muestreo en dos fases

Muestreo II

Licenciatura en Estadística

2023

Queremos seleccionar una muestra s para estimar distintos parámetros θ de U y:

- el marco de muestreo (F) no tiene información auxiliar ($\mathbf{x}_i$) para seleccionar una muestra bajo un diseño "inteligente".
  - estratificar
  - asignar π_i proporcional al peso relativo de la unidad (e.g. PPS)
- **tampoco** tenemos información a nivel agregado de la población ($\mathbf{X}$) para poder utilizar estimadores de regresión/calibración (e.g. $\hat{Y}^{\text{RA}} = [X/\hat{X}^{\text{HT}}] \times \hat{Y}^{\text{HT}}$)

ejemplo

en una encuesta a empresas, queremos estimar el total de las ventas (y),

$$Y = \sum_{i \in U} y_i$$

y no tenemos información útil (e.g. cantidad de empleados, remuneraciones) ni el marco muestral ni totales provenientes de otras fuentes.

- el **muestreo en dos fases** nos proporciona una "solución".

El muestreo en dos fases es útil cuando:

1. la variable de interés y es relativamente cara de relevar, pero una variable x que se encuentra correlacionada con y puede ser relevada de forma fácil y barata.
2. para el tratamiento de la no respuesta
3. para muestrear poblaciones raras (i.e. con una prevalencia baja en la población)
4. para "mejorar" los marcos muestrales

seleccionamos una muestra en dos fases de selección:

1. seleccionamos una muestra aleatoria (e.g. bajo un SI) de $n^{(1)}$ elementos de U a la cual llamamos **fase 1**.
   - Recolectamos información de $\mathbf{x}_i$ para todos los individuos incluidos en la muestra de la fase 1.
   - $n^{(1)}$ debe ser lo suficientemente grande para poder "captar" la distribución de $\mathbf{x}_i$
   - Obviamente asumimos que relevar datos de $\mathbf{x}_i$ es "barato".
2. Asumimos que la muestra de la primera fase es nuestro marco muestral y seleccionamos una muestra aleatorio de tamaño $n^{(2)}$, a la cual llamamos fase 2 y recolectamos información de la variable de interés y solo para los individuos de la fase 2.

dado que estamos tratando la muestra de la fase 1 como si fuera nuestro marco muestral, podemos utilizar la información recolectada en la fase 1 para diseñar la muestra de la segunda fase.

- la información $\mathbf{x}_i$ recolectada en la fase 1 puede ser utilizada para:
  1. construir estratos
  2. definir probabilidades de inclusión en la muestra de la segunda fase
  3. utilizar estimadores de regresión/calibración.

no importa los esfuerzos que se hagan van a existir individuos i incluidos en la muestra s de la cual no se va a poder obtener información, es decir, va a existir **no respuesta**

- una muestra aleatoria s es seleccionada de U bajo un diseño $p(s)$ cualquiera.
- los individuos de s son clasificados en dos estratos: respondentes (R) y no respondentes (NR)
- la muestra de la primera fase es la muestra original s.

La variable

$$x_i = \begin{cases} 1 & \text{si el individuo i responde} \\ 0 & \text{si el individuo i no responde} \end{cases} \tag{1}$$

es observada para todos individuos de la fase 1 (muestra original).

Luego, la información acerca de x_i es utilizada para la muestra de la segunda fase.

- la variable de interés y_i es observada para todos los individuos donde $x_i = 1$.
- una submuestra es seleccionada para aquellos individuos donde $x_i = 0$

- supongamos que tenemos una población U y nuestro interés es un sobconjunto (dominio) U_d, de tamaño N_d .
- U_d representa una prevalencia pequeña en la población $P_d = N_d/N$.
- no conocemos (a priori) que individuos de U pertenecen a U_d
- en la primera fase seleccionamos una muestra $s^{(1)}$ de U para identificar a individuos de U_d.
- en una segunda fase se selecciona una muestra $s^{(2)}$ únicamente teniendo en cuenta a todos los individuos de la primera fase que pertenecen a U_d

- en las encuestas a hogares usualmente se utilizan como marco de muestreo F los censos de población.
- a lo largo del tiempo el marco del censo tiende a perder calidad y cobertura; y es inviable poder actualizarlo.
- se selecciona una muestra del marco del censo a nivel de UPMs y la misma es "actualizada" y luego utilizada para seleccionar distintas muestras de hogares y personas.
- la primera fase es la selección de las UPMs y es denominada marco maestro o master frame.

- sea $s^{(1)}$ la muestra de la primera fase, la cual, es seleccionada de U.
- las unidades incluidas en $s^{(1)}$ son determinadas por las siguientes variables aleatorias:

$$Z_i = \begin{cases} 1 & \text{si i es seleccionado en muestra de primera fase} \\ 0 & \text{si i NO es seleccionado en muestra de primera fase} \end{cases}$$

- sea $w_i^{(1)}$ el ponderador original de la muestra de la primera fase

$$w_i^{(1)} = \frac{1}{P[Z_i = 1]} = \frac{1}{\pi_i^{(1)}}$$

teoría para el muestreo en dos fases

- observamos el set de variables auxiliares $\mathbf{x}_i$ para cada uno de los individuos incluidos en $s^{(1)}$ y podemos estimar dichos totales como:

$$\hat{\mathbf{X}}^{(1)} = \sum_{i \in s^{(1)}} w_i^{(1)} \mathbf{x}_i = \sum_{i \in U} Z_i w_i^{(1)} \mathbf{x}_i$$

- la variable aleatoria indicadora de pertenecía a la muestra de la segunda fase $s^{(2)}$ es:

$$D_i = \begin{cases} 1 & \text{si i es seleccionado en muestra de segunda fase} \\ 0 & \text{si i NO es seleccionado en muestra de segunda fase} \end{cases}$$

- la probabilidad de selección de un individuo en la segunda fase depende de si el individuo fue seleccionado en la primera fase y puede llegar a depender de la información auxiliar $\mathbf{x}_i$ recolectada en la primera fase
- denotamos esta dependencia como $P(D_i = 1|\mathbf{Z})$, es decir, solo asumimos dependencia de $\mathbf{Z}$ y asumimos que la información auxiliar relevada en la primera fase conocida.
- el ponderador de la segunda fase depende de cuales individuos fueron seleccionados en la primera fase

$$w_i^{(2)} = \begin{cases} \frac{1}{P(D_i=1|\mathbf{Z})} = \frac{1}{\pi_i^{(2)|(1)}} & \text{si } Z_i = 1 \\ 0 & \text{si } Z_i = 0 \end{cases}$$

- el análogo del estimador HT en el muestreo en dos fases es:

$$\hat{Y}^{(2)} = \sum_{i \in s^{(2)}} w_i^{(1)} w_i^{(2)} y_i = \sum_{i \in U} Z_i D_i w_i^{(1)} w_i^{(2)} y_i$$

- el estimador anterior se le denomina "el estimador de expansión doble o HT*" dado que "expande" los datos y_i por el producto de los dos ponderadores muestrales
- se puede demostrar que:

$$V(^{(2)}) = V(\hat{Y}^{(1)}) + E(V[\hat{Y}^{(2)}|\mathbf{Z}])$$

donde $Y^{(1)} = \sum_{i \in s^{(1)}} w_i^{(1)} y_i$

- el primero término corresponde a la varianza que hubieramos obtenido si los valores de y_i hubieran sido observados para todos los individuos de la primera fase
- el segundo término es la varianza adicional por el hecho de realizar un sub-muestreo en la fase 2.
- la varianza en el muestreo en dos fases es SIEMPRE más grande que si hubiéramos recolectado la información de la variable y_i para todos los $n^{(1)}$ individuos seleccionados en la primera fase.
- esperamos que el segundo término sea pequeño en comparación con el estimador HT de tamaño $n^{(2)}$ que no utiliza ningun tipo de información auxiliar.