no respuesta

Muestreo II

Licenciatura en Estadística

2023

dos enfoques:

- **deterministico:** la elección de una unidad no es aleatoria, i.e, podemos separar previamente a las unidades en dos estratos: respondentes (R) y no respondentes (NR).
- **estocástico:** cada unidad i tiene una probabilidad mayor que cero. La unidad realiza una elección aleatoria de responder o no.

bajo el enfoque deterministico, el sesgo de la media de una variable y cualquiera es:

$$\text{sesgo.NR}(\hat{\bar{Y}}_R) = \frac{M}{N}(\bar{Y}_R - \bar{Y}_{NR})$$

donde

- $\hat{\bar{Y}}_R$ es la media estimada utilizando a los respondentes de la muestra
- $\bar{Y}_R$ es la verdadera media de los respondentes en la población
- $\bar{Y}_{NR}$ es la media de los no respondentes en la población
- M es el tamaño de los NR en la población

enfoque + utilizado en la práctica

$$I_i = \begin{cases} 1 & \text{si la unidad i es seleccionada en la muestra} \\ 0 & \text{si la unidad i no es seleccionada en la muestra} \end{cases}$$

$$R_i = \begin{cases} 1 & \text{si responde dado que salió en la muestra} \\ 0 & \text{si no responde dado que salió en la muestra} \end{cases}$$

- $\text{Prob}[I_i] = \pi_i$
- $\text{Prob}[R_i = 1 | I_i = 1] = \phi_i$, en donde ϕ_i se le denomina propensity-score.
- las unidades que tienen $\phi_i = 0$ se les denomina **hard-core**, i.e, no importa los intentos que se hagan las mismas nunca van a responder.

bajo este enfoque el sesgo por NR se puede describir como:

$$\text{sesgo.NR}(\hat{\bar{Y}}_R) = \frac{1}{N\bar{\phi}} \sum (y_i - \bar{Y}_U)(\phi_i - \bar{\phi})$$

- el sesgo depende de la covarianza entre y y ϕ_i
- si ϕ no esta relacionada con la variable y entonces no habría sesgo (difícil!)

si **conocemos** las **verdaderas** probabilidades de responder ϕ_i podemos utilizar la "doble expansión" para obtener estimaciones insesgadas:

$$\hat{Y} = \sum_{i \in R} w_i^{nr} y_i$$

donde

$$w_i^{nr} = \frac{1}{\pi_i \times \phi_i}$$

obviamente $\phi_i > 0$ **siempre**

problema!

en la práctica las probabilidades o propensiones de responder ϕ_i son **desconocidas**

- nos vamos a centrar en intentar estimar (aproximar) las propensiones desconocidas ϕ_i
- vamos a tener que (lamentablemente) asumir cosas (i.e. un modelo).
- debemos abandonar el parádigma de la inferencias basadas en el diseño

inferencia provenientes de muestras aleatorias en la teoría

$$\left.\begin{array}{c}\text{respondieron}\\[1em]\text{conocimiento del proceso de selección}\end{array}\right\}\textbf{estimaciones}$$

---

inferencia provenientes de muestras aleatorias en la práctica

$$\left.\begin{array}{c}\text{respondieron (algunos)}\\[1em]\underbrace{\textbf{estimación}\text{ del proceso de selección}}_{\text{información auxiliar + asumir}}\end{array}\right\}\textbf{estimaciones}$$

- **Missing Completely at Random (MCAR)**: ocurre cuando la probabilidad de responder ϕ_i no depende ni de y_i ni de x_i, i.e., todos tienen la misma probabilidad de responder.
- **Missing at Random (MAR)**: ocurre cuando la probabilidad de responder ϕ_i no depende de la variable de interés y_i pero si depende de las variables auxiliares x_i.
- **NonIgnorable Nonresponse(NINR)**: ocurre cuando la probabilidad de responder ϕ_i depende de las variables de interés y_i. Debido a que no conocemos datos de y_i para los NR el sesgo es imposible de detectar y obviamente, imposible de eliminar.

ajuste creando clases de NR

- la idea es crear clases/grupos en donde todos tienen la misma probabilidad de responder o los mismos valores de la variable y
- si lo anterior se cumple y bajo el enfoque **deterministico**, eliminaríamos el sesgo ocasionado por la NR

problemas

1. no tenemos los valores de la y para los NR
2. las clases generalmente se basan en la probabilidad de responder
3. si usamos covariables $\mathbf{x}$ que sean buenas predictoras de la y es un plus!

creamos G clases y **asumimos** que todas las unidades dentro de la misma clase, tienen la misma probabilidad de responder.

- la probabilidad estimada para una clase g queda definida como:

$$\hat{\phi}_{i,g} = \mathrm{TR_w} = \frac{\sum_{i \in R} w_i}{\sum_{i \in s} w_i}$$

cómo se hace en la práctica?

- vamos probando distintas formas crear las clases en busca de aquellas que presenten distintas TR
- para crear las calases debemos tener información para toda la muestra (i.e. R y NR)
- generalmente esa información proviene del marco muestral F (e.g. estratos, UPM, etc.)

- asumimos que la NR es MAR, i.e. $\phi_i = \phi_i(\mathbf{x}_i)$
- si $\phi_i = \phi_i(y_i)$ estamos en problemas! ya que no tenemos datos de y para los NR
- otra situación es cuando $\phi_i = \phi_i(\mathbf{U}_i)$, es decir, la NR depende de información auxiliar $\mathbf{U}_i$ omitida o que no tenemos para construir el modelo.

intentamos estimar las propensiones $\hat{\phi}_i$ por medio de un modelo/algoritmo (e.g. logit/probit) usando covariables x conocidas tanto para R y NR una vez hecho esto, i.e. computados los $\hat{\phi}_i$ hay dos opciones:

1. ajuste por propensiones simples

$$w_i^{nr} = \frac{1}{\pi_i \times \hat{\phi}_i}$$

2. **propensiones estratificadas**: se utilizan $\hat{\phi}_i$ para crear clases de NR y luego se aplica un factor de ajuste común, el cual, es computado utilizando alguna métrica de resumen (e.g. la media o mediana).