Sociedad Ecuatoriana de Estadística

"Análisis de Encuestas por Muestreo con R"

Muestreo de conglomerados

Andrés Peña M.

a.pena@rusersgroup.com

Diciembre 2021

X Seminario Internacional de Estadística Aplicada

Tabla de contenidos

Muestreo de conglomerados

Muestreo por conglomerados (una etapa)

Un conglomerado es un conjunto de elementos de la población.

Una muestra de conglomerados es una muestra aleatoria en la cual cada unidad muestral es un conglomerado de elementos.

El uso de conglomerados se debe a dos razones principalmente:

1. No existen marcos de los elementos de la población, o son muy caros de construir, o es imposible construirlos.

2. Muestrear conglomerados es menos costoso que un m.a.s. de elementos, sobre todo cuando el costo de obtener la información se incrementa al aumentar la distancia entre los elementos.

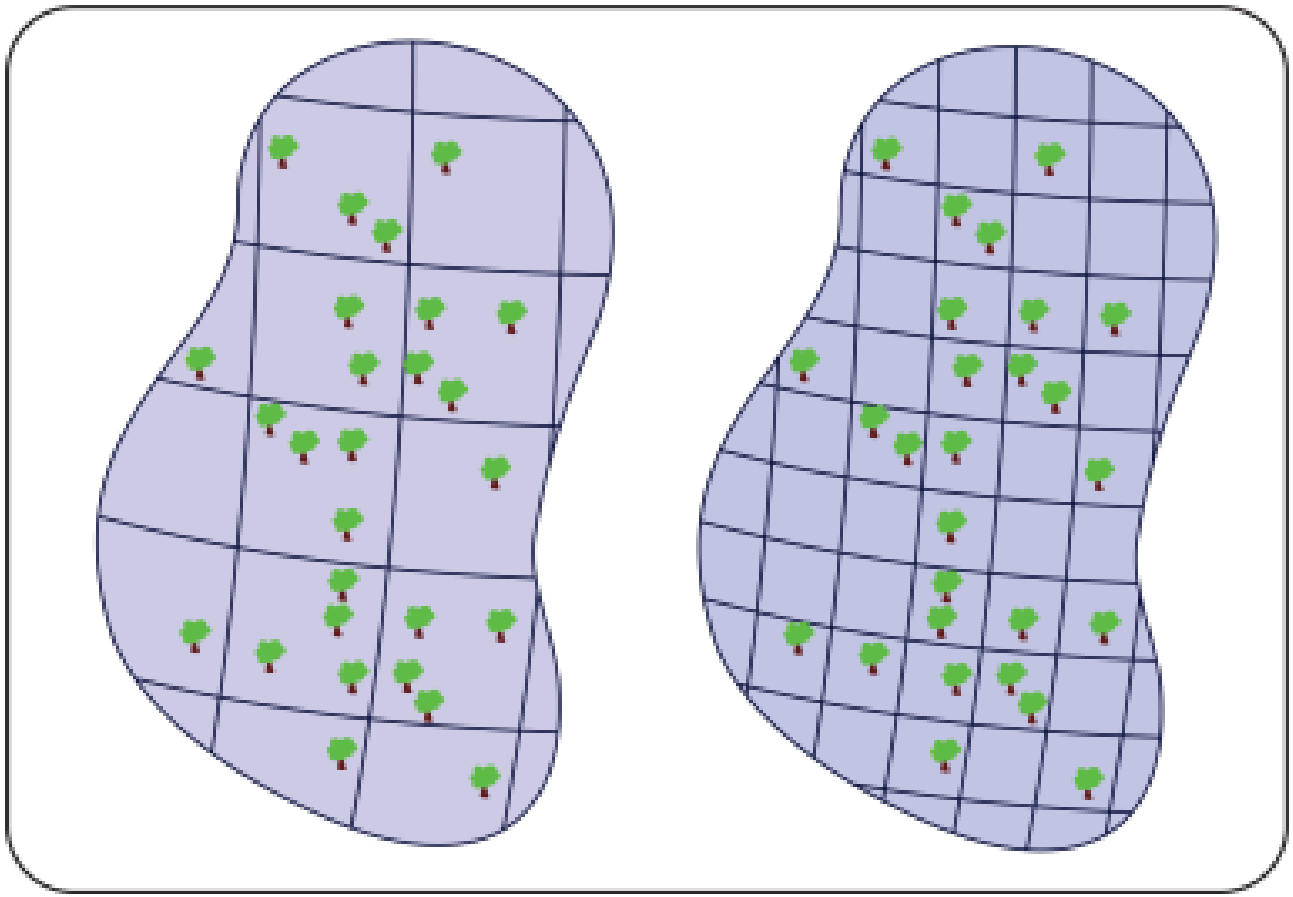
Tamaño de los conglomerados

En algunas situaciones el tamaño de los conglomerados (número de elementos que los componen) está dado. Por ejemplo, si los conglomerados son las casillas electorales y los elementos de cada conglomerado son los votantes de la casilla, el tamaño está fijo.

En otros casos, nosotros definimos el tamaño de los conglomerados, por ejemplo, si queremos estimar la proporción de árboles muertos en cierto bosque, debemos definir el área de bosque de cada conglomerado.

Si existe variabilidad en la densidad de árboles muertos a lo largo del bosque, entonces, sería deseable muestrear áreas pequeñas seleccionadas al azar o sistemáticamente.

Tamaño de los conglomerados



Tamaño de los conglomerados

- Muchas áreas pequeñas $\Rightarrow$ control variabilidad.
- Pocas áreas grandes $\Rightarrow$ economía.
- Elementos dentro del conglomerado pueden estar correlacionados.
- Balance entre tamaño y número de conglomerados.
- Pruebas piloto con varios tamaños de conglomerado.

Muestreo por conglomerados

- En muestreo estratificado queremos que los estratos contengan unidades muy homogéneas dentro y heterogéneas entre estratos.
- En muestreo por conglomerados queremos que los conglomerados contengan unidades muy heterogéneas dentro y homogéneas entre ellos.

Notación

A nivel poblacional:

N No. de conglomerados en la población
n No. de conglomerados en muestra
M_i No. de elementos en el conglomerado $i, i = 1, \ldots, N$
$M = \sum_{i=1}^{N} M_i$ Total de elementos en la población
Y_{ij} Valor de la medición del elemento j del conglomerado i (a veces no lo tenemos)
$Y_i = \sum_{j=1}^{M_i} Y_{ij}$ Total del conglomerado i (a veces es lo que tenemos)

Notación

$\bar{Y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} Y_{ij}$ Promedio del conglomerado

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

$\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$ Promedio de totales de conglomerados (generalmente no interesa)

$\bar{Y}_e = \frac{Y}{M} = \frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} M_i}$ Promedio por elemento (es el que interesa)

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

Estimador del total poblacional

Suponga que tenemos una m.a.s. de n conglomerados.
El estimador del promedio por conglomerado es:

$$\hat{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

donde $y_i = \sum_{j=1}^{M_i} y_{ij}$ es el total observado del conglomerado i.
El estimador del total poblacional Y es:

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

Estimador del total poblacional

Con varianza y estimador de varianza:

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

$$\hat{V}(\hat{Y}) = N^2 \left(1 - \frac{n}{N}\right) \frac{\hat{S}_b^2}{n}$$

donde

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

Estimador de la Media poblacional (por elemento)

Si se conoce M, el total de elementos en la población, entonces, el estimador de la Media poblacional por elemento es:

$$\hat{\hat{Y}}_e = \frac{\hat{Y}}{M} = \frac{N}{Mn} \sum_{i=1}^{n} y_i$$

Con varianza y estimador de varianza:

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

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

Características de estos estimadores

Estos dos estimadores, el del total poblacional y de la media poblacional por elemento, son insesgados, pero frecuentemente tienen varianzas grandes, ya que si el número de elementos en los conglomerados (M_i) es muy diferente, genera variabilidad entre los totales de los conglomerados.

Si el tamaño del conglomerado M_i está fuertemente relacionado con el total del conglomerado, lo que generalmente sucede, entonces se prefieren estimadores de razón.

Estimador de la Media poblacional por elemento. (Razón)

$$\hat{Y}_e = \frac{\hat{Y}}{\hat{M}} = \frac{\frac{N}{n}\sum_{i=1}^{n} y_i}{\frac{N}{n}\sum_{i=1}^{n} M_i} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} M_i}$$

con varianza:

$$V\left(\hat{Y}_e\right) = \left(1 - \frac{n}{N}\right)\frac{1}{n}\frac{1}{\bar{M}^2}\sum_{i=1}^{N}\frac{\left(Y_i - \bar{Y}_e M_i\right)^2}{N-1}$$

donde, $\bar{M} = \frac{M}{N}$ es el tamaño promedio de los conglomerados

Estimador de la Media poblacional por elemento. (Razón)

Estimador de la varianza del estimador:

$$\hat{V}\left(\hat{\hat{Y}}_e\right) = \left(1 - \frac{n}{N}\right) \frac{1}{n} \frac{1}{\hat{\hat{M}}^2} \sum_{i=1}^{n} \frac{\left(y_i - \hat{\hat{Y}}_e M_i\right)^2}{n-1}$$

donde

$$\hat{\hat{M}} = \frac{\hat{M}}{N} = \frac{\frac{N}{n}\sum_{i=1}^{n} M_i}{N} = \sum_{i=1}^{n} \frac{M_i}{n}$$

Estimador del Total poblacional. (Razón)

$$\hat{Y} = M\hat{\bar{Y}}_e$$

con M conocida.

Con varianza y estimador de varianza:

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

Estimador de una Proporción poblacional. (Razón)

Sea

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

El estimador de la proporción de unidades con la característica es:

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

con varianza estimada:

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

Tamaño de muestra

Se fijan la precisión δ y la confianza $1-\alpha$

$$\delta = z_{1-\alpha/2}\sqrt{V\left(\hat{\bar{Y}}_e\right)}$$

$$\delta^2 = z^2_{1-\alpha/2}\left(\frac{1}{n}-\frac{1}{N}\right)\frac{1}{\bar{M}^2}S_b^2$$

Despejando n :

$$n = \frac{Nz^2_{1-\alpha/2}S_b^2}{N\delta^2\bar{M}^2 + z^2_{1-\alpha/2}S_b^2} = \frac{NS_b^2}{\frac{N\delta^2\bar{M}^2}{z^2_{1-\alpha/2}} + S_b^2}$$

Gracias!!!