Conceptos básicos

Muestreo II

Licenciatura en Estadística

2023

Notación

- $U = \{1, 2, ..., N\}$ = población o universo de interés
- N = tamaño de la población U
- y = variable de interés (target variable)
- y_i = valor que toma la variable y en el individuo i
- $\theta = f(\mathbf{y})$ = parámetro de interés (e.g. total, media, proporción, ratio, etc.)
- $\mathbf{x}_i^T$ = set de variables auxiliares (covariables)
- s = muestra representativa (?) de la población U seleccionada bajo una estrategia de selección (diseño muestral) cualquiera $p(s)$
- n = tamaño de la muestra (i.e. cantidad de unidades seleccionadas)

estimación

En un principio nos centramos en un **parámetro sencillo** de la población U, como ser el total de una variable y cualquiera

$$Y = \sum_{i=1}^{N} y_i = \sum_{i \in U} y_i$$

El objetivo es intentar aproximar (i.e. **estimar**) el valor desconocido de Y utilizando los datos relevados (generalmente por medio de una encuesta) de la variable y en la muestra s .

estimación

El **problema** se centra en como hacer para **extrapolar/expandir** la muestra para estimar (representar) a los no encuestados (i.e. $i \in U - s$)

$$Y = \sum_{i \in s} y_i + \sum_{i \in U - s} y_i$$

muestra representativa

Definición

toda unidad de la población U tiene una probabilidad de inclusión en la muestra s mayor que cero

$$\pi_i = \text{prob}[i \in s] > 0 \ \forall i \in U$$

A tener en cuenta:

- no todas las muestras son como una versión **"miniatura"** de la población.
- unidades generalmente tienen probabilidades de inclusión π_i distintas.
- las π_i son elegidas de forma arbitraria o no por parte de el/la investigador/a
- si los datos se "**ponderan**" de forma correcta se pueden obtener estimaciones insesgadas

estimador Horvitz-Thompson (1952)

Las unidades incluidas en la muestra (s) son divididas por su probabilidad de selección π_i

$$\hat{Y} = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in s} w_i \times y_i$$

buenas propiedades teóricas:

- insesgado (i.e $E(\hat{Y}) = \text{mean}[\hat{Y}] = Y$)
- correlación positiva entre la variables de interés y y π_i mejora las precisiones (e.g. reducción del error estándar $< SE >$)

ponderadores w_i

- los ponderadores w_i provienen del diseño y son computados como el inverso de la probabilidad de selección $w_i = \pi_i^{-1}$
- son denominamos como ponderadores originales o basados en el diseño
- representan la cantidad de unidades de la población U que no fueron incluidas en la muestra menos 1 (ella misma). Por ejemplo si $w_i = 100$ significa que esa unidad se representa así misma y a otras 99 unidades no incluidas en s

inferencia basada en el diseño

únicamente se basan el proceso de selección (diseño muestral) que da lugar a la muestra s

- repetir lo mismo muchas veces. (i.e. las propiedades de los estimadores están basadas en la distribución de probabilidad del diseño muestral)
- las variables se consideran **fijas** (i.e. no son variables aleatorias)

inferencia basada en el diseño

En el caso del estimador HT:

- **Insesgado**: si promediamos las estimaciones $\hat{Y}$ de todas las muestras posibles coincide con el verdadero valor del parámetro Y
- $SE(\hat{Y})$cuantifica el spread, es decir, cómo varían las estimaciones $\hat{Y}$ entre muestra y muestra.

Varianza del estimador HT

la varianza teórica del estimador HT para el estimador del total Y es:

$$\text{var}(\hat{Y}) = \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$$

donde $\pi_{ij} = \text{P}(i \in s \ \& \ j \in s)$

y un estimador insesgado de la varianza es:

$$\widehat{\text{var}}(\hat{Y}) = \sum_{i \in s} \sum_{j \in s} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$$

Observación: si todos los pares de elementos tiene probabilidad mayor a cero de pertenecer a la muestra, decimos que el diseño es medible y el estimador de la varianza es insesgado.

errores estándar (SE)

El error estándar ($\sqrt{\text{var}}$) depende de varios factores:

- que tan distintas son las unidades de la población (varianza de y)
- tamaño de la muestra (n)
- tamaño de la población (N).
  - Aunque solo tiene impacto cuando la tasa de muestreo $f = n/N$ es grande (e.g. mayor a 0.10)
- diseño muestral (**muestreo I**)
- método del estimación (**muestreo II**)

errores estándar (SE)

- Una vez seleccionada la muestra y recolectada la información se procede a computar estimaciones PUNTUALES para distintos parámetros.
- Es una una buena práctica computar los errores estándar (SE)
- Los SE*s* son utilizados para computar medidas descriptivas de calidad de las estimaciones:

$$\widehat{cv}(\hat{\theta}) = \frac{\widehat{\mathrm{SE}}(\hat{\theta})}{\hat{\theta}}$$

- También son utilizados para realizar inferencias por medio de la construcción de margenes de error $moe = \text{cte} \times \widehat{\mathrm{SE}}(\hat{\theta})$ y/o intervalos de confianza $\hat{\theta} \pm moe$

otros parámetros

- **media**

$$\bar{Y} = N^{-1} \sum_{i \in U} y_i = \frac{Y}{N}$$

  - **obs:** si la variable y es una variable binaria/dicotómica la media es una proporción $\bar{Y} = p$

- **ratio:** cociente entre los totales de dos variables

$$R = \frac{\sum_{i \in U} y_{1i}}{\sum_{i \in U} y_{2i}} = \frac{Y_1}{Y_2}$$