CUBO, Vol.3. Octubre 1987
Págs.56-70

FORMULACION METODOLOGICA DE LA TEORIA DEL MUESTREO DE CONGLOMERADOS DE TAMAÑOS DESIGUALES

Manuel López R. *

RESUMEN:

El tema central de esta comunicación trata sobre Muestreo de Conglomerados Desiguales, seleccionados con probabilidades iguales y con probabilidades proporcionales a los tamaños de los mismos.

* Depto. Matemática y Estadística Universidad de la Frontera.

En la teoría sobre muestreo por conglomerados se distinguen dos situaciones:

a) El muestreo de conglomerados iguales, que se refiere al caso en que cada uno de los conglomerados tiene el mismo número de elementos, y;

b) El muestreo de conglomerados desiguales, donde los conglomerados difieren entre sí, en cuanto al número de elementos.

En este estudio se presenta un desarrollo sistemático de los aspectos principales de la Teoría de Muestreo por conglomerados de tamaños desiguales.

El desarrollo propiamente tal, se hizo en base a una metodología que se basa en las siguientes etapas:

i) Especificación de la población a estudiar;
ii) Definición de los parámetros a estimar;
iii) Elección del plan de muestreo;
iv) Proposición de estimadores para los parámetros;
v) Propiedades de los estimadores;
vi) Definición de las medidas de variabilidad de los estimadores;
vii) Proposición de los estimadores para las medidas de variabilidad.

En primer lugar, se trata el muestreo de conglomerados con probabilidades iguales de selección y a continuación, se presenta el muestreo de conglomerados seleccionados con probabilidades desiguales con reempla

zo y sin reemplazo.

En la selección con probabilidades desiguales se exponen los métodos siguientes:

a) El Método de Hansen y Hurwitz
b) El Método de Lahiri
c) El Método de Horvitz y Thompson
e) El Método de Murthy
f) Métodos relacionados con el muestreo sistemático.
g) El Método de Rao, Hartley y Cochran

En el estudio sobre los diferentes métodos mencionados anteriormente, se hace un exhaustivo desarrollo matemático, fundamentando cada uno de los resultados obtenidos. Además, se dan algunos ejemplos numéricos, para visualizar en mejor forma, la parte teórica de los diferentes métodos.

La notación que se usará es la siguiente:

N : Nº total de conglomerados desiguales en la población.
n : Nº de conglomerados desiguales en la muestra.
Y : Total poblacional de la característica en estudio.
M_i : Número de elementos en el i-ésimo conglomerado.
y_{ij} : Valor de la característica del elemento j-ésimo perteneciente al conglomerado i-ésimo.
y_i : Valor total de la característica en el conglomerado i-ésimo.

$$y_i = \sum_{j=1}^{M_i} y_{ij}$$

$\bar{y}_i$: Valor promedio por elemento de la característica en el conglomerado i-ésimo.

$$\bar{y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij}$$

$\bar{Y}$: Valor promedio de la característica por conglomerado en la población.

$\bar{\bar{Y}}$: Valor promedio de la característica por elemento en la población.

M_o : N° total de elementos en la población.

$$M_o = \sum_{i=1}^{N} M_i$$

METODOS DE SELECCION CON PROBABILIDADES IGUALES

Primer método, de acuerdo a la metodología.

i) La población consta de N conglomerados desigua - les.

ii) Y : Valor total de la característica en la pobla ción.

$$Y = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij} \quad o \quad Y = \sum_{i=1}^{N} y_i$$

iii) M.A.S. sin reemplazo de n conglomerados desiguales, con igual probabilidad de selección.

iv) $\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} \sum_{j=1}^{M_i} y_{ij}$ o $\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} y_i$

v) $\hat{Y}$ es insesgado y consistente.

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

vii) $v(\hat{Y}) = \frac{N^2(1-f)}{n} \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$, $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$

Segundo método, de acuerdo a la metodología.

i) La población consta de N conglomerados desiguales.

ii) Y : Valor total de la característica en la población.

$$Y = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}$$

iii) M.A.S. sin reemplazo de n conglomerados desiguales.

iv) $\hat{Y}_R = M_o \cdot \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} M_i} = M_o \cdot \frac{\bar{y}}{\bar{m}} = M_o \cdot \hat{R}$, $\bar{m} = \frac{\sum_{i=1}^{n} M_i}{n}$, $\hat{R} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} M_i}$

v) $\hat{Y}_R$ es sesgado y consistente.

vi) Si el número de conglomerados seleccionados es grande.

$$V(\hat{Y}_R) \doteq \frac{N^2(1-f)}{n} \frac{\sum_{i=1}^{N} (y_i - M_i R)^2}{N-1}$$

$$\doteq \frac{N^2(1-f)}{n} \frac{\sum_{i=1}^{N} M_i^2 (\bar{y}_i - \bar{\bar{Y}})^2}{N-1}, \quad R = \frac{Y}{X} = \frac{Y}{M_o} = \bar{\bar{Y}}$$

vii) $$v(\hat{Y}_R) \doteq \frac{N^2(1-f)}{n} \frac{\sum_{i=1}^{n} M_i^2 (\bar{y}_i - \hat{\bar{\bar{Y}}}_R)^2}{n-1}$$

$$\hat{\bar{\bar{Y}}}_R = \frac{\hat{Y}_R}{m_o} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} M_i} = \hat{R}$$

METODOS DE SELECCION CON PROBABILIDADES DESIGUALES.

METODO DE HANSEN Y HURWITZ.

La técnica consiste en seleccionar las unidades muestrales con probabilidad proporcional a su tamaño (pps).

Unidad u_i	Tamaño M_i	Suma acumulada de M_i	Intervalos de Selección
u_1	M_1	M_1	$1 \ldots\ldots\ldots M_1$
u_2	M_2	$M_1 + M_2$	$M_2 + 1 \ldots . M_1 + M_2$
u_3	M_3	$M_1 + M_2 + M_3$	$M_1 + M_2 + 1 \ldots M_1 + M_2 + M_3$
.	.	.	.
.	.	.	.
.	.	.	.
u_N	M_N	$\sum_{i=1}^{N} M_i = M_o$	$\sum_{i=1}^{N-1} M_i + 1 \ldots M_o$

Se elige un número aleatorio entre 1 y M_o. Luego se determina a que intervalo de selección pertenece este número y la unidad perteneciente a este intervalo es la seleccionada.

Así, la probabilidad que tiene cualquier unidad de ser seleccionada, es proporcional a su tamaño, es decir,

$$P(u_i) = \frac{M_i}{M_o}, \text{ para todo } i = 1, \ldots, N.$$

METODO DE LAHIRI.

Sea M_a el valor máximo (o algo mayor) de las N medidas de tamaño M_i.

Se seleccionan dos números aleatorios. El primero entre 1 y N, el cual señala en forma provisoria la unidad seleccionada. El segundo se elige entre 1 y M_a.

Ahora, si el segundo número es menor o igual que el tamaño M_i de la unidad seleccionada en forma provisoria, esta unidad es definitivamente incluida en la muestra. En caso contrario, se repite el proceso.

Este procedimiento mantiene la probabilidad de selección proporcional al tamño de la unidad.

De acuerdo a la metodología, veremos la selección con probabilidades desiguales con reemplazo, para estimar el total poblacional.

i) La población está compuesta por conglomerados de tamaños desiguales. Aquí

$$P(u_i) = \frac{M_i}{M_o} \text{ , o}$$

$$z_i = P(u_i) = \frac{M'_i}{M'_o} \text{ , donde } M'_i \text{ es aproximadamente igual a } M_i.$$

ii) Y: Total poblacional.

iii) La selección de las unidades es con probabilidades desiguales y con reemplazo.

iv) $$\hat{Y}_{pps} = \frac{M_o}{n} \sum_{i=1}^{n} \bar{y}_i$$

v) $\hat{Y}_{pps}$ es insesgado y consistente.

vi) $$V(\hat{Y}_{pps}) = \frac{M_o}{n} \sum_{i=1}^{N} M_i (\bar{y}_i - \bar{\bar{Y}})^2$$

$$V(\hat{Y}_{ppz}) = \frac{1}{n} \sum_{i=1}^{N} z_i \left(\frac{y_i}{z_i} - Y \right)^2$$

vii) $$v(\hat{Y}_{pps}) = M_o^2 \sum_{i=1}^{n} (\bar{y}_i - \bar{\bar{y}})^2 / n(n-1), \text{ para todo } n > 1$$

$$v(\hat{Y}_{ppz}) = \sum_{i=1}^{n} \left(\frac{y_i}{z_i} - \hat{Y}_{ppz} \right)^2 / n(n-1), \text{ para todo } n > 1.$$

ACURACIDAD RELATIVA DE LOS TRES METODOS DE ESTIMACION.

Probabilidades iguales:

i) $\hat{Y}$ denotado por $\hat{Y}_u$.

ii) $\hat{Y}_R$

Probabilidades proporcionales al tamaño:

iii) $\hat{Y}_{pps}$

Comparemos las varianzas de los tres estimadores.

$$(N-1) \doteq N \text{ y } E(y_i - \bar{Y})^2 = \frac{\sum_{i=1}^{N} (y_i - \bar{Y})^2}{N}$$

i) $n\, V(\hat{Y}_u) = (1-f)\, E\, (Ny_i - Y)^2$

ii) $n\, V(\hat{Y}_R) \doteq (1-f)\, E\, \left(\frac{M_i}{\bar{M}}\right)^2 (M_o\bar{y}_i - Y)^2$

iii) $n\, V(\hat{Y}_{pps}) = E\, \left(\frac{M_i}{\bar{M}}\right)\, (M_o\bar{y}_i - Y)^2$

$V(\hat{Y}_u)$ depende de la acuracidad $Ny_i = NM_i\bar{y}_i$ como estimador de Y.

$V(\hat{Y}_R)$ y $V(\hat{Y}_{pps})$ dependen de la acuracidad

$M_o\bar{y}_i = \frac{M_o y_i}{M_i}$ como estimador de Y.

ii) da mayor ponderación a las unidades grandes que iii) $\hat{Y}_u$ y $\hat{Y}_R$ se benefician por el fcp.

ii) Se cumple solamente en muestras grandes.

También se ha trabajado con el modelo de regresión:

$$y_i = \alpha + \beta M_i + e_i$$

donde $E(e_i/M_i) = 0$

$$V(e_i) = V(y_i/M_i) \doteq M_i S^2 [M_i \rho + (1-\rho)]$$

$$V(e_i) = c.M_i^g \quad , \quad 1 < g < 2$$

Del modelo se obtiene:

$$\bar{Y} = \alpha + \beta . \bar{M} + \bar{e}_N \; ; \quad \bar{Y} = \frac{\alpha}{\bar{M}} + \beta + \frac{\bar{e}_N}{\bar{M}}$$

En este caso $\bar{e}_N$ es ínfimo.

Bajo el modelo, las varianzas a comparar son:

$$\frac{nV_u}{N^2} = \beta^2 V(M_i) + c\, E(M_i^g)$$

$$\frac{nV_R}{N^2} \doteq \frac{\alpha^2 V(M_i)}{\bar{M}^2} + c\, E\,(M_i^g)$$

$$\frac{nV_{pps}}{N^2} \doteq \frac{\alpha^2 V(M_i)}{\bar{M}^2} + c\, \bar{M}\, E\,(M_i^{g-1})$$

Si $\alpha = 0$, entonces $V_R < V_u$.

Si β es grande y $\alpha = 0$, $V_{pps} < V_u$, para $g \geq 1$

Si $g > 1$, entonces $V_{pps} < V_R$. Esto se cumple en la mayoría de los casos, en que $\beta \neq 0$.

Si $g < 1$, entonces $V_{pps} > V_R$.

Como conclusión, $\hat{Y}_{pps}$, generalmente es más preciso.

Ahora, en el muestreo con probabilidades desi - guales sin reemplazo, definamos lo siguiente:

$z_i = \frac{M_i}{M_o}$ es la probabilidad de seleccionar la i-ésima unidad en la primera extracción.

$\sum_{j \neq i}^{N} z_j \frac{M_i}{M_o - M_j} = z_i \sum_{j \neq i}^{N} \frac{z_j}{1 - z_j}$, es la probabilidad de seleccionar la i-ésima unidad en la segunda extracción.

$\pi_i = z_i + z_i \sum_{j \neq i}^{N} \frac{z_j}{1 - z_j}$, es la probabilidad de seleccionar la i-ésima unidad en cualquiera de las dos extrac - ciones.

En este caso, HORVITZ y THOMPSON propusieron un estimador para el total poblacional que lleva su nombre. De acuerdo a la metodología, se tiene:

i) La población está agrupada en conglomerados de tama - ños desiguales, que se han estratificado.

ii) Y: Total poblacional.

iii) Una cantidad pequeña de unidades se extraen de cada estrato, con probabilidades desiguales sin reemplazo.

iv) $\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$, y_i es el total de la i-ésima unidad.

$\pi_i > 0$, para todo i = 1,..., N.

v) $\hat{Y}_{HT}$ es insesgado y consistente.

vi) $V(\hat{Y}_{HT}) = \sum_{i=1}^{N} \frac{(1-\pi_i)}{\pi_i} y_i^2 + 2 \sum_{i=1}^{N} \sum_{j>1}^{N} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} y_i y_j$

vii) $v_1(\hat{Y}_{HT}) = \sum_{i=1}^{n} \frac{(1-\pi_i)}{\pi_i^2} y_i^2 + 2 \sum_{i=1}^{n} \sum_{j>i}^{n} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j \pi_{ij}} y_i y_j$

$$v_2(\hat{Y}_{HT}) = \sum_{i=1}^{n} \sum_{j>i}^{n} \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

Otros métodos propuestos son:

METODO DE BREWER.

Para n = 2, este método de selección muestral mantiene $\pi_i = 2z_i$ y usa el estimador de HORVITZ y THOMPSON:

$$\hat{Y}_{HT} = \frac{y_i}{\pi_i} + \frac{y_j}{\pi_j} = \frac{y_i}{2z_i} + \frac{y_j}{2z_j} = \tfrac{1}{2}\left(\frac{y_i}{z_i} + \frac{y_j}{z_j}\right)$$

Puesto que este método usa el estimador de HORVITZ y THOMPSON los puntos vi) y vii) de la metodología, proveen fórmulas para la varianza y la varianza estimada de $\hat{Y}_B$. BREWER mostró que la varianza es siempre menor que la del estimador $\hat{Y}_{ppz}$ en muestreo con reemplazo.

METODO DE MURTHY.

El estimador propuesto es:

$$\hat{Y}_M = \sum_{i=1}^{n} \frac{P(S/i).y_i}{P(S)}$$

donde:

P(S/i) : Probabilidad condicional de obtener el conjunto de unidades que fueron extraídas, dado que la i-ésima unidad fue extraída primero.

P(S) : Probabilidad de obtener el conjunto de unidades que fueron extraídas.

Método relacionado con el muestreo sistemático, propuesto por MADOW.

El estimador propuesto es:

$$\hat{Y}_{sys} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{z_i}$$

METODO DE RAO, HARTLEY y COCHRAN.

El estimador propuesto es:

$$\hat{Y}_{RHC} = \sum_{g=1}^{n} Z_g \frac{y_g}{z_g} = \sum_{g=1}^{n} \hat{Y}_g \text{ , donde}$$

Z_g : total del tamaño para el grupo g.

y_g y z_g, se refieren a la unidad extraìda del grupo g.

Importante:

En la elección de un método en la práctica, se debe considerar lo siguiente:

- La facilidad con la cual la muestra puede ser extraída.
- La simplicidad del estimador.
- La acuracidad del estimador.
- Aprovechar la estimación de la varianza del estimador.

REFERENCIAS:

COCHRAN, WILLIAM G. Sampling techiques. 1977.
KONIJN, H.S. Statistical theory of sample survey design and analysis. 1973.
RAJ, DES.; The design of sample surveys. 1972.
RAJ, DES.; Sampling theory. 1968.
SANCHEZ-CRESPO, J.L. Curso intensivo de muestreo en poblaciones finitas. 1980.
SUKHATME, B.V. and P.V. SUKHATME. Sampling theory of surveys with aplications. 1970.