Sociedad Ecuatoriana de Estadística

"Análisis de Encuestas por Muestreo con R"

Muestreo bietápico

Andrés Peña M.

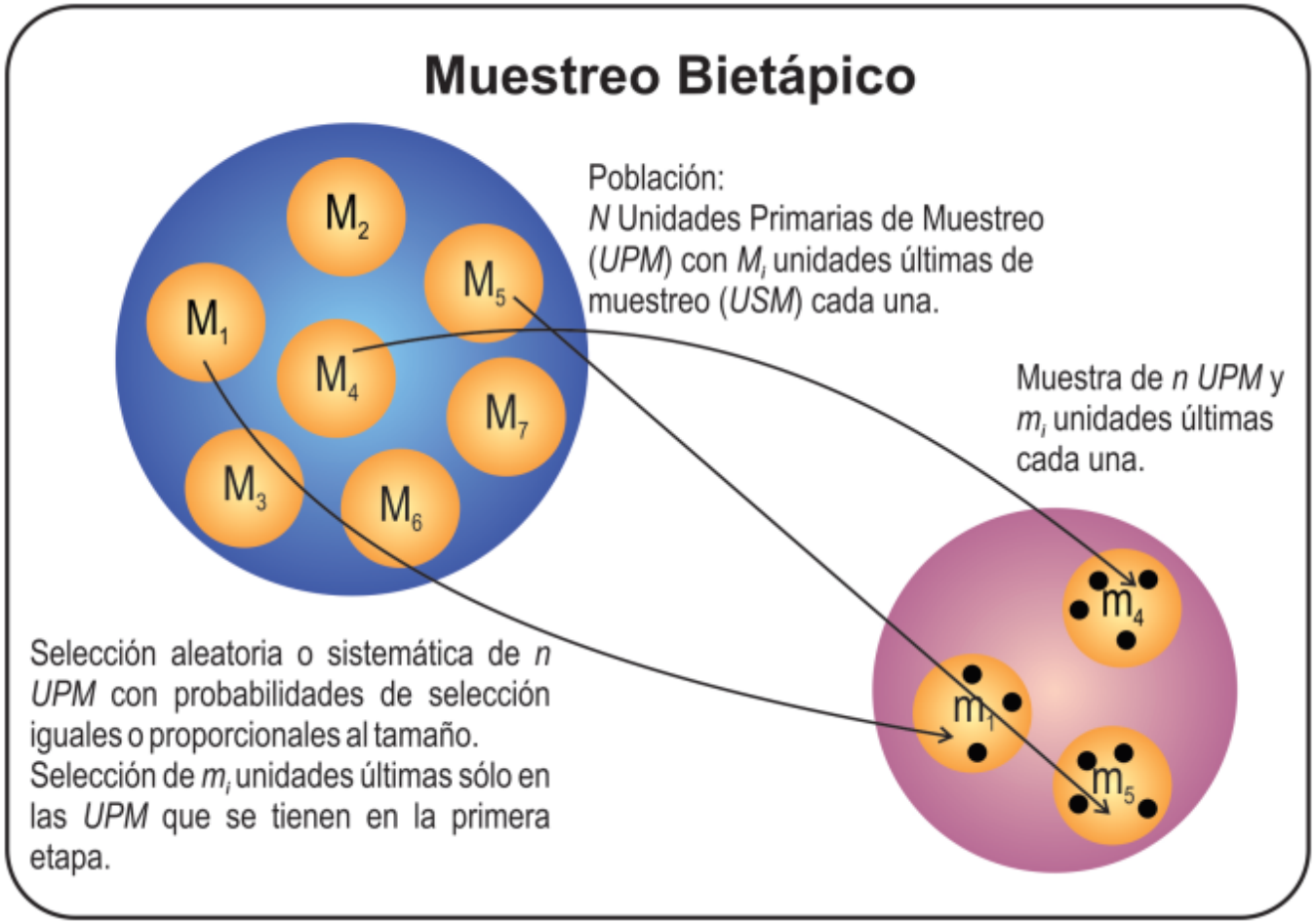
agpena@colmex.mx

Mayo 2023

Tabla de contenidos

Muestreo bietápico

Muestreo bietápico



Muestreo bietápico

No se censan los conglomerados en muestra, sino que se toma una muestra de sus elementos.

Por ejemplo, se quiere estimar el número de personas "desocupadas" en la Ciudad de México. La población es el conjunto de personas en edad productiva, de la cual no hay marco.

Si tuvieramos el marco y seleccionáramos una M.A.S. de personas, sería muy costoso que la muestra quedara dispersa en toda la ciudad.

Para remediar esto, se forman nuevas unidades de muestreo llamadas **Unidades Primarias de Muestreo** (UPM). Para el ejemplo, las UPM podrían ser las manzanas, de las cuales si se tiene marco (mapas de la ciudad).

Tamaño de los conglomerados

Se selecciona al azar (M.A.S.) cierto número de manzanas y de cada manzana seleccionada se construye el marco de viviendas, del cual se selecciona una muestra (M.A.S.) de viviendas que serán las **Unidades de Segunda Etapa** (USM) para, posteriormente, censar las personas en edad productiva de estas viviendas seleccionadas.

También se puede combinar con muestreo estratificado, por ejemplo, las UPM se pueden agrupar en colonias/barrios o sectores según nivel socioeconómico.

Notación

A nivel poblacional:

N Número de UPM (se cuenta con un marco)

M_i Número de USM en la UPM_i

$$M = \sum_{i=1}^{N} M_i$$ Total de USM (generalmente no se conoce)

Y_{ij} Valor de la medición en la USM_j de la UPM_i

$$Y_i = \sum_{j=1}^{M_i} Y_{ij}$$ Total de la UPM_i

$$\bar{Y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} Y_{ij}$$ Promedio de la UPM_i

$\bar{Y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} Y_{ij}$ Promedio de la UPM

$Y = \sum_{i=1}^{N} Y_i = \sum_{i=1}^{N} \sum_{j=1}^{M_i} Y_{ij}$ Total poblacional

Notación

$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ Promedio de los totales de UPM

$\bar{Y}_e = \frac{Y}{M} = \frac{Y}{\sum_{i=1}^{N} M_i}$ Media por elemento

$S_{wi}^2 = \frac{\sum_{j=1}^{M_i}\left(Y_{ij} - \bar{Y}_i\right)^2}{M_i - 1}$ Varianza entre USM de la UPM_i

$S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2$ Varianza entre totales de UPM

$$S_b^2 >> S_{wi}^2$$

Notación

Si se considera una M.A.S. para UPM y una M.A.S. para USM:

A nivel muestral:

n Número de UPM en muestra

m_i Número de USM muestreadas en la UPM_i

y_{ij} Medición de la USM_j en muestra de la UPM_i en muestra

$\hat{Y}_i = \bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}$ Promedio muestral de las USM de la UPM_i

$\hat{Y}_i = M_i \hat{\bar{Y}}_i$ Total estimado de la UPM_i, M_i es conocido ya que se refiere a la UPM_i en muestra

Notación

$\hat{S}_b^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{\hat{Y}} \right)^2$ Varianza estimada entre *UPM*

$\hat{S}_{wi}^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2$ Varianza estimada entre USM dentro de la UPM_i

$\overline{\hat{Y}} = \frac{1}{n} \sum_{i=1}^{n} \hat{Y}_i = \frac{1}{n} \sum_{i=1}^{n} M_i \overline{\hat{Y}}_i$ Promedio de totales estimados de UPM

Estimador del total poblacional

$$\hat{Y} = N\overline{\hat{Y}} = \frac{N}{n}\sum_{i=1}^{n}\hat{Y}_i$$

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

$$= \sum_{i=1}^{n}\sum_{j=1}^{m_i}\frac{N}{n}\frac{M_i}{m_i}y_{ij}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m_i} f_{ij}y_{ij}$$

Estimador del total poblacional

Donde, f_{ij} es el factor de expansión.

Recordando el ejemplo anterior,

$$\begin{aligned} P(\text{ vivienda } j \text{ de la manzana } i) =& P(\text{ vivienda } j \mid \text{ manzana } i) \times \\ & P(\text{ manzana } i) \\ =& \frac{m_i}{M_i} \frac{n}{N} \end{aligned}$$

Si $m_i \propto M_i$, es decir, $\frac{M_i}{m_i} = k$ el diseño es **autoponderado**, es decir, los factores de expansión son iguales $f_{ij} = f = \frac{N}{n} k, \forall j, \forall i$.

Varianza del estimador del Total

$$V(\hat{Y}) = \underbrace{N^2 \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2}_{(1)} + \frac{N}{n} \sum_{i=1}^{N} M_i^2 \underbrace{\left(\frac{1}{m_i} - \frac{1}{M_i}\right)}_{(2)} S_{wi}^2$$

(1) Es el 90%-95% del valor de $V(\hat{Y})$.

(2) Es cero si $m_i = M_i$, es decir, si se censan las UPM. Es el caso del muestreo de conglomerados.

Es común que los valores de Y_{ij} sean semejantes dentro de cada UPM. Esto hace que los S_{wi}^2 sean pequeños. Los totales Y_i de UPM difieren mucho si los números M_i de USM dentro de cada UPM son diferentes. Además, S_b^2 es una varianza entre totales, no entre valores individuales. Todo esto hace que la primera parte de $V(\hat{Y})$ constituya gran parte de su valor.

Varianza del estimador del Total

Como los valores de las Y_{ij} tienden a ser parecidos dentro de cada una de las UPM, entonces se genera una correlación, llamada correlación intraconglomerado.

Esta correlación hace que la información tenga cierta redundancia, lo que se refleja en varianza de los estimadores mayor que la que se obtendría con un muestreo directo unietápico de las unidades.

Estimador de la varianza

La varianza del estimador del Total se estima con:

$$\hat{V}(\hat{Y}) = N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \hat{S}_b^2 + \frac{N}{n} \sum_{i=1}^{n} M_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) \hat{S}_{wi}^2$$

Donde:

$$\hat{S}_b^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{\hat{Y}}\right)^2$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} \left[M_i \hat{\bar{Y}}_i - \frac{1}{n} \sum_{i=1}^{n} M_i \bar{y}_i\right]^2$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} \left[M_i \bar{y}_i - \frac{1}{n} \sum_{i=1}^{n} M_i \bar{y}_i\right]^2$$

El intervalo aproximado del $(1-\alpha)100\%$ de confianza para Y :

$$\hat{Y} \pm z_{1-\alpha/2} \sqrt{\hat{V}(\hat{Y})}$$

Estimador de la Media por elemento (Razón)

$$
\begin{aligned}
\hat{\bar{Y}}_e &= \frac{\hat{Y}}{\hat{M}} \\
&= \frac{\frac{N}{n}\sum_{i=1}^{n}\hat{Y}_i}{\frac{N}{n}\sum_{i=1}^{n}M_i} \\
&= \frac{\sum_{i=1}^{n}\hat{Y}_i}{\sum_{i=1}^{n}M_i} \\
\hat{\bar{Y}}_e &= \frac{\sum_{i=1}^{n}M_i\bar{y}_i}{\sum_{i=1}^{n}M_i}
\end{aligned}
$$

Varianza del estimador de la Media por elemento

$$V\left(\hat{\bar{Y}}_e\right) = \left(1 - \frac{n}{N}\right) \frac{1}{n\bar{M}^2} \sum_{i=1}^{N} \frac{M_i^2 \left(\bar{Y}_i - \bar{Y}_e\right)^2}{N-1} + \frac{1}{nN\bar{M}^2} \sum_{i=1}^{N} M_i^2 \left(1 - \frac{m_i}{M_i}\right) \frac{S_{wi}^2}{m_i}$$

Con estimador

$$\hat{V}\left(\hat{\bar{Y}}_e\right) = \left(1 - \frac{n}{N}\right) \frac{1}{n\hat{\bar{M}}^2} \sum_{i=1}^{n} \frac{M_i^2 \left(\bar{y}_i - \hat{\bar{Y}}_e\right)^2}{n-1} + \frac{1}{nN\hat{\bar{M}}^2} \sum_{i=1}^{n} M_i^2 \left(1 - \frac{m_i}{M_i}\right) \frac{\hat{S}_{wi}^2}{m_i}$$

donde $\hat{\bar{M}} = \sum_{i=1}^{n} \frac{M_i}{n}$.

Estimador de la Media por elemento

Si se conoce M, el total de USM en la población, otra forma de estimar la media por elemento es:

$$\hat{\bar{Y}}_e = \frac{\hat{Y}}{M} = \frac{N}{Mn} \sum_{i=1}^{n} M_i \bar{y}_i$$

Con varianza y estimador de varianza:

$$V\left(\hat{\bar{Y}}_e\right) = \frac{1}{M^2} V(\hat{Y})$$

$$\hat{V}\left(\hat{\bar{Y}}_e\right) = \frac{1}{M^2} \hat{V}(\hat{Y})$$

Estimador de una Proporción

$$\hat{P} = \frac{\sum_{i=1}^{n} M_i \hat{p}_i}{\sum_{i=1}^{n} M_i}$$

donde, $\hat{p}_i$ es la proporción en la UPM_i, es decir, $\hat{p}_i = \sum_{j=1}^{m_i} \frac{y_{ij}}{m_i}$ y

$$y_{ij} = \begin{cases} 1 & U_{ij} \text{ tiene la característica A} \\ 0 & U_{ij} \text{ no tiene la característica A} \end{cases}$$

El estimador de la varianza del estimador de la proporción es:

$$\hat{V}(\hat{P}) = \left(1 - \frac{n}{N}\right) \frac{1}{n\hat{\bar{M}}^2} \frac{\sum_{i=1}^{n} M_i^2 \left(\hat{p}_i - \hat{P}\right)^2}{n-1} + \frac{1}{nN\hat{\bar{M}}^2} \sum_{i=1}^{n} M_i^2 \left(1 - \frac{m_i}{M_i}\right) \left(\frac{\hat{p}_i \left(1 - \hat{p}_i\right)}{m_i - 1}\right)$$

Tamaño de muestra

Una forma de calcular el tamaño de muestra, que se utiliza en la práctica es la siguiente:

Si se desprecia la variación entre USM dentro de las UPM y se fija la precisión δ y la confianza $1-\alpha$ entonces,

$$\delta = z_{1-\alpha/2}\sqrt{V(\hat{Y})} = z_{1-\alpha/2}\sqrt{N^2\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2}$$

despejando n :

$$n = \frac{Nz_{1-\alpha/2}^2 S_b^2}{N\delta^2 + z_{1-\alpha/2}^2 S_b^2}$$

n es el número de UPM a muestrear.

Cuántas USM? Lo menos posible (de 2 a 5). En la ENEMDU de Ecuador se toman 7 viviendas con criterios técnicos y operativos (INEC, 2023; pp. 14).

Gracias!!!