Inferencia con muestras no probabilísticas

Muestreo II

Licenciatura en Estadística

2023

estimaciones utilizando muestras en la práctica

base rectangular con los datos de la muestra en donde una columna corresponde a los ponderadores muestrales (w)

ID	$y_{(1)}$	$y_{(2)}$	...	$y_{(J)}$	w
1	20	A	...	1	20
2	15	B	...	0	45
⋮	⋮	⋮	...	⋮	⋮
n	35	C	...	1	17

- **uni-weight** enfoque usualmente utilizado para producir estimaciones para distintos parámetros y dominios
- enfoque predominante en encuestas de gran escala

enfoque uni-weight

- facil para computar estimaciones puntuales. Las mismas son llevadas acabo anexando un ponderador a cada una de las observaciones
- crea economias de escala
- util en encuestas continuas
- genera consistencias entre las estimaciones de la encuesta
- los ponderadores w_i son computados generalmente utilizando estimadores calibrados/regresión teniendo en cuenta las necesidades de la encuesta
- NO es la estrategia óptima para todos los parámetros y/o dominios

enfoque basado en el diseño

- ha jugado un rol predominante en la producción de estadísticas oficiales
- hermosas propiedades teóricas (i.e. Alicia en el mundo de las maravillas)

enfoque basado en el diseño (en la teoria)

- $U = \{1, 2, ..., i, ...N\}$
- nuestro objetivo es realizar buenas estimaciones de

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

- definimos el diseño muestral de acuero a nuestras necesidades (e.g. estratos, distintas tasas de muestreo, etc.)

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

- seleccionamos la muestra (s) y todos los individuos están contentos de proporcionarnos los datos

enfoque basado en el diseño

estimador Horvitz-Thompson

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

buenas propiedades teóricas

- insesgado $\rightarrow \mathrm{E}(\hat{Y}^{\mathrm{HT}}) = Y$
- varianza "simple"

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

donde $\pi_{ij} = \mathrm{Prob}(i \in s \ \& \ j \in s)$

estimadores asistidos por modelos

- utilizamos un modelo superpoblacional m para predecir el valor de y para cada una de las unidades de U. Estimados el modelo $\hat{m}(\mathbf{x}_i)$ utilizando los datos de la muestra
- el rol del modelo m es de asistir al estimador

$$\hat{Y}^{\text{PRED}} = \sum_{i \in U} \hat{m}(\mathbf{x}_i) + \sum_{i \in s} w_i (y_i - \hat{m}(\mathbf{x}_i))$$

- es asintoticamente insesgado $\rightarrow \text{E}(\hat{Y}^{\text{PRED}}) = Y$

estimadores asistidos por modelos

no es importante que el modelo elegido sea verdadero

PERO su **precisión** dependerá del poder predictivo del modelo!

$$\widehat{\text{var}}(\hat{Y}^{\text{PRED}}) = \sum_{i \in s} \sum_{j \in s} \pi_{ij}^{-1} (\pi_{ij} - \pi_i \pi_j) \frac{e_i}{\pi_i} \frac{e_j}{\pi_j}$$

donde $e_i = y_i - \hat{m}(\mathbf{x}_i)$

GREG

si $m(.)$ es un modelo de regresión lineal:
$y_i = m(\mathbf{x}) + \epsilon_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i$ donde $\epsilon_i \sim (0, \sigma_i^2)$

el **estimador de regresión** es

$$\hat{Y}^{\text{GREG}} = \sum_{i \in U} \hat{y}_i + \sum_{i \in s} w_i (y_i - \hat{y}_i) = \sum_{i \in U} \mathbf{x}_i^T \hat{\mathbf{B}} + \sum_{i \in s} w_i (y_i - \mathbf{x}_i^T \hat{\mathbf{B}})$$

$$= \sum_{i \in s} \left[1 + \left(\sum_{i \in U} \mathbf{x}_i - \sum_{i \in s} w_i \mathbf{x}_i \right)^T \left(\sum_{i \in s} w_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \mathbf{x}_i \right] w_i y_i$$

si x es una variable categórica, el estimador GREG es igual al estimador Post-estratificado.

estimador asistido por arboles de regresión

los arboles de regresión utilizan un algoritmo para realizar una partición recursiva sobre el espacio de predicción en cajas/nodos $B1, ..B_l.., B_L$ de tal forma que las unidades dentro de cada caja son homogéneas respecto a la variable de interés y

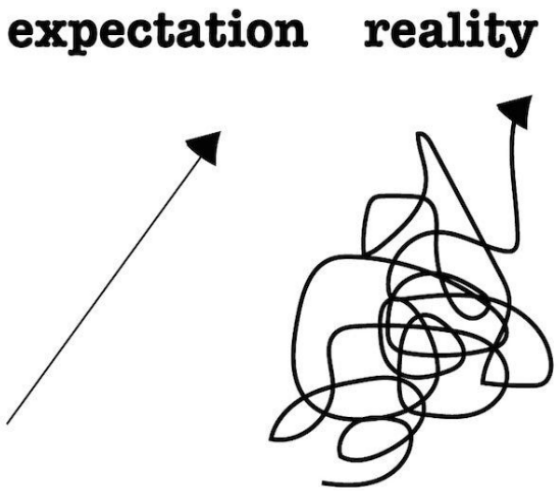
- es un estimador **post-estratificado** donde los post-estratos (celdas) son elegidas de forma **automática** por el algoritmo

$$\hat{Y}^{\text{RT}} = \sum_{l=1}^{L} \frac{N_l}{\sum_{i \in s_l} w_i} (\sum_{i \in s_l} w_i y_i)$$

realidad

bajo el enfoque de la inferencia basada en el diseño la calidad de las estimaciones es cuantificada por medio del error estándar (SE)

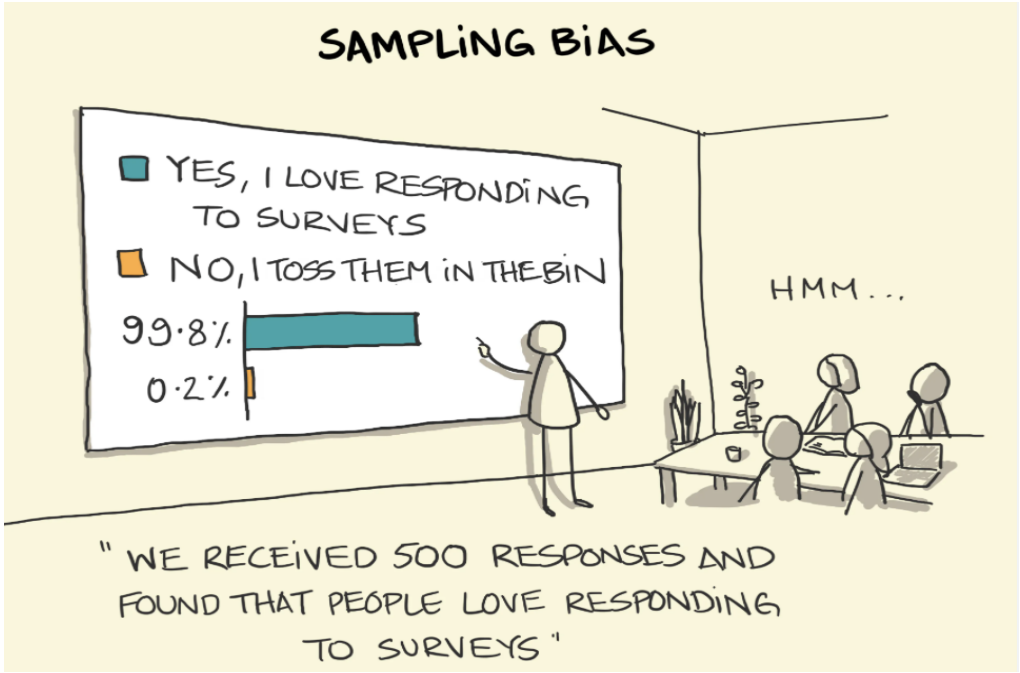
nuestras "herramientas" en teoría son insesgadas...pero la realidad dice otra cosa



realidad

- **problemas de cobertura** $\rightarrow$ algunas unidades tienen $\pi_i = 0 \rightarrow$ **sesgo**
- **existe no respuesta**$\rightarrow$ **sesgo**
  - utilizamos modelos para reducir el sesgo y realizamos supuestos respecto al mecanismo de no respuesta (**MAR**)
- los estimadores GREG deberían ser considerados model-based en lugar de model-assisted, debido a que la elección del modelo es crucial para poder reducir el sesgo

muestras no probabilísticas



es posible realizar inferencias!

pero necesitamos tener disponible mucha información auxiliar ($\mathbf{x}$) y asumir "cosas"(e.g. modelos y mecanismos de respuesta). No es libre de modelos!

diferentes enfoques

quasi-randomization

- modelar la probabilidad de aparecer en la muestra (B)
- necesitamos información **x** para casos los encuestados y no encuestados para estimar un modelo asumiendo MAR
- necesitamos $\mathbf{x}_i \ \forall i \in U$ o podemos utilizar una muestra de referencia (probabilistica) A
- si el modelo es VERDADERO, podemos realizar estimaciones insesgadas utilizando el estimador HT

$$\hat{Y} = \sum_{i \in B} \frac{y_i}{\hat{\phi}_i}$$

diferentes enfoques

modelos superpoblacionales

- utilizamos un modelo m para predecir los valores y de todas las unidades no encuestadas
- estimamos el modelo utilizando la muestra B. Si el modelo se mantiene, entonces:

$$\hat{Y} = \sum_{i \in B} y_i + \sum_{i \in U-B} \hat{m}(\mathbf{x}_i) \doteq \sum_{i \in U} \hat{m}(\mathbf{x}_i)$$

también podemos utilizar la muestra de referencia A en el caso de que no se tenga información auxiliar disponible en U

$$\hat{Y} = \sum_{i \in B} y_i + \sum_{i \in A} w_i^A \hat{m}(\mathbf{x}_i) \doteq \sum_{i \in A} w_i^A \hat{m}(\mathbf{x}_i)$$

estimador doble robusto (DR)

- utilizado las muestras A y B estimamos las propensiones asumiendo MAR
- usando la muestra B y las propensiones estimadas $\hat{\phi}_i$ estimamos el método $\hat{m}(\mathbf{x}_i)$

$$\hat{Y}^{\text{DR}} = \sum_{i \in A} w_i^A \hat{m}(\mathbf{x}_i) + \left[\sum_{i \in B} \frac{(y_i - \hat{m}(\mathbf{x}_i))}{\hat{\phi}_i} \right]$$

si el modelo superpoblacional m es verdadero y/o el modelo asumido para estimar las propensiones de participación es verdadero, entonces el estimador DR será insesgado

distintas elecciones

las probabilidades de participación pueden ser estimadas utilizando distintos métodos pero SIEMPRE asumiendo que el mecanismo de participación es MAR

- métodos paramétricos (e.g. logit)
- métodos no paramétricos (e.g. arboles de regresión, bosques aleatorios, etc.)

de igual forma con el modelo m para predecir la variable de interés y

- métodos paramétricos (e.g. linear model)
- métodos no paramétricos(e.g. **arboles de regresión**, bosquees aleatorios, etc.)

en esos casos, la elección dependerá de la información auxiliar disponible (i.e. microdata o conteos poblacionales)

DR utilizando arboles de regresión

si el método m es un árbol de regresión (RT), el estimador DR es:

$$\hat{Y}^{\text{DR}} = \sum_{i \in A} w_i^A \hat{m}(\mathbf{x}_i) + \left[\sum_{i \in B} \frac{(y_i - \hat{m}(\mathbf{x}_i))}{\hat{\phi}_i} \right]$$

$$= \sum_{l=1}^{L} \frac{\hat{N}_l^A}{\hat{N}_l^B} \left(\sum_{i \in B_l} \hat{\phi}^{-1} y_i \right) = \sum_{i \in B} w_i^* y_i$$

donde

- $\hat{N}_l^A = \sum_{i \in A_l} w_i^A$ es el estimador del tamaño poblacional de la caja l utilizando la muestra de referencia A
- $\hat{N}_l^B = \sum_{i \in B_l} \hat{\phi}_i^{-1}$ utilizando la muestra no probabilística B
- $w_i^* = (\hat{N}_l^A / \hat{N}_l^B) \hat{\phi}_i^{-1}$

Encuesta de las actitudes de los uruguayos hacia la población inmigrante

condición de elegibilidad: personas de 18 0 +

método de muestreo: Random digit dialing (RDD) a teléfonos celulares. No se realizó muestreo por cuotas.

- tamaño de muestra teórico: 30,000
- tasa de respuesta: 3.5% $\Longrightarrow$ muestra de voluntarios (B) de tamaño $n_B = 1050$

pregunta: cree que la inmigración es favorable para Uruguay?

- $y = 1$ si está de acuerdo, y $y = 0$ si no está de acuerdo.

parámetro de interés: proporción de las personas que apoyan la llegada de inmigrantes a Uruguay

estimación de las propensiones de participación

- muestra de referencia (A) = Encuesta Continua de Hogares llevada a cabo por el Instituto Nacional de Estadística.
  - una linda muestra aleatoria de $n_A = 90,000$
- juntamos A con B
- covariables $\mathbf{x}$ dispobibles en ambas muestras (A and B)
  - edad
  - sexo
  - educación
- $\hat{\phi}_i(\mathbf{x}_i)$ son estimadas utilizando un modelo logit.

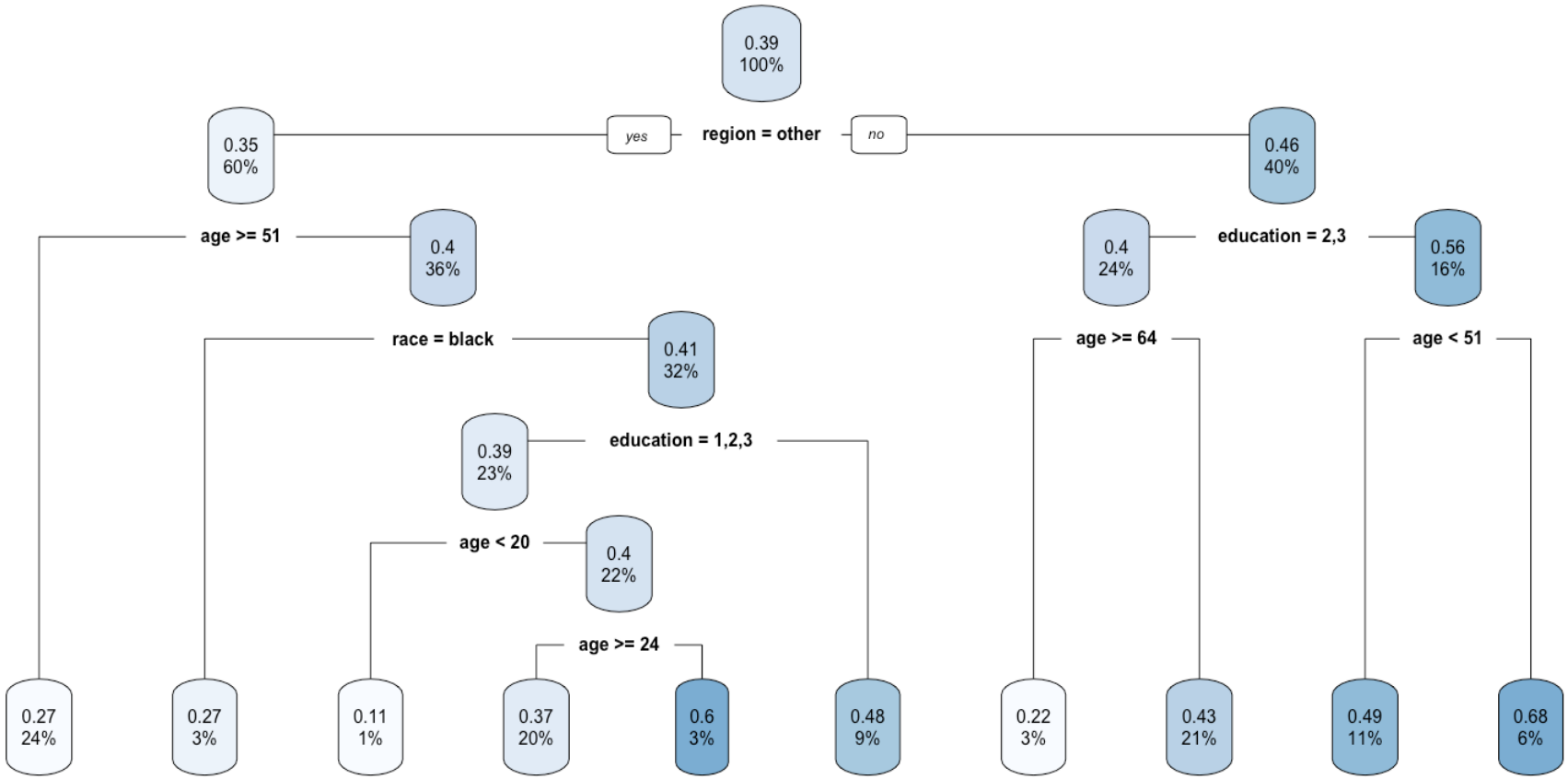
modelo superpoblacional

el método m elegido es un árbol de regresión

la estimación de m es llevado a cabo utilizando la muestra B y utilizando las probabilidades de participación estimadas $\hat{\phi}_i(\mathbf{x}_i)$

- $y =$ variable de interés y las covariables $\mathbf{x}$ son:
  - edad
  - sexo
  - raza
  - educación
  - región
- el tamaño mínimo de muestra permitido en las cajas (nodo terminal) es 10

el árbol de regresión



ponderadores finales de la muestra B

luego computamos las estimaciones de los tamaños poblacionales de las cajas N_l utilizando la muestra de referencia A y la muestra B:

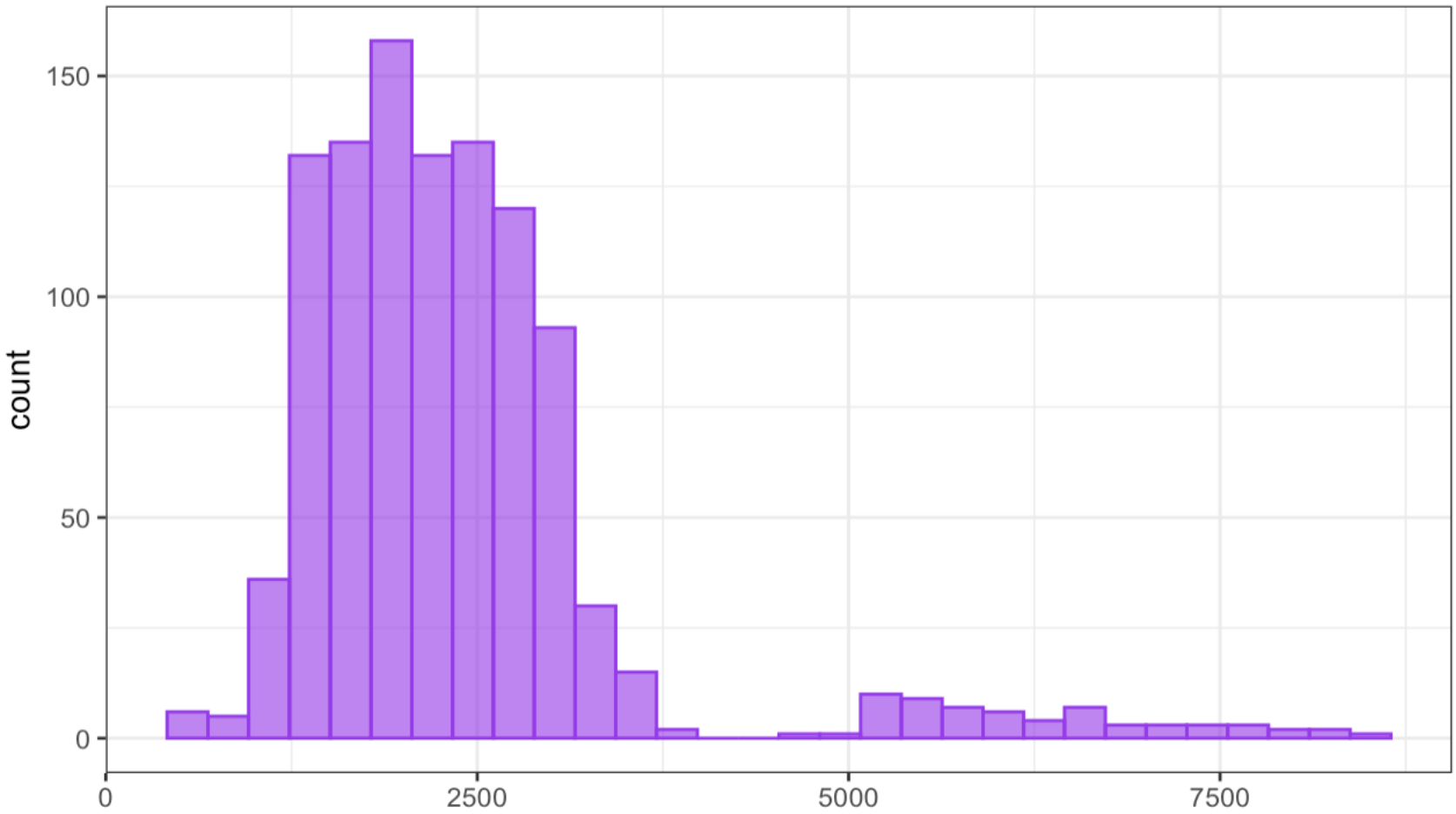
$$\hat{N}_l^A = \sum_{i \in A_l} w_i^A, \qquad \hat{N}_l^B = \sum_{i \in B_l} \hat{\phi}_i^{-1}$$

los ponderadores finales de la muestra B son:

$$w_i^* = (\hat{N}_l^A / \hat{N}_l^B)\hat{\phi}_i^{-1}$$

- **Importante:** podemos utilizar los ponderadores w_i^* pra estimar parámetros construidos con la variable no modelada. La solución no es óptima, pero crea economías de escala

distribución final de los ponderadores w_i^*



comparación de las estimaciones

- estimación pura (sin ponderar):

$$\widehat{\bar{Y}} = \frac{\sum_{i \in B} y_i}{n_B} = 0.432$$

- estimación utilizando las propensiones estimadas:

$$\widehat{\bar{Y}} = \frac{\sum_{i \in B} \hat{\phi}_i^{-1} y_i}{\sum_{i \in B} \hat{\phi}_i^{-1}} = 0.389$$

- estimaciones utilizando el estimador DR:

$$\widehat{\bar{Y}} = \frac{\sum_{i \in B} w_i^* y_i}{\sum_{i \in B} w_i^*} = 0.397$$

conclusiones

- necesidad de continuar realizando muestras "aleatorias" (A) para producir estimaciones de indicadores claves
- las muestras no probabilísticas (B) son una fuente barata para obtener estimaciones y las inferencias pueden ser llevadas a cabo utilizando la muestra de referencia (A)
- hay que ser HONESTOS. Se debe indicar si los modelos asumidos no son correctos, las inferencias serán invalidadas.