline{x}_i}{n} = 64200,$$

pues $\overline{x}_1 = 10$,$\overline{x}_2 = 12{,}4$ y $\overline{x}_3 = 11{,}2$.

Mediante el mtodo de los conglomerados últimos, consideramos las unidades últimas pertenecientes a cada conglomerado y calculamos los estimadores del total obtenidos a partir de cada conglomerado, $\widehat{\theta}_i = NM_i\overline{x}_i$, con lo que $\widehat{\widehat{X}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\theta}_i$.
Entonces,

$$\widehat{V}(\widehat{\widehat{X}}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\widehat{\theta}_i - \widehat{\widehat{X}})^2 =$$

$$= \frac{1}{n(n-1)}\left(\sum_{i=1}^{n}\widehat{\theta}_i^2 - n\widehat{\widehat{X}}^2\right) = \frac{7981870}{6} = 1330311{,}6,$$

pues $\widehat{\theta}_1^2 = 25\cdot 10^6$, $\widehat{\theta}_2^2 = 55{,}3536\cdot 10^6$, y $\widehat{\theta}_3^2 = 80{,}2816\cdot 10^6$ y por tanto $\widehat{\mathrm{em}} = 1153{,}3913$.