# Presentación de tesina

## Estimacion en diseños muestrales complejos en presencia de no respuesta

Tomás Capretto

Facultad de Ciencias Económicas y Estadística
Universidad Nacional de Rosario

30 de mayo 2019

# Introducción

- En Argentina, como en la mayoría de los países, existe un organismo público encargado de producir estadísticas oficiales (INDEC).
- Estadísticas referentes al territorio, población, economía y sociedad.
  - Planificación de políticas públicas.
  - Insumo de investigaciones del ámbito académico y privado.
- IPC, ENDI 2018, ENVI 2017.

# Diseños muestrales complejos

- Los parámetros de interés corresponden poblaciones distribuidas en un territorio geográfico muy extenso.
- Generalmente requiere la utilización de un diseño muestral complejo.

Las unidades que componen la muestra son el resultado de realizar una seleccion que presenta alguna de las siguientes características

- Múltiples etapas de muestreo.
- Selección de conglomerados en las etapas iniciales.
- Estratificación de unidades de al menos una de las etapas.
- Distintas probabilidades de selección.

# Muestreo en poblaciones finitas

El muestreo consiste en la selección de una parte de la población con el fin de obtener resultados para la misma.
Sea $U = \{1, \cdots, k, \cdots, N\}$ de tamaño $N$ la población de interés, donde $k$ representa a la unidad $u_k$, con $k = 1, \cdots, N$.

## Muestreo probabilístico

1. Permite definir el conjunto de muestras posibles, $\mathcal{S} = \{s_1, s_2, \cdots, s_M\}$.
2. Cada muestra $s$ tiene asociada una probabilidad de selección $p(s)$.
3. El procedimiendo da a cada unidad de la población una probabilidad no nula y conocida de selección, $\pi_k$.
4. Un mecanismo aleatorio selecciona una de las muestras posibles respetando $p(s)$.

Este carácter probabilístico posibilita la utilización de la teoría de inferencia estadística para realizar **estimaciones puntuales** y **cuantificar su imprecisión**.

# Errores presentes en un estudio por muestreo

Los estudios por muestreo están expuestos a diferentes tipos de errores, muestrales y no muestrales.

- El **error muestral** se produce por estimar características de interés en una población utilizando solo a una parte de esta.
- El **error no muestral** se divide en trés grandes áreas:
  - Error de cobertura
  - Error de medición
  - **Error de no respuesta**

# La problemática de la no respuesta

Los problemas de no respuesta aparecen cuando se obtienen cuestionarios no completos en su totalidad.

**La no respuesta refleja un intento no exitoso de obtener la información deseada de una unidad elegible para el estudio.**

La falta de respuesta es un factor que incide en la precisión de las estimaciones de un producto estadístico.

- Disminuye el tamaño de muestra efectivo.
  - Incrementa error de muestreo.
  - Limitación en la generalización de las conclusiones.
- Puede introducir sesgo en la estimaciones.

# La problemática de la no respuesta

Existen diversas sugerencias para tratar la no respuesta mediante un incremento de la tasa de respuesta.

- Mayor difusión de los operativos.
- Nuevos intentos de contacto.
- Personalización del enfoque de la encuesta.

Pero...

- En general, los Institutos de Estadística vienen reportando decrecimientos en la tasa de respuesta.
- A pesar de los esfuerzos realizados la respuesta nunca será completa.

# Tratamientos para lidiar con la no respuesta

Una solución viable y popular en la práctica es realizar algún tratamiento para lidiar con la no respuesta.

- **Imputación**
  Procedimiento mediante el cual valores faltantes son reemplazados por *valores artificiales*.
- **Ajuste de pesos muestrales**
  Incrementa los pesos muestrales de los encuestados de forma tal que se compense por la falta de respuesta cuando se realizan las estimaciones.

Generalmente, el primero es utilizado para tratar la no respuesta a nivel ítem, mientras que el segundo es utilizado para tratar la no respuesta a nivel unidad.

Esta tesina se reduce a estudiar la **no respuesta a nivel unidad**.

# Mecanismos de no respuesta

Sean $y_k$ y $\mathbf{x}_k$ una característica de interés y un vector de variables auxiliares de la unidad $k$ $\forall k \in U$.

Bajo el supuesto que todas las unidades tienen una probabilidad no nula de responder se pueden describir escenarios tres mecanismos que la generan.

- **Missing completely at random (MCAR):** Todas las unidades tienen la misma probabilidad de responder. Esta, no está relacionada con la variable en estudio, $y_k$, ni con ninguna variable auxiliar, $x_k$.
- **Missing at random (MAR):** La probabilidad que tiene la unidad $k$ de responder no está relacionada con $y_k$, pero si depende de $\mathbf{x_k}$.
- **Not missing at random (NMAR):** La probabilidad de responder no solo que puede estar relacionada a $\mathbf{x_k}$, sino que está relacionada a la variable de interés $y_k$.

# Sesgo por no respuesta

Previo a estudiar el sesgo por no respuesta necesitamos de algunas definiciones

Sean

$$I_k = \begin{cases} 1 & \text{si la unidad } k \text{ está en la muestra} \\ 0 & \text{caso contrario} \end{cases}$$

$$R_k = \begin{cases} 1 & \text{si la unidad } k \text{ responde dado que está en la muestra} \\ 0 & \text{si la unidad } k \text{ no responde} \end{cases}$$

La probabilidad que la unidad $k$ esté en la muestra, conocida como probabilidad de inclusión de primer orden, $P(I_k = 1) = \pi_k$, mientras que la probabilidad de que la unidad $k$ responda dado que está en la muestra es $P(R_k = 1 \mid I_k = 1) = \phi_k$.

# Sesgo por no respuesta

El estudio del sesgo que puede introducir la no respuesta suele hacerse en términos del estimador de la media de Hájek.

## Estimador de Hájek para la media $\bar{y}_U$

Sea $y$ una característica de interés y $s$ una muestra obtenida según el diseño muestral $p(s)$. El estimador de la media de Hájek se define

$$\bar{y}_{HJ} = \frac{\hat{t}_y}{\hat{N}} = \frac{\sum_{k \in s} d_k y_k}{\sum_{k \in s} d_k}$$

donde

- $d_k = \frac{1}{\pi_k}$ es el peso muestral de la unidad $k \in s$.
- $\pi_k$ es la probabilidad de inclusión de la unidad $k$, $\pi_k = \sum_{s \ni k} p(s)$.

$\bar{y}_{HJ}$ resulta sesgado de $\bar{y}_{k \in U} = \frac{\sum_U y_k}{N}$, pero su sesgo es depreciable si el coeficiente de variacion del denominador es pequeño.

# Sesgo por no respuesta

En presencia de no respuesta y considerando al estimador $\bar{y}_{HJ}$ calculado con el subconjunto de unidades que responde $s_r$, es posible demostrar que

$$B(\bar{y}_{HJ}) \approx \frac{1}{N\bar{\phi}} \sum_{k=1}^{N} (y_k - \bar{y}_U)(\phi_k - \bar{\phi}) = \frac{1}{\bar{\phi}} \sigma_\phi \sigma_y \rho_{\phi,y} = CV_\phi \sigma_y \rho_{\phi,y}$$

donde:

- $\bar{\phi}$ media poblacional de la propensión a responder.
- $\sigma_\phi$ desvío estándar poblacional de la propensión a responder.
- $\sigma_y$ desvío estándar poblacional de la variable en estudio.
- $\rho_{\phi,y}$ correlación poblacional entre la propensión a responder y la variable en estudio.
- $CV_\phi$ coeficiente de variación de la propensión a responder.

# Sesgo por no respuesta

Teniendo presente la expresión del sesgo

$$B(\bar{y}_{HJ}) \approx \frac{1}{N\bar{\phi}} \sum_{k=1}^{N} (y_k - \bar{y}_U)(\phi_k - \bar{\phi}) = \frac{1}{\bar{\phi}} \sigma_\phi \sigma_y \rho_{\phi,y} = CV_\phi \sigma_y \rho_{\phi,y}$$

Se observa que $\bar{y}_{HJ}$ es aproximadamente insesgado cuando se cumple al menos una de las siguientes condiciones

- La correlación entre la variable de interés y la probabilidad de responder es nula.
- Todas las unidades tienen la misma probabilidad de responder.

# Motivación del ajuste por no respuesta

Sea $d_k^*$ el peso que le corresponde a la unidad $k$ despues de realizar algún ajuste por no respuesta, y sea la siguiente estimador del total $t_y$

$$\hat{t}_y = \sum_{s \in s_r} d_k^* y_k$$

Si consideramos a la respuesta como una fase mas del muestreo, podemos calcular la esperanza de este estimador primero sobre el muestreo, $E_I$, y luego sobre el mecanismo de respuesta $E_R$

$$E_I E_R(\hat{t}_y) = E_I E_R \left( \sum_{k \in U} I_k R_k d_k^* y_k \right) = \sum_{k \in U} d_k^* y_k E_I E_R(I_k R_k)$$

Si fuera posible que $d_k^* = 1/E_I E_R(I_k R_k)$ la expresión se reduce al total poblacional y se obtiene un estimador insesgado de $t_y$.

# Motivación del ajuste por no respuesta

Notando que $E_I E_R(I_k R_k) = E_I[I_k E_R(R_k \mid Ik)] = \pi_k \phi_k$ se concluye que el estimador del total planteado es insesgado de $t_y$ cuando

$$d_k^* = (\pi_k \phi_k)^{-1}$$

Bajo las siguientes condiciones se obtiene un estimador insesgado del total:

- El muestreo es probabilístico, es decir, $\pi_k > 0 \ \forall k \in U$.
- Todas las unidades tienen una probabilidad no nula y conocida de responder en el caso de ser seleccionadas en la muestra, es decir, $\phi_k > 0 \ \forall k \in U$.

**Problema:** $\phi_k$ es desconocido.
En nuestro intento de obtener estimadores al menos aproximadamente insesgados tendremos que asumir que la respuesta es un fenómeno estocástico y elaborar alguna estrategia para ajustar los $d_k$.

# Ajuste por no respuesta

El principal objetivo de esta práctica es reducir el sesgo que pueda causar la no respuesta.

Se suelen distinguir dos categorías de métodos para ajustar los pesos del diseño $d_k$.

- **Enfoque de propensity score:** Esta clase de métodos busca lidiar con la no respuesta a través de la estimación directa de $\phi_k$ mediante algún modelo y lo utiliza de la forma antes expuesta.
- **Enfoque de calibración:** Se ajustan los pesos muestrales de los que responden de forma tal que las estimaciones de variables auxiliares coincidan con sus totales poblacionales conocidos.

# Ajuste mediante la propensión a responder

Supongamos que la probabilidad de responder responde al siguiente modelo

$$\phi_k = m(\mathbf{x_k}, \gamma)$$

donde $m(\cdot)$ es una función dada, $\mathbf{x_k}$ es un vector de variables auxiliares conocido tanto para los que responden como para los que no, y $\gamma$ es un vector de parámetros desconocidos.

Dada una estimación $\hat{\gamma}$ se obtiene $\hat{\phi}_k = m(\mathbf{x_k}, \hat{\gamma})$. Esto da lugar a un conjunto de pesos $\{\tilde{w}_k\}_{k \in s_r}$ donde $\tilde{w}_k = d_k \hat{\phi}_k^{-1}$.

Un estimador que utiliza este conjunto de pesos recibe le nombre de estimador ajustado por la propension a responder (PSA). Por ejemplo, para el total

$$\hat{t}_{PSA} = \sum_{k \in s_r} d_k \hat{\phi}_k^{-1} y_k = \sum_{k \in s_r} \tilde{w}_k y_k$$

# Ajuste mediante la propensión a responder

**Lo bueno:**
Si la relación funcional $m(\cdot)$ está correctamente especificada $\hat{t}_{PSA}$ es asintóticamente insesgado y consistente de $t_y$.

**Lo no tan bueno:**
No se suelen utilizar modelos paramétricos porque

- Pueden generar $\hat{\phi}_k$ muy pequeños $\implies$ $\{\tilde{w}_k\}_{k \in s_r}$ muy dispersos.
- Suelen ser vulnerables ante incorrecta especificación de $m(\cdot)$ (desconocida).

La práctica habitual es utilizar un procedimiento no paramétrico, que se espera que provea cierta protección ante una incorrecta especificación de $m(\cdot)$.

# Ajuste mediante clases de ponderación

Si se pueden crear clases donde las unidades tengan una probabilidad similar de responder o valores similares de la variable respuesta, el sesgo por no respuesta puede ser reducido drásticamente.

**Algunas alternativas**

- Estimar $\phi_k$ mediante un modelo paramétrico. Crear clases de unidades con $\hat{\phi}_k$ similares, por ejemplo, basados en percentiles. Calcular un factor de ajuste común para todas las unidades de una misma clase.
- Utilizar los niveles o una combinación de los niveles de una o mas variables categóricas como clases de ajuste. Por ejemplo, el estrato al que la unidad pertenece en alguna etapa de muestreo.

Independientemente del método elegido, una de las formas posibles de calcular el factor de ajuste para la clase $c$ con $c = 1, \cdots, C$ es

$$\hat{\phi}_k^{-1} = a_c = \frac{\sum_{k \in s_c} d_k}{\sum_{k \in s_{r,c}} d_k}$$

# El enfoque de calibración

- Clase de estimadores presentados originalmente por Särndall y Deville (1992).
- No fue pensado originalmente como un método para corregir por no respuesta.

Sea $\mathbf{x_k}$ el vector de $p$ variables auxiliares para la unidad $k$ y $\mathbf{t_x} = \sum_U \mathbf{x_k}$ el vector de totales poblacionales conocidos. El estimador de calibración para el total $t_y$ es uno de la forma

$$\hat{t}_y^{CAL} = \sum_{k \in s} w_k y_k$$

donde el conjunto de pesos $\{w_k\}_{k \in s}$ satisface la ecuación de calibración, $\hat{\mathbf{t}}_\mathbf{x} = \sum_{k \in s} w_k \mathbf{x_k} = \mathbf{t_x}$

# Determinación de los pesos de calibración

Sea $G$ una función de distancia de argumento $r_k = w_k/d_k$, que mide la distancia entre $\{w_k\}_{k\in s}$ y $\{d_k\}_{k\in s}$. $G$ debe ser positiva, convexa y satisfacer $G(1) = G'(1) = 0$.

$$\min_{w_k} \sum_{k\in s} d_k G(w_k/d_k)$$
$$\sum_{k\in s} w_k \mathbf{x_k} = \mathbf{t_x}$$

Este problema de búsqueda de extremo condicional en $p$ ecuaciones no lineales con $p$ incógnitas se resuelve introduciendo un vector de multiplicadores de Lagrange $\lambda$ y la solución general es de la forma

$$w_k = d_k F(\mathbf{x_k'}\lambda)$$

donde $F$ es la inversa de la función $G$.

# Elección de la función de distancia

El **método lineal** es muy sencillo y en ciertos casos presenta ecuaciones de calibración con forma cerrada. En el caso que se deban utilizar métodos numéricos converge en muy pocas iteraciones. En este caso

$$F(\mathbf{x}_k'\lambda) = 1 + \mathbf{x}_k'\lambda \text{ con } \mathbf{x}_k'\lambda \in \mathbb{R}$$

Pero este no ofrece ningún control sobre el rango de los factores de ajuste obtenidos.

El **método logit** permite asignar una cota inferior $L$ y una cota superior $U$ a $F(\mathbf{x}_k'\lambda)$. En este método la función $F$ toma la forma de una función logística y satisface

$$F(\mathbf{x}_k'\lambda) \in (L, U)$$

expresado de otra forma

$$Ld_k < w_k < Ud_k$$

# Calibración en conteos conocidos de tablas de frecuencia

Los Institutos de Estadística suelen calibrar en base a totales conocidos de una tabla de frecuencias (celdas o márgenes).

Supongamos una tabla de dos dimensiones con $n_{row}$ filas y $n_{col}$ columnas,

| | $X_{2,1}$ | $X_{2,2}$ | $\cdots$ | $X_{2,n_{col}}$ | |
|---|---|---|---|---|---|
| $X_{1,1}$ | $N_{11}$ | $N_{12}$ | $\cdots$ | $N_{1n_{col}}$ | $N_{1+}$ |
| $X_{1,2}$ | $N_{21}$ | $N_{22}$ | $\cdots$ | $N_{2n_{col}}$ | $N_{2+}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $X_{1,n_{row}}$ | $N_{n_{row}1}$ | $N_{n_{row}2}$ | $\cdots$ | $N_{n_{row}n_{col}}$ | $N_{n_{row}+}$ |
| | $N_{+1}$ | $N_{+2}$ | $\cdots$ | $N_{+n_{col}}$ | $N$ |

- Calibración en totales $N_{ij}$, entendido como **post-estratificación completa**. Cada una de las $n_{row} \times n_{col}$ celdas es un post-estrato.
- Calibración en totales marginales, entendido como *post-estratificación incompleta* o **estimor de raking**.

# Calibración en los márgenes (raking)

En el caso de calibración en los márgenes de una tabla de dos dimensiones se tiene que

$$\mathbf{x_k} = (\delta_{1\cdot k}, \delta_{2\cdot k}, \cdots, \delta_{n_{row}\cdot k}, \delta_{\cdot 1k}, \delta_{\cdot 2k}, \cdots, \delta_{\cdot n_{col} k})$$

donde $\delta_{i\cdot k} = 1$ si la unidad $k$ pertenece a la fila $i$ y 0 en el caso contrario, y $\delta_{\cdot jk} = 1$ si la unidad $k$ pertenece a la columna $j$ y 0 en el caso contrario. Luego

$$\mathbf{t_x} = (N_{1+}, \cdots, N_{n_{row}+}, N_{+1}, \cdots, N_{+n_{col}})'$$

Es frecuente que las variables auxiliares utilizadas sean conteos de personas según categorías sociodemográficas. Por ejemplo, intervalos de edad y sexo.

# Calibración como método para corregir por no respuesta

- Puede **reducir el error muestral y el error por no respuesta**, siempre y cuando se cuente con *buena* información auxiliar.
  - Las variables del vector auxiliar están relacionadas con la no respuesta y la característica de interés.
- Aplicable para cualquiera de los diseños muestrales conocidos y cualquier vector auxiliar.

El conjunto de pesos que se obtiene es $\{w_k\}_{k \in s_r}$ donde $w_k = d_k F(\mathbf{x}_\mathbf{k}' \lambda_\mathbf{r})$ y $\lambda_r$ es determinado de forma tal que cumpla con las ecuaciones de calibración. El estimador de calibración para estimar el total de $y$

$$\hat{t}_y^{CAL} = \sum_{k \in s_r} w_k y_k = \sum_{k \in s_r} d_k F(\mathbf{x}_\mathbf{k}' \lambda_\mathbf{r}) y_k$$

- Cuando se utilizan tablas conteos de tablas de frequencias subyace la idea de que dentro de las clases formadas por los niveles de las variables auxiliares los valores de $y$ son similares $\implies$ es factible *inflar* los pesos de los que responden para representar a los que no responden dentro de cada clase.

# Sesgo del estimador de calibración por no respuesta

Recordemos que $\hat{t}_{PSA}$ es asintóticamente insesgado y consistente si $m(\cdot)$ está bien especificada.

El estimador $\hat{t}_y^{CAL}$ goza de las mismas priopedades si se cumple

$$F(\mathbf{x}_\mathbf{k}'\lambda_\mathbf{r}) = \hat{\phi}_k^{-1}$$

Se puede demostrar que el sesgo asintótico de $\hat{t}_y^{CAL}$ es

$$B\left(\hat{t}_y^{CAL}\right) \approx \sum_{k\in U}(\phi_k F_k - 1)(y_k - \mathbf{x}_k^T \mathbf{B}_{\phi f})$$

donde

$$\mathbf{B}_{\phi f} = \left(\sum_{k\in U}\phi_k f_k \mathbf{x}_k \mathbf{x}_x^T\right)^{-1} \sum_{k\in U}\phi_k f_k \mathbf{x}_\mathbf{k} y_k$$

con $F_k \equiv F(\lambda_N^T \mathbf{x}_k)$, $f_k \equiv F'(\lambda_N^T \mathbf{x}_k)$ y $\lambda_N$ es el valor al que $\lambda_r$ converge en probabilidad.

# Sesgo del estimador de calibración por no respuesta

$$B\left(\hat{t}_y^{CAL}\right) \approx \sum_{k \in U} (\phi_k F_k - 1)(y_k - \mathbf{x}_k^T \mathbf{B}_{\phi f})$$

El estimador $\hat{t}_y^{CAL}$ es aproximadamente insesgado asintóticamente cuando

1. La covariancia poblacional entre los residuos $e_k = (y_k - \mathbf{x}_k^T \mathbf{B}_{\phi f})$ y $\delta_k = (\phi_k F_k - 1)$ es nula.
   Esta condición se satisface si $y_k = \mathbf{x}_k^T \beta + \epsilon_k$ con $E(\epsilon_k \mid \mathbf{x}_k) = 0$ y si $\phi_k$ no está relacionada con $y_k$ luego de condicionar en $\mathbf{x}_k$.

2. O cuando $F_k = \phi_k^{-1}$.
   Una función de calibración $F(\cdot)$ que satisfaga esta condición asegura que el estimador $\hat{t}_y^{CAL}$ es asintóticamente insesgado sin importar la característica de interés, incluso si no se cumple la condición anterior.

# Ajuste por no respuesta en uno y dos pasos.

1. **Ajuste en dos pasos:**

- Se multiplica al peso original $d_k$ por un factor de ajuste por no respuesta, $\hat{\phi}_k^{-1}$. El objetivo es **reducir el sesgo por no respuesta**.
- Calibración del peso ajustado. Se busca consistencia entre las estimaciones obtenidas y los totales poblacionales conocidos. Además, si $\mathbf{x}_k$ está relacionado a $y_k$ el estimador calibrado tiende a ser mas eficiente. $w_k^{C,2} y_k = d_k \hat{\phi}_k^{-1} F(\mathbf{x}_k^{-1} \lambda)$.

2. **Ajuste en un paso:**

- Solo se calibran los pesos originales. No estima explicitamente probabilidades de respuesta. $w_k^{C,1} = d_k F(\mathbf{x}_k^{-1} \lambda)$.
- Mismos objetivos: Reducir sesgo por no respuesta, asegurar consistencia entre estimacioes y totales conocidos, y posiblemente, reducir variancia de estimadores puntuales.
- La elección de $F(\cdot)$ tiene mayor relevancia.

# Estimación de variancia

Se debe considerar su precisión, tiempo de cómputo, simplicidad e incluso procedimientos administrativos.

- Necesaria para indicar precisión del estimador y calcular intervalos de confianza.
- Abunda la teoría para **diseños simples** y **estimadores lineales**.

El ajuste de pesos muestrales tiene el propósito de disminuir el sesgo por no respuesta, pero produce estimadores no lineales.

**Alternativas**

- Linealización.
  - Aproxima un estimador no lineal con uno lineal y estimar la variancia de este último.
- Replicación.
  - Obtener submuestras. Estimar el parámetro como si fuera la muestra completa. La variancia se estima a partir de las estimaciones obtenidas en cada replicación.

# Bootstrap

- Presentado por Efron (1979) en el contexto de estadística clásica.
- No existe una única alternativa en el muestreo de poblaciones finitas.

**Algunas ventajas**

- Permite reflejar implícitamente en la variancia los efectos de los ajustes por no respuesta y la calibración.
- Bootstrap cuenta con la bondad que no solo permite obtener estimaciones de variancia, sino que también permite estimar la distribución en el muestreo del estimador.

# Bootstrap según Rao, Wu, Yue (1992)

Realiza escalado sobre los pesos muestrales.

Sea $\theta$ un parámetro de interés de una población finita y $\hat{\theta}$ su estimador basado en la muestra original.

Dado un diseño estratificado multietápico, repetir un número grande de veces B, generalmente entre 200 y 500:

- Seleccionar *srswr* de $m_h$ UPM de las $n_h$ UPM en la muestra, independientemente para cada estrato $h$, con $h = 1, \cdots, H$.
- Calcular el peso de la replicación $b$ para la unidad $k$ mediante

$$w^*_{hik} = \left[ \left\{ 1 - \sqrt{\frac{m_h}{n_h - 1}} \right\} + \sqrt{\frac{m_h}{n_h - 1}} m^*_{hi} \right] w_{hik}$$

  donde $m^*_{hi}$ es el número de veces que la UPM $i$ del estrato $h$ es seleccionada en la muestra con reemplazo.
- Calcular $\hat{\theta}^*_b$, réplica $b$ de $\hat{\theta}$, utilizando $w^*_{hik}$ en vez de $w_{hik}$.

# Bootstrap según Rao, Wu, Yue (1992)

La variancia Bootstrap del estimador $\hat{\theta}$ viene dada por

$$\hat{V}_{BS} = \frac{1}{B} \sum_{b=1}^{B} (\hat{\theta}_b^* - \hat{\theta})^2$$

donde $\hat{\theta}$ podría ser reemplazado por $\hat{\theta}_{(\cdot)}^* = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b^*$

**Notas**

- En general se utiliza $m_h = n_h - 1$.
- Estima la variancia mediante la variabilidad de las unidades de primera etapa.
- Si las fracciones de muestreo en los estratos de primera etapa son lo suficientemente pequeñas.
  - La contribución a la variancia de las siguientes etapas son depreciables.
  - El sesgo del estimador es depreciable también.

# Resumiendo. . .

- Los estudios por muestreo a gran escala suelen recurrir a diseños complejos.
- La no respuesta es un fenómeno siempre presente.
  - Sea la tríada $(\phi_k, y_k, \mathbf{x}_k)\ \forall k \in U$
  - Si $\mathbf{x}_k$ y $\phi_k$ están fuertemente relacionados, y además existe asociación entre $\mathbf{x}_k$ y $y_k \implies$ suponer que no existe relación entre $\phi_k$ y $y_k$ va contra la intuición.
- Diferentes alternativas para realizar ajustes por no respuesta.
  - Pero obtenemos estimadores no lineales.
- Métodos de replicación para obtener estimaciones de variancia.
  - En este caso hablamos de un tipo de Bootstrap, pero no es la única opción.

# Aplicación

## Introducción

- Estudio por simulación que busca imitar un diseño muestral complejo a gran escala con presencia de no respuesta, similar a los que puedan darse en un Instituto de Estadística de un país.
- Diversos escenarios de no respuesta y diversas estrategias para realizar el ajuste por no respuesta.
- Evaluación del sesgo de los diferentes estimadores y del desempeño de Bootstrap según Rao, Wu, Yue como método general para estimar la variancia en todos los casos.
- Se utiliza una población sintética de viviendas del Estado de Texas correspondiente al año 2010.

## Ubicación de Texas

# Elaboración del marco muestral

## Materiales

- Población sintética compuesta por 8.921.047 viviendas que albergan a 24.512.643 personas, agrupadas a nivel de condado.
  - Variables de la vivienda: Ingreso anual y cantidad de personas.
  - Variables de la persona: Edad y sexo.
- Viviendas georreferenciadas mediante latitud y longitud.
- Conjunto de polígonos para las localidades del Estado, al 1/2018.

## Asignación de localidad

- Intersección punto-polígono. 79.68% de las viviendas asignadas.
- Las viviendas remanentes fueron asignadas a la localidad más cercana.

## Construcción de áreas de muestreo (USM)

- Compuestas por aproximadamente 200 viviendas contiguas de la misma localidad.

# Elaboración del marco muestral

- Se utilizó un algoritmo que es una adaptación del **método iterativo del vecino más cercano**.
- Permite crear clusters de tamaños pre-definidos, supongamos un tamaño **m**.

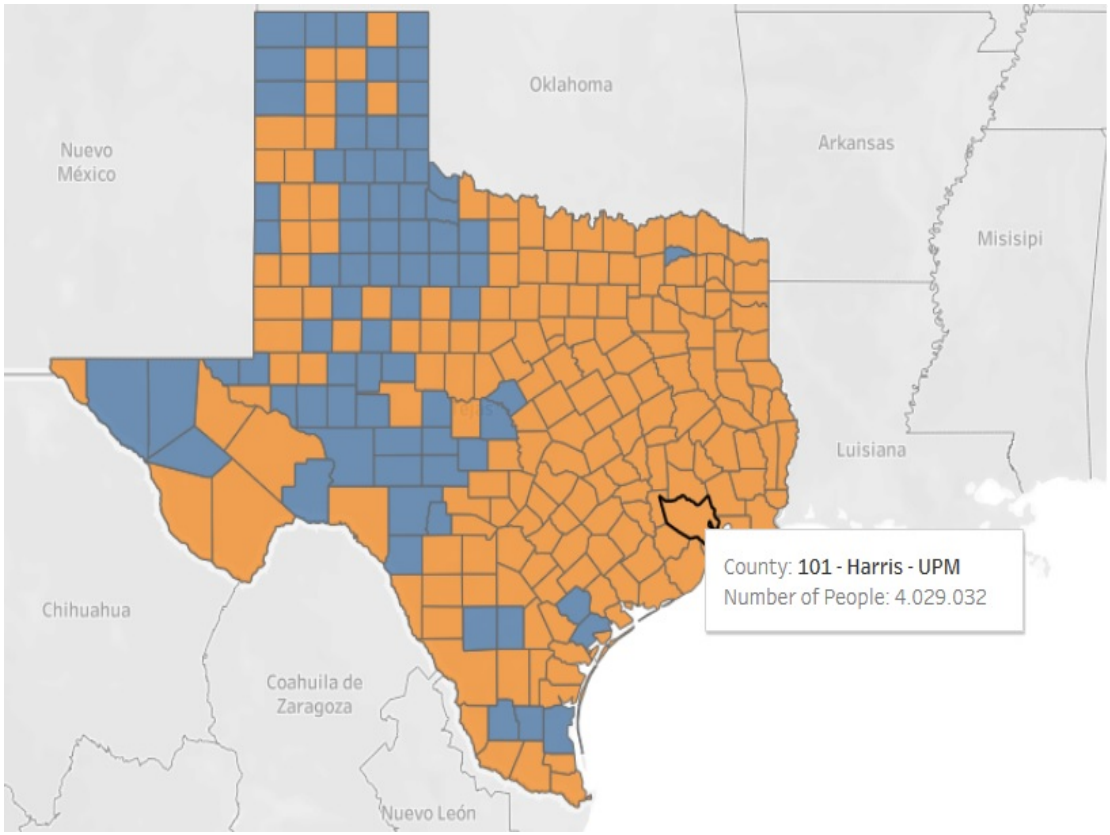
Mientras haya puntos sin asignar:

- Se selecciona un punto. En el primer ciclo se toma un punto al azar y en los restantes puede ser al azar o seguir una regla específica.
- Se calcula distancia a cada uno de los puntos. El punto seleccionado y los **m-1** vecinos mas cercanos son asignados a un nuevo cluster.
- Estos puntos son removidos del listado, evitando entrar en un bucle y/o asignar puntos a multiples grupos.

# Elaboración del marco muestral

## Unidades Primarias de Muestreo (UPM)

Las UPM se corresponden con condados siempre y cuando la población de éste sea de 7500 habitantes o mayor. Caso contrario, se corresponde con un conjunto de condados contiguos.




Los 254 condados del estado de Texas conformaron un total de 206 Unidades Primarias de Muestreo.

# Elaboración del marco muestral

**Filtros operativos y estratificación**

Pensando en un hipotético operativo de campo...

1. Eliminación de todas aquellas áreas de muestreo con menos de 50 viviendas.
2. Eliminación de las UPM con menos de 10 áreas.

La población del marco quedó conformada por 8.087.602 viviendas que albergan 21.349.048 personas. La cantidad de UPM pasó de 206 a 180.

3. Estratificación de UPM: Se conformaron 4 estratos según cantidad de habitantes. Uno de ellos agrupa a las UPM que contienen aglomerados urbanos más importantes y será incluido con probabilidad 1 en las muestras de la simulación.

# Características del estudio de simulación

Hipotético estudio por muestreo que tiene por objetivo estimar el ingreso medio de los hogares del Estado.

**Diseño muestral**

Muestreo probabilístico en 3 etapas, con las siguientes características:

- **Etapa 1:** Seleccionar una muestra $s_I$ de UPM con probabilidad proporcional a la cantidad de viviendas, independientemente en cada estrato . Se incluyen todas las UPM del estrato de inclusión forzosa.
- **Etapa 2:** Para cada UPM $\in s_I$ se selecciona una muestra $s_{II_i}$ de áreas de muestreo mediante *srswor*.
- **Etapa 3:** Dentro de cada USM $\in s_{II_i}$ se selecciona una muestra $s_{iq}$ de viviendas mediante *srswor*.

# Características del estudio de simulación

**Tamaños muestrales por etapa y estrato de UPM**

| Etapa | Estrato de UPM | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| I | 30 | 20 | 8 | 7 |
| II | 10 | 30 | 60 | 250 |
| III | 10 | 10 | 10 | 10 |
| Total de viviendas | 3000 | 6000 | 4800 | 17500 |

El tamaño muestral en término de unidades elementales es $n = 31300$.

# Características del estudio de simulación

Las fracciones de muestreo en la primera etapa fueron $f_1 = 0.29$, $f_2 = 0.38$ y $f_3 = 0.5$ para los primeros 3 estratos de UPM. El estrato de inclusión forzosa tiene 7 UPMs. Las fracciones de muestro dentro de cada una se encuentran entre 0.035 y 0.188.

**Asignación de valores faltantes**

- **MCAR:** Se le asignó a cada hogar la misma probabilidad de responder, igual al porcentaje de respuesta deseado.
- **MAR:** La probabilidad de responder asignada a cada hogar es proporcional a la cantidad de personas en el hogar.
- **NMAR:** Cada hogar tiene una probabilidad de responder inversamente propocional su ingreso.

Todos los casos fueron simulados con dos porcentajes de respuesta: 60% y 80%.

# Características del estudio de simulación

## Ajuste de pesos

1. **Sin ajuste:** Se utilizan los pesos del diseño, $w_k^* = d_k$. $\longrightarrow \hat{\bar{y}}_{HJ_r}$.
2. **Ajuste por no respuesta 1:** Se ajusta utilizando clases de ponderación. La variable de clasificación es el número de personas en el hogar, $w_k^* = d_k \phi_k^{-1} \longrightarrow \hat{\bar{y}}_{NR1}$.
3. **Ajuste por no respuesta 2:** Se ajusta utilizando clases de ponderación. La variable de clasificación es el estrato de UPM, $w_k^* = d_k \phi_k^{-1} \longrightarrow \hat{\bar{y}}_{NR2}$.
4. **Calibración:** Se calibran los pesos a conteos poblacionales marginales según intervalos de edad y sexo, $w_k^* = d_k F(\mathbf{x}_k' \lambda) \longrightarrow \hat{\bar{y}}_{CAL}$.
5. **Ajuste por no respuesta 1 + Calibración:** Se ajustan los pesos del diseño de igual manera que en **2.** y luego se calibran de igual manera que en **4.**, $w_k^* = d_k \phi_k^{-1} F(\mathbf{x}_k' \lambda) \longrightarrow \hat{\bar{y}}_{NR1+CAL}$.
6. **Ajuste por no respuesta 2 + Calibración:** Se ajustan los pesos del diseño de igual manera que en **3.** y luego se calibran de igual manera que en **4.**, $w_k^* = d_k \phi_k^{-1} F(\mathbf{x}_k' \lambda) \longrightarrow \hat{\bar{y}}_{NR2+CAL}$.

# Características del estudio de simulación

**Forma del estimador y ajuste de pesos utilizados**

En todos los escenarios simulados el estimador del ingreso medio de los hogares toma la forma del estimador de Hájek, calculado con el conjunto que responde

$$\hat{\bar{y}}_{HJ_r} = \frac{\sum_{k \in s_r} w_k^* y_k}{\sum_{k \in s_r} w_k^*}$$

- Se estima la variancia del estimador puntual utilizando Bootstrap según Rao, Wu, y Yue (1992).

# Características del estudio de simulación

**Simulación y evaluación de los estimadores**

- Se obtienen 10.000 muestras según el diseño muestral presentado.
- Con cada una de las muestras se obtienen 3 subconjuntos de respondentes, uno para cada mecánica de respuesta.
- En cada uno de los 3 conjuntos de respondentes se obtienen los 6 conjuntos de pesos detallados.
- Se estima el ingreso medio en los 18 escenarios resultantes.
- Además, con las primeras 1.000 muestras se estima la variancia del estimador puntual utilizando $B = 200$ réplicas.

El procedimiento descripto se realiza una vez para cada tasa de respuesta.

A continuación se presentan las medidas resumenes calculadas.

# Características del estudio de simulación

**Sesgo relativo del estimador puntual**

$$RB_{SIM}(\hat{\bar{y}}) = \frac{E_{SIM}[\hat{\bar{y}}] - \bar{y}_U}{\bar{y}_U}$$

donde $E_{SIM}[\hat{\bar{y}}] = \frac{1}{10000} \sum_{j=1}^{10000} \hat{\bar{y}}_{(j)}$.

**Sesgo relativo del estimador de variancia**

$$RB_{SIM}[\hat{V}(\hat{\bar{y}})] = \frac{E_{SIM}[\hat{V}(\hat{\bar{y}})] - V_{SIM}(\hat{\bar{y}})}{V_{SIM}(\hat{\bar{y}})}$$

donde
$E_{SIM}[\hat{V}(\hat{\bar{y}})] = \frac{1}{1000} \sum_{j=1}^{1000} \hat{V}(\hat{\bar{y}})_{(j)}$ esperanza simulada de la variancia y
$V_{SIM}(\hat{\bar{y}}) = \frac{1}{10000} \sum_{j=1}^{10000} \left[\hat{\bar{y}}_{(j)} - E_{SIM}[\hat{\bar{y}}]\right]^2$ variancia simulada del estimador.

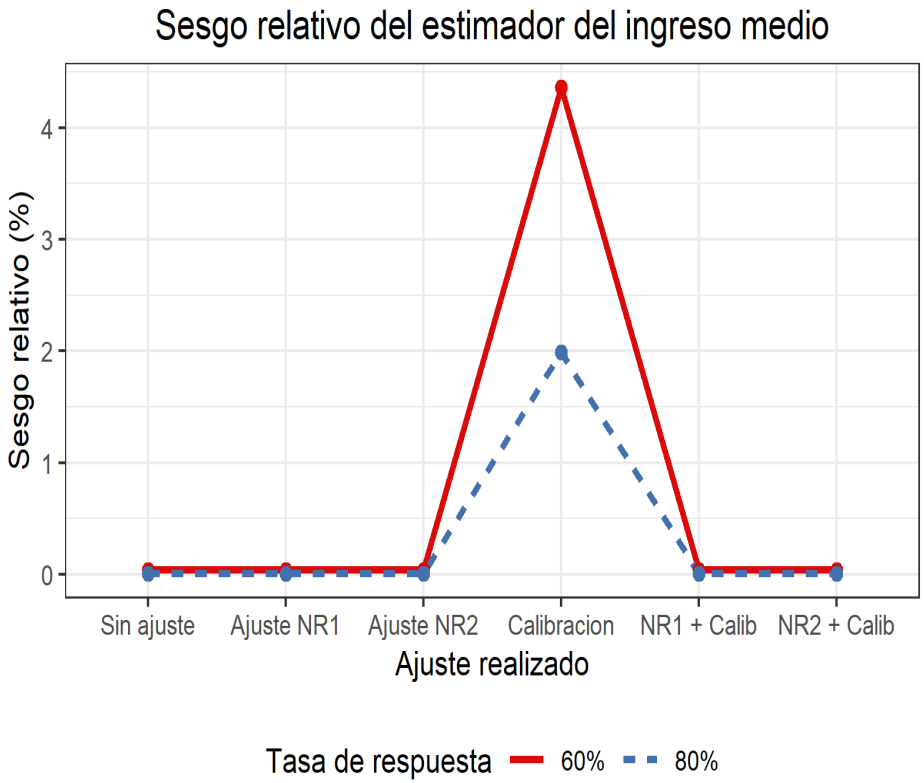
# Características del estudio de simulación

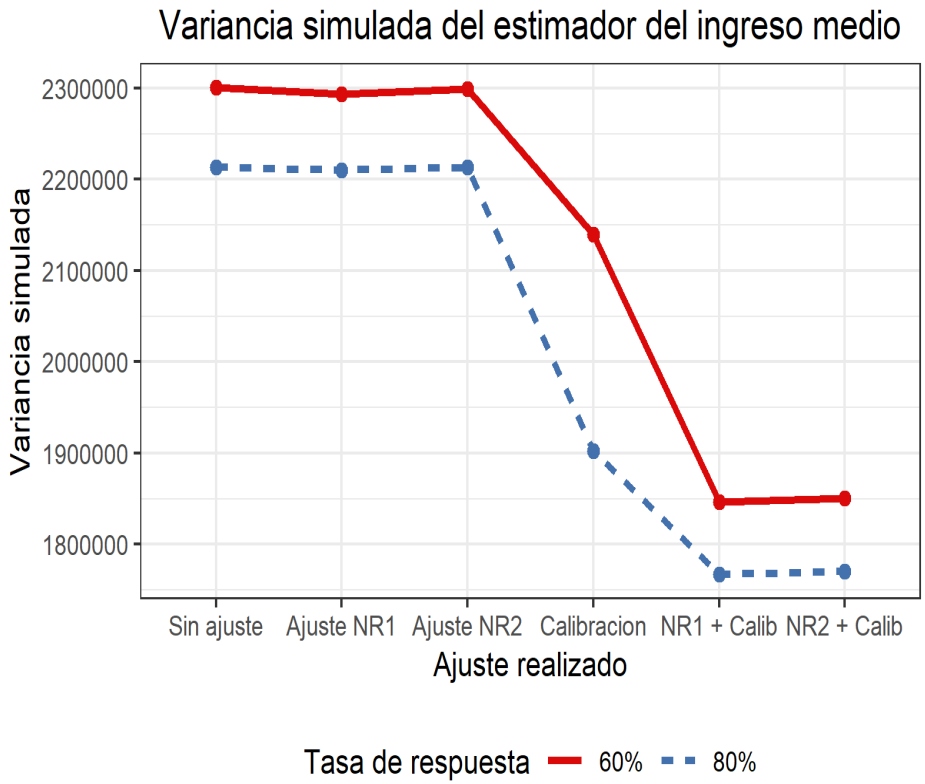
Por último, si bien no se compara el estimador de variancia propuesto con otros estimadores, es interesante ver si su dispersión varía con los ajustes de pesos propuestos.

$$V_{SIM}[\hat{V}(\hat{\bar{y}})] = \frac{1}{1000}\sum_{j=1}^{1000}\left[\hat{V}(\hat{\bar{y}})_{(j)} - E_{SIM}[\hat{V}(\hat{\bar{y}})]\right]^2$$

## Sesgo relativo y variancia simulada del estimador puntual

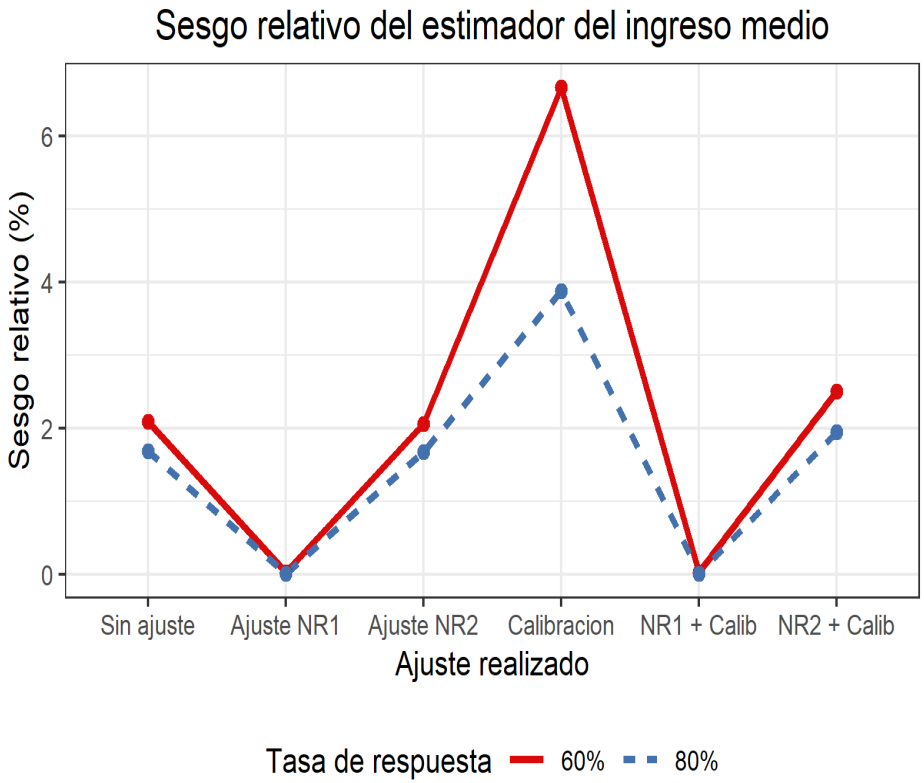
### Missing completely at random

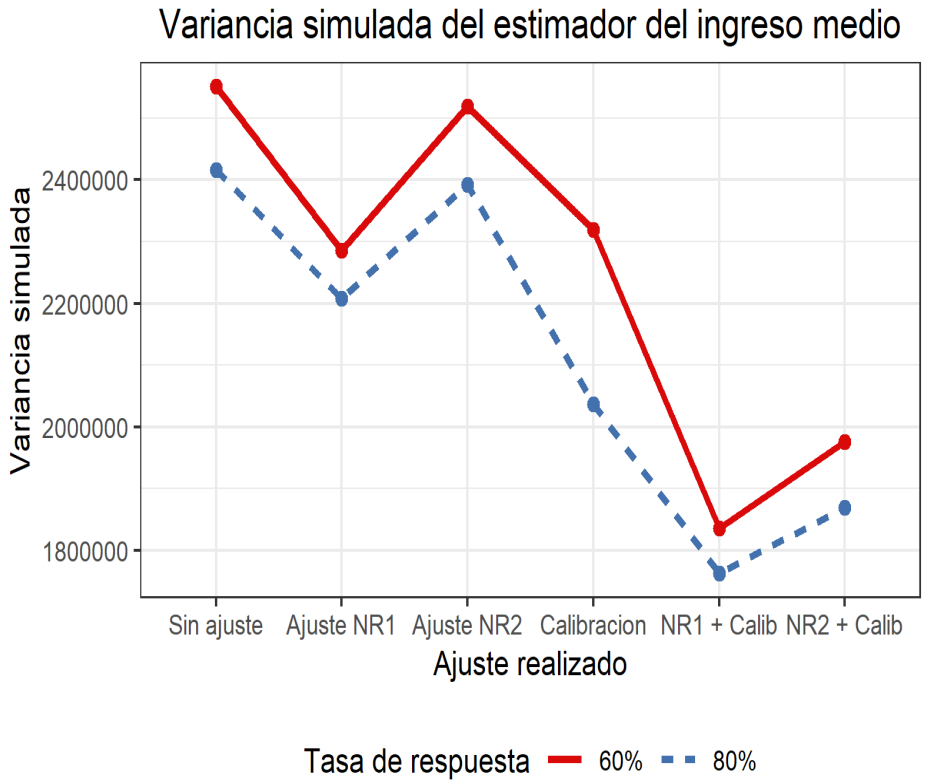


## Sesgo relativo y variancia simulada del estimador puntual

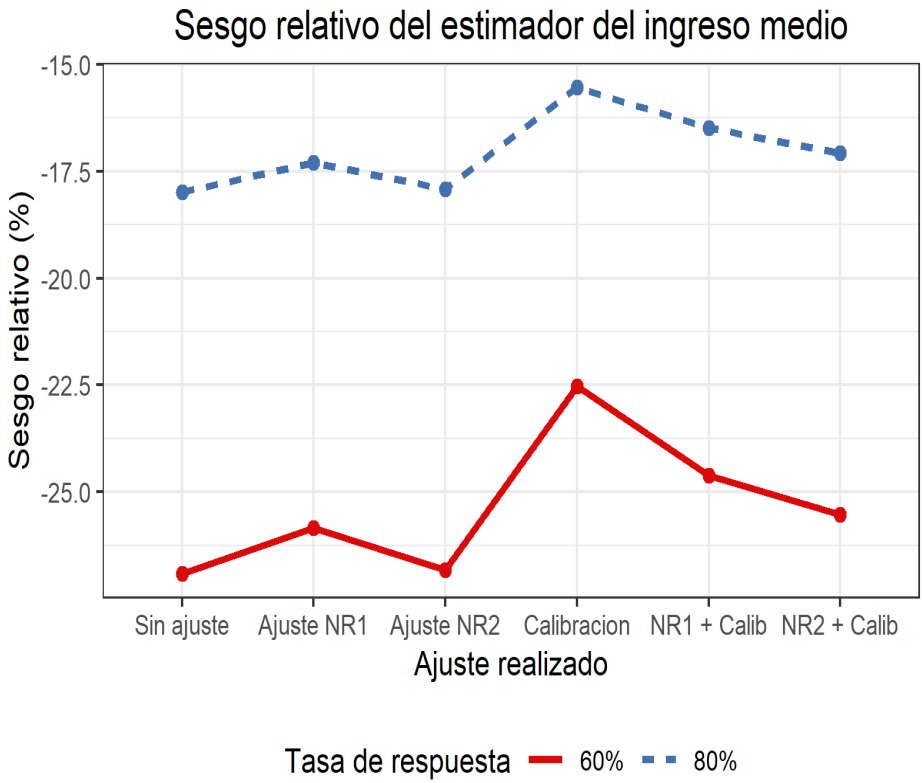
### Missing at random

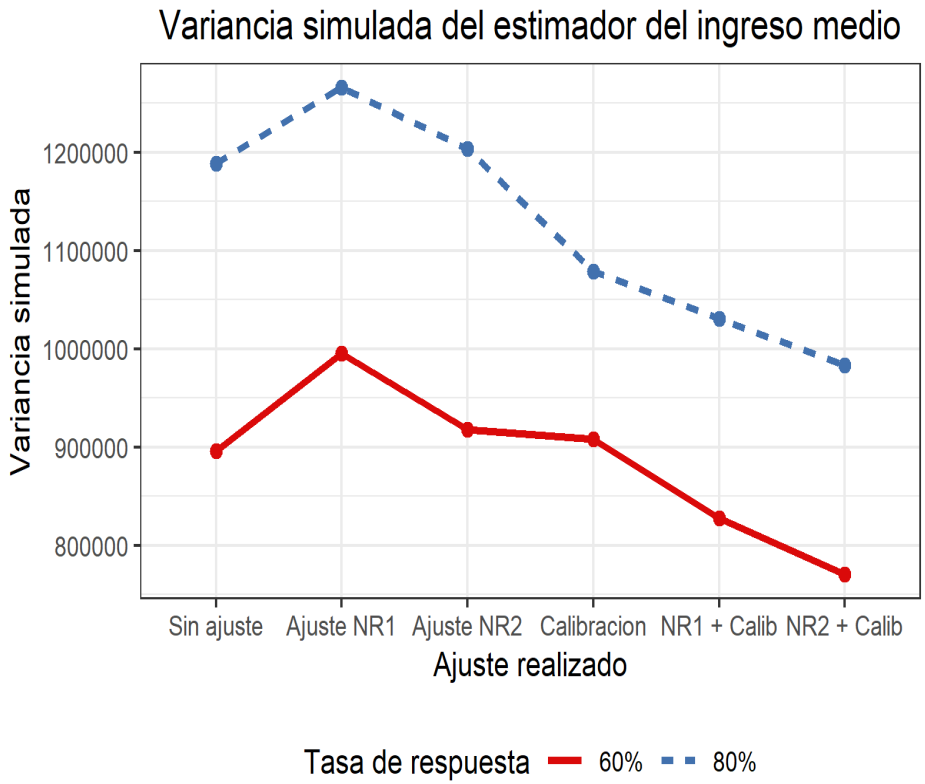


## Sesgo relativo y variancia simulada del estimador puntual

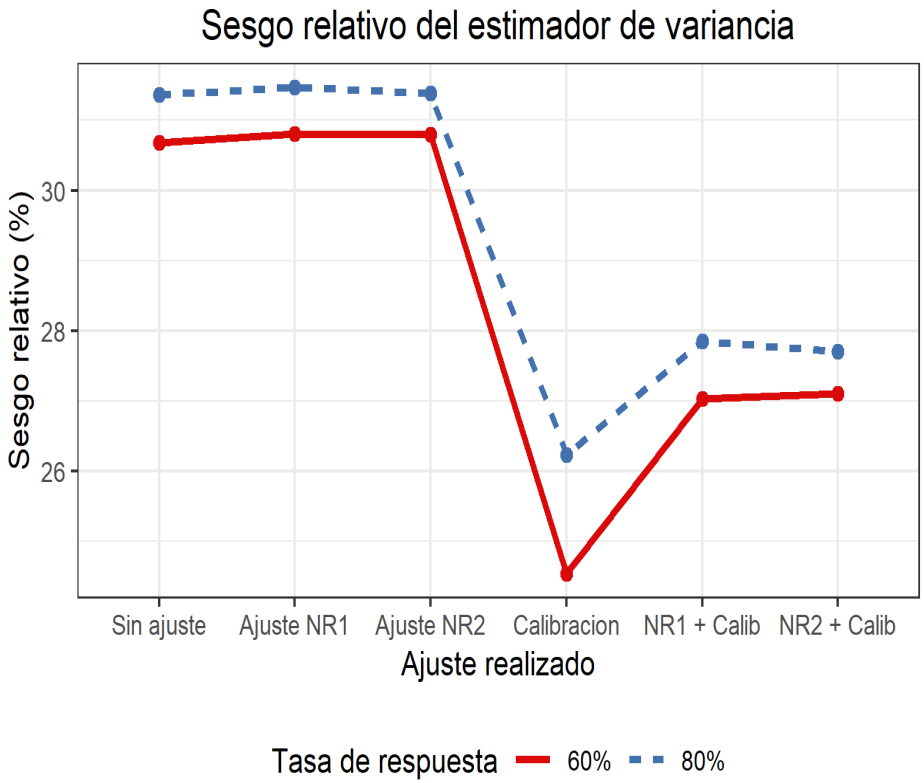
### Not missing at random

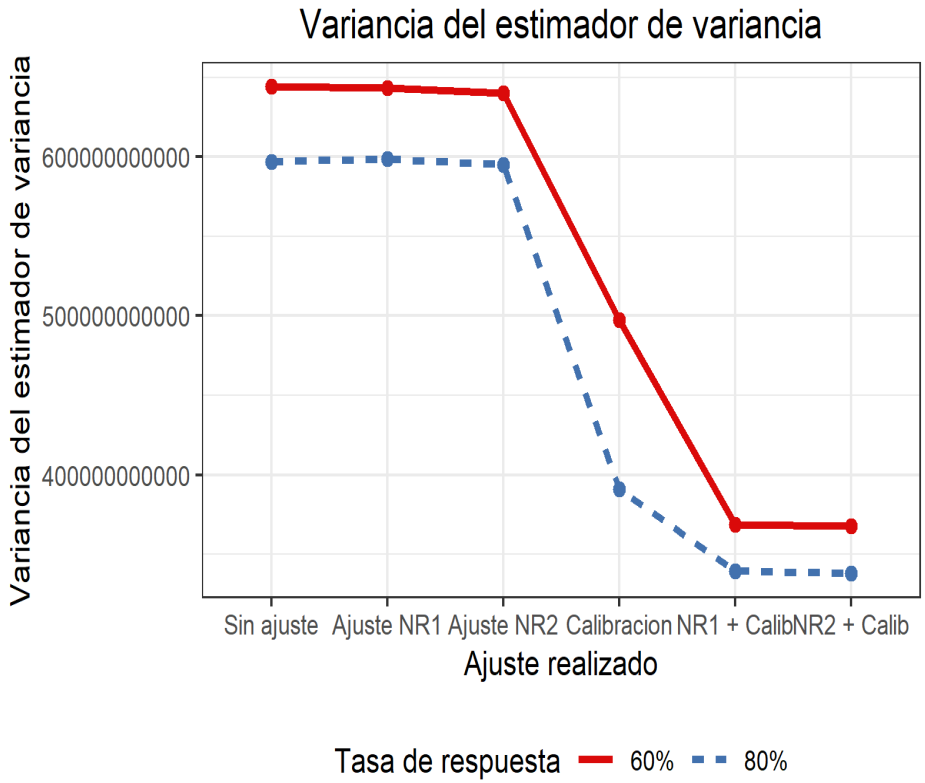


## Sesgo relativo y variancia simulada del estimador de variancia

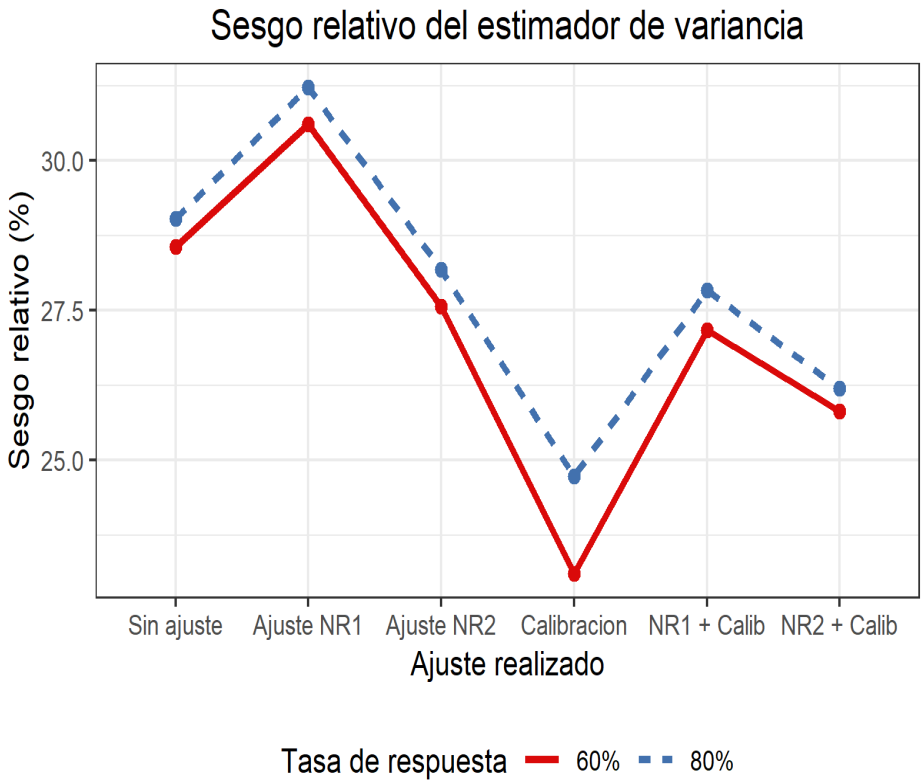
### Missing completely at random

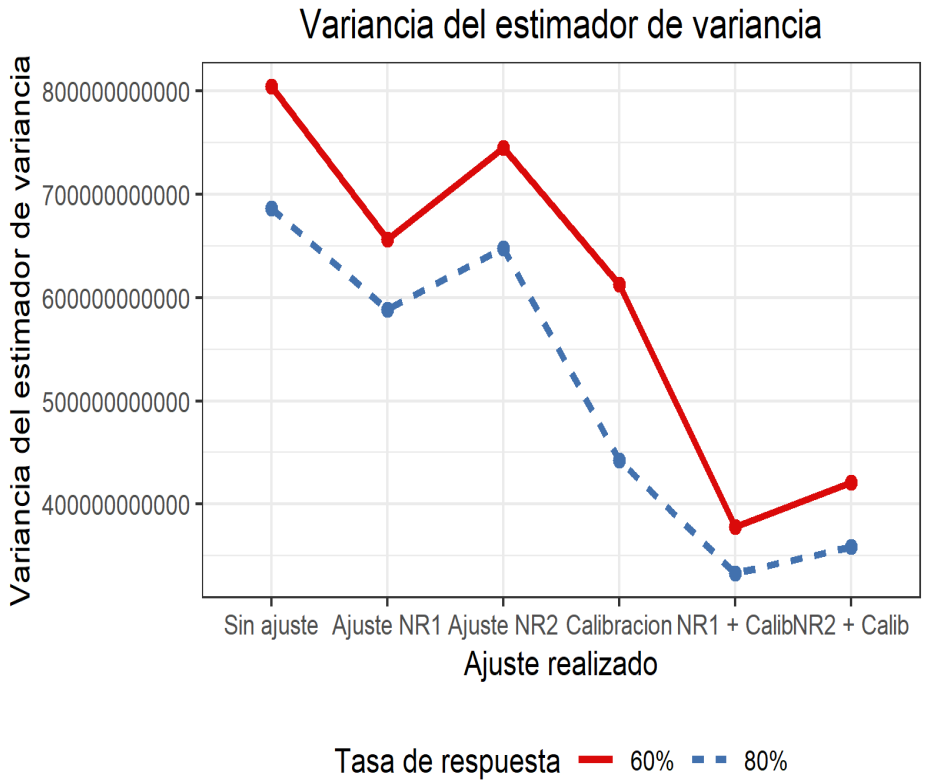


## Sesgo relativo y variancia simulada del estimador de variancia

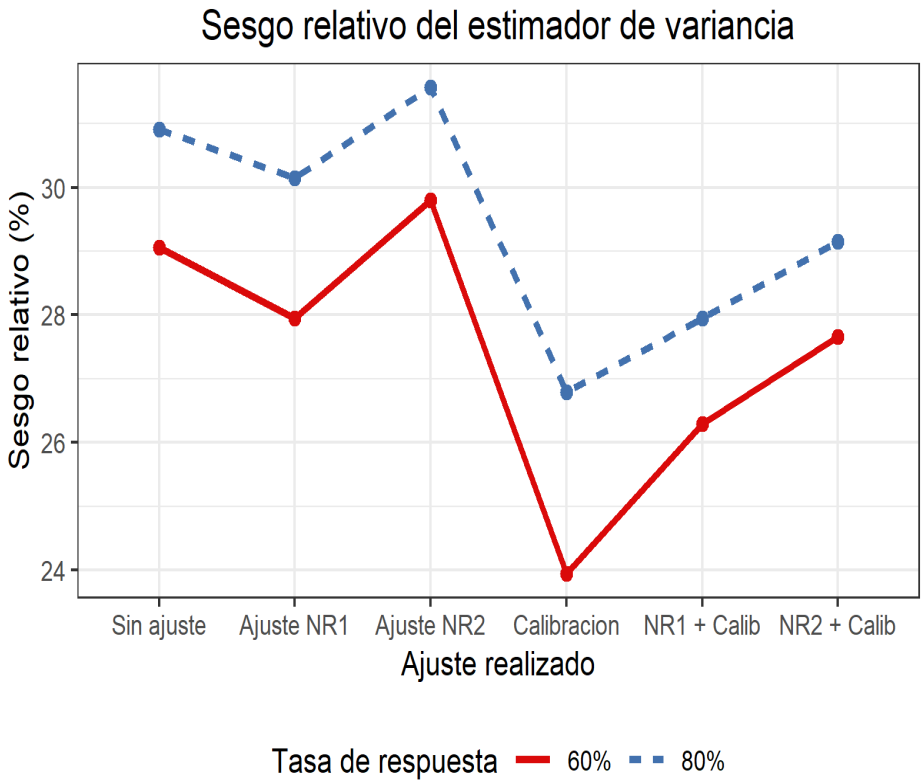
### Missing at random

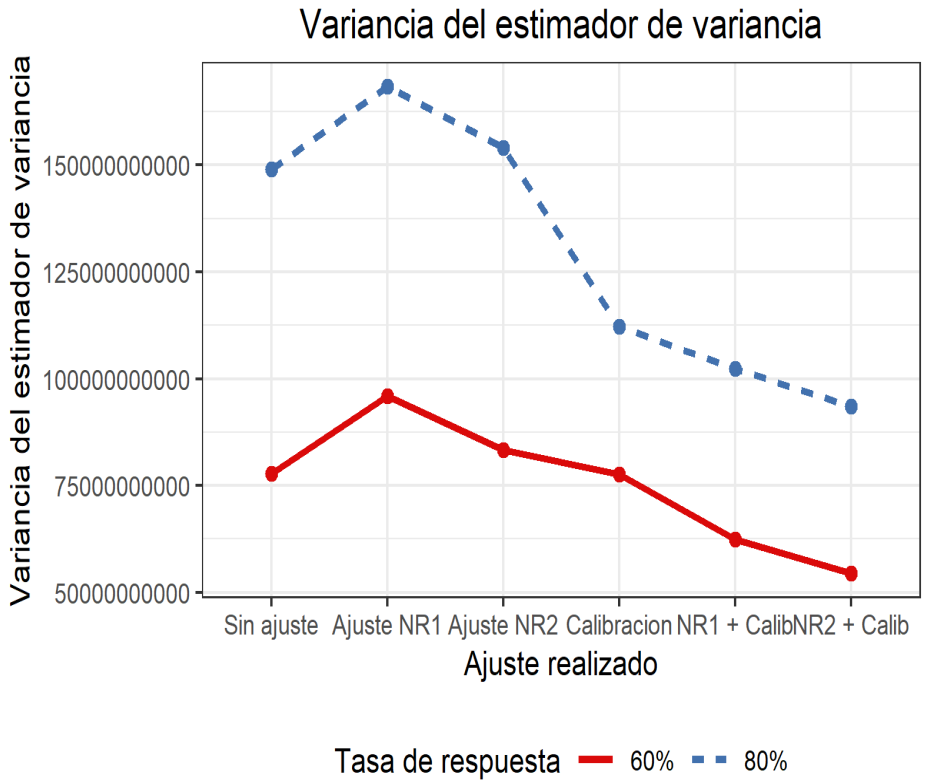


## Sesgo relativo y variancia simulada del estimador de variancia

### Not missing at random



# Conclusiones

**En términos generales**

- Presentamos los mecanismos subyacentes que pueden dar origen a la no respuesta en un estudio por muestreo.
- Se mencionaron las consecuencias en términos del sesgo.
- Vimos algunas de las opciones disponibles en la literatura para tratar la no respuesta una vez que ha ocurrido.
- Tenemos el problema de los estimadores no lineales. Propusimos métodos de replicación.

# Conclusiones

**En términos del sesgo del estimador puntual**

- Si la respuesta es MCAR observamos que no se introduce sesgo.
  - Igualmente hay incremento en el error de muestreo.
  - El estimador calibrado si presenta sesgo. Porque?
- Si la respuesta es MAR y conocemos las variables relacionadas a la no respuesta, podemos eliminar el sesgo casi completamente.
  - Aunque en diseños a gran escala es muy improbable.
  - El escenario simulado es muy simplista.
- Si la no respuesta está exclusivamente relacionada a la variable en estudio el sesgo es proporcional al porcentaje de respuesta y no se puede mitigar (a diferencia de MAR).
  - En la realidad, la respuesta puede estar también relacionada a variables auxiliares utilizadas en algún modelo para mitigar el sesgo.

# Conclusiones

**En términos de la variancia**

- El ajuste a un paso (solo calibración) logra reducir la variancia del estimador puntual y del estimador de variancia.
  - Pero no olvidemos que introdujo sesgo en el estimador puntual.
- El ajuste a dos pasos parece respetar la disminución del sesgo por el ajuste mediante $\hat{\phi}_k^{-1}$ y la reducción en su variancia asociada a la calibración.
  - La asociación a una menor variancia es independiente de que la no respuesta introduzca sesgo o no.
  - Se corresponde con recomendaciones halladas en la bibliografía.

# Conclusiones

**Respecto a Bootstrap como estimador de variancia**

- Las estimaciones de variancia son muy conservadoras.
  - Nos previene de IC exageradamente estrechos... pero es mucho sesgo.
- Pueden ser varios los motivos que causen el sesgo.
  - El muestreo utilizado sin reemplazo es mas eficiente que si fuera realizado con reemplazo.
  - Fracciones de muestreo en estratos de UPM no depreciables.
- Otros estudios muestran resultados similares cuando la fracción de muestreo y el sesgo.
- Existen alternativas que buscan utilizar las fracciones de muestreo en el proceso de estimación.
  - No muy estudiadas aún.