PPS y muestreo por conglomerados

Muestreo II

Licenciatura en Estadística

2023

Introducción PPS

- en algunos casos la variable para algunos individuos de la población posee valores extremadamente altos.
- para la selección de la muestra aleatoria s podemos definir probabilidades de inclusión π_i que dependan del tamaño del individuo.
- más precisamente del tamaño o peso relativo.

Introducción PPS

- el tamaño viene dado por una variable auxiliar (covariable) x
- los valores de la variable x deben ser conocidos a nivel de la unidad para toda la población U.
- el tamaño relativo del individuo i queda definido como:

$$\text{peso relativo}_i = p_i = \frac{x_i}{\sum\limits_{i \in U} x_i} = \frac{x_i}{X}$$

muestreo PPS

- en el muestreo PPS las probabilidades de inclusión son definidas en base al tamaño o peso relativo del individuo p_i.
- si el muestreo se realiza con reemplazo, las probabilidades de inclusión quedan definidas como

$$p_i = \frac{x_i}{\sum_{i \in U} x_i}$$

muestreo PPS

si el muestreo se realiza sin reemplazo, las probabilidades de inclusión quedan definidas como

$$\pi_i = \text{prob}[i \in s] = n \times p_i = n \times \frac{x_i}{\sum_{i \in U} x_i}$$

si la variable de interés se encuentra correlacionada con el tamaño del individuo, la reducción de la varianza del estimador será considerable en comparación a un muestreo EPSEM directo (e.g. simple)

consideraciones PPS

- si existen individuos con valores x_i extremadamente altos, las probabilidades de inclusión π_i pueden llegar a ser mayores que uno.
- se asignan una probabilidad de inclusión $\pi_i = 1$ para los individuos que cumplan que

$$n \times x_i > \sum_{i \in U} x_i$$

- a estos individuos se les denomina de inclusión forzosa
- para el resto de los individuos se les asigna la probabilidad de forma proporcional. Obviamente, se sacan los forzosos del total y del tamaño de muestra.

Estimación del SE

La estimación viene dada como

$$\widehat{\mathrm{SE}}^2(\hat{Y}) = \widehat{\mathrm{var}}(\hat{Y}) = \frac{1}{n(n-1)} \times \sum_{i \in s} (n w_i y_i - \hat{Y})^2$$

- todos los programas aproximan los SE (por defecto) utilizando la expresión anterior.
- asumen que el muestreo se realizó con probabilidades distintas y **con remplazo!!**.

Estimación del SE

Si la tasa de muestreo $f = n/N$ es relativamente grande (e.g más de 0.10) se le aplica un factor de corrección

$$\text{fpc} \times \widehat{\text{Var}}_{PPS}(\hat{t})$$

muestreo por conglomerados en una o varias etapas

el muestreo directo de elementos no es siempre posible

- no disponemos de un marco muestral y el costo de crear uno es prohibitivo.
- los individuos de la población se encuentran muy dispersos, lo cual, implica costos prohibitivos de relevamiento de los datos.

introducción (1)

En los diseños por conglomerados en una primera etapa se seleccionan grupos (clusters o conglomerados) de elementos. Los clusters los llamamos unidades primarias de muestreo (UPM)

Luego:

- podemos censar el conglomerado
- sacar una muestra del conglomerado (dos etapas)

introducción (2)

La población U es particionada en conglomerados (grupos de elementos) generalmente conformados de forma natural

unidad de análisis	conglomerados
niñas/os que asisten a educación primaria	centros educativos
hogares	manzanas

muestreo por conglomerados

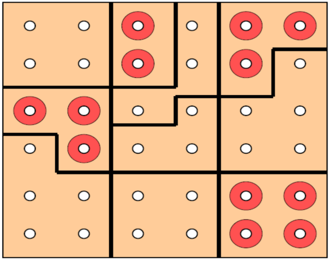
seleccionamos una muestra aleatoria de conglomerados

- varias estrategias: diseño simple, sistemático, PPS, estratificado, etc.

Todos los elementos pertenecientes a los conglomerados seleccionados son encuestados (i.e.censo del conglomerado)

muestreo por conglomerados

población $N = 35$ individuos conglomerados (agrupados) en $M = 10$ clusters. Seleccionamos muestra de $m = 4$ clusters. Tamaño de muestra total $n = 12$



ventajas

- reducción de los costos de relevamiento producto de tener una muestra menos dispersa geográficamente
- más fácil de implementar en poblaciones conglomeradas naturalmente (hogares, escuelas)

desventajas

- generalmente menos eficiente que un diseño simple. Los individuos dentro de los conglomerados tienden a ser homogéneos respecto a las variables de interés.
- tamaño de muestra final desconocido a priori debido a que no sabemos el tamaño de los conglomerados seleccionados en la muestra (a priori)

muestreo simple de conglomerados

- la probabilidad de selección de un conglomerado j es

$$\pi_j = \frac{m}{M}$$

- la probabilidad de selección de un individuo i que pertenece al conglomerado j es $\pi_{ij} = \pi_j = m/M$ y el ponderador original es $w_{ij} = \pi_{ij}^{-1} = M/m$
- la estimación del total de la variable y es

$$\hat{Y} = \sum_{i \in s} w_i \times y_i = \sum_{j=1}^{M} \frac{M}{m} \times Y_j$$

estimación del SE

$$\widehat{\mathrm{SE}}^2(\hat{Y}) = \widehat{\mathrm{var}}(\hat{Y}) = M^2(1-m/M) \times \frac{\sum_{j=1}^{m} (Y_j - \mathrm{mean}[Y_j])^2/(m-1)}{m}$$

i.e. en vez de utilizar los datos de la variable y, utilizamos los totales de los clusters.

muestreo en dos etapas

- en la primera etapa seleccionamos una muestra de conglomerados (UPM) bajo un diseño Simple, Sistemático, PPS, estratificado, etc.
- en una segunda etapa seleccionamos individuos (con las estrategias que hemos visto) dentro de las UPM seleccionadas en la primera etapa.

.

observaciones en dos etapas

- muy utilizados cuando los conglomerados son muy grandes
- podemos controlar el tamaño de muestra
- permite aumentar cantidad de UPM a ser seleccionadas (m)

ventajas en dos etapas

- reducción considerable de costos de relevamiento de las encuestas debido a que tenemos una muestra menos dispersa geográficamente.
- generalmente más eficiente en comparación con el muestreo en una etapa cuando los conglomerados son homogéneos respecto a las variables de interés.
- el tamaño de muestra puede ser controlado si la cantidad de individuos a sortear en la segunda etapa es fijo.
- no necesitamos un marco muestral completo (solo para las UPMs de la muestra). Únicamente hacemos el marco en campo para UPMs incluidas en la muestra.

desventajas en dos etapas

- Generalmente no es más eficiente que un diseño aleatorio simple
- las fórmulas de los SEs son extremadamente difíciles o imposibles de calcular . Se tiene que recurrir a aproximaciones (lineales o remuestreo/réplicas, e.g. Bootstrap, Jackknife, etc.)

estimación del SE en dos etapas

- es la suma de la varianzas en cada una de las etapas de muestreo

$$\widehat{\mathrm{SE}}^2(\hat{t}) = \widehat{\mathrm{var}}(\hat{Y}) = \widehat{\mathrm{var}}_{UPM}(\hat{Y}) + \widehat{\mathrm{var}}_{USM}(\hat{Y})$$

- generalmente para la estimación aproximamos únicamente con la varianza de la primera etapa (**método del ultimo conglomerado**).

ejemplo 1 ponderadores w_i en dos etapas

escuela	# de estudiantes	probabilidad selección escuela	escuela seleccionada	estudiantes seleccionados bajo un m.a.s	probabilidad condicional de seleccionar al estudiante	probabilidad total de seleccionar al estudiante	ponderador del estudiante (w)
1	50	0.50					
2	30	0.50	x	10	0.33	0.167	6
3	20	0.50					
4	100	0.50	x	10	0.10	0.050	20
total	200	2.00					

- muestreo aleatorio simple de $m = 2$ escuelas
- muestreo aleatorio simple de 10 estudiantes por escuela
- tamaño de muestra $n = 2 \times 10 = 20$
- **ponderadores muy variables**

$$\hat{N} = \sum_{i \in s} w_i = (6 \times 10) + (20 \times 10) = 260 \neq 200$$

ejemplo 2 ponderadores w_i en dos etapas

escuela	# de estudiantes	peso proporcional (pi)	probabilidad selección escuela	escuela seleccionada	estudiantes seleccionados bajo un m.a.s	probabilidad condicional de seleccionar al estudiante	probabilidad total de seleccionar al estudiante	ponderador del estudiante (w)
1	50	0.25	0.50					
2	30	0.15	0.30	x	10	0.33	0.100	10
3	20	0.1	0.20					
4	100	0.5	1.00	x	10	0.10	0.100	10
total	200	1	2.00					

- muestreo aleatorio PPS de $m = 2$ escuelas
- muestreo aleatorio simple de 10 estudiantes por escuela
- tamaño de muestra $n = 2 \times 10 = 20$
- **diseño autoponderado(EPSEM)**

$$\hat{N} = \sum_{i \in s} w_i = (20 \times 10) = 200$$