

# **TESINA**

**“Estimación en diseños muestrales complejos en presencia  
de no respuesta”**

Para optar por el título de  
Licenciado en Estadística

Alumno: Tomás Capretto

Director: M. Sc. Gonzalo Pablo Marí



Facultad de Ciencias Económicas y Estadística

Universidad Nacional de Rosario

## **Agradecimientos**

Es enorme la felicidad que se me presenta en este momento cúlmine. No es solo consecuencia de un gran esfuerzo realizado o del fruto de tantas horas de dedicación. Es consecuencia, principalmente, de lo afortunado que me siento al haber sido acompañado por semejantes personas no solo durante la realización de esta tesina, sino a lo largo de toda mi etapa como estudiante. Es a ellos, a quienes quiero agradecerles.

En primer lugar, a Dios y a mi familia, por la vida y ser siempre soporte fundamental.

A mis amigos, por su amistad invaluable.

A Juli, por su amor y el aguante incondicional que me tiene.

A mi Director, Gonzalo. Nada de este trabajo hubiera sido posible sin su dedicación, conocimiento y compromiso.

A todos los profesores de la carrera, porque supieron transmitirme mucho más que solo conocimiento.

A mis compañeros, por los momentos de estudio compartidos, los mates, el camino transitado.

## **Resumen**

Los estudios por muestreo están expuestos a diferentes tipos de errores, muestrales y no muestrales. Uno de los tipos de errores no muestrales es la no respuesta. Esta refleja un intento no exitoso de obtener toda la información deseada de una unidad elegible para el estudio.

Existe consenso en que es necesario estudiar la falta de respuesta como uno de los factores que incide en la precisión de las estimaciones de un producto estadístico. La misma introduce sesgo si la población de los que no responden presenta características distintas a quienes si responden, respecto de las variables de interés. Esto también ocasiona una disminución en el tamaño de la muestra. Dicha disminución incrementa el error de muestreo, lo cual conlleva a una limitación en la generalización de las conclusiones al conjunto de la población.

Una solución viable y popular en la práctica es realizar algún tratamiento para lidiar con la no respuesta. El ajuste de pesos muestrales incrementa los pesos de los encuestados de forma tal que se compense por la no respuesta cuando se calculan estimaciones.

Esta tesina busca evaluar empíricamente las propiedades de diversas estrategias de ajuste de pesos muestrales, con principal foco en la capacidad para reducir el sesgo por no respuesta. Se utilizan ajustes de pesos muestrales basados en la ponderación mediante la propensión a responder, ajustes bajo el enfoque de calibración, y una estrategia de ajuste que surge de una combinación de ambos.

Por otro lado, el problema de construir expresiones para la variancia no solo depende del diseño muestral, sino que también es influenciado por el método utilizado para corregir la no respuesta. La elección del estimador de variancia a utilizar debe basarse también en cuestiones relativas al tiempo, costo y simplicidad.

Este trabajo propone al estimador de variancia Bootstrap presentado por Rao, Wu y Yue (1992) como método general para estimar la variancia en el contexto de estudios por muestreo a gran escala. Particularmente, se busca evaluar su desempeño como estimador de variancia de los estimadores que surgen de los diferentes ajustes por no respuesta.

A fin de evaluar el desempeño de los diferentes estimadores puntuales y el estimador de variancia, se realiza un estudio por simulación donde se 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. Para realizar las simulaciones necesarias se utiliza una población sintética de viviendas del Estado de Texas correspondiente al año 2010.

## Tabla de contenidos

Agradecimientos.....	1
Resumen .....	2
Tabla de contenidos .....	4
1. Introducción .....	6
2. Objetivos .....	13
3. Métodos .....	14
3.1 Métodos muestrales en poblaciones finitas.....	14
3.1.1 Conceptos básicos .....	14
3.1.2 Estimadores y sus propiedades estadísticas básicas .....	19
3.1.3 Muestreo por conglomerados .....	22
3.1.3.1 Muestreo de conglomerados en una etapa.....	23
3.1.3.2 Muestreo por conglomerados en dos etapas .....	26
3.1.3.3 Muestreo multietápico .....	30
3.2 El problema de la no respuesta .....	32
3.2.1 Sesgo por no respuesta .....	32
3.2.2 Mecanismos de no respuesta.....	35
3.3 Ajuste por no respuesta .....	36
3.3.1 Motivación del ajuste por no respuesta .....	37
3.3.2 Ajuste mediante la propensión a responder.....	38
3.4 Estimador de calibración.....	39
3.4.1 Estimador de calibración para el caso de respuesta completa .....	39
3.4.1.1 Elección de la función $G$ .....	41
3.4.1.2 Calibración en conteos conocidos de tablas de frecuencia .....	42
3.4.2 Estimador de calibración en presencia de no respuesta .....	45
3.4.2.1 Sesgo del estimador de calibración en presencia de no respuesta .....	46
3.4.3 Ajuste por no respuesta en uno y en dos pasos.....	47
3.4.4 Notas sobre el sesgo, el vector auxiliar y la propensión a responder .....	49
3.5 Estimación de variancia.....	51
4. Aplicación .....	56
4.1 Elaboración del marco muestral.....	56
4.1.1 Materiales .....	56
4.1.2 Asignación de localidad.....	56
4.1.3 Construcción de áreas de muestreo.....	57
4.1.4 Unidades Primarias de Muestreo .....	58
4.1.5 Filtros operativos y estratificación .....	59
4.2 Características del estudio de simulación .....	59

4.2.1 Diseño muestral.....	59
4.2.2 Asignación de valores faltantes .....	61
4.2.3 Ajuste de pesos muestrales.....	61
4.2.4 Estimadores utilizados.....	63
4.2.5 Evaluación de los estimadores .....	63
4.2.6 Resultados .....	65
5. Conclusiones.....	72
Bibliografía .....	76
Anexo 1: Descripción de la población sintética .....	80
Anexo 2: Descripción de Condados y UPM .....	81
Anexo 3: Descripción de estratos de UPM .....	88
Anexo 4: Gráficos .....	89
Anexo 5: Código utilizado .....	104

## **1. Introducción**

Los actores de la sociedad moderna son cada vez más conscientes de cuán importante es contar con información relevante, confiable y precisa. El Estado, privados y la sociedad en general toman decisiones con la asistencia de la información disponible sobre problemas de su incumbencia. El primero, podría preguntarse sobre el nivel de empleo y su evolución a lo largo del último año, a fin de evaluar y diseñar sus políticas laborales. El segundo, puede querer mejorar un producto basado en la opinión de sus clientes, y el último, podría negociar el ajuste de su contrato de alquiler en base al índice de precios de los últimos 6 meses.

En el intento de obtener respuesta a estas incógnitas de la manera más precisa posible, se podría considerar encuestar a todas las personas de la población activa para saber si trabajan o no, para el segundo caso, preguntarle a cada cliente qué le mejoraría al producto de la empresa, y en el último, recabar información sobre el precio de todos los productos y servicios comercializados en el país durante el periodo relevante.

El común denominador entre estas posibles soluciones es que asumen que la forma de obtener una respuesta *correcta* a su pregunta es mediante la obtención de información sobre todas las unidades de interés. Sin embargo, estas alternativas son impracticables por cuestiones logísticas, económicas y temporales. Aún más, obtener información de la totalidad de las unidades en estudio no garantiza que el resultado esté libre de errores.

Uno de los métodos de obtención de datos más utilizados en la práctica es el estudio por muestreo. En vez de indagar exhaustivamente sobre el conjunto finito de unidades que se desea estudiar, se hace una investigación utilizando una parte de este. Un estudio por muestreo cuesta menos que una enumeración completa, usualmente consume menos tiempo, e incluso puede ser más preciso (Särndal, Swensson, & Wretman, 1992).

En Argentina, como sucede generalmente en otros países, existe un organismo oficial de estadística que se encarga de producir estadísticas sobre características y actividades a nivel nacional. El Instituto Nacional de Estadística y Censos (INDEC) provee estadísticas referentes al

territorio, población, economía y sociedad. Es el único organismo a nivel nacional dotado de capacidad logística, técnica y legal para realizar tales tareas.

Uno de los operativos de recolección de información más importantes que lleva a cabo el INDEC, es sin dudas el Censo de Población, Hogares y Viviendas que se realiza cada diez años. Sin embargo, indicadores tan relevantes, y con un requerimiento temporal inferior, como el índice de precios al consumidor (IPC), la tasa de desocupación o el porcentaje de personas bajo la línea de pobreza se obtienen mediante estudios por muestreo.

Un diseño muestral elegido criteriosamente, debe asignar a cada unidad de la población de interés una probabilidad no nula y conocida de ser seleccionada en la muestra. Este carácter probabilístico de los diseños muestrales utilizados es lo que posibilita inferir sobre medidas de interés en la población a través de la utilización de la teoría de inferencia estadística.

Los diseños muestrales utilizados por INDEC para estimar características de interés son en general complejos. En otras palabras, las unidades que componen la muestra son el resultado de realizar una selección de unidades en varias etapas, donde generalmente se estratifica a las unidades de al menos una de las etapas y se utilizan diferentes probabilidades de selección. Por ejemplo, en la primera etapa se podrían seleccionar aglomerados urbanos, luego áreas conformadas por viviendas dentro de estos aglomerados, y finalmente un subconjunto de esas viviendas como tercera etapa.

La naturaleza de estos diseños muestrales exige contar con un listado de viviendas exhaustivo, actualizado y con información de contacto precisa, cumplimentando los niveles de desagregación geográfica que requiera el diseño. En la práctica la elaboración de estos marcos muestrales requiere de mucho esfuerzo, ya que generalmente son el producto de un censo de población y luego deben ser actualizados periódicamente para asegurar su fiabilidad como elemento de listado de las unidades en la población de interés.

Los estudios por muestreo están expuestos a diferentes tipos de errores, muestrales y no muestrales. El error muestral es el que se produce por estimar características de interés de una

población solo con una parte de esta. Por otro lado, Groves (1989) cataloga al error no muestral en tres grandes áreas. Una es la cobertura, donde el error resulta si algunos miembros de la población en estudio tienen probabilidad nula de ser incluidos en la muestra. En segundo lugar, se refiere al efecto de la medición, donde el instrumento o partes del instrumento son construidos de manera que producen datos inválidos o no fidedignos. Por último, el efecto de la no respuesta, donde las unidades de la muestra que no responden difieren de aquellas que si lo hacen en cuanto a variables de interés en la encuesta.

Los problemas de no respuesta aparecen cuando se obtienen cuestionarios no completos en su totalidad, ya sea porque no se contesta alguno o todos los ítems. En otras palabras, la no respuesta refleja un intento no exitoso de obtener toda la información deseada de una unidad elegible para el estudio.

Existe consenso en que es necesario estudiar la falta de respuesta como uno de los factores que incide en la precisión de las estimaciones de un producto estadístico. La misma introduce sesgo si la población de los que no responden presenta características distintas a quienes si lo hacen, respecto de las variables de interés. Esto también ocasiona una disminución en el tamaño de la muestra. Dicha disminución incrementa el error de muestreo, lo cual conlleva a una limitación en la generalización de las conclusiones al conjunto de la población.

Se cuenta con alternativas para lidiar con la no respuesta, como por ejemplo realizar un nuevo intento de contacto con los no respondientes o, como sugieren Groves y Couper (1998), considerar las observaciones que el encuestador pueda hacer sobre la vivienda, el jefe de hogar, las condiciones de vida de las personas, el barrio, etc. para ser utilizados en la personalización del enfoque de la encuesta a cierto perfil de hogares propensos a participar en el estudio y así mejorar la tasa de respuesta.

A pesar de los métodos existentes para mejorar la respuesta o contactar a los que no responden, la respuesta nunca será completa, ni será un fenómeno completamente al azar. Generalmente se obtiene respuesta solo de un subconjunto de la muestra que además difiere con el conjunto de

unidades que no responden en características de interés de la encuesta, introduciendo sesgo en las estimaciones.

Una solución viable y popular en la práctica es realizar algún tratamiento para lidiar con la no respuesta. Las dos técnicas principales son la imputación y el ajuste de pesos muestrales. La imputación es el procedimiento mediante el cual valores artificiales de una variable se utilizan para sustituir valores faltantes genuinos. El ajuste de pesos muestrales incrementa los pesos de los encuestados de forma tal que se compense por la no respuesta cuando se calculan estimaciones. El primero de estos métodos se utiliza cuando una unidad de la muestra no responde algunos ítems de la encuesta, mientras que el segundo se utiliza cuando la unidad muestral no responde a ninguna pregunta de la encuesta. Estos fenómenos se conocen como no respuesta a nivel ítem y no respuesta a nivel unidad, respectivamente.

Se pueden mencionar diversos métodos para ajustar por no respuesta a nivel unidad a los pesos de los encuestados. Estos difieren en la forma que calculan el factor de ponderación y pueden ser clasificados en dos grupos. Por un lado, se tienen los métodos bajo el enfoque de calibración, que ajustan los pesos muestrales de forma tal que las estimaciones de variables auxiliares coincidan con sus totales poblacionales conocidos, bajo la restricción de que los pesos calibrados y los pesos originales difieran lo menos posible. Por otro lado, otro enfoque que ha ganado popularidad es el de *propensity score*. Este utiliza probabilidades estimadas de respuesta, o de propensión a responder, para corregir los pesos originales y no pertenece al enfoque de calibración.

Sin embargo, el dilema de la no respuesta es que las técnicas estándar utilizadas para obtener estimadores insesgados no funcionan y no existe método perfecto para realizar inferencia estadística válida. Los métodos de uso popular son intentos más o menos exitosos para reducir el sesgo por no respuesta (Särndal, Swensson, & Wretman, 1992).

Särndal y Lundström (2005) realizan un análisis exhaustivo del sesgo del estimador del total, bajo el enfoque de calibración. Derivan una expresión aproximada del sesgo independiente del diseño muestral y mencionan un conjunto de condiciones que deben cumplir las variables auxiliares para que diferentes estimadores sean *casi insesgados*. En otras palabras, los autores enfatizan en que

ante la presencia de no respuesta el mejor de los escenarios es uno donde se puede disminuir el sesgo considerablemente, pero sabiendo que no se lo elimina en su totalidad.

Otra cuestión muy relevante es que en todo estudio por muestreo cada estimación puntual que se realice debe ser correspondida con una estimación de su dispersión. La estimación de variancia es necesaria por dos motivos: para indicar la precisión del estimador y para calcular intervalos de confianza. La variancia de un estadístico muestral es función de la forma del estadístico y de la naturaleza del diseño muestral.

En diseños muestrales complejos es muy importante considerar el diseño utilizado para derivar la expresión de la variancia. Es muy común utilizar expresiones que se corresponden con diseños simples, cuando en realidad la expresión que se corresponde con el diseño es otra. Uno de los principales problemas de los estimadores de variancia es que deben ser construidos e implementados para cada diseño en particular, siendo esta una tarea a veces muy dificultosa.

La utilización de métodos para corregir la no respuesta introduce otro componente de aleatoriedad. Además de contar con las probabilidades de selección de las unidades, dadas por el diseño, también se introduce a las probabilidades de respuesta de cada unidad, generalmente desconocidas. Es decir, los elementos de la muestra se corresponden con una probabilidad de inclusión conocida y una probabilidad de responder desconocida. El problema de construir expresiones para la variancia no solo depende del diseño muestral, sino que también es influenciado por el método utilizado para corregir la no respuesta.

Särndal y Lundström (2005) dedican el capítulo 11 de su libro a estimadores de variancia en el contexto de calibración. Sin embargo, la derivación de estos estimadores a diferentes diseños muestrales complejos es difícil y desincentiva su utilización en la práctica. La elección del estimador de variancia a utilizar debe basarse también en cuestiones relativas al tiempo, costo y simplicidad. Algunas alternativas relevantes son los estimadores de variancia mediante método de linealización por series de Taylor o los métodos de replicación, que reutilizan la muestra para obtener estimaciones.

El principal beneficio de los métodos de replicación es que utilizan una única fórmula de estimación de variancia para todas las estadísticas y existen variantes que aplican a estadísticos no suaves, a diferencia del método de linealización que requiere la derivación de una fórmula separada para cada estadística y no es aplicable para estadísticos no suaves por inexistencia de derivadas primeras continuas.

Dentro de los métodos de replicación se pueden mencionar, Jackknife Balanced Repeated Replications y Bootstrap. Una bondad que comparten estos métodos es que son aplicables a estimadores no lineales y sus estimadores coinciden con los estimadores de la teoría clásica para estimadores lineales.

En particular, Rao, Wu y Yue (1992) proponen un método de estimación de variancia Bootstrap que no solo conserva las propiedades mencionadas, sino que también produce estimaciones consistentes para estadísticos no suaves. Desafortunadamente sus propiedades en diseños muestrales complejos no han sido muy estudiadas, aunque se conoce que es más inestable en ciertas ocasiones que los estimadores producidos mediante Jackknife o Linealización.

En base a todo lo mencionado anteriormente, considerando un diseño muestral complejo y diferentes escenarios de no respuesta, esta tesis busca evaluar empíricamente las propiedades de diversas estrategias de ajuste de pesos muestrales, con principal foco en la capacidad para reducir el sesgo por no respuesta. Se utilizan ajustes de pesos muestrales basados en la ponderación mediante la propensión a responder, ajustes bajo el enfoque de calibración, y una estrategia de ajuste que surge de una combinación de ambos.

Además, dada su popularidad en la práctica y su simplicidad, es de interés realizar las estimaciones de variancia necesarias mediante el método propuesto por Rao, Wu y Yue (1992) a fin de conocer su desempeño en un escenario que se asemeja a los estudios por muestreo realizados por los organismos oficiales de estadística, donde la no respuesta es un problema siempre presente y se utiliza algún procedimiento para tratar de corregirla. Particularmente, se estudia empíricamente el sesgo y la variabilidad de este estimador bajo las diferentes alternativas de ajuste de pesos muestrales.

Para realizar las simulaciones necesarias, se utiliza una población sintética de viviendas del Estado de Texas correspondiente al año 2010. La misma es uno de los componentes de una población sintética elaborada para todo Estados Unidos por *RTI International*<sup>1</sup> que incluye personas, escuelas, instituciones, etc. (Wheaton, 2014). Cada una de las viviendas se encuentra geolocalizada y esto permite agruparlas en los niveles de desagregación geográficos requeridos por los diseños muestrales usuales.

Si bien la elaboración de áreas de muestreo no es uno de los temas principales de la presente tesina, se comenta en términos generales la metodología utilizada para la elaboración del marco muestral a partir de la población sintética obtenida.

---

<sup>1</sup> *RTI International* es un nombre comercial de *Research Triangle Institute*.

## **2. Objetivos**

El presente trabajo de tesina tiene por objetivos:

- i. Estudiar y evaluar empíricamente el sesgo que puede implicar la no respuesta en un estudio por muestreo.
- ii. Estudiar y aplicar la calibración y el ajuste por clases de ponderación en un diseño muestral complejo con presencia de no respuesta.
- iii. Evaluar empíricamente la capacidad del ajuste de pesos muestrales para reducir el posible sesgo por no respuesta ante diferentes mecanismos que la generan.
- iv. Estudiar y aplicar el método de variancia Bootstrap en todos los escenarios de simulación. Estudiar empíricamente su bondad como estimador ante los diversos ajustes de pesos muestrales mediante el sesgo relativo y su dispersión.

## 3. Métodos

### 3.1 Métodos muestrales en poblaciones finitas

#### 3.1.1 Conceptos básicos

El muestreo en poblaciones finitas considera un conjunto de técnicas dentro de la Estadística que tiene como objetivo obtener estimaciones de ciertos parámetros de una población, los cuales son fijos y desconocidos. El muestreo consiste en la selección de una parte de la población de forma tal que permite obtener resultados para la misma.

En el muestreo probabilístico, el método utilizado para seleccionar la muestra debe asignar a cada unidad de la población de interés una probabilidad no nula y conocida de ser seleccionada en la muestra. Esta característica es lo que posibilita inferir sobre medidas de interés en la población a través de la utilización de la teoría de inferencia estadística, que brinda reglas para realizar las estimaciones puntuales y cuantificar su imprecisión. Por el contrario, en el muestreo no probabilístico no se tiene conocimiento de la probabilidad de selección de las unidades de interés y consecuentemente no se puede obtener estimación de la precisión de los estimadores ni asegurar la representatividad de la muestra.

Sea el conjunto  $U = \{u_1, \dots, u_k, \dots, u_N\}$  de tamaño  $N$ , conocido, la población de interés. Idealmente éste se corresponde con el conjunto representado en el marco muestral, es decir, no existen ni duplicados ni faltantes. A fines de no desviar la atención a este tipo de error, en el resto del presente trabajo se asume que tal condición es satisfecha. Es frecuente representar a las unidades del conjunto  $U$  a través del subíndice correspondiente,  $U = \{1, \dots, k, \dots, N\}$ .

Un muestreo probabilístico es el que cumple las siguientes condiciones.

1. Se puede definir el conjunto de las muestras que se pueden obtener con el procedimiento de muestreo,  $\mathcal{S} = \{s_1, s_2, \dots, s_M\}$ .
2. Cada muestra  $s$  tiene asociada una probabilidad conocida de selección  $p(s)$ .
3. El procedimiento le da a cada elemento en la población una probabilidad no nula de selección.

4. Un mecanismo aleatorio selecciona una muestra de las posibles respetando las probabilidades  $p(s)$ .

La función  $p(\cdot)$ , que define una distribución de probabilidad en  $\mathcal{S} = \{s_1, s_2, \dots, s_M\}$ , es llamada *diseño muestral*. Dado un diseño muestral  $p(\cdot)$ , se puede pensar a cualquier muestra  $s$  como el resultado de una variable aleatoria  $S$  cuya distribución de probabilidad está especificada por  $p(\cdot)$ .

Sea  $\mathcal{S}$  el conjunto de todas las muestras posibles (luego  $\mathcal{S}$  es el conjunto de  $2^N$  subconjuntos de  $U$ , incluyendo al conjunto vacío y a  $U$  mismo). Se tiene que  $P(S = s) = p(s)$  para cada  $s \in \mathcal{S}$  y dado que  $p(s)$  es una distribución de probabilidad en  $\mathcal{S}$ , se tiene:

$$\begin{aligned} p(s) &\geq 0 \quad \forall s \in \mathcal{S} \\ \sum_{s \in \mathcal{S}} p(s) &= 1 \end{aligned} \quad (1)$$

El tamaño muestral  $n_s$  es la cantidad de elementos en la muestra  $s$ . Existen diseños muestrales donde el tamaño muestral puede variar de muestra en muestra, pero esta tesina se restringe a diseños con tamaño muestral fijo, es decir,  $n_s = n \quad \forall s \in \mathcal{S}$ .

Dado un diseño muestral particular, la inclusión de un elemento  $k$  en la muestra es un evento aleatorio indicado por la variable aleatoria  $I_k$ , definida como:

$$I_k = \begin{cases} 1 & \text{si } k \in S \\ 0 & \text{si } k \notin S \end{cases}. \quad (2)$$

$I_k = I_k(S)$  es función de la variable aleatoria  $S$  y se conoce como indicador de pertenencia del elemento  $k$  a la muestra. La probabilidad de que una unidad  $k$  pertenezca a una muestra es denominada probabilidad de inclusión de primer orden de la unidad  $k$  y se obtiene mediante:

$$\pi_k = P(\text{unidad } k \in s) = P(I_k = 1) = \sum_{s \ni k} p(s), \quad (3)$$

donde el último término representa la suma de la probabilidad de cada una de las muestras que contienen a la unidad  $k$ , y dado que se consideran diseños de tamaño fijo,  $\sum_U \pi_k = n$ . La probabilidad de que dos unidades  $k$  y  $l$  pertenezcan a una muestra es denominada probabilidad de inclusión de segundo orden de las unidades  $k$  y  $l$  se obtiene mediante

$$\pi_{kl} = P(\text{unidades } k \text{ y } l \in s) = P(I_k I_l = 1) = \sum_{s \ni k \& l} p(s), \quad (4)$$

notar que  $\pi_{kk} = \pi_k$  y  $\pi_{kl} = \pi_{lk}$ .

La condición para que un muestreo sea probabilístico se corresponde con  $\pi_k > 0 \forall k \in U$ . Si además se cumple que  $\pi_{kl} > 0 \forall k, l \in U$  con  $k \neq l$ , el diseño muestral es medible y permite obtener estimaciones de variancias e intervalos de confianza válidos basados en los datos observados.

La mayoría de las encuestas tiene por objetivo estimar uno o más parámetros de interés. Para ello se debe elegir, por un lado, el diseño muestral y el esquema de selección, y por otro, la fórmula para calcular la estimación de cada parámetro (estimador). Una *estrategia* de muestreo es la combinación de un diseño muestral y un estimador. Estas decisiones no son independientes, ya que la elección de un estimador generalmente depende del diseño aplicado.

Un parámetro poblacional es una medida resumen, fija y desconocida, de una característica sobre la población objetivo. Sea  $y_k$  el valor de la característica de interés correspondiente a la unidad  $k$  de la población con  $k = 1, \dots, N$ . Notar que es una característica fija y no variable, ya que la aleatoriedad se introduce por el mecanismo de selección.

Luego un parámetro poblacional puede ser descripto como una función de  $y_k \forall k = 1, \dots, N$ ,

$$\theta = f(y_1, y_2, \dots, y_N). \quad (5)$$

Por ejemplo, el total y la media poblacional se pueden escribir como:

$$\theta = t_y = \sum_{k \in U} y_k = y_1 + y_2 + \dots + y_N, \quad (6)$$

$$\theta = \bar{y}_U = \frac{1}{N} \sum_{k \in U} y_k = \frac{y_1 + y_2 + \dots + y_N}{N}. \quad (7)$$

O una medida de la variabilidad de la característica bajo estudio, como es la variancia, se presenta como:

$$\theta = S_{yU}^2 = \frac{\sum_{k \in U} (y_k - \bar{y}_U)^2}{N-1} = \frac{\sum_{k \in U} y_k^2}{N-1} - \frac{(\sum_{k \in U} y_k)^2}{N(N-1)}. \quad (8)$$

Un parámetro también puede ser función de más de una variable de interés:

$$\theta = \frac{t_y}{t_z} = \frac{\sum_{k \in U} y_k}{\sum_{k \in U} z_k}. \quad (9)$$

Un estimador, es una función de la muestra que se espera que para la mayoría de las muestras produzca valores cercanos al parámetro de interés. Sea una muestra particular,  $s$ , el estimador viene dado por:

$$\hat{\theta} = g(y_{i1}, y_{i2}, \dots, y_{in}) = \hat{\theta}(s). \quad (10)$$

Por ejemplo, bajo un muestreo simple al azar, que asigna a cada muestra de tamaño  $n$  una misma probabilidad, un estimador usual de la media poblacional es la media muestral:

$$\hat{\theta} = \bar{y}_s = \frac{\sum_{k \in s} y_k}{n}. \quad (11)$$

Cada muestra posible origina una estimación particular y estos valores permiten definir la distribución muestral o probabilística del estimador. Esto posibilita estudiar propiedades relevantes como su posición y dispersión.

Conocer la distribución muestral de  $\hat{\theta}$  requiere, para cada valor posible  $c$  de  $\hat{\theta}$ , una especificación de la probabilidad

$$P(\hat{\theta} = c) = \sum_{s \in \mathcal{S}_c} p(s), \quad (12)$$

donde  $\mathcal{S}_c$  es el conjunto de muestras  $s$  para los cuales  $\hat{\theta} = c$ .

Si bien en la práctica es frecuente que no se pueda especificar exactamente la distribución en el muestreo de un estadístico, se recurre a diversas medidas resúmenes para comparar el desempeño de los estimadores. Por un lado, es de interés estudiar el valor esperado (o esperanza) y el sesgo de un estimador:

$$E(\hat{\theta}) = \sum_{s \in S} p(s) \hat{\theta}(s), \quad (13)$$

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta. \quad (14)$$

La esperanza no es más que el promedio ponderado de los valores posibles del estimador, donde los pesos de cada uno están dados por su probabilidad de ocurrencia. En conjunto con el sesgo permiten el estudio de la estabilidad de un estimador. Además, brindan conocimiento sobre el valor alrededor del cual se debe esperar que se encuentre el estimador, y cuanto este nivel difiere del parámetro de interés. Un estimador es insesgado cuando su esperanza coincide con el parámetro de interés, luego,  $B(\hat{\theta}) = 0$ .

Por otro lado, generalmente se utiliza la variancia para medir la dispersión de un estimador respecto a su valor medio:

$$V(\hat{\theta}) = \sum_{s \in S} p(s) [\hat{\theta}(s) - E(\hat{\theta})]^2, \quad (15)$$

y suele ser útil considerar al error cuadrático medio, que combina el sesgo y la variancia para brindar un índice de precisión de un estimador bajo un diseño dado, o sea

$$ECM(\hat{\theta}) = E[\hat{\theta} - \theta]^2 = \sum_{s \in S} p(s) [\hat{\theta}(s) - \theta]^2 = V(\hat{\theta}) + B^2(\hat{\theta}). \quad (16)$$

Lo ideal sería que en la práctica se pueda elegir aquel estimador insesgado y de menor variancia. Sin embargo, esto rara vez es posible en diseños muestrales complejos y la elección suele ser tolerar un poco de sesgo a cambio de mayor precisión.

Por último, otra medida muy utilizada para evaluar la precisión de un estimador que tiene la ventaja de ser independiente de la unidad de medida es el coeficiente de variación. Éste es comúnmente expresado como un porcentaje y refleja el tamaño relativo del desvío estándar respecto de la media del estimador.

$$CV(\hat{\theta}) = \frac{\sqrt{[V(\hat{\theta})]}}{E(\hat{\theta})}. \quad (17)$$

### 3.1.2 Estimadores y sus propiedades estadísticas básicas

Sea un diseño probabilístico que asigna a cada elemento de la población una probabilidad de inclusión  $\pi_k$ , y sea  $d_k = \frac{1}{\pi_k}$  el peso muestral que le corresponde a cada uno de esos elementos. El estimador de Horvitz-Thompson para el total es:

$$\hat{t}_\pi = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in s} d_k y_k = \sum_{k \in s} \check{y}_k = \sum_{k \in U} \check{y}_k I_k, \quad (18)$$

y para el caso de la media es:

$$\hat{\bar{y}}_\pi = \frac{\hat{t}_\pi}{N} = \frac{1}{N} \sum_{k \in s} d_k y_k. \quad (19)$$

Es importante notar que los valores  $\check{y}_k = \frac{y_k}{\pi_k}$  son constantes fijas y la aleatoriedad se introduce por el método de selección de la muestra  $s$ , o equivalentemente mediante las variables indicadoras  $I_k \forall k \in U$ .

Horvitz y Thompson (1952) utilizaron el principio de expandir a los valores de acuerdo a la inversa de su probabilidad de selección. La expansión por el factor  $1/\pi_k$  tiene el efecto de incrementar la importancia de los elementos en la muestra. Dado que la muestra contiene menos elementos que los que hay en la población, se necesita expandir los valores observados de forma tal que en la agregación se alcancen los niveles de esta última. Bajo esta lógica, el elemento  $k$  representa  $d_k = \frac{1}{\pi_k}$  elementos en la población. Por esta razón, este estimador también recibe el nombre de  $\pi$ -estimador.

La presente tesina se reduce solo a mencionar algunas propiedades del estimador de Horvitz-Thompson. La intención es brindar un marco teórico y constituir puntos de partida y comparación para cuando se presenten los estimadores utilizados en la aplicación.

El estimador de Horvitz-Thompson del total poblacional es un estimador lineal que emplea las probabilidades de inclusión de primer orden,  $\pi_k$ . Para el caso de la media, este requiere conocer el tamaño de la población  $N$ .

Una alternativa para estimar parámetros poblacionales como la media cuando se desconoce  $N$ , es el estimador de Hájek (1971):

$$\hat{\bar{y}}_{HJ} = \frac{\sum_{k \in s} y_k / \pi_k}{\sum_{k \in s} 1 / \pi_k} = \frac{\hat{t}_\pi}{\hat{N}}, \quad (20)$$

que conociendo  $N$  puede ser fácilmente extendido para estimar un total

$$\hat{t}_{HJ} = N \bar{y}_{HJ}. \quad (21)$$

A continuación, se presentan las fórmulas correspondientes a la variancia y una estimación de la misma para un diseño general (Särndal, Swensson, & Wretman, 1992).

La variancia para el  $\pi$  – estimador del total es:

$$V(\hat{t}_\pi) = \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) \frac{y_k y_l}{\pi_k \pi_l}. \quad (22)$$

Siempre que el diseño sea medible, es decir,  $\pi_{kl} > 0 \forall k, l \in U$ , un estimador insesgado viene dado por:

$$\hat{V}(\hat{t}_\pi) = \sum_{k \in s} \sum_{l \in s} \frac{(\pi_{kl} - \pi_k \pi_l)}{\pi_{kl}} \frac{y_k y_l}{\pi_k \pi_l}. \quad (23)$$

Para el caso de la media la extensión es sencilla, ya que se verifica que:

$$V(\hat{\bar{y}}_\pi) = V\left(\frac{\hat{t}_\pi}{\hat{N}}\right) = \frac{V(\hat{t}_\pi)}{N^2}. \quad (24)$$

Aplica el mismo razonamiento para su estimador.

Es frecuente que se utilice la siguiente expresión de la variancia del estimador de Horvitz-Thompson debida a Sen, Grundy y Yates (SGY) para diseños muestrales de tamaño fijo:

$$V_{SGY}(\hat{t}_\pi) = -\frac{1}{2} \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) \left( \frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2. \quad (25)$$

Siempre que  $\pi_{kl} > 0 \forall k, l \in U$ , un estimador insesgado de  $V_{SGY}(\hat{t}_\pi)$  es:

$$\hat{V}_{SGY}(\hat{t}_\pi) = -\frac{1}{2} \sum_{l \in k} \sum_{l \in s} \frac{(\pi_{kl} - \pi_k \pi_l)}{\pi_{kl}} \left( \frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2. \quad (26)$$

Cuando el tamaño es fijo, la variancia **(25)** y **(24)** coinciden. Los estimadores no son necesariamente idénticos, pero son ambos insesgados. Se observa también que en este caso el estimador de variancia  $\hat{V}_{SGY}(\hat{t}_\pi)$  es no negativo siempre que  $\pi_{kl} - \pi_k \pi_l < 0 \forall k, l \in U$ .

Por otro lado, se puede notar que el estimador **(20)** es una razón entre dos estimadores para el total de  $y$  y  $z$ , donde  $z_k = 1 \forall k \in U$ . Dada la no linealidad de este, se puede realizar una aproximación por series de Taylor y obtener una expresión de la variancia (Särndal, Swensson, & Wretman, 1992):

$$AV(\hat{y}_{HJ}) = \frac{1}{N^2} \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) \frac{y_k - \bar{y}_U}{\pi_k} \frac{y_l - \bar{y}_U}{\pi_l}, \quad (27)$$

cuya estimación es:

$$\hat{V}(\hat{y}_{HJ}) = \frac{1}{\hat{N}^2} \sum_{k \in s} \sum_{l \in s} \frac{(\pi_{kl} - \pi_k \pi_l)}{\pi_{kl}} \frac{y_k - \bar{y}_{HJ}}{\pi_k} \frac{y_l - \bar{y}_{HJ}}{\pi_l}. \quad (28)$$

Análogamente al caso del  $\pi$  – estimador, cuando se considera el estimador del total **(21)**,  $AV(\hat{t}_{HJ}) = N^2 AV(\hat{y}_{HJ})$  y  $\hat{V}(\hat{t}_{HJ}) = \hat{N}^2 \hat{V}(\hat{y}_{HJ})$ .

La falta de conocimiento del tamaño poblacional  $N$  imposibilita el uso del  $\pi$  – estimador de la media, por lo que se debe optar por el estimador de Hájek. Por el contrario, cuando  $N$  es conocido, se debe optar entre **(19)** y **(20)**.

En general **(20)** resulta sesgado de  $\bar{y}_U$ , pero su sesgo es depreciable si el coeficiente de variación del denominador ( $\sum_{k \in s} 1/\pi_k$ ) es pequeño. Además, por los motivos que se listan debajo, este suele ser mejor que  $\bar{y}_\pi$  independientemente de que  $N$  sea conocido o no (Särndal, Swensson, & Wretman, 1992). Si bien el estimador de Hájek es originalmente un estimador de razón, para algunos diseños los estimadores de Horvitz-Thompson y Hájek son idénticos, como por ejemplo muestreo simple al azar.

En primer lugar, la forma de la variancia (25) indica que  $\hat{y}_{HJ}$  es preferible cuando los valores  $y_k - \bar{y}_U$  son todos cercanos a cero. En segundo lugar,  $\hat{y}_{HJ}$  presenta menor variabilidad en diseños donde el tamaño muestral no es constante. Si el tamaño muestral obtenido  $n_s$  es mayor que su valor esperado, la suma del numerador y la suma del denominador tendrán relativamente más términos. Análogamente, si el tamaño de la muestra es menor que su valor esperado, ambas sumas tendrán una cantidad menor de términos. La razón conserva de este modo una cierta estabilidad. Por el contrario,  $\hat{y}_\pi$  tiene denominador fijo, luego, carece de esta propiedad. En otras palabras, el  $\pi$  – estimador es más susceptible que el estimador de Hájek ante la introducción de variabilidad que ocasiona el hecho de que el tamaño de muestra no sea fijo igual a  $N$ . Y finalmente en tercer lugar, en casos donde  $\pi_k$  no está correlacionado o lo está en forma negativa con los valores  $y_k$ . Si la muestra contiene un elemento con un valor muy grande de  $y_k$  pero un valor pequeño de  $\pi_k$ , la suma del numerador de  $\hat{y}_{HJ}$  es muy grande. Sin embargo, esto es compensado, hasta cierto punto, por grandes valores de  $1/\pi_k$  en el denominador. Luego,  $\hat{y}_{HJ}$  goza de una cierta insensibilidad ante muestras poco probables, que no está presente en  $\hat{y}_\pi$  por considerar un denominador fijo.

### **3.1.3 Muestreo por conglomerados**

Los diseños muestrales más sencillos suelen considerar la selección de elementos. Es decir, los elementos de la población de interés pueden ser utilizados como unidades muestrales en una única etapa de muestreo. Sin embargo, en muchas encuestas de mediana o larga escala, el muestreo directo de elementos no es utilizado por alguna de las siguientes razones:

- i. No existe un marco muestral que identifique a cada elemento de la población, y su construcción resulta imposible.
- ii. La población de elementos está distribuida en un área muy grande, lo que resultaría en una muestra muy dispersa geográficamente. Esto se asocia con altos costos de posibles viajes si la encuesta es en persona y la dificultad para lograr una supervisión eficiente, lo que podría *incrementar la tasa de no respuesta* y agravar errores de medición.

Existe una variedad de diseños muestrales que dan solución a estos problemas: desde muestreo por conglomerados hasta diseños muestrales estratificados multietápicos con distintas probabilidades de selección en cada etapa.

En el muestreo por conglomerados se seleccionan grupos de unidades llamados conglomerados, que pueden estar determinados por áreas geográficas, establecimientos, escuelas u otro tipo de unidad de agrupamiento. Se selecciona una muestra probabilística de conglomerados y todas las unidades elementales del mismo son encuestadas. Precisamente, el método mencionado es conocido como muestreo por conglomerados en una etapa. Es frecuente que se muestreen unidades dentro de los conglomerados seleccionados en la primera etapa, lo que conforma otra etapa de muestreo. La extensión de este procedimiento a tres o más etapas es lo que se conoce como muestreo multietápico.

Por ejemplo, si es de interés conocer alguna característica referente a las personas de una ciudad, pero se cuenta solo con un listado de viviendas, se puede pensar que el conjunto de personas dentro de cada vivienda compone un conglomerado y cada persona dentro de este conglomerado es una unidad elemental.

### **3.1.3.1 Muestreo de conglomerados en una etapa**

En este diseño la población  $U = \{1, \dots, k, \dots, N\}$  es particionada en  $N_I$  subpoblaciones llamados conglomerados y representados por  $U_1, U_2, \dots, U_{N_I}$ . A su vez, estos forman el conjunto  $U_I = \{1, \dots, i, \dots, N_I\}$ , que representa a la población de conglomerados de la cual será seleccionada la muestra. La notación es tomada de Särndal, Swensson y Wretman (1992), quienes indican que ésta no solo es útil porque  $I$  identifica a las unidades que son conglomerados, sino que también indica la etapa de selección en cuestión. Luego,  $N_i$  representa el número de unidades de la población que pertenecen al conglomerado  $U_i$ , que puede ser variable. La partición de  $U$  puede ser representada de la siguiente manera:

$$U = \bigcup_{i \in U_I} U_i, \quad N = \sum_{i \in U_I} N_i. \tag{29}$$

El muestreo por conglomerados en una etapa se define:

1. Una muestra probabilística  $s_I$  de conglomerados es seleccionada de  $U_I$  de acuerdo con el diseño  $p_I(\cdot)$ . El tamaño de  $s_I$  es  $n_I$  para un diseño de tamaño fijo, o  $n_{s_I}$  para un diseño de tamaño variable.
2. Se encuesta a todos los elementos de la población que pertenecen a los conglomerados seleccionados.

La unidad de muestreo no son las unidades de interés en la población, sino que son los conglomerados. El método de selección de muestras puede ser cualquiera de los métodos conocidos y los conglomerados pueden estar estratificados. Por último, si el tamaño de los conglomerados es variable, el tamaño muestral en término de unidades elementales será variable, aunque se utilice un diseño muestral fijo en términos de conglomerados, ya que éste depende de cuáles son los conglomerados seleccionados en la muestra.

Las probabilidades de inclusión de primer y segundo orden para los conglomerados son:

$$\begin{aligned}\pi_{Ii} &= \sum_{s_I \ni i} p_I(s_I) \text{ para conglomerado } i, \\ \pi_{Iij} &= \sum_{s_I \ni i \wedge j} p_I(s_I) \text{ para conglomerados } i \text{ y } j, \\ \pi_{Iii} &= \pi_{Ii}.\end{aligned}\tag{30}$$

Las probabilidades de inclusión de primer y segundo orden para los elementos son:

$$\begin{aligned}\pi_k &= P(k \in s) = P(i \in s_I) = \pi_{Ii} \text{ si } k \in U_i, \\ \pi_{kl} &= P(k, l \in s) = P(i \in s_I) = \pi_{Ii} \text{ si } k, l \in U_i, \\ \pi_{kl} &= P(k, l \in s) = P(i, j \in s_I) = \pi_{Iij} \text{ si } k \in U_i \text{ y } l \in U_j, \\ \pi_{kk} &= \pi_k.\end{aligned}\tag{31}$$

Esto amerita introducir una nueva notación. Sea el total del conglomerado  $i$ :

$$t_i = \sum_{k \in U_i} y_k,\tag{32}$$

y sea el total de la población a estimar:

$$t = \sum_{k \in U} y_k = \sum_{k \in U_I} t_i.\tag{33}$$

El  $\pi$  – estimador del total es:

$$\hat{t}_\pi = \sum_{i \in S_I} \frac{t_i}{\pi_{Ii}} \quad (34)$$

Su variancia viene dada por:

$$V(\hat{t}_\pi) = \sum_{i \in U_I} \sum_{j \in U_I} (\pi_{Iij} - \pi_{Ii}\pi_{Ij}) \frac{t_i}{\pi_{Ii}} \frac{t_j}{\pi_{Ij}}, \quad (35)$$

y en diseños de tamaño fijo puede ser escrita como:

$$V_{SGY}(\hat{t}_\pi) = -\frac{1}{2} \sum_{i \in U_I} \sum_{j \in U_I} (\pi_{Iij} - \pi_{Ii}\pi_{Ij}) \left( \frac{t_i}{\pi_{Ii}} - \frac{t_j}{\pi_{Ij}} \right)^2. \quad (36)$$

Los estimadores insesgados para respectivos casos son:

$$\hat{V}(\hat{t}_\pi) = \sum_{i \in S_I} \sum_{j \in S_I} \frac{(\pi_{Iij} - \pi_{Ii}\pi_{Ij})}{\pi_{Iij}} \frac{t_i}{\pi_{Ii}} \frac{t_j}{\pi_{Ij}}, \quad (37)$$

y

$$\hat{V}_{SGY}(\hat{t}_\pi) = -\frac{1}{2} \sum_{i \in S_I} \sum_{j \in S_I} \frac{(\pi_{Iij} - \pi_{Ii}\pi_{Ij})}{\pi_{Iij}} \left( \frac{t_i}{\pi_{Ii}} - \frac{t_j}{\pi_{Ij}} \right)^2. \quad (38)$$

Mientras que el muestreo estratificado crea agrupamientos homogéneos y aumenta la precisión al compararlo con muestreo simple al azar (MSA), en el muestreo por conglomerados los grupos son heterogéneos y generalmente se tiene menos precisión que en MSA. A pesar de la ineficiencia del muestreo por conglomerados respecto al muestreo simple al azar, el primero es frecuentemente utilizado en encuestas a gran escala porque suele ser conveniente en términos operativos y de costo al encuestar unidades agrupadas en un área determinada en vez de encuestar unidades de una muestra muy esparcida, como la que puede resultar del muestreo simple al azar.

La eficiencia del muestreo por conglomerados puede ser mejorada sustancialmente cuando se cuenta con información auxiliar. La elección de la estrategia depende de la información disponible. Un caso es cuando se cuenta con una medida del tamaño para cada conglomerado,  $u_i$ . Si esta es proporcional al total de la variable en estudio  $t_i$ , se puede reducir la variancia del  $\pi$  – estimador utilizando un muestreo de conglomerados donde la probabilidad de selección de cada conglomerado es proporcional a  $u_i$ . Alternativamente, se pueden formar estratos de

conglomerados de forma tal que la variabilidad de la medida del tamaño sea pequeña dentro de cada estrato.

### **3.1.3.2 Muestreo por conglomerados en dos etapas**

Es posible demostrar que el muestreo por conglomerados simple al azar es más eficiente que un muestreo simple al azar en términos de variancia si y solo si el promedio de las variancias dentro de cada conglomerado es mayor que la variancia global,  $S_{yU}^2$ , con conglomerados de igual tamaño. Por el contrario, (y frecuente en la práctica) si los conglomerados son homogéneos dentro y/o de tamaño variable, un muestreo simple al azar de conglomerados es inficiente. (Särndal, Swensson, & Wretman, 1992, Sección 4.2.2).

En estudios por muestreo a gran escala, como los que lleva a cabo el Instituto de Estadística de un país, los conglomerados suelen corresponderse con áreas geográficas. Consecuentemente, las unidades dentro un mismo conglomerado tienden a ser similares, resultando en un bajo nivel de heterogeneidad intra conglomerado.

La variabilidad del  $\pi$ -estimador bajo muestreo por conglomerados puede ser mejorada si se selecciona una cantidad mayor de conglomerados. Sin embargo, esto implica un costo mayor y puede no ser posible debido al presupuesto asignado para el estudio.

Una forma de controlar los costos y aumentar el número de conglomerados, es seleccionar una muestra de unidades dentro de los conglomerados seleccionados. Esto implica tener que estimar la característica de interés en cada unidad primaria de muestreo, lo que introduce una nueva fuente de variabilidad.

El muestreo por conglomerados en dos etapas puede ser llevado a cabo de la siguiente manera:

- i. Se divide a la población de elementos en subpoblaciones disjuntas llamadas unidades primarias de muestreo (UPM). Se selecciona una muestra probabilística de UPM y esto compone la primera etapa de muestreo.
- ii. Las UPM pueden estar compuestas por unidades o por conglomerados, llamados unidades secundarias de muestreo (USM). En la segunda etapa, se selecciona una

muestra probabilística de USM dentro de cada UPM. Cuando las USM están compuestas por conglomerados, se encuesta a todas las unidades elementales que componen a las USM seleccionadas.

En este caso se tienen dos fuentes de variabilidad. Esto puede ser extendido a los diseños muestrales multietápicos, que consisten en tres o más etapas de muestreo. Existe una jerarquía entre las unidades de muestreo: unidades primarias de muestreo (UPM), unidades secundarias de muestreo dentro de UPM (USM), unidades terciarias muestreo dentro de USM (UTM), etc.

Es necesario introducir la notación utilizada en estos diseños, ya que a medida que se van agregando etapas, esta se vuelve más compleja.

La población de elementos  $U = \{1, \dots, k, \dots, N\}$  es partitionada en  $N_I$  UPM  $U_1, \dots, U_i, \dots, U_{N_I}$ . Este conjunto es  $U_I = \{1, \dots, i, \dots, N_I\}$  donde el tamaño de cada  $U_i$  es  $N_i$  y vale  $N = \sum_{i \in U_I} N_i$ . Luego, un muestreo por conglomerados en dos etapas consta de

- i. Primera etapa: una muestra  $s_I$  de UPM se selecciona de  $U_I$  de acuerdo con el diseño  $p_I(\cdot)$ .
- ii. Segunda etapa: para cada  $i \in s_I$ , una muestra  $s_i$  de elementos es seleccionada de  $U_i$  siguiendo un diseño  $p_i(\cdot | s_I)$ .

Se observa que se puede utilizar cualquier diseño en la primera y segunda etapa. El diseño aplicado en la segunda etapa puede depender de la muestra seleccionada en la primera, e incluso variar de una muestra de primera etapa a otra. Además, submuestrear en  $U_i$  no tiene que ser necesariamente independiente de submuestrear en  $U_j$ . Sin embargo, la teoría de muestreo presentada en esta tesina se corresponde con diseños invariantes e independientes. La propiedad de invariancia significa que cada vez que la UPM  $i$  es incluida en la muestra de primera etapa, el mismo diseño de submuestreo  $p_i(\cdot)$  debe ser utilizado. La independencia se refiere a que el submuestreo en una UPM particular es independiente del submuestreo en cualquier otra UPM.

Se presenta el muestreo por conglomerados en dos etapas donde las USM son elementos, y no conglomerados.

Los tamaños muestrales se presentan de la siguiente manera. El número de UPM en la muestra  $s_I$  es representado por  $n_{s_I}$ , o simplemente  $n_I$  para diseños fijos. El número de elementos en  $s_i$  es notado  $n_{s_i}$ , o análogamente  $n_i$  para diseños fijos. Finalmente, el número total de elementos en la muestra  $s$  es  $n_s = \sum_{i \in s_I} n_{s_i}$ .

Las probabilidades de inclusión asociadas al muestreo en dos etapas se expresan de la siguiente forma:

- i.  $\pi_{Ii}$  y  $\pi_{Iij}$  son las probabilidades de inclusión de primer y segundo orden correspondientes a  $p_I(\cdot)$ .
- ii.  $\pi_{k|i}$  y  $\pi_{kl|i}$  son las probabilidades de inclusión de primer y segundo orden correspondientes a  $p_i(\cdot | s_I)$ .

Para utilizar los estimadores vistos hasta el momento y sus respectivas variancias y estimadores de variancia se deben obtener las probabilidades de inclusión respecto de los elementos. Utilizando las propiedades de invariancia e independencia

$$\pi_k = \pi_{Ii} \pi_{k|i} \quad \text{si } k \in U_i \quad (39)$$

y

$$\pi_{kl} = \begin{cases} \pi_{Ii} \pi_{k|i} & \text{si } k = l \in U_i \\ \pi_{Ii} \pi_{kl|i} & \text{si } k \neq l \in U_i, k \neq l \\ \pi_{Ii} \pi_{k|i} \pi_{l|j} & \text{si } k \in U_i \text{ y } l \in U_j, i \neq j \end{cases}. \quad (40)$$

En muestreo por conglomerados en dos etapas el total de cada UPM debe ser estimado, y puede ser hecho utilizando el  $\pi$ -estimador de la siguiente forma:

$$\hat{t}_{i\pi} = \sum_{k \in s_i} \frac{y_k}{\pi_{k|i}}. \quad (41)$$

Esto permite expresar al  $\pi$ -estimador de la variable de interés como:

$$\hat{t}_\pi = \sum_{i \in s_I} \frac{\hat{t}_{i\pi}}{\pi_{li}} = \sum_{i \in s_I} \sum_{k \in s_i} \frac{y_k}{\pi_{li} \pi_{k|i}} = \sum_{i \in s_I} \sum_{k \in s_i} \frac{y_k}{\pi_k} = \sum_{i \in s_I} \sum_{k \in s_i} d_k y_k. \quad (42)$$

Existen expresiones tanto para la variancia como para los estimadores insesgados bajo ciertas condiciones. A su vez, estas permiten diferenciar la contribución a la variabilidad total de cada una de las etapas de muestreo:

$$V_{2E}(\hat{t}_\pi) = V_{UPM} + V_{USM}, \quad (43)$$

$$V_{UPM} = \sum_{i \in U_I} \sum_{j \in U_I} (\pi_{lij} - \pi_{li} \pi_{lj}) \frac{t_i}{\pi_{li}} \frac{t_j}{\pi_{lj}}, \quad (44)$$

$$V_{USM} = \sum_{i \in U_I} \frac{V_i}{\pi_{li}}, \quad (45)$$

donde  $V_i$  es la variancia del estimador de Horvitz-Thompson respecto a la segunda etapa, o simplemente  $V(\hat{t}_{i\pi})$ . Esta puede ser expresada como:

$$V_i = \sum_{k \in U_i} \sum_{l \in U_i} (\pi_{kl|i} - \pi_{k|i} \pi_{l|i}) \frac{y_k}{\pi_{k|i}} \frac{y_l}{\pi_{l|i}}. \quad (46)$$

Se puede notar que los componentes de variancia serán nulos en las siguientes situaciones:

- i.  $s_I = U_I$  con probabilidad 1. Luego  $\pi_{li} = \pi_{lij} = 1 \forall i, j$ . Luego,  $V_{UPM} = 0$  y  $V_{USM} = \sum_{U_I} V_i$  es la variancia de un  $\pi$ -estimador para muestreo estratificado con  $N_I$  UPM como el conjunto de estratos.
- ii.  $s_i = U_i$  con probabilidad 1  $\forall i$ . Luego  $V_{USM} = 0$  y  $V_{UPM}$  es la variancia del  $\pi$ -estimador para muestreo por conglomerados en una etapa.

El primer componente de (43) puede ser estimado de forma insesgada mediante:

$$\hat{V}_{UPM} = \sum_{i \in s_I} \sum_{j \in s_I} \left( \frac{\pi_{lij} - \pi_{li} \pi_{lj}}{\pi_{lij}} \right) \frac{\hat{t}_{i\pi}}{\pi_{li}} \frac{\hat{t}_{j\pi}}{\pi_{lj}} - \sum_{i \in s_I} \frac{1}{\pi_{li}} \left( \frac{1}{\pi_{li}} - 1 \right) \hat{V}_i, \quad (47)$$

donde  $\hat{V}_i$  es un estimador insesgado de (46). Este puede ser escrito como:

$$\hat{V}_i = \sum_{k \in U_i} \sum_{l \in U_i} \left( \frac{\pi_{kl|i} - \pi_{k|i}\pi_{l|i}}{\pi_{kl|i}} \right) \frac{y_k}{\pi_{k|i}} \frac{y_l}{\pi_{l|i}}. \quad (48)$$

El segundo componente de (43) puede ser estimado de forma estimada mediante:

$$\hat{V}_{USM} = \sum_{i \in S_I} \frac{\hat{V}_i}{\pi_{Ii}^2}. \quad (49)$$

Luego, un estimador insesgado de  $V_{2E}(\hat{t}_\pi)$  es:

$$\hat{V}_{2E}(\hat{t}_\pi) = \sum_{i \in S_I} \sum_{j \in S_I} \left( \frac{\pi_{Iij} - \pi_{Ii}\pi_{Ij}}{\pi_{Iij}} \right) \frac{\hat{t}_{i\pi} \hat{t}_{j\pi}}{\pi_{Ii} \pi_{Ij}} + \sum_{j \in S_I} \frac{\hat{V}_i}{\pi_{Iij}}. \quad (50)$$

Se puede notar que si bien es posible obtener un estimador insesgado para la variancia  $V_{2E}(\hat{t}_\pi)$ , el cálculo del estimador de variancia puede ser complicado porque requiere estimar  $V_i$  para cada  $i \in S_I$ .

### 3.1.3.3 Muestreo multietápico

A pesar de su complejidad, los diseños muestrales en tres o más etapas son usualmente utilizados en encuestas a gran escala. Por ejemplo, el Estudio Nacional sobre el Perfil de las Personas con Discapacidad realizado por el INDEC en el año 2018 utilizó un muestreo en tres etapas que recurría a estratificación y diferentes probabilidades de selección para las primeras dos etapas. Se presenta este tipo de diseños en forma general para  $r \geq 2$  etapas.

La población de elementos  $U = \{1, \dots, k, \dots, N\}$  es particionada en  $N_I$  UPM  $U_1, \dots, U_i, \dots, U_{N_I}$ . Este conjunto es  $U_I = \{1, \dots, i, \dots, N_I\}$  donde el tamaño de cada  $U_i$  es  $N_i$  y vale  $N = \sum_{i \in U_I} N_i$ .

Sean también:

- $s_I$ : La muestra de conglomerados de primera etapa, obtenida a través del diseño  $p_I(\cdot)$ .
- $\pi_{Ii}$  y  $\pi_{Iij}$ : Las probabilidades de inclusión de primer y segundo orden correspondientes a  $p_I(\cdot)$ .
- $t_i$ : El total poblacional del conglomerado  $i$  de primera etapa.

- $\hat{t}_i$ : Un estimador del total poblacional del conglomerado  $i$  de primera etapa. En este caso se asume que es un estimador insesgado de  $t_i$  para las  $r - 1$  últimas etapas de selección.
- Luego:

$$E(\hat{t}_i | s_I) = t_i. \quad (51)$$

Sea también  $V_i = V(\hat{t}_i | s_I)$  la variancia de  $\hat{t}_i$  debido a las  $r - 1$  últimas etapas de selección, y  $\hat{V}_i$  un estimador insesgado de  $V_i$  dado  $s_I$ , es decir,  $E(\hat{V}_i | s_I) = V_i$ .

Considerando las reglas de invariancia e independencia, en un muestreo en  $r$  etapas, un estimador insesgado del total poblacional  $t_y$  viene dado por:

$$\hat{t} = \sum_{i \in s_I} \frac{\hat{t}_i}{\pi_{li}}. \quad (52)$$

Luego, la variancia viene dada por:

$$V_{Mcr}(\hat{t}) = \sum_{i \in U_I} \sum_{j \in U_I} (\pi_{lij} - \pi_{li}\pi_{lj}) \frac{t_i}{\pi_{li}} \frac{t_j}{\pi_{lj}} + \sum_{i \in U_I} \frac{V_i}{\pi_{li}}. \quad (53)$$

donde el primer término representa a la variancia contribuida por la primera etapa de selección y la segunda sumatoria combina la variancia contribuida por las etapas subsiguientes. La misma puede ser estimada de forma insesgada mediante:

$$\hat{V}_{Mcr}(\hat{t}) = \sum_{i \in s_I} \sum_{j \in s_I} \left( \frac{\pi_{lij} - \pi_{li}\pi_{lj}}{\pi_{lij}} \right) \frac{\hat{t}_i}{\pi_{li}} \frac{\hat{t}_j}{\pi_{lj}} + \sum_{i \in s_I} \frac{\hat{V}_i}{\pi_{li}}. \quad (54)$$

Si bien esto dice poco de la naturaleza exacta de las últimas  $r - 1$  etapas de selección, tiene interés práctico por la siguiente razón: El esfuerzo computacional asociado al muestreo multietápico es frecuentemente considerable. En particular, esto se presenta en la estimación de variancia. Calcular las  $\hat{V}_i$  necesarias en el estimador de variancia puede ser muy trabajoso. Por ejemplo, si  $\hat{t}_i$  es el  $\pi$ -estimador relativo a las últimas  $r - 1$  últimas etapas de selección, el cálculo de  $\hat{V}_i$  requiere a las probabilidades de inclusión de segundo orden para todos los  $r - 1$  diseños.

El muestreo por conglomerados en tres etapas es un caso particular del muestreo multietápico. En este caso, la estimación insesgada de la variancia de un estimador del total requiere las probabilidades de inclusión de segundo orden para los tres diseños utilizados además de todos los componentes de variancia introducidos por las etapas de muestreo.

Es de gran necesidad contar con estimadores de variancia más sencillos. Existen alternativas que, a costa de tolerar un pequeño sesgo, logran simplicidad (Särndal, Swensson, & Wretman, 1992). Sin embargo, estos estimadores son presentados en un contexto donde se asume que todas las unidades en la muestra responden, y tampoco existen otros errores ajenos al muestreo.

En el marco de esta tesina, se busca estimar variancias en situaciones más realistas, como puede ser un estudio por muestreo complejo llevado a cabo por un Instituto Nacional de Estadística. En éstos la no respuesta es una problemática siempre presente. Los estimadores puntuales y de dispersión no solo deben considerarla, sino que idealmente deben ser útiles al propósito de mitigar los posibles sesgos que esta implica. Luego, es de gran importancia no solo conocer las propiedades de tales estimadores, sino también las consecuencias que puede tener diferentes mecanismos de no respuesta en ellos.

### **3.2 El problema de la no respuesta**

#### **3.2.1 Sesgo por no respuesta**

Existe consenso en que es necesario estudiar la falta de respuesta como uno de los factores que incide en la precisión de las estimaciones de un producto estadístico. La misma introduce sesgo si la población de los que no responden presenta características distintas a quienes si lo hacen, respecto de las variables de interés. Esto también ocasiona una disminución en el tamaño de la muestra. Dicha disminución incrementa el error de muestreo, lo cual conlleva a una limitación en la generalización de las conclusiones al conjunto de la población.

Para analizar el sesgo que puede introducir la no respuesta se presentan dos enfoques respecto al mecanismo que genera la respuesta. La misma puede ser pensada como determinística o estocástica (Kalton & Maligalig, 1991).

- 1. Determinística** – Cada unidad elegible en la población está predeterminada a responder, o no, en el caso de ser invitado a participar. La elección no es aleatoria. Bajo este razonamiento las unidades podrían ser clasificadas en los que responden y los que no responden de antemano.
- 2. Estocástica** – Cada elemento de la población tiene una probabilidad desconocida y no nula de responder. Luego, cada unidad toma una decisión al azar de cooperar o no cuando es invitado a participar.

Dados estos dos modelos sobre el fenómeno de la no respuesta, la comparación suele hacerse en términos del sesgo del estimador de la media de Hájek  $\hat{y}_{HJ}$  calculado con el conjunto de unidades de la muestra que responden  $s_r$ .

Luego, bajo el enfoque determinístico:

$$B(\hat{y}_{HJ}) = \frac{N_m}{N} (\bar{y}_{Ur} - \bar{y}_{Um}), \quad (55)$$

donde  $\bar{y}_{Ur}$  es la media verdadera para la población de unidades que responde,  $\bar{y}_{Um}$  es la media verdadera para la población de quienes no responden,  $N_m/N$  es la tasa de no respuesta poblacional calculada como la razón del tamaño de la población de aquellos que no responden  $N_m$ , sobre el tamaño poblacional total  $N$ .

Es claro que, según este enfoque, la no respuesta introduce sesgo si la media poblacional de las unidades que responden es diferente de la media poblacional de las que no responden.

Bajo una lógica similar, se puede estudiar el sesgo por no respuesta asumiendo que la respuesta es un mecanismo estocástico. Sean las siguientes variables indicadoras:

$$I_k = \begin{cases} 1 & \text{si la unidad } k \text{ está en la muestra,} \\ 0 & \text{caso contrario} \end{cases}, \quad (56)$$

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

La probabilidad de que la unidad  $k$  esté en la muestra no es más que la probabilidad de inclusión de primer orden, es decir  $P(I_k = 1) = \pi_k$ , mientras que la probabilidad de que la unidad responda

dado que está en la muestra es  $P(R_k = 1 | I_k = 1) = \phi_k$ . Rosenbaum P. y Rubin D.B. (1983) lo llaman *propensity score*<sup>2</sup> de la unidad  $k$ .

Este enfoque trata al hecho de que las unidades de la muestra respondan la encuesta como un mecanismo probabilístico, y luego la respuesta puede ser vista como una etapa más en el proceso de muestreo. Para ello es necesario asumir que cada elemento en la población tiene una probabilidad positiva de responder en el caso que sea incluida en la muestra,  $\phi_k > 0 \forall k \in s$ .

El sesgo que tiene el estimador  $\bar{y}_{HJ}$  viene aproximado por la siguiente fórmula:

$$B(\hat{y}_{HJ}) \cong \frac{1}{N\bar{\phi}} \sum_{k=1}^N (y_k - \bar{y}_U)(\phi_k - \bar{\phi}) = \frac{1}{\bar{\phi}} \sigma_\phi \sigma_y \rho_{\phi,y} = CV_\phi \sigma_y \rho_{\phi,y}, \quad (57)$$

donde:

- $\bar{\phi} = \sum_{k=1}^N \phi_k / N$  es la media de la propensión a responder de todas las unidades de la población.
- $\sigma_\phi$  es el desvío estándar poblacional de la propensión a responder.
- $\sigma_y$  es el desvío estándar poblacional de la variable en estudio.
- $\rho_{\phi,y}$  es la correlación poblacional entre la propensión a responder y la variable en estudio.
- $CV_\phi$  es el coeficiente de variación de la propensión a responder.

Se observa que bajo un enfoque estocástico el estimador de la media Hájek calculado sobre el conjunto que responde tiene un sesgo que depende de la correlación entre la variable de interés,  $y$ , y la probabilidad de responder,  $\phi$ . Además, se puede notar que el sesgo se relaciona con el coeficiente de variación de la probabilidad de responder y la variabilidad de la característica de interés. Luego,  $\hat{y}_{HJ}$  es aproximadamente insesgado cuando se cumple al menos una de las siguientes condiciones

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

---

<sup>2</sup> Este concepto puede ser pensado como la *propensión a participar* de la unidad.

### **3.2.2 Mecanismos de no respuesta**

Previo a discutir la metodología utilizada para mitigar los posibles efectos de la no respuesta es necesario introducir la terminología para describir los mecanismos de ocurrencia de la misma. La capacidad de los métodos para mitigar el sesgo que puede introducir la no respuesta depende de la naturaleza del mecanismo que la genera. Además, en esta tesis se plantean los diferentes escenarios de simulación a partir de ellos. La siguiente terminología es tomada de Little R. y Rubin D. (2002)

- i. **Missing completely at random (MCAR):** La probabilidad que tiene una unidad de responder no está relacionada con la variable en estudio,  $y_k$ , ni con ninguna variable auxiliar,  $x_k$ . Si cada unidad tiene la misma probabilidad de responder  $\phi_k$ , el mecanismo de respuesta es MCAR. Este escenario respeta las condiciones de insesgamiento en forma aproximada planteadas para  $\bar{y}_{HJ}$ .
- ii. **Missing at Random (MAR):** La probabilidad que tiene una unidad de responder no está relacionada con  $y_k$ , pero sí depende de  $x_k$ , que es un vector de variables auxiliares conocido tanto para los que responden como para los que no. En este caso, es factible elaborar un modelo para la probabilidad de responder basado en  $x_k$ .
- iii. **Not missing at random (NMAR):** La probabilidad de responder de la unidad no solo que puede estar relacionada con  $x_k$ , sino que está relacionada con la variable de interés  $y_k$ . En este caso, la dependencia que tiene la respuesta en  $y_k$  no puede ser mitigada mediante un modelo que utilice a  $x_k$ . Como  $y_k$  es desconocida para las unidades que no responden, no es posible utilizarla de alguna forma para reducir el sesgo de no respuesta.

Una consecuencia evidente de la no respuesta es la disminución del tamaño muestral efectivo, que se reduce de  $n$  a  $n_r$ , con  $n_r \leq n$ . Una alternativa posible es calcular las estimaciones con el subconjunto de  $n_r$  unidades tal como se lo hubiera hecho con la muestra completa. Sin embargo, esta opción solo es válida en el caso que el mecanismo de respuesta sea MCAR.

Es demasiado optimista pensar que el mecanismo de respuesta no depende de ninguna característica asociada a la unidad de estudio, ya que, además, éste suele estar relacionado con variables auxiliares e incluso con la variable en estudio. Por otro lado, no solo que se desconoce la relación funcional que pueda existir entre la probabilidad de responder y un conjunto de variables auxiliares, sino que no se puede saber de antemano cuales son las variables que afectan a la probabilidad de respuesta, y aun si se supiera, tener acceso a ellas puede ser muy dificultoso o imposible.

Tal como mencionan Särndal y Lundström (2005), ante la presencia de no respuesta el mejor de los escenarios es aquel en el cual se disminuye considerablemente el sesgo ocasionado, pero sabiendo que no es posible eliminarlo en su totalidad.

### **3.3 Ajuste por no respuesta**

El ajuste de no respuesta se refiere los diversos intentos realizados para tratar la no respuesta una vez que ha ocurrido, es decir, después de aceptar el hecho de que no todas las unidades de la muestra responden. Como lo sugiere la palabra *ajuste*, se realizan modificaciones a un procedimiento de estimación original o *ideal*, que hubiera sido utilizado en el caso de que todas las unidades respondiesen. El principal propósito de esta práctica es reducir el sesgo que pueda causar la no respuesta.

Los principales métodos para el ajuste por no respuesta son la imputación de valores faltantes y el ajuste de pesos muestrales. Generalmente, el primero de estos métodos se utiliza cuando una unidad de la muestra no responde algunos ítems de la encuesta, mientras que el segundo se utiliza cuando la unidad muestral no responde a ninguna pregunta de la encuesta. Estos fenómenos se conocen como no respuesta a nivel ítem y no respuesta a nivel unidad, respectivamente.

El ajuste de pesos muestrales implica alterar los pesos de las unidades que responden, en comparación con los pesos que se hubieran utilizado en el caso que se hubiese obtenido una respuesta total. Dada la pérdida de observaciones provocada por la no respuesta, la ponderación que se realiza conlleva un aumento de los pesos originales para aquellas unidades que responden. Por otro lado, la imputación se utiliza cuando hay no respuesta a nivel de ítem, y ésta consiste en asignar valores sustitutos a los datos faltantes de una unidad. Es frecuente que en un estudio por

muestreo se imputen los valores faltantes de las unidades que presentan no respuesta a nivel ítem y se ajusten los pesos muestrales de los respondientes para contemplar la no respuesta a nivel unidad.

Dado que en el presente trabajo solo se estudia la no respuesta a nivel unidad, los términos *ajuste por no respuesta y ponderación de pesos muestrales* son utilizados como sinónimos.

En el caso de estudios por muestreo que presentan no respuesta, es deseable que un estimador cumpla con las siguientes propiedades:

- i. Sesgo pequeño.
- ii. Variancia pequeña.
- iii. Esté basado en un conjunto de pesos que cuando sean utilizados para estimar cantidades poblacionales conocidas de variables auxiliares, el resultado obtenido se corresponda con éstos.
- iv. El conjunto de pesos muestrales utilizado puede ser utilizado para estimar el total de cualquier variable de interés de forma precisa.

### **3.3.1 Motivación del ajuste por no respuesta**

Es de central importancia que se utilicen todos los recursos disponibles para reducir en la mayor medida posible el sesgo de un estimador a utilizar. Los esfuerzos pueden centrarse en el sesgo relacionado al diseño utilizado y en el sesgo relacionado al mecanismo de respuesta.

Sea  $d_k^*$  el peso que le corresponde a la unidad  $k$  luego de que se realice algún ajuste por no respuesta y sea el siguiente estimador del total:

$$\hat{t}_y = \sum_{k \in s_r} d_k^* y_k. \quad (58)$$

La teoría de muestreo en dos fases, y considerando a la respuesta como una fase más del muestreo, permite calcular a la esperanza de (59) primero sobre el muestreo,  $E_I$ , y luego sobre el mecanismo de respuesta,  $E_R$ , o sea:

$$E_I E_R(\hat{t}_y) = E_I E_R \left( \sum_{k \in U} I_k R_k d_k^* y_k \right) = \sum_{k \in U} d_k^* y_k E_I E_R(I_k R_k). \quad (59)$$

Si fuese posible que  $d_k^* = 1/E_I E_R(I_k R_k)$  la expresión se reduce al total poblacional. Dado que  $E_I E_R(I_k R_k) = E_I[I_k E_R(R_k | I_k)] = \pi_k \phi_k$ , este peso debería ser  $d_k^* = (\pi_k \phi_k)^{-1}$ .

Luego, las condiciones necesarias para obtener un estimador insesgado del total son:

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

Ante la presencia de no respuesta y con el objetivo de obtener estimadores insesgados, no solo se requiere que se utilice un muestreo probabilístico, sino que además todas las unidades deben tener una probabilidad positiva de responder en el caso de ser parte de la muestra. Luego, es necesario asumir que la respuesta es un mecanismo estocástico y no determinístico.

### 3.3.2 Ajuste mediante la propensión a responder

Existen diversos métodos que buscan lidiar con la no respuesta a través de la estimación directa de  $\phi_k$  para su utilización en la forma expuesta en la sección anterior. Para ello se asume que  $\phi_k > 0 \forall k$  y que las unidades responden independientemente. Es posible postular un modelo de la siguiente forma:

$$\phi_k = m(\mathbf{x}_k, \boldsymbol{\gamma}), \quad (60)$$

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  $\boldsymbol{\gamma}$  es un vector de parámetros desconocidos.

Una posible estimación de  $\phi_k$  es  $\hat{\phi}_k = m(\mathbf{x}_k, \hat{\boldsymbol{\gamma}})$  donde  $\hat{\boldsymbol{\gamma}}$  es una estimación de  $\boldsymbol{\gamma}$ . Luego se obtiene el conjunto de pesos  $\{\tilde{w}_k\}_{k \in S_r}$  donde  $\tilde{w}_k = d_k / \hat{\phi}_k$ . Un estimador que utiliza este conjunto de pesos recibe el nombre de estimador ajustado por la propensión a responder o PSA por sus siglas en inglés (Propensity-Score Adjusted). Por ejemplo, para el total

$$\hat{t}_{PSA} = \sum_{k \in S_r} d_k \hat{\phi}_k^{-1} y_k = \sum_{k \in S_r} \tilde{w}_k y_k \quad (61)$$

Este tipo de estimadores responde a la lógica del muestreo en dos fases expresada también en la sección anterior.

Existen formas paramétricas para la estimación de las  $\phi_k$ , como por ejemplo una regresión logística. Kim & Kim (2007) demuestran que el estimador **(61)** es asintóticamente insesgado y consistente de  $t_y$  sin importar cual sea la característica en estudio, siempre y cuando la relación funcional **(60)** esté correctamente especificada. Sin embargo, no es frecuente la utilización de modelos paramétricos porque estos pueden generar  $\hat{\phi}_k$  muy pequeños, lo que resulta en un conjunto de pesos  $\{\tilde{w}_k\}_{k \in S_r}$  muy dispersos. Además, estos métodos son vulnerables ante una incorrecta especificación de  $m(\cdot)$ , que en la práctica es desconocida.

Generalmente se prefiere utilizar métodos no paramétricos porque se espera que estos provean de cierta protección si la forma de  $m(\cdot)$  es especificada incorrectamente. Un método popular, llamado *método score* (Haziza & Beaumont, 2007) consiste en obtener estimaciones preliminares de la propensión a responder utilizando un modelo paramétrico, y luego particionar la muestra en conjuntos homogéneos respecto a  $\hat{\phi}_k$  llamados clases de ponderación. Finalmente, el peso de una unidad que responde es ajustado mediante un factor de ajuste común para todos los elementos de la clase a la que esta unidad pertenece. Valliant, Dever, & Kreuter (2018) sugieren diversas alternativas como por ejemplo la tasa de respuesta observada para la clase.

Por último, es necesario notar que en la práctica las probabilidades de respuesta nunca son conocidas, y el estudio del efecto de la no respuesta requiere que se realicen supuestos sobre éstas.

### 3.4 Estimador de calibración

#### 3.4.1 Estimador de calibración para el caso de respuesta completa

Los estimadores de calibración para el caso de respuesta completa son descriptos en Särndal y Deville (1992) y en Deville et. al. (1993). Su utilización para el ajuste de no respuesta es presentada por Lundström y Särndal (1999) y extensamente trabajado en Särndal y Lundström (2005).

A diferencia de estimadores clásicos como el estimador de Horvitz-Thompson o el estimador de Hájek, el método de calibración constituye un enfoque general y sistemático para la utilización de información auxiliar. Una gran variedad de estimadores utilizados para ajustar por no respuesta están comprendidos bajo este enfoque, como por ejemplo los estimadores de post-estratificación, raking, y los estimadores de regresión lineal generalizada. Haziza y Lesage (2016) muestran que

existe una relación entre los estimadores de calibración y los modelos paramétricos para ajustar por propensión a responder. Esta relación se da a través de la función de distancia utilizada en el proceso de calibración; cada una de estas implica un modelo de no respuesta distinto.

Para llevar a cabo el proceso de calibración se debe considerar la información auxiliar disponible. Esta puede consistir en una sola variable o bien un conjunto de ellas y el vector que la contiene se denomina vector auxiliar. En el contexto de estudios por muestreo a gran escala, es frecuente que las variables utilizadas refieran a conteos de personas según diversas clasificaciones sociodemográficas.

Sea  $\mathbf{x}_k$  el vector de variables auxiliares para la unidad  $k$  y sea  $\mathbf{t}_x = \sum_{k \in U} \mathbf{x}_k$ , vector de totales poblacionales conocidos. Generalmente se tiene que  $\mathbf{x}_k$  es conocido para todo  $k \in U$  pero basta con que sea conocido para las unidades de la muestra.

Un estimador de calibración para el total  $t_y$  es uno de la forma:

$$\hat{t}_y^{CAL} = \sum_{k \in s} w_k y_k, \quad (62)$$

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

La elección del conjunto de pesos calibrados se hace de forma tal que una medida de distancia entre los vectores  $(d_1, d_2, \dots, d_n)'$  y  $(w_1, w_2, \dots, w_n)'$  sea mínima, sujeto a la ecuación de calibración.

Sea una función de distancia  $G$  de argumento  $r_k = w_k/d_k$ , que mide la distancia entre  $\{w_k\}_{k \in s}$  y  $\{d_k\}_{k \in s}$ . La función  $G$  debe ser positiva, convexa y satisfacer  $G(1) = G'(1) = 0$  (Särndal & Deville, 1992).

Una vez que la función es elegida, el problema consiste en determinar los pesos  $\{w_k\}_{k \in s}$  que son solución del siguiente problema de búsqueda de extremo condicional:

$$\begin{aligned} \min_{w_k} & \sum_{k \in s} d_k G(w_k/d_k) \\ \text{s.t. } & \sum_{k \in s} w_k \mathbf{x}_k = \mathbf{t}_x \end{aligned} \quad (63)$$

Se minimiza una suma ponderada de distancias entre los pesos originales y los pesos de calibración, sujeto a la ecuación de calibración. Este problema se resuelve introduciendo un vector de multiplicadores de Lagrange  $\lambda' = (\lambda_1, \dots, \lambda_p)$ , donde el lagrangiano es igual a:

$$L = \sum_{k \in S} d_k G(w_k/d_k) - \lambda' \left( \sum_{k \in S} w_k \mathbf{x}_k - \mathbf{t}_x \right), \quad (64)$$

lo que conduce a:

$$w_k = d_k F(\mathbf{x}'_k \lambda) \quad (65)$$

donde  $F$  es la inversa de la derivada de la función  $G$ . El vector  $\lambda$  está determinado por la resolución de un sistema de  $p$  ecuaciones no lineales con  $p$  incógnitas que resultan de las ecuaciones de calibración  $\sum_{k \in S} F(\mathbf{x}'_k \lambda) \mathbf{x}_k = \mathbf{t}_x$ . Para la resolución del problema, generalmente se utilizan métodos iterativos como el método de Newton.

### 3.4.1.1 Elección de la función $G$

Existen varias funciones que pueden ser utilizadas para realizar la calibración. Éstas se diferencian en la complejidad matemática, el consecuente tiempo de cómputo, y sobre todo en la capacidad que tienen de controlar cuán extremos pueden llegar a ser los factores de ajuste producidos. Los métodos más populares son el lineal, exponencial, logit y truncado. Algunos métodos, como por ejemplo el lineal y exponencial, pueden llegar a producir pesos extremadamente grandes, y en el caso lineal, incluso negativos. Otros, como por ejemplo el método logit y el método truncado, permiten establecer una cota inferior y una superior para el valor de  $r_k = w_k/d_k$ . A continuación, se presentan las funciones  $G$  y  $F$  (donde  $r_k = w_k/d_k$  y  $u_k = \mathbf{x}'_k \lambda$ ) para el método lineal y el método logit.

#### Método lineal

$$\begin{aligned} G(r_k) &= \frac{1}{2} (r_k - 1)^2 & r_k \in \mathbb{R} \\ F(u_k) &= 1 + u_k & u_k \in \mathbb{R} \end{aligned} \quad (66)$$

La forma lineal de  $F$  le da su nombre a este método, que puede verse como equivalente a un método de estimación clásico que usa información auxiliar, llamado estimador de regresión. Si  $\sum_{k \in S} d_k \mathbf{x}_k \mathbf{x}_k^T$  existe, las ecuaciones de calibración tienen forma cerrada. Caso contrario se utilizan métodos numéricos. Este método tiene la principal ventaja de converger muy rápidamente (en general dos iteraciones son suficiente), pero no permite especificar límites para la razón  $r_k$ , por lo que podría arrojar pesos muy elevados o incluso negativos.

### Método logit

$$\begin{aligned} G(r_k) &= \left[ (r_k - L) \log\left(\frac{r_k - L}{1 - L}\right) + U - r_k \log\left(\frac{U - r_k}{U - 1}\right) \right] \frac{1}{A} \quad L < r_k < U \\ F(u_k) &= \frac{L(U - 1) + U(1 - L)\exp(Au_k)}{U - 1 + (1 - L)\exp(Au_k)} \in (L, U) \\ \text{con } A &= \frac{U - L}{(1 - L)(U - 1)} \end{aligned} \tag{67}$$

La forma logística de  $F$  le da nombre a este método. Si bien el aspecto computacional es más complejo, este permite especificar una cota inferior  $L$  y una cota superior  $U$  para la razón  $r$ . Sin embargo, no existe un conjunto de valores predefinidos para todos los problemas. Estos valores dependen de la estructura de los datos en la muestra y los márgenes de calibración en la población. Cuanto más diferente es la estructura de la muestra de la estructura de la población, más distante se encuentra  $r_k$  de la unidad.

La elección de un método de calibración depende principalmente de los recursos computacionales disponibles y la preferencia del investigador. Särndal y Deville (1992) muestran que el estimador de calibración (62) es aproximadamente insesgado por diseño y consistente por diseño de manera asintótica de  $t_y$  sin importar la característica de interés, y que todos los métodos de calibración son asintóticamente equivalentes en el sentido de que todos conducen al estimador de calibración basado en el método lineal.

#### 3.4.1.2 Calibración en conteos conocidos de tablas de frecuencia

Una aplicación importante de la técnica mencionada se presenta en relación con la calibración en totales conocidos de una tabla de frecuencia (celdas o márgenes) de cualquier número de

dimensiones. En este caso la discusión se reduce solo a tablas de dos dimensiones, con  $n_{row}$  filas y  $n_{col}$  columnas, donde cada uno de los elementos poblacionales es clasificado dentro de una de las  $n_{row} \times n_{col}$  celdas. Por ejemplo, individuos clasificados en rango de edad y categoría socio-profesional. Supóngase que la celda poblacional  $U_{ij}$  contiene  $N_{ij}$  elementos  $i = 1, \dots, n_{row}; j = 1, \dots, n_{col}$  de forma tal que  $N = \sum_{i=1}^{n_{row}} \sum_{j=1}^{n_{col}} N_{ij}$ . Se pueden distinguir dos casos, (a) calibración en los totales de las celdas  $N_{ij}$ , que puede ser entendido como una *post-estratificación completa* y (b) calibración en los totales marginales, que puede ser conocida como *post-estratificación incompleta*. En el caso (a), el estimador de calibración coincide con el popular estimador de post-estratificación donde cada una de las  $n_{row} \times n_{col}$  celdas es un post-estrato. El caso (b) se corresponde con una clase de estimadores conocidos como procedimientos de raking generalizados. Para este último caso se mencionan dos de las situaciones donde se aplica en la práctica. La primera se da cuando se conocen los conteos marginales, pero se desconocen los conteos por celda  $N_{ij}$ . Esto puede suceder cuando los conteos marginales provienen de distintas fuentes de datos, y naturalmente se carece de los conteos por celda. Aquí la calibración se hace en los márgenes por una cuestión de necesidad. La segunda situación se da cuando se cuenta con los conteos por celda, pero algunas de estas presentan valores nulos o muy pequeños. Si bien parece posible calibrar utilizando los valores de las celdas, en la práctica, es usual optar por calibrar en los márgenes ya que en esta situación se lo considera más seguro (Särndal & Deville, 1992). Aún más, a mayor número de dimensiones, mayor riesgo de presentar celdas con conteos pequeños o nulos, y consecuentemente mayor es la preferencia por calibrar en los márgenes en vez de celdas individuales.

Para calibrar en los márgenes se debe contar con un vector auxiliar  $\mathbf{x}_k \forall k \in U$  y el total auxiliar asociado a ese vector,  $\mathbf{t}_x$ . Se puede ver que  $\mathbf{x}_k = (\delta_{1 \cdot k}, \delta_{2 \cdot k}, \dots, \delta_{n_{row} \cdot k}, \delta_{\cdot 1 k}, \delta_{\cdot 2 k}, \dots, \delta_{\cdot n_{col} k})'$  donde  $\delta_{i \cdot k} = 1$  si el elemento  $k$  pertenece a la fila  $i$  y 0 en el caso contrario, y  $\delta_{\cdot j k} = 1$  si el elemento  $k$  pertenece a la columna  $j$  y 0 en el caso contrario. En consecuencia se tiene  $\sum_{k \in U} \mathbf{x}_k = (N_{1+}, \dots, N_{n_{row}+}, N_{+1}, \dots, N_{+n_{col}})'$  donde  $N_{i+} = \sum_{j=1}^{n_{col}} N_{ij}$ ,  $N_{+j} =$

$\sum_{j=1}^{n_{col}} N_{ij}$ . Si se expresa al vector  $\lambda'$  de la forma  $\lambda' = (u_1, \dots, u_{n_{row}}, v_1, \dots, v_{n_{col}})$ , se tiene que  $x'_k \lambda' = u_i + v_j$  cuando  $k$  pertenece a la celda  $ij$ . Siendo  $\widehat{N}_{ij} = \sum_{k \in s_{ij}} 1/\pi_k$  las ecuaciones de calibración resultan:

$$\begin{aligned} \sum_{j=1}^{n_{col}} \widehat{N}_{ij} F(u_i + v_j) &= N_{i+} \quad i = 1, \dots, n_{row} \\ \sum_{i=1}^{n_{row}} \widehat{N}_{ij} F(u_i + v_j) &= N_{+j} \quad j = 1, \dots, n_{col} \end{aligned} \tag{68}$$

La solución a estas ecuaciones puede encontrarse mediante algún procedimiento iterativo, utilizando la función  $F$  especificada por el investigador. Es relevante notar que una de las  $n_{row} + n_{col}$  ecuaciones es redundante, por lo que es posible fijar una de las componentes en 0, por ejemplo  $u_1$  o  $v_{n_{col}}$ , y luego resolver el sistema para  $i = 2, \dots, n_{row}$  y  $j = 1, \dots, n_{col}$  o  $i = 1, \dots, n_{row}$  y  $j = 1, \dots, (n_{col} - 1)$  respectivamente. Notar que  $u_i + v_j$  permanece invariante ante la eliminación de una de las ecuaciones.

Una vez obtenidos  $u_i$  y  $v_j$  se calculan los factores  $F(u_i + v_j)$ , luego las estimaciones calibradas de los conteos por celda  $\widehat{N}_{ij}^{CAL} = \widehat{N}_{ij} F(u_i + v_j)$  y los pesos calibrados  $w_k = d_k \widehat{N}_{ij}^{CAL} / \widehat{N}_{ij}$ . Finalmente, el estimador de calibración para el total es

$$\widehat{t}_y^{CAL} = \sum_{k \in s} w_k y_k = \sum_{i=1}^{n_{row}} \sum_{j=1}^{n_{col}} \widehat{N}_{ij}^{CAL} \tilde{y}_{s_{ij}} \tag{69}$$

donde  $\tilde{y}_{s_{ij}} = (\sum_{k \in s_{ij}} d_k y_k) \widehat{N}_{ij}$ . Las estimaciones de los conteos por celda  $\widehat{N}_{ij}^{CAL}$  suelen ser una mejora sustancial en los estimadores  $\widehat{N}_{ij}$ . De hecho, (69) puede ser tan eficiente como el estimador post-estratificado ( $\sum_{i=1}^{n_{row}} \sum_{j=1}^{n_{col}} N_{ij} \tilde{y}_{s_{ij}}$ ) cuando los  $N_{ij}$  son conocidos. Si el efecto de las filas y las columnas en  $y$  es aditivo, es decir, su interacción es depreciable, luego (69) y el estimador post-estratificado tienen prácticamente la misma variancia (Särndal & Deville, 1992)

### 3.4.2 Estimador de calibración en presencia de no respuesta

El enfoque de calibración ayuda a cumplimentar el doble objetivo de reducir el error muestral y el error por no respuesta, siempre y cuando se cuente con buena información auxiliar. Además, puede ser aplicado para cualquiera de los diseños muestrales utilizados frecuentemente y para cualquier vector auxiliar de información.

Särndal y Lundström (2005) distinguen entre dos niveles de información auxiliar que se pueden utilizar para calibrar pesos muestrales ante la presencia de no respuesta. En ambos casos, se asume que las variables auxiliares están relacionadas con la no respuesta, y posiblemente, con la característica de interés. El rol de estas variables es contribuir a la reducción del sesgo que introduce la no respuesta.

- i. Nivel poblacional ( $U$  – level): El vector de variables auxiliares  $\mathbf{x}_k^*$  está disponible al menos para todas las unidades de  $s_r$  y se conoce al vector de totales poblacionales  $\mathbf{t}_{\mathbf{x}^*} = \sum_{k \in U} \mathbf{x}_k^*$ .
- ii. Nivel muestral ( $s$  – level): El vector de variables auxiliares  $\mathbf{x}_k^o$  está disponible para todas las unidades en la muestra, pero se desconocen los totales poblacionales,  $\sum_{k \in U} \mathbf{x}_k^o$ . En cambio, se utiliza el vector de totales muestrales  $\hat{\mathbf{t}}_{\mathbf{x}^o} = \sum_{k \in s} d_k \mathbf{x}_k^o$ .

Sea el vector de variables auxiliares para la unidad  $k$   $\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^o \end{pmatrix}$  y el correspondiente vector de totales  $\mathbf{t}_x = \begin{pmatrix} \mathbf{t}_{\mathbf{x}^*} \\ \hat{\mathbf{t}}_{\mathbf{x}^o} \end{pmatrix}$ . El conjunto de pesos obtenidos es  $\{w_k\}_{k \in s_r}$  donde:

$$w_k = d_k F(\mathbf{x}_k' \boldsymbol{\lambda}_r), \quad (70)$$

$F$  es nuevamente la derivada de la función de distancia elegida para resolver el problema de búsqueda de extremo y  $\boldsymbol{\lambda}_r$  es determinado de forma tal que cumpla con las ecuaciones de calibración:

$$\sum_{k \in s_r} w_k \mathbf{x}_k = \mathbf{t}_x. \quad (71)$$

Finalmente, el estimador de calibración para estimar el total de una característica de interés es:

$$\hat{t}_y^{CAL} = \sum_{k \in S_r} w_k y_k = \sum_{k \in S_r} d_k F(\mathbf{x}'_k \boldsymbol{\lambda}_r) y_k. \quad (72)$$

Al igual que en el caso de respuesta completa, la calibración a totales poblacionales asegura que la distribución de las variables auxiliares en el conjunto que responde es igual a la distribución de estas variables en la población. Cuando la distribución de las variables auxiliares es calibrada a los valores poblacionales, también se está asumiendo que la distribución ponderada de la variable de interés se asemeja a la distribución de ésta en la población. Esto refleja que se supone que el mecanismo de no respuesta es *missing at random* (MAR). Es decir, dentro de las clases formadas por los niveles de las variables auxiliares las unidades tienen valores similares de la variable de interés. Luego, en esas clases los que responden se asemejan a los que no responden y es factible *inflar* el peso de los primeros para representar el total de la clase.

### 3.4.2.1 Sesgo del estimador de calibración en presencia de no respuesta

Recordando que el estimador  $\hat{t}_{PSA}$  es asintóticamente insesgado y consistente de  $t_y$  sin importar cual sea la característica en estudio, siempre y cuando la relación funcional (60) esté bien especificada, se puede notar que  $\hat{t}_y^{CAL}$  en (72) goza de las mismas propiedades si se cumple:

$$F(\mathbf{x}'_k \boldsymbol{\lambda}_r) = \hat{\phi}_k^{-1}. \quad (73)$$

Esta expresión sugiere que el factor de ajuste  $F(\mathbf{x}'_k \boldsymbol{\lambda}_r)$  puede ser visto como un estimador implícito de  $\hat{\phi}_k^{-1}$ .

Luego, se estudia el sesgo de  $\hat{t}_y^{CAL}$  en presencia de no respuesta. Éste se define como  $B(\hat{t}_y^{CAL}) = E_I E_R(\hat{t}_y^{CAL}) - t_y$ , donde  $E_I$  refiere a la esperanza respecto del diseño muestral y  $E_R$  respecto del mecanismo de respuesta. Utilizando expansión por series de Taylor e ignorando términos de mayor orden, el sesgo de  $\hat{t}_y^{CAL}$  se aproxima por (Haziza & Lesage, 2016):

$$B(\hat{t}_y^{CAL}) \approx - \sum_{k \in U} (1 - \phi_k F_k)(y_k - \mathbf{x}'_k \mathbf{B}_{\phi f}), \quad (74)$$

donde

$$\mathbf{B}_{\phi f} = \left( \sum_{k \in U} \phi_k f_k \mathbf{x}_k \mathbf{x}_k^T \right)^{-1} \sum_{k \in U} \phi_k f_k \mathbf{x}_k y_k, \quad (75)$$

con  $F_k \equiv F(\mathbf{x}'_k \boldsymbol{\lambda}_N)$ ,  $f_k \equiv F'(\mathbf{x}'_k \boldsymbol{\lambda}_N)$  y  $\boldsymbol{\lambda}_N$  es el valor al que  $\hat{\boldsymbol{\lambda}}_r$  converge en probabilidad.

La expresión (75) permite observar condiciones en las cuales el estimador  $\hat{t}_y^{CAL}$  es aproximadamente insesgado asintóticamente:

- i. Por un lado, el sesgo asintótico (74) desaparece si la covariancia poblacional entre los residuos  $e_k = (y_k - \mathbf{x}'_k \mathbf{B}_{\phi f})$  y  $\delta_k = (\phi_k F_k - 1)$  es igual a cero.

Esta condición se satisface si:

$$y_k = \mathbf{x}'_k \boldsymbol{\beta} + \epsilon_k, \quad (76)$$

con

$$E(\epsilon_k | \mathbf{x}_k) = 0, \quad (77)$$

y si la probabilidad de responder  $\phi_k$  no está relacionada con  $y_k$  luego de condicionar en los valores de  $\mathbf{x}_k$ . Esta última condición es el supuesto de que la respuesta es *missing at random* (MAR).

- ii. Por otro lado, el sesgo asintótico de  $\hat{t}_y^{CAL}$  es nulo si:

$$F_k = \phi_k^{-1}. \quad (78)$$

Luego, la selección de una función de calibración  $F(\cdot)$  que satisfaga (78) asegura que el estimador de calibración (72) es asintóticamente insesgado sin importar la característica de interés, incluso si no se cumplen (76) y (77).

En otras palabras, esta última condición explica que la elección de una función particular  $F(\cdot)$  implica un modelo paramétrico para la probabilidad de responder. Por ejemplo, en el caso del método lineal,  $\phi_k^{-1} = 1 + \mathbf{x}'_k \boldsymbol{\lambda}$  para todo  $k \in U$ .

### 3.4.3 Ajuste por no respuesta en uno y en dos pasos

En esta sección se introducen algunas de las prácticas usuales para realizar el ajuste de pesos muestrales. Hay dos tipos de procedimientos comúnmente utilizados para tratar la no respuesta a nivel unidad: (i) ajuste por la propensión a responder seguido de calibración, conocido como

enfoque a dos pasos (*two-step approach*) y (ii) ajuste de no respuesta mediante calibración, conocido como enfoque a un paso (*one-step approach*) (Haziza & Lesage, 2016).

El enfoque a dos pasos consiste primero en modificar los pesos muestrales multiplicándolos por un factor de ajuste por no respuesta, usualmente el inverso de una estimación de la propensión a responder. Luego estos pesos ajustados son nuevamente modificados mediante calibración, de forma tal que la estimación de variables auxiliares coincide con totales conocidos.

El objetivo del primer paso de ajuste es reducir el sesgo por no respuesta. Que se logre reducir el sesgo de manera eficiente depende de la información auxiliar disponible (Särndal & Lundström, 2005). En el segundo paso se busca consistencia entre las estimaciones obtenidas con los pesos de la encuesta y totales poblacionales conocidos. Si las variables de calibración están relacionadas a las características de interés, el estimador de calibración resultante tiende a ser más eficiente que un estimador no calibrado.

Los objetivos del ajuste a un paso son los mismos que a dos pasos: reducir el sesgo por no respuesta, asegurar la consistencia entre estimaciones de la encuesta y totales conocidos, y posiblemente, reducir la variancia de los estimadores puntuales. Sin embargo, la mayor diferencia es que el ajuste de pesos se realiza en un solo paso y no requiere la estimación explícita de probabilidades de respuesta.

La elección de la función de calibración tiene mayor relevancia en el caso del ajuste a un paso. En el ajuste a dos pasos los pesos ya fueron modificados cuando se realiza la calibración, por lo que la elección se puede hacer utilizando los mismos criterios que en un contexto de respuesta completa. Por otro lado, en el enfoque a un paso, diferentes funciones de calibración pueden conducir a estimadores de calibración con propiedades muy distintas en términos de sesgo y error cuadrático medio (Haziza & Lesage, 2016).

Si bien el enfoque a un paso no utiliza explícitamente probabilidades de responder en la construcción de los estimadores, la expresión (78) sugiere que cada función de calibración se corresponde con un modelo para  $\phi_k$ . Una incorrecta especificación de la función  $F(\cdot)$  puede conducir a mayor sesgo en las estimaciones, incluso en presencia de una fuerte asociación entre

la variable en estudio y el conjunto de variables de calibración. Haziza y Lesage (2016) tratan con mayor profundidad este problema y presentan resultados empíricos que exponen esta situación. La obtención del conjunto de pesos que se corresponde con el ajuste a un paso es la expuesta en la sección 3.4.2. Por una cuestión de notación, se puede expresar a  $\hat{t}_y^{CAL}$  como  $\hat{t}_y^{C,1} = \sum_{k \in s_r} w_k^{C,1} y_k$ . Por otro lado, un estimador del total  $t_y$  bajo un ajuste en dos pasos es de la forma:

$$\hat{t}_y^{C,2} = \sum_{k \in s_r} w_k^{C,2} y_k = \sum_{k \in s_r} d_k \hat{\phi}_k^{-1} F(\mathbf{x}'_k \boldsymbol{\lambda}) y_k. \quad (79)$$

La capacidad que tenga este estimador para cumplir los objetivos planteados depende del método utilizado para estimar propensión a responder y de cuan relacionada esté  $y_k$  con  $\mathbf{x}_k$ .

Si se pueden crear clases o grupos donde las unidades tengan una probabilidad similar de responder o valores similares de la variable respuesta, el sesgo por no respuesta puede ser prácticamente eliminado. Haziza y Beaumont (2007) presentan un método para la creación de clases donde la probabilidad de responder es homogénea. También es frecuente que en la práctica se utilicen los niveles de alguna variable categórica para conformar estas clases.

Independientemente del método utilizado para agrupar a las unidades, una manera sencilla de calcular un factor de corrección por propensión a responder para las unidades de la clase  $c$  con  $c = 1, \dots, C$  es

$$\hat{\phi}_k^{-1} = a_c = \frac{\sum_{k \in s_c} d_k}{\sum_{k \in s_{r,c}} d_k} \quad (80)$$

para todo  $k \in s_r \wedge k \in c$ .

Luego, el factor de corrección por no respuesta es la suma de los pesos de las unidades de la muestra que pertenecen a la clase  $c$ , divididos por la suma de los pesos de las unidades de esa clase que responden (Valliant, Dever, & Kreuter, 2018).

#### **3.4.4 Notas sobre el sesgo, el vector auxiliar y la propensión a responder**

En las secciones previas se ha notado en reiteradas oportunidades que, ante la presencia de no respuesta, el sesgo del estimador tiene que ver con las relaciones que puedan existir entre: la probabilidad de responder, la variable en estudio y el vector auxiliar.

Los términos MCAR y MAR antes descriptos son catalogados como respuesta ignorarable. Por el contrario, cuando la no respuesta es un mecanismo NMAR se dice que ésta es no ignorarable.

En el caso de MCAR, como el fenómeno de no respuesta no está relacionado con ninguna variable auxiliar ni con la variable en estudio, se puede considerar que el conjunto de unidades que responden conforma una submuestra aleatoria de la muestra original. Esto no se cumple de la misma manera cuando la no respuesta es MAR, ya que aquí hay una relación entre la probabilidad de responder y el conjunto de variables auxiliares. Sin embargo, las unidades que responden conforman una submuestra aleatoria dentro de las clases dadas por el vector auxiliar. Luego, las variables auxiliares pueden ser utilizadas para corregir el sesgo que provoca la no respuesta.

Särndal y Lundström (2005) sugieren que las interpretaciones posibles de los fenómenos MAR y MCAR están sujetas a ciertas ambigüedades. Para cada elemento  $k$  de una población finita  $U$  existe una triada  $(\phi_k, y_k, x_k)$ . Si la propensión a responder  $\phi_k$  está fuertemente relacionada con  $x_k$  pero no así con  $y_k$ , es improbable que exista una relación considerable entre  $x_k$  y  $y_k$ . Tal situación no se corresponde con el requerimiento de una fuerte asociación entre  $x_k$  y  $y_k$  para controlar el sesgo por no respuesta y reducir la variancia del estimador  $\hat{t}_y^{CAL}$ .

Por otro lado, si de hecho sucede que  $x_k$  y  $\phi_k$  están fuertemente relacionados, no es inusual que también exista una asociación entre  $x_k$  y  $y_k$ . Luego, suponer que no existe relación entre  $\phi_k$  y  $y_k$  va contra la intuición y es improbable de ocurrir en una población finita (Särndal & Lundström, 2005).

En un estudio por muestreo, casi sin excepción, existe alguna relación entre la propensión a responder y alguna de las variables de interés. La forma que se ha encontrado de lidiar con esto en los estimadores es mediante la incorporación de información auxiliar.

Como define Lohr (1999), el concepto de no respuesta ignorarable significa que ‘un modelo puede explicar el mecanismo de no respuesta y la no respuesta puede ser ignorada luego de que el modelo la toma en cuenta, no que la no respuesta puede ser completamente ignorada y utilizar los métodos para datos completos’. El punto central reside en la capacidad que el modelo elegido tenga para explicar el mecanismo de no respuesta.

La posición defendida por Särndal y Lundström (2005) es que probablemente  $\phi_k$  y  $y_k$  siempre estén relacionados en algún punto, y que una explicación completa de  $\phi_k$  en términos de un vector  $x_k$  es prácticamente imposible. Todas las situaciones son no ignorables. Nunca se debe esperar contar con un vector auxiliar que corrija con total efectividad al sesgo que introduzca la no respuesta. En otras palabras, ningún modelo puede explicar por completo la variabilidad de la propensión a responder, sin importar cual sea el vector auxiliar utilizado. Es posible estimar y explotar la relación que pueda existir entre  $x_k$  y  $\phi_k$ , pero es utópico esperar una explicación completa de  $\phi_k$ .

### **3.5 Estimación de variancia**

La teoría presentada hasta esta sección trata principalmente a estimadores puntuales y los diversos intentos para mitigar el sesgo que puede introducir la no respuesta. Sin embargo, tal como se menciona en la introducción, en todo estudio por muestreo cada estimación puntual que se realice debe ser correspondida con una estimación de su error muestral. La estimación de variancia es necesaria por dos motivos: para indicar la precisión del estimador y para calcular intervalos de confianza

Mucha de la teoría desarrollada para los estudios por muestreo, y particularmente para la estimación de variancia, tiene que ver con diseños muestrales simples y estimadores lineales, mientras que las encuestas en la actualidad generalmente tienen que ver con diseños muestrales complejos y estimadores no lineales. Un estudio complejo presenta características como la estratificación, muestreo multietápico, probabilidades de selección desiguales, ajustes por no respuesta y/o por falta de cobertura, etc. La elección de un estimador de variancia apropiado no es sencilla. Esta debe considerar la precisión del estimador, tiempo, costo, simplicidad e incluso procedimientos administrativos.

El fenómeno de la no respuesta es central en el presente trabajo. Los elementos de la muestra se corresponden con una probabilidad de inclusión conocida y una probabilidad de responder desconocida. El problema de construir expresiones para la variancia no solo depende del diseño muestral, sino que también es influenciado por el método utilizado para corregir la no respuesta.

En la sección anterior se presentaron alternativas para incorporar información auxiliar con la finalidad de mitigar el sesgo por no respuesta, obtener estimaciones consistentes de variables auxiliares conocidas y, en lo posible, reducir la variabilidad del estimador. El problema que presentan estos ajustes de pesos muestrales es que producen estimadores no lineales. Esto dificulta la obtención de una expresión de la variancia del estimador y su consecuente estimación.

Una alternativa para aproximar la variancia de un estimador no lineal es el método de linealización por series de Taylor. Pero este requiere la derivación de una fórmula específica para cada estimador. En un estudio por muestreo a gran escala esto puede ser inviable porque suele haber demasiados parámetros de interés distintos y de variada complejidad. Un estudio más completo de este método se puede encontrar en Krewski y Rao (1981) y en Wolter (2007).

Otro método general para la estimación de variancia de estimadores no lineales son los métodos de replicación. La idea general es obtener una serie de submuestras (réplicas), y luego utilizar cada una de ellas para estimar al parámetro con el mismo procedimiento que en la muestra completa. La variancia luego es estimada a partir de la variabilidad entre las estimaciones que surgen de las replicaciones.

Uno de los métodos de replicación más populares es Bootstrap. Este fue presentado por Efron (1979) en el contexto de la estadística clásica, donde las observaciones se asumen independientes e igualmente distribuidas. En el contexto de estudios por muestreo, se han realizado diversos intentos para adaptar este método y utilizarlo en estimaciones de poblaciones finitas. Un listado exhaustivo y actualizado es presentado en Mashreghi, Haziza y Léger (2016).

El buen desempeño de Bootstrap en diseños sencillos, donde se tienen estimadores cerrados para la variancia, es uno de los factores que motiva la utilización de Bootstrap como método de replicación para obtener estimaciones de variancia en diseños más complejos, donde una estimación de variancia puede no estar disponible.

Rao y Wu (1988) proponen un método de estimación de variancia Bootstrap para diseños estratificados multietápicos que realiza un escalado en las observaciones previo a obtener las muestras de replicación y así una estimación valida de la variancia. Sin embargo, no puede

aplicarse a estimadores no suaves (como la mediana). Posteriormente, Rao, Wu y Yue (1992) sugieren una modificación al mismo que realiza un escalado sobre los pesos en vez de a las observaciones y permite aplicarse tanto a estimadores suaves como no suaves.

Sea  $\theta$  un parámetro de interés de una población finita y  $\hat{\theta}$  su estimador basado en toda la muestra. Dado un diseño estratificado multietápico, la obtención de los pesos Bootstrap y la estimación de la variancia de  $\hat{\theta}$  acorde a la propuesta de Rao, Wu y Yue (1992) consiste en repetir un número grande de veces  $B$ , generalmente entre 200 y 500: seleccionar una muestra simple al azar con reemplazo de  $m_h$  UPM de las  $n_h$  UPM en la muestra, independientemente para cada estrato  $h$ , con  $h = 1, \dots, H$ . Luego, el peso Bootstrap en la replicación  $b$  para la unidad  $k$  se calcula:

$$w_{hik}^* = \left[ \left\{ 1 - \sqrt{\frac{m_h}{n_h - 1}} \right\} + \sqrt{\frac{m_h}{n_h - 1}} \frac{n_h}{m_h} m_{hi}^* \right] w_{hik}, \quad (81)$$

donde  $m_{hi}^*$  el número de veces que la UPM  $i$  del estrato  $h$  es seleccionada en la muestra con reemplazo. Finalmente, se debe calcular el estimador de interés  $\hat{\theta}_b^*$ , la réplica  $b$  de  $\hat{\theta}$ , utilizando  $w_{hik}^*$  en vez de  $w_{hik}$  en la fórmula del estimador para el que se quiere obtener la variancia.

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

$$\hat{V}_{BS}(\hat{\theta}) = \frac{1}{B} \sum_{b=1}^B (\hat{\theta}_b^* - \hat{\theta})^2, \quad (82)$$

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

La gran ventaja de este método, y de los métodos de replicación en general, es que permiten reflejar implícitamente los efectos en la variancia de los ajustes por no respuesta y calibración. Para ello, cada conjunto de pesos Bootstrap debe ser modificado con los mismos procedimientos que se utilicen para la muestra original. Es decir, si los pesos originales son modificados mediante algún ajuste por no respuesta y/o calibración, esto mismo debe realizarse a cada conjunto de pesos replicados previo a utilizarlos para estimar el parámetro de interés y su variancia.

Además, 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. Así, este método facilita la obtención de intervalos de confianza. Valliant, Dever, y Kreuter (2018) listan en el capítulo 15 algunas alternativas para realizar tal tarea, ya que no existe un único procedimiento. Es una práctica común utilizar un muestreo de conglomerados con probabilidad proporcional al tamaño sin reemplazo para incrementar la eficiencia de los estimadores respecto de un muestreo con probabilidad proporcional al tamaño con reemplazo y también evitar la posibilidad de seleccionar un conglomerado más de una vez en la muestra.

Particularmente, el escalado de los pesos en (81) se corresponde con un diseño donde el muestreo de UPM es realizado con reemplazo. La utilidad de este método en la práctica se debe a que sobreestima la variancia verdadera para ciertos diseños muestrales y esto implica que en esos casos  $\hat{V}_{BS}(\hat{\theta})$  es un estimador conservador de la variancia.

En el caso de diseños que implican la selección de unidades con probabilidad proporcional al tamaño, Wolter (2007) demuestra para el caso particular  $\hat{\theta} = \hat{t}_y$ , que el estimador de variancia Bootstrap  $V_{BS}(\hat{\theta})$  incluye correctamente el componente de variancia dentro de cada UPM y que el sesgo de este estimador surge del componente de variancia entre UPM. Éste viene dado por:

$$B(V_{BS}(\hat{\theta})) = \sum_{h=1}^H \frac{n_h}{n_h - 1} \left( V_{pps} \left\{ \frac{1}{n_h} \sum_k^{n_h} \frac{Y_{hk}}{p_{hk}} \right\} - V_{\pi ps} \left\{ \frac{1}{n_h} \sum_k^{n_h} \frac{Y_{hk}}{\pi_{hk}} \right\} \right), \quad (83)$$

donde  $V_{pps}$  representa la variancia en un muestro con probabilidad proporcional al tamaño con reemplazo, y  $V_{\pi ps}$  representa a la variancia en un muestro con probabilidad proporcional al tamaño sin reemplazo. Luego, el estimador  $V_{BS}(\hat{\theta})$  tiene un sesgo positivo siempre que la variancia del estimador sea menor bajo muestro  $\pi ps$  que bajo muestro  $pps$ . Rao, Wu y Yue (1992) mencionan que, si bien se puede incurrir en una sobreestimación de la verdadera variancia, es probable que el sesgo relativo sea pequeño si la fracción de muestro en cada estrato es pequeña.

Matemáticamente, los estimadores de variancia para totales muestrales basados en diseños muestrales complejos donde la selección de las UPM es sin reemplazo son función de los factores de corrección por población finita (FPC), quienes representan a la proporción de la población

finita que no es incluida en la muestra, y las probabilidades de selección conjuntas. Mayores tamaños muestrales conducen a FPCs menores, cuyo efecto es una reducción en las estimaciones de variancia (en el caso de muestreo sin reemplazo). Por el contrario, tamaños muestrales pequeños relativos al tamaño de la población resultan en fracciones de muestreo muy pequeñas y un FPC cercano a la unidad. Estos suelen ser ignorados en diseños muestrales complejos.

Cuando se asume que la selección de las unidades en la primera etapa de selección es con reemplazo, también se está eliminando al factor de corrección por población finita de los estimadores de variancia. En una muestra multietápica, compleja, estratificada y cuyas unidades de primera etapa son seleccionadas sin reemplazo, las etapas posteriores de un muestreo por conglomerados contribuyen a la variancia de una estimación del total mediante una suma, a través de los estratos de primera etapa, de las fracciones de muestreo de cada estrato multiplicadas por la variancia adicional de las etapas posteriores (Cañette, 2010) y (Cochran, 1977, pp. 278-279). Luego, si las fracciones de muestreo en los estratos de primera etapa son lo suficientemente pequeñas, las contribuciones a la variancia que puedan tener las siguientes etapas de muestreo son depreciables y pueden ser ignoradas. Además, el sesgo del estimador de variancia que asume un muestreo de conglomerados (UPM) en una etapa con reemplazo, es depreciable también.

## **4. Aplicación**

### **4.1 Elaboración del marco muestral**

#### **4.1.1 Materiales**

El estudio por simulación realizado en esta tesina utiliza una población sintética de viviendas del Estado de Texas correspondiente el año 2010. La misma es uno de los componentes de una población sintética elaborada para todo Estados Unidos por RTI International que incluye personas, escuelas, instituciones, etc. (Wheaton, 2014). En el Estado de Texas, esta población está compuesta por 8.921.047 viviendas que albergan a 24.512.643 personas y están agrupadas a nivel de condado. Se cuenta con otras características referentes a cada vivienda listadas exhaustivamente en el Anexo 1.

Cada una de las viviendas en el listado se encuentra representada espacialmente mediante un punto dado por sus coordenadas en latitud y longitud. Complementariamente, se utiliza un conjunto de polígonos que se corresponden con los límites de las localidades de Texas efectivos al 1 de enero de 2018 (U.S. Census Bureau, 2018). Estos polígonos son utilizados para identificar la localidad a la que cada vivienda del listado pertenece.

Se utilizó al software de programación R en su versión 3.5.3 para realizar todos los cómputos necesarios. El enlace que conduce al repositorio con el código se encuentra en el Anexo 5.

#### **4.1.2 Asignación de localidad**

La asignación de viviendas a localidades, proceso llamado geocodificación inversa, fue realizada mediante la intersección punto-polígono. Una vivienda es asignada a un polígono si ésta está contenida dentro del último. Sin embargo, no todas las viviendas se encuentran ubicadas en regiones delimitadas por los polígonos provistos, por lo que este método no logra asignarlas completamente. Esto puede suceder porque puede haber viviendas que se encuentran ubicadas fuera del área determinada para cualquiera de las localidades, por el hecho de que la población de viviendas es sintética, y también porque se está trabajando con dos fuentes de datos distintas elaboradas en dos momentos temporales diferentes. En esta aplicación, el método de asignación

a través de la intersección punto-polígono permitió asignar localidad a 7,108,290 (79.68%) de las viviendas.

Para el conjunto de viviendas que no fueron asignadas a ninguna localidad mediante la geocodificación inversa, se calculó la distancia punto-polígono entre cada una de las viviendas no asignadas y cada una de las localidades dentro del condado al que pertenecen<sup>3</sup>. Finalmente, se asignó cada vivienda a la localidad más cercana.

#### 4.1.3 Construcción de áreas de muestreo

El nivel de agregación más bajo de unidades elementales son las áreas de muestreo intra-localidad conformadas por viviendas. Estas se corresponden con las USM del diseño muestral en la simulación. En general, resulta deseable que el tamaño de las áreas seleccionadas para realizar el muestreo final de viviendas sea uniforme. En este trabajo se optó por elaborar áreas cuyo tamaño sea de 200 viviendas, que está sujeto a la cantidad de viviendas en cada localidad. La elección de este tamaño para las áreas se basa en que este es el tamaño generalmente utilizado en los marcos del Instituto Nacional de Estadística y Censos de Argentina (INDEC).

Supongamos que se cuenta con un conjunto de  $n$  unidades georreferenciadas, y se desea distribuirlos en grupos de igual tamaño,  $m$ . La cantidad de grupos está dada por:

$$g = \max\{x \in \mathbb{N} \mid x \times m \leq n\}. \quad (84)$$

Esto significa que se obtienen grupos de igual tamaño solo cuando el resto de la división  $\frac{n}{g}$  es nulo.

Caso contrario, es decir  $n \bmod g \neq 0$ , esta cantidad es distribuida lo más uniformemente posible entre los  $g$  grupos.

---

<sup>3</sup> Se utilizaron los paquetes *lwgeom* versión 0.1-5 y *sf* versión 0.7-2

Si el cociente entre el número de unidades a agrupar sobre la cantidad de grupos  $\frac{n}{g}$  es entero, cada grupo está compuesto por  $m + \frac{n}{g}$  unidades. Caso contrario, hay  $g_1$  grupos cuyo tamaño es  $m + \left\lceil \frac{n}{g} \right\rceil$ , y  $g_2$  grupos cuyo tamaño es  $m + \left\lceil \frac{n}{g} \right\rceil + 1$ .

Los valores de  $g_1$  y  $g_2$  son aquellos que satisfacen la siguiente ecuación

$$g_1 \left\lceil \frac{n}{g} \right\rceil + g_2 \left( \left\lceil \frac{n}{g} \right\rceil + 1 \right) = n \quad (85)$$

lo que conduce a

$$g_1 = g \left( \left\lceil \frac{n}{g} \right\rceil + 1 \right) - n \wedge g_2 = g - g_1 \quad (86)$$

El algoritmo utilizado para la elaboración de áreas de muestreo es una adaptación del método iterativo del vecino más cercano (Monlong, 2018). Este permite obtener clusters que respetan tamaños asignados, pero su carga computacional se vuelve impracticable a medida que aumenta el tamaño del conjunto a agrupar. Por tal motivo, en aquellas localidades con más de 15000 viviendas se realizó una desagregación previa en grupos mediante el algoritmo de k-medias y se aplicó el método iterativo del vecino más cercano en cada una de estas grandes áreas. Además, en aquellas localidades con 300 o menos viviendas, el tamaño deseado para las áreas fue de 150 viviendas en vez de 200. En estos casos mantener un tamaño de 200 implicaría que una de las áreas tenga un tamaño muy inferior a la restante. De esta forma, se busca evitar grandes disparidades entre el tamaño de las áreas dentro de una misma localidad.

#### **4.1.4 Unidades Primarias de Muestreo**

De acuerdo con el *US Bureau of Statistics* las Unidades Primarias de Muestreo se corresponden con un condado, siempre y cuando la población de este sea de 7500 habitantes o mayor, o con un conjunto de condados contiguos cuando alguno de estos tenga una población menor a 7500 habitantes (U.S. Census Bureau, 2006).

Con la asistencia del software Tableau (versión 2018.3) se creó un mapa de condados que permitió conformar de forma manual las UPM. En el caso que hubiera varios condados contiguos cuya

población no sea superior a 7500 habitantes, estos fueron unidos en una misma UPM. En cambio, cuando los condados de estas características no estuvieran contiguos, se los anexó a un condado ya existente. Los 254 condados del estado de Texas conformaron un total de 206 Unidades Primarias de Muestreo. Un listado exhaustivo de ambos se encuentra en el Anexo 2.

#### **4.1.5 Filtros operativos y estratificación**

Considerando un hipotético operativo de campo y el diseño muestral a utilizar, se aplicaron otros dos filtros previos a la conformación del marco muestral definitivo. En primer lugar, se eliminaron todas aquellas áreas de muestreo con menos de 50 viviendas. Esto, en realidad está eliminando localidades que solo contienen un área y ésta es demasiado pequeña como para considerarla. En segundo lugar, se eliminaron las UPM con menos de 10 áreas. Luego de estos filtros, la población del marco resultante quedó conformada por 8.087.602 viviendas que albergan a 21.349.048 personas, y la cantidad de UPM se redujo de 206 a 180.

El último paso en la construcción del marco muestral fue la estratificación de las Unidades Primarias de Muestreo. Se conformaron 4 estratos según la cantidad de habitantes, donde uno de ellos agrupa a las UPM que contienen a los aglomerados urbanos más relevantes del Estado y es un estrato de inclusión forzosa en diseño utilizado en el estudio por simulación. La definición y composición de los estratos se encuentra en el Anexo 3.

### **4.2 Características del estudio de simulación**

#### **4.2.1 Diseño muestral**

En esta tesina se propone un hipotético estudio por muestreo que tiene por objetivo estimar el ingreso medio de los hogares del Estado de Texas. Para ello, se selecciona una muestra probabilística de viviendas. A fin de emular los diseños frecuentemente utilizados para encuestas a hogares en Argentina, se selecciona la muestra mediante un diseño por conglomerados estratificado en tres etapas.

- i. **Etapa 1:** Seleccionar una muestra estratificada  $s_l$  de Unidades Primarias de Muestreo con probabilidad proporcional a la cantidad de viviendas mediante el método

sistemático de Madow (Madow, 1949). Aquellos aglomerados urbanos más importantes conforman un estrato de inclusión forzosa.

- ii. **Etapa 2:** Para cada UPM  $\in s_l$  se selecciona una muestra  $s_{ll_i}$  de áreas de muestreo mediante una selección simple al azar sin reemplazo.
- iii. **Etapa 3:** Dentro de cada USM  $\in s_{ll_i}$  se selecciona una muestra  $s_{iq}$  de viviendas mediante muestreo simple al azar sin reemplazo.

A continuación, se brinda un detalle de la cantidad de unidades seleccionadas en cada etapa de muestreo según estrato de UPM.

**Tabla 1:** Cantidad de unidades seleccionadas por etapa y por 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 para este estudio en término de unidades elementales es de  $n = 31300$  viviendas.

Las fracciones de muestreo para los estratos de UPM fueron  $f_1 = \frac{30}{104} = 0.29$ ,  $f_2 = \frac{20}{53} = 0.38$  y  $f_3 = \frac{8}{16} = 0.50$  para los primeros tres estratos, respectivamente. El estrato de inclusión forzosa contiene 7 Unidades Primarias de Muestreo. Las fracciones de muestreo dentro de cada una de esas UPM fueron  $f_{4,1} = 0.083$ ,  $f_{4,2} = 0.18$ ,  $f_{4,3} = 0.058$ ,  $f_{4,4} = 0.199$ ,  $f_{4,5} = 0.035$ ,  $f_{4,6} = 0.077$ , y  $f_{4,7} = 0.128$ .

#### **4.2.2 Asignación de valores faltantes**

Dada una muestra obtenida según el diseño presentado en la sección anterior y en base a la información auxiliar disponible, se asigna un porcentaje de valores faltantes de forma que se emulen las mecánicas de no respuesta MCAR, MAR, y NMAR introducidas en la sección **3.2.2**.

- i. **Missing completely at random (MCAR):** Para simular este escenario se le asigna a cada unidad la misma probabilidad de responder, igual a la tasa de respuesta deseada.
- ii. **Missing at Random (MAR):** En este caso cada unidad muestreada tiene una probabilidad de responder relacionada a la cantidad de personas por hogar. La idea subyacente al utilizar esta variable es que, a menor cantidad de personas en el hogar, menor chance de responder.
- iii. **Not missing at random (NMAR):** Este escenario se emula asignando a cada vivienda una probabilidad de responder que depende del ingreso del hogar, que es la variable de interés. Aquí, la probabilidad de responder tiene una relación negativa con el ingreso de la vivienda y está dada por un modelo logístico.

Todos los casos mencionados anteriormente son simulados con dos porcentajes de respuesta, 60% y 80%.

#### **4.2.3 Ajuste de pesos muestrales**

En cada uno de los escenarios de no respuesta se prueban diversos ajustes de pesos previo a realizar las estimaciones lo que permite evaluar su capacidad de mitigar el sesgo de no respuesta y mejorar la precisión de los estimadores. Sea  $w_k^*$  el peso definitivo para la unidad  $k$ , con  $k \in s_r$ .

- i. Sin ajuste: Se utilizan los pesos de diseño, es decir,  $w_k^* = d_k$  y el estimador recibe el nombre  $\hat{y}_{HJ_r}$ .
- ii. Ajuste por no respuesta 1: Se ajustan los pesos según el método de Class Weighting Adjustment de la sección **3.4.3** utilizando a la cantidad de personas por hogar como variable de clasificación. Aquí  $w_k^* = d_k \hat{\phi}_k^{-1}$  y el estimador es  $\hat{y}_{NR1}$ .

- iii. Ajuste por no respuesta 2: Se ajustan los pesos según el método de Class Weighting Adjustments de la sección **3.4.3** utilizando al estrato de UPM como variable de clasificación. Aquí  $w_k^* = d_k \hat{\phi}_k^{-1}$  y el estimador es  $\hat{y}_{NR2}$ .
- iv. Calibración<sup>4</sup>: Los pesos de diseño son calibrados a conteos poblacionales marginales de personas por intervalos de edad y sexo. Este ajuste se corresponde con lo realizado en una gran cantidad de Institutos de Estadística y es el introducido en la sección **3.4.1.2**. En este caso  $w_k^* = d_k F(\mathbf{x}'_k \boldsymbol{\lambda})$  y el estimador es  $\hat{y}_{CAL}$ .
- v. Ajuste por no respuesta 1 + Calibración: En primer lugar, los pesos de diseño son ajustados de la misma forma que en (ii) y luego son calibrados de la misma forma que en (iv). Es decir,  $w_k^* = d_k \hat{\phi}_k^{-1} F(\mathbf{x}'_k \boldsymbol{\lambda})$  y el estimador es  $\hat{y}_{NR1}$
- vi. Ajuste por no respuesta 2 + Calibración: En primer lugar, los pesos de diseño son ajustados de la misma forma que en (iii) y luego son calibrados de la misma forma que en (iv). Es decir,  $w_k^* = d_k \hat{\phi}_k^{-1} F(\mathbf{x}'_k \boldsymbol{\lambda})$  y el estimador es  $\hat{y}_{NR2}$ <sup>5</sup>

Vale la pena realizar la siguiente aclaración sobre el ajuste por no respuesta según el método de *Class Weighting Adjustment*. En un primer lugar este se realiza con la variable que se sabe que está relacionada a la probabilidad de responder en el mecanismo de respuesta MAR: cantidad de personas en el hogar. No se espera que esto sea un fiel reflejo de lo que ocurre en la práctica, ya que no solo que es imposible saber cuáles son las variables relacionadas a la propensión a responder, sino que, de saberlo, es improbable tener acceso a ellas. Particularmente, es improbable que se pueda conocer la cantidad de personas en aquellos hogares que no responden. Sin embargo, en este caso, se busca evaluar las propiedades de este ajuste en un escenario *ideal*. Luego, también se realiza este tipo de ajuste utilizando como clase de ajuste al estrato de UPM, que se asemeja más al tipo de ajuste que se puede realizar en la práctica.

---

<sup>4</sup> La calibración de pesos se realizó en todos los casos con la función de distancia logit y el intervalo para la razón entre los pesos calibrados y los pesos de diseño fue [0,4; 3,85].

<sup>5</sup> El ajuste de pesos en (iv) se corresponde con lo que Haziza y Lesage (2016) llaman ajuste en un paso, mientras que los ajustes (v) y (vi) son llamados ajustes en dos pasos.

#### 4.2.4 Estimadores utilizados

El parámetro de interés a estimar en este estudio por muestreo es el ingreso medio por hogar del Estado de Texas, representado por  $\bar{y}_U$ . Un estimador ampliamente utilizado es el estimador de Hájek  $\hat{y}_{HJ}$  cuya fórmula es presentada en (20). En esta aplicación el estimador es calculado sobre el conjunto de las unidades que responden y sus pesos vienen dados por alguno de los ajustes anteriormente planteados:

$$\hat{y}_{HJ_r} = \frac{\sum_{k \in S_r} w_k^* y_k}{\sum_{k \in S_r} w_k^*}, \quad (87)$$

donde  $w_k^*$  representa al peso de la unidad  $k$ , con  $k \in s_r$ , correspondiente al método de ajuste utilizado.

Se utiliza el método Bootstrap de Rao, Wu y Yue para estimar la variancia de (87) en todos los escenarios dados por la combinación de la mecánica de no respuesta y la metodología de ajuste de corrección de pesos muestrales.

En el estrato de inclusión forzosa la selección de UPM contemplada en el método Bootstrap debe ser modificada. En este caso, en vez de seleccionar al azar y con reemplazo a las UPM para calcular el factor de ajuste de los pesos, se seleccionan a las USM dentro de cada UPM y estas últimas son las que conforman los estratos en términos del procedimiento Bootstrap.

#### 4.2.5 Evaluación de los estimadores

En el estudio por simulación se obtienen 10,000 muestras según el diseño presentado y en cada una de ellas se obtienen subconjuntos de unidades que responden bajo cada una de las condiciones dadas por la combinación de tasa y mecanismo de respuesta. Luego, y utilizando cada una de las muestras, se estima el ingreso medio con cada uno de los 6 conjuntos de pesos anteriormente mencionados. Además, con las primeras 1,000 muestras obtenidas se estima la variancia según el método de Rao, Wu y Yue con  $B = 200$ .

Sea  $\hat{y}$  uno de los seis estimadores utilizados en este trabajo. Para cada conjunto de unidades que responden  $j = 1, 2, \dots, 10000$  se estima  $\bar{y}$  y con los primeros 1000 también se calcula  $\hat{V}(\bar{y})$ . Las

medidas resúmenes calculadas para evaluar la bondad de los estimadores puntuales y de dispersión son las siguientes

- i. Sesgo relativo del estimador puntual:

$$RB_{SIM}(\hat{y}) = \frac{E_{SIM}[\hat{y}] - \bar{y}_U}{\bar{y}_U}, \quad (88)$$

donde

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

es la esperanza simulada del estimador puntual, e  $\bar{y}_U$  es el ingreso medio poblacional de los hogares en el Estado de Texas.

- ii. Sesgo relativo del estimador de variancia:

$$RB_{SIM}[\hat{V}(\hat{y})] = \frac{E_{SIM}[\hat{V}(\hat{y})] - V_{SIM}(\hat{y})}{V_{SIM}(\hat{y})}, \quad (89)$$

donde

$$E_{SIM}[\hat{V}(\hat{y})] = \frac{1}{1000} \sum_{j=1}^{1000} \hat{V}(\hat{y})_{(j)},$$

es la esperanza simulada del estimador de variancia,

$$V_{SIM}(\hat{y}) = \frac{1}{10000} \sum_{j=1}^{10000} [\hat{y}_{(j)} - E_{SIM}[\hat{y}]]^2,$$

es la variancia simulada del estimador puntual.

- iii. Error cuadrático medio del estimador puntual:

$$ECM_{SIM}(\hat{y}) = (E_{SIM}[\hat{y}] - \bar{y}_U)^2 + V_{SIM}(\hat{y}) = B_{SIM}(\hat{y})^2 + V_{SIM}(\hat{y}). \quad (90)$$

Además, si bien no se compara el estimador de variancia Bootstrap con otros estimadores de variancia, es interesante calcular la variancia simulada del estimador de variancia para evaluar si la dispersión del estimador de variancia difiere para los ajustes de pesos propuestos, o sea:

$$V_{SIM}[\hat{V}(\hat{y})] = \frac{1}{999} \sum_{j=1}^{1000} \left[ \hat{V}(\hat{y})_{(j)} - E_{SIM}[\hat{V}(\hat{y})] \right]^2. \quad (91)$$

#### 4.2.6 Resultados

A continuación, se presentan las tablas 2 y 3 que resumen los resultados de las simulaciones realizadas para ambas tasas de respuesta, junto con gráficos que reflejan el comportamiento del sesgo relativo y de la variancia del estimador puntual. En el Anexo 4 se encuentra un conjunto más exhaustivo de gráficos que resumen las distribuciones empíricas de los estimadores utilizados y permiten realizar una comparación visual de las medidas de interés.

**Tabla 2:** Tabla resumen de resultados de la simulación. Tasa de respuesta = 80%

Mecanismo de Respuesta	Estimador Puntual	$RB_{SIM}(\hat{y}) \times 10^2$	$V_{SIM}(\hat{y}) \times 10^{-6}$	$ECM_{SIM}(\hat{y}) \times 10^{-6}$	$RB_{SIM}[\hat{V}(\hat{y})] \times 10^2$	$V_{SIM}[\hat{V}(\hat{y})] \times 10^{-11}$
MCAR	$\hat{y}_{HJ_r}$	0.005%	2.213	2.213	31.37%	5.971
	$\hat{y}_{NR1}$	0.005%	2.209	2.209	31.47%	5.989
	$\hat{y}_{NR2}$	0.005%	2.213	2.213	31.38%	5.952
	$\hat{y}_{CAL}$	1.99%	1.902	3.735	26.23%	3.911
	$\hat{y}_{NR1+CAL}$	0.009%	1.766	1.766	27.85%	3.394
	$\hat{y}_{NR2+CAL}$	0.009%	1.770	1.766	27.70%	3.377
MAR	$\hat{y}_{HJ_r}$	1.69%	2.416	3.744	29.03%	6.865
	$\hat{y}_{NR1}$	0.01%	2.207	2.208	31.22%	5.885
	$\hat{y}_{NR2}$	1.68%	2.392	3.691	28.18%	6.475
	$\hat{y}_{CAL}$	3.88%	2.036	8.989	24.73%	4.423
	$\hat{y}_{NR1+CAL}$	0.01%	1.761	1.762	27.83%	3.327
	$\hat{y}_{NR2+CAL}$	1.96%	1.869	3.636	26.20%	3.585
NMAR	$\hat{y}_{HJ_r}$	-17.99%	1.188	15.088	30.91%	1.490
	$\hat{y}_{NR1}$	-17.29%	1.265	13.954	30.15%	1.684
	$\hat{y}_{NR2}$	-17.92%	1.203	14.961	31.57%	1.540
	$\hat{y}_{CAL}$	-15.54%	1.078	11.272	26.79%	1.123
	$\hat{y}_{NR1+CAL}$	-16.47%	1.030	12.675	27.94%	1.024
	$\hat{y}_{NR2+CAL}$	-17.07%	0.983	13.564	29.16%	0.935

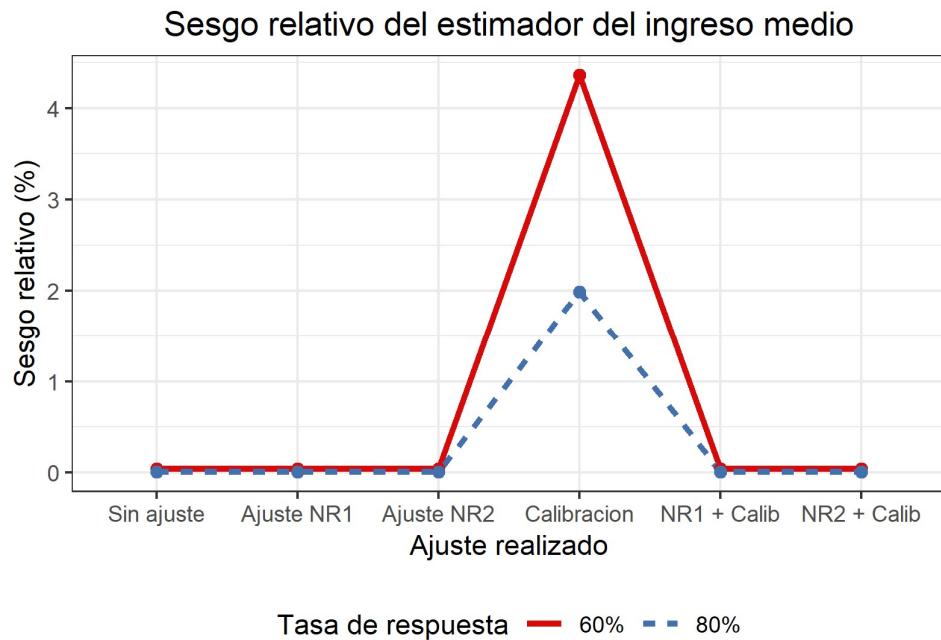
En un primer lugar, se puede observar que, en el caso que la no respuesta sea un fenómeno MCAR, el sesgo relativo del estimador de la media es prácticamente nulo en todos los casos, con excepción del estimador de calibración a un paso,  $\hat{y}_{CAL}$ , donde el estimador tiene un sesgo relativo de 1.99% cuando la tasa de respuesta es 80% y 4.36% cuando esta es del 60%.

**Tabla 3:** Tabla resumen de resultados de la simulación. Tasa de respuesta = 60%

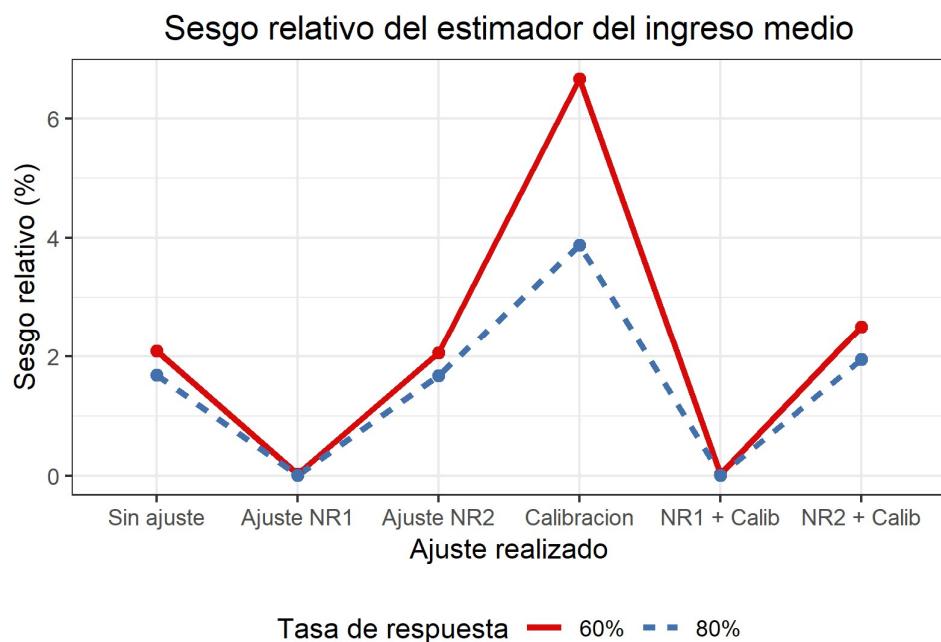
Mecanismo de Respuesta	Estimador Puntual	$RB_{SIM}(\hat{y}) \times 10^2$	$V_{SIM}(\hat{y}) \times 10^{-6}$	$ECM_{SIM}(\hat{y}) \times 10^{-6}$	$RB_{SIM}[\hat{V}(\hat{y})] \times 10^2$	$V_{SIM}[\hat{V}(\hat{y})] \times 10^{-11}$
MCAR	$\hat{y}_{HJ_r}$	0.004	2.300	2.301	30.68	6.443
	$\hat{y}_{NR1}$	0.004	2.293	2.295	30.80	6.433
	$\hat{y}_{NR2}$	0.40	2.299	2.300	30.80	6.402
	$\hat{y}_{CAL}$	4.36	2.139	10.927	24.54	4.972
	$\hat{y}_{NR1+CAL}$	0.04	1.846	1.847	27.04	3.685
	$\hat{y}_{NR2+CAL}$	0.04	1.850	1.851	27.11	3.675
MAR	$\hat{y}_{HJ_r}$	2.10	2.551	4.581	28.57	8.044
	$\hat{y}_{NR1}$	0.03	2.285	2.286	30.60	6.565
	$\hat{y}_{NR2}$	2.07	2.518	4.499	27.56	7.448
	$\hat{y}_{CAL}$	6.67	2.319	22.877	23.10	6.131
	$\hat{y}_{NR1+CAL}$	0.04	1.835	1.836	27.16	3.782
	$\hat{y}_{NR2+CAL}$	2.51	1.975	4.893	25.81	4.211
NMAR	$\hat{y}_{HJ_r}$	-26.92	0.896	33.594	29.05	0.777
	$\hat{y}_{NR1}$	-25.83	0.995	30.958	27.94	0.959
	$\hat{y}_{NR2}$	-26.82	0.918	33.350	29.80	0.834
	$\hat{y}_{CAL}$	-22.53	0.908	23.564	23.94	0.776
	$\hat{y}_{NR1+CAL}$	-24.61	0.827	28.083	26.29	0.625
	$\hat{y}_{NR2+CAL}$	-25.54	0.770	30.226	27.66	0.544

Por otro lado, cuando el mecanismo de no respuesta es MAR, solo aquellos estimadores que se valen de pesos corregidos con un modelo que utiliza a la variable auxiliar que está relacionada con la propensión a responder logran mitigar el sesgo que introduce la no respuesta. En este caso, nuevamente el estimador de calibración a un paso tiene un sesgo mayor que el estimador que utiliza el conjunto de pesos originales, para ambas tasas de respuesta.

**Gráfico 1:** Sesgo relativo del estimador del ingreso medio según ajuste y tasa de respuesta – MCAR.

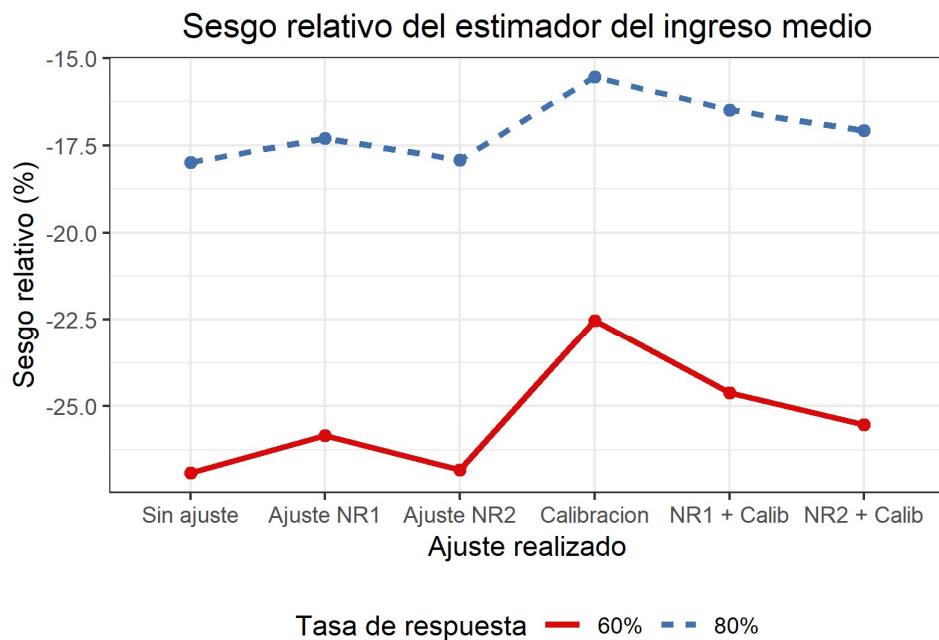


**Gráfico 2:** Sesgo relativo del estimador del ingreso medio según ajuste y tasa de respuesta – MAR.



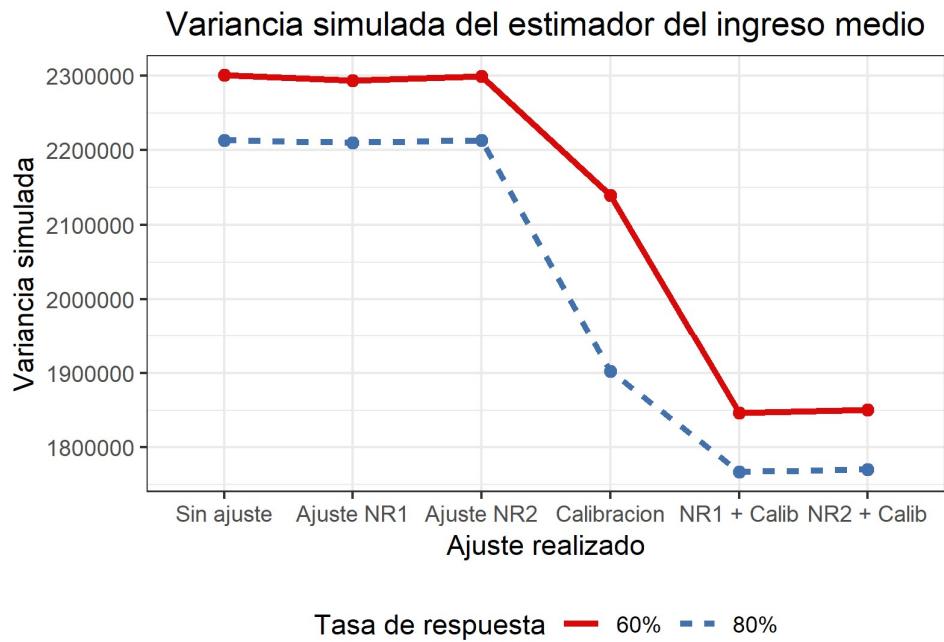
Si la probabilidad de responder está relacionada a la variable en estudio (NMAR), se observa que la no respuesta introduce sesgo en todos los casos, sin importar que conjunto de pesos se utilice. Además, a menor tasa de respuesta, mayor es la magnitud del sesgo. El sesgo relativo del estimador puntual está entre -15% y -18% cuando la tasa de respuesta es del 80% y entre -22 y -27% cuando la tasa de respuesta es del 60%. En todos estos casos, el sesgo es negativo ya que la probabilidad de responder es inversamente proporcional al ingreso del hogar.

**Gráfico 3:** Sesgo relativo del estimador del ingreso medio según ajuste y tasa de respuesta – NMAR.

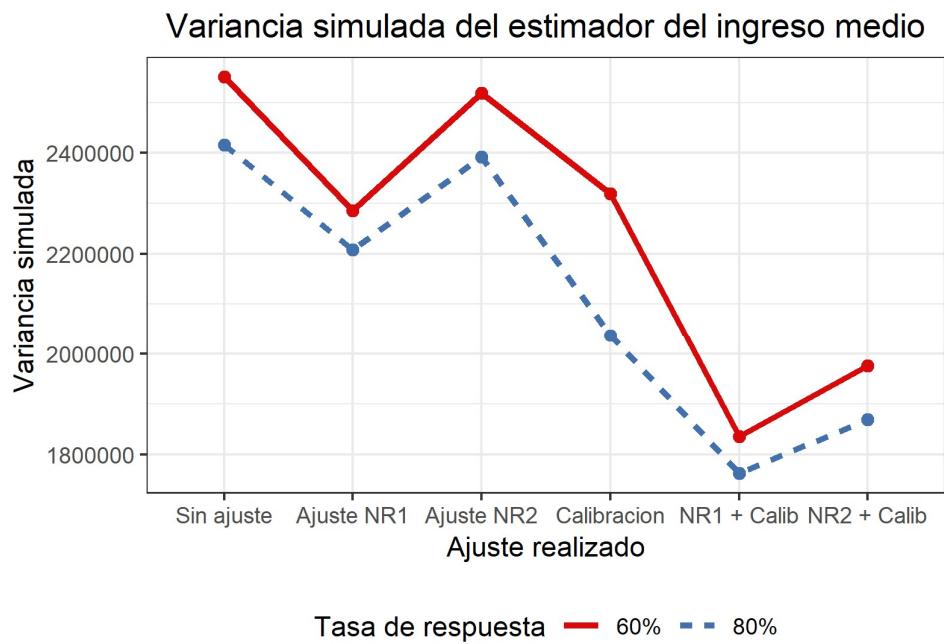


En todos los escenarios de simulación se puede observar que, independientemente del sesgo relativo, los estimadores que utilizan pesos calibrados presentan una reducción en la variancia del estimador respecto al resto. Esto se evidencia particularmente en el caso que se realiza un ajuste de pesos a dos pasos.

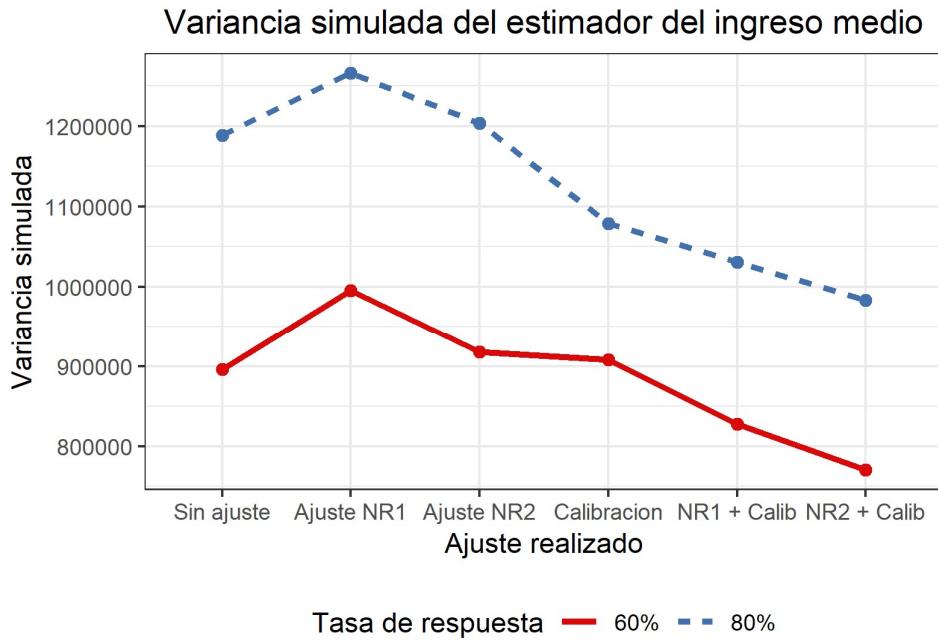
**Gráfico 4:** Variancia simulada del estimador del ingreso medio según ajuste y tasa de respuesta – MCAR.



**Gráfico 5:** Variancia simulada del estimador del ingreso medio según ajuste y tasa de respuesta – MAR.



**Gráfico 6:** Variancia simulada del estimador del ingreso medio según ajuste y tasa de respuesta – NMAR.



Considerando todos los escenarios y ajustes de pesos planteados, se observa que el estimador de la media presenta su menor error cuadrático medio cuando se utiliza al conjunto de pesos ajustado en dos pasos que se vale de la variable auxiliar relacionada a la propensión a responder para el modelo de ajuste,  $\hat{y}_{NR1+c}$ .

Si bien la menor variancia simulada se observa en el caso que la respuesta sea un fenómeno NMAR, los estimadores presentan los valores más altos del ECM como consecuencia del sesgo. La disminución en la variabilidad de los estimadores bajo este fenómeno de respuesta se debe a que la menor probabilidad de responder para hogares de altos ingresos genera un truncamiento en la cola derecha de la distribución del ingreso.

Por otro lado, el sesgo relativo del estimador de variancia es positivo y no depreciable en todos los escenarios de no respuesta y para todos los conjuntos de pesos utilizados. En todos los casos el sesgo relativo mencionado es mayor al 20%, resultando en estimadores muy conservadores.

Dado un mecanismo y tasa de respuesta, se puede observar que el sesgo relativo del estimador de variancia es en general un poco menor para aquellos estimadores que utilizan pesos calibrados:

$$\hat{y}_{CAL}, \hat{y}_{NR1+CAL} \text{ y } \hat{y}_{NR2+C}$$

Si bien es preferible utilizar estimadores conservadores de la variancia, la magnitud observada en el sesgo relativo del estimador de variancia es demasiado alta. Esto sugiere que en el futuro se evalúen otros estimadores que mejoren esta situación.

Finalmente, se observa que, dado un mecanismo de respuesta, la dispersión del estimador de variancia es menor para aquellos estimadores que utilizan pesos calibrados. Esta disminución en la dispersión se acentúa en aquellos estimadores que ajustan sus pesos por no respuesta previo a realizar la calibración.

Cuando la respuesta es un fenómeno MCAR o MAR, a pesar de esta reducción en la variancia, el error cuadrático medio de  $\hat{y}_{CAL}$  es mayor que todos los demás como consecuencia de su mayor sesgo. Este incremento en el error cuadrático medio respecto de los otros estimadores se acentúa cuando la tasa de respuesta es menor. Por ejemplo, crece de  $3.735 \times 10^6$  a  $10.927 \times 10^6$  cuando la respuesta es MCAR y de  $8.989 \times 10^6$  a  $22.877 \times 10^6$  cuando la respuesta es MAR.

## **5. Conclusiones**

En esta tesina se presentaron los mecanismos subyacentes que pueden dar origen a la no respuesta en un estudio por muestreo y se describieron cuáles son las posibles consecuencias de ésta en los estimadores a utilizar. Principalmente, se describió como este fenómeno puede introducir sesgo en las estimaciones.

Por otro lado, se mencionaron diversos ajustes que se pueden realizar sobre los pesos muestrales con el fin de mitigar el sesgo y en la medida de lo posible mejorar la precisión de los estimadores. Dado su frecuente uso en los Institutos de Estadística de diversos países, se introdujo al enfoque de calibración como un procedimiento sistemático para la utilización de información auxiliar y se describieron sus propiedades como método de ajuste por no respuesta. Se estudiaron las condiciones en las cuales ayuda a reducir el sesgo de los estimadores y mejorar su precisión. Al mismo tiempo, se nota que la bondad de este método está sujeta a la calidad de la información auxiliar disponible y la idoneidad de la función de distancia utilizada para relacionarla con la probabilidad de respuesta.

Se remarcó que no es factible obtener un modelo para la no respuesta que logre mitigar totalmente el sesgo que ésta introduce. Solo en escenarios muy simples se puede lograr una disminución casi total del sesgo del estimador puntual. Sin embargo, esto requiere conocer cuáles son las variables relacionadas al fenómeno de la no respuesta y tener acceso a ellas, una tarea imposible en la práctica.

En esta tesina también se observó que la tarea de obtener estimadores de variancia se dificulta a medida que el diseño muestral utilizado se vuelve más complejo. El ajuste por no respuesta, necesario ante la presencia de unidades que no responden, dificulta aún más la obtención de estimadores de variancia porque resulta en estimadores no lineales. Se propuso al estimador Bootstrap de Rao, Wu y Yue (1992) como un método general para estimar la variancia en el contexto de estudios por muestreo a gran escala. Este permite incorporar en la estimación de variancia los ajustes realizados a los pesos del diseño siguiendo el mismo procedimiento empleado en cada uno de los pesos correspondientes a las réplicas.

A través de un estudio por simulación se buscó evaluar el desempeño de diversos estimadores puntuales, dados por los diferentes ajustes por no respuesta descriptos, y del método Bootstrap mencionado. Para ello se replicó un estudio por muestreo a gran escala basado en un diseño muestral complejo junto a diversos escenarios de no respuesta. Esto se realizó sobre una población sintética para la cual se construyeron unidades muestrales similares a las que se utilizan en encuestas a gran escala en el país.

En un primer lugar se pudo observar que, si el mecanismo de respuesta es MCAR, la no respuesta no introduce sesgo en el estimador puntual. Luego, se observó que en el caso que la probabilidad de responder está relacionada con una variable auxiliar, solo aquellos estimadores que se valen de pesos corregidos con un modelo que utiliza a esta variable auxiliar logran mitigar el sesgo introducido. Finalmente, si la probabilidad de respuesta está exclusivamente relacionada a la variable en estudio, el sesgo que introduce la no respuesta es inversamente proporcional a la tasa de respuesta.

En general, en aquellos casos que la no respuesta no es un mecanismo MCAR, o no se utiliza un modelo que se valga de las variables que se relacionan a probabilidad de responder bajo un mecanismo, la magnitud del sesgo aumenta conforme se reduce la tasa de respuesta.

Los resultados de la simulación muestran que la utilización del estimador que solo contempla la calibración de los pesos,  $\hat{y}_{CAL}$ , logra reducir la variabilidad del estimador puntual y del estimador de variancia, pero puede introducir sesgo en la estimación puntual. Haziza y Lesage (2016) realizan un estudio más profundo y mencionan que una de las posibles causas es la utilización de una función de distancia errónea. Sin embargo, en el ámbito de este trabajo no se puede aseverar tal conclusión. La evaluación de las propiedades del estimador de calibración ante diferentes funciones de distancia excede al presente trabajo, y constituye materia de estudios futuros.

Por otro lado, el sesgo positivo hallado en las estimaciones de variancia indica que el estimador Bootstrap estudiado produce sistemáticamente estimaciones conservadoras bajo las condiciones utilizadas, lo que puede ser considerado positivo si se piensa que esto previene de intervalos de

confianza exageradamente estrechos para los parámetros de interés. Sin embargo, la magnitud del sesgo relativo observado es demasiado elevada. Esto sugiere que se debe seguir investigando el comportamiento del estimador utilizado en diferentes contextos y evaluar alternativas que reduzcan su sesgo, o bien proponer otros estimadores.

Uno de los posibles motivos del alto sesgo es el hecho de que el diseño utilizado es más eficiente en términos de variancia que un diseño que realiza una selección con reemplazo y las fracciones de muestreo en los estratos de primera etapa no son depreciables. Mach, Dumais, y Robidou (2005) realizan un estudio por simulación y obtienen resultados similares en cuanto al sesgo relativo cuando la fracción de muestreo no es depreciable. Estos autores sugieren que el sesgo también podría estar relacionado con la fracción de muestreo en la segunda etapa.

Exceptuando la gran magnitud del sesgo del estimador puntual cuando la no respuesta es un fenómeno NMAR, el ajuste por no respuesta a dos pasos es el método que en general arroja mejores resultados. Esto se observa particularmente para el estimador  $\hat{y}_{NR1+c}$ . Por un lado, en los casos que la no respuesta introduce sesgo en el estimador puntual, se observa una reducción del mismo. Luego, independientemente de que la no respuesta introduzca sesgo o no, este estimador presenta una variancia menor que los demás. Por último, no solo se observa que bajo este ajuste el estimador de variancia presenta un sesgo levemente menor, sino que también se nota una reducción en su dispersión.

Del párrafo anterior se desprende que los resultados hallados en el estudio por simulación se corresponden con la recomendación de Haziza y Lesage (2016) respecto a la utilización de un ajuste por no respuesta en dos pasos por sobre aquel que contempla un solo paso.

Finalmente, es importante remarcar que más allá de la metodología que se utilice para ajustar por no respuesta, la bondad de los resultados depende de la calidad de la información auxiliar disponible. En un estudio por muestreo a gran escala los parámetros de interés suelen ser de una complejidad mucho mayor al expuesto en este trabajo. Lo mismo sucede con las relaciones entre las variables auxiliares, la probabilidad de responder y las variables de interés. En tales casos, es

utópico pensar que se puede desarrollar un modelo que permita explicar la probabilidad de respuesta en su totalidad. Luego, en el mejor de los casos, se podrá conseguir una reducción sustancial del sesgo que introduce la no respuesta, pero no una eliminación total del mismo.

## Bibliografía

- Cañette, I. (2010). Analysis of Complex Survey Data in Stata. *Mexican Stata Users Group Meeting.*
- Capretto, T., & Marí, G. (2018). Método de agrupamiento geoespacial para la segmentación de una población de viviendas. *Investigaciones en la Facultad de Ciencias Económicas y Estadísticas - UNR.*
- Cobben, F. (2009). *Nonresponse in sample surveys: methods for analysis and adjustment.* The Hague: Statistics Netherlands.
- Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.). John Wiley & Sons, Inc.
- Deville, J.-C., Särndal, C.-E., & Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association*, 88(423), 1013-1020.
- Durbin, J. (1953). Some Results in Sampling Theory when the Units are Selected with Unequal Probabilities. *Journal of the Royal Statistical Society. Series B (Methodological)*, 15(2), 262-269.
- Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1), 1-26.
- Groves, R. M. (1989). *Survey Errors and Survey Costs.*
- Groves, R. M., & Couper, M. P. (1998). *Nonresponse in Household Interview Surveys.* John Wiley & Sons, Inc.
- Hájek, J. (1971). Comment on An essay on the logical foundations of survey sampling by Basu, D. (R. a. Holt, Ed.) *Foundations of Statistical Inference*, 236.
- Haziza, D., & Beaumont, J.-F. (2007). On the Construction of Imputation Classes in Surveys. *International Statistical Review*, 75(1), 25-43.

- Haziza, D., & Lesage, É. (2016). A Discussion of Weighting Procedures for Unit Nonresponse. *Journal of Official Statistics*, 32(1), 129-145.
- Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663-685.
- INDEC. (2018). *Estudio Nacional sobre el Perfil de las Personas con Discapacidad: resultados definitivos 2018*. Ciudad Autónoma de Buenos Aires: INDEC.
- Kalton, G., & Maligalig, D. S. (1991). A comparison of methods of weighting adjustment for nonresponse. *Proceedings of the US Bureau of the Census Annual Research Conference*, 409-428.
- Kim, J. K., & Kim, J. J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics*, 35(4), 501-514.
- Krewski, D., & Rao, J. N. (1981). Inference From Stratified Samples: Properties of the Linearization, Jackknife and Balanced Repeated Replication Methods. *The Annals of Statistics*, 1010-1019.
- Little, R. J., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2nd ed.). John Wiley & Sons, Inc.
- Lohr, S. L. (1999). *Sampling: design and analysis*. Pacific Grove, CA: Duxbury Press.
- Lundström, S., & Särndal, C.-E. (1999). Calibration as a Standard Method for Treatment of Nonresponse. *Journal of Official Statistics*, 15(2), 305-327.
- Mach, L., Dumais, J., & Robidou, L. (2005). A study of the properties of a bootstrap variance estimator under sampling without replacement. *Federal Committee on Statistical Methodology (FCSM) Research Conference*. Arlington, VA.
- Madow, W. G. (1949). On the Theory of Systematic Sampling, II. *The Annals of Mathematical Statistics*, 20(3), 333-354. doi:10.1214/aoms/1177729988

- Mashreghi, Z., Haziza, D., & Léger, C. (2016). A survey of bootstrap methods in finite population sampling. *Statistics Surveys*, 10, 1-52.
- Monlong, J. (9 de Junio, 2018). *Clustering into same size clusters*. Recueprado de Hippocamplus: <http://jmonlong.github.io/Hippocamplus/2018/06/09/cluster-same-size/>
- Rao, J. N., & Wu, C. F. (1988). Resampling Inference With Complex Survey Data. *Journal of the American Statistical Association*, 83(401), 231-241.
- Rao, J. N., Wu, C. F., & Yue, K. (1992). Some Recent Work on Resampling Methods for Complex Surveys. *Survey Methodology*, 18(2), 209-217.
- Rosenbaum, P. R., & Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1), 41-55.
- Särndal, C.-E., & Deville, J.-C. (1992). Calibration Estimators in Survey Sampling. (L. Taylor & Francis, Ed.) *Journal of the American Statistical Association*, 87(418), 376-382.
- Särndal, C.-E., & Lundström, S. (1999). Calibration as Standard Method for Treatment of Nonresponse. *Journal of Official Statistics*, 15(2), 305-327.
- Särndal, C.-E., & Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Springer.
- Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer.
- U.S. Census Bureau. (2006). *Current Population Survey Design and Methodology*.
- U.S. Census Bureau. (Septiembre, 2018). *TIGER/Line Shapefiles*. Recuperado de TIGER/Line Shapefiles: <https://www.census.gov/cgi-bin/geo/shapefiles/index.php>
- Valliant, R., Dever, J. A., & Kreuter, F. (2018). *Practical Tools for Designing and Weighting Survey Samples* (2nd ed.). New York: Springer.
- Wheaton, W. D. (Mayo, 2014). *2010 U.S. Synthetic Population*. RTI International. Recuperado de <https://www.epimodels.org/midas/Rpubsyntdata1.do>

Wolter, K. M. (2007). *Introduction to Variance Estimation* (2nd ed.). Springer.

## **Anexo 1: Descripción de la población sintética**

En el presente anexo se listan y describen a las variables de los conjuntos de datos utilizados en esta tesina. Tanto para el caso de viviendas como el de individuos, los datos venían originalmente en archivos en formato .csv por separado para cada condado, por lo que este no aparece como campo de las tablas, sino que fue elaborado luego.

### **i. Conjunto sintético de viviendas**

El conjunto de variables referente a las viviendas utilizado en esta tesina se describe en la siguiente tabla.

**Tabla A1.1: Descripción de las variables del conjunto sintético de viviendas**

<b>Campo</b>	<b>Descripción</b>
sp_id	Identificador numérico único de cada vivienda.
hh_income	Ingreso anual del hogar.
hh_size	Cantidad de personas en el hogar.
hh_age	Edad del jefe de hogar.
latitude	Latitud de la vivienda.
longitude	Longitud de la vivienda.

### **ii. Conjunto sintético de personas**

Este conjunto de datos fue utilizado para obtener la composición de personas por hogar y además calcular los totales poblacionales utilizados en la calibración de pesos.

**Tabla A1.2: Descripción de las variables del conjunto sintético de personas**

<b>Campo</b>	<b>Descripción</b>
sp_id	Identificador numérico único de cada persona.
sp_hh_id	Identificador del hogar en el cual reside la persona. Este campo se corresponde con el campo <b>sp_id</b> de la <b>Tabla A1.1</b> .
age	Edad de la persona
sex	Sexo de la persona, donde 1 = masculino y 2 = femenino

## Anexo 2: Descripción de Condados y UPM

En la siguiente tabla se presentan a los condados y unidades primarias de muestreo de forma exhaustiva junto con la cantidad de personas y de hogares por condado.

**Tabla A2.1: Composición de Unidades Primarias de Muestreo**

<b>UPM</b>	<b>Condado</b>	<b>Cantidad de Personas</b>	<b>Cantidad de Hogares</b>
1	Culberson	2350	908
1	Hudspeth	3297	1174
1	Jeff Davis	2282	1034
2	Bailey	7067	2468
2	Cochran	3048	1113
3	Dallam	6643	2448
3	Hartley	4695	1771
3	Oldham	1767	691
4	Hansford	5435	2006
4	Sherman	2967	1081
5	Hemphill	3866	1382
5	Lipscomb	3352	1263
5	Roberts	923	359
6	Collingsworth	3029	1179
6	Wheeler	5303	2181
7	Armstrong	1744	751
7	Carson	6055	2452
7	Donley	3413	1517
8	Briscoe	1611	692
8	Swisher	7295	2762
9	Childress	5684	2326
9	Hall	3293	1372
10	Cottle	1534	677
10	Foard	1254	573
10	Hardeman	4128	1722
10	King	288	113
10	Knox	3529	1506
11	Floyd	6319	2402
11	Motley	1204	542
12	Crosby	5788	2237
12	Dickens	2085	930
13	Garza	4242	1671
13	Kerr	47938	20550
14	Fisher	3839	1668
14	Haskell	5419	2297

14	Stonewall	1433	642
15	Borden	678	264
15	Kent	746	350
15	Scurry	15192	5838
16	Baylor	3760	1669
16	Shackelford	3320	1367
16	Throckmorton	1649	721
17	Denton	651811	240288
17	Lamar	49168	19829
18	Mills	4945	1975
18	San Saba	5420	2257
19	Concho	2480	1041
19	Mason	4086	1754
19	Menard	2231	994
20	Edwards	1990	839
20	Kinney	3290	1350
20	Real	3311	1374
21	Pecos	13376	4894
21	Terrell	1017	430
22	Kimble	4631	2016
22	Sutton	4037	1550
23	Crockett	3630	1422
23	Irion	1658	653
23	Schleicher	3515	1182
24	Crane	4362	1471
24	Upton	3323	1256
25	Loving	70	39
25	Ward	10546	3995
25	Winkler	6968	2578
26	Goliad	7217	2868
26	Refugio	7473	2841
27	La Salle	5253	1931
27	Live Oak	10321	4257
27	McMullen	743	310
28	Brooks	7131	2642
28	Jim Hogg	5360	1902
28	Kenedy	404	147
29	Coke	3244	1466
29	Mitchell	7021	2809
29	Sterling	1067	440
30	Glasscock	1176	441
30	Martin	4653	1649
30	Reagan	3271	1156
31	Anderson	44256	17218

32	Andrews	14738	5259
33	Angelina	83691	31090
34	Aransas	22508	9795
35	Archer	8936	3538
36	Atascosa	44541	15246
37	Austin	28559	10837
38	Bandera	19962	8564
39	Bastrop	71533	25840
40	Bee	24461	9042
41	Bell	301710	114035
42	Bexar	1669355	608917
43	Blanco	10522	4309
44	Bosque	17701	7254
45	Bowie	85207	34669
46	Brazoria	301464	106542
47	Brazos	181594	71738
48	Brewster	9041	4207
49	Brown	36476	14778
50	Burleson	17046	6822
51	Burnet	41475	16511
52	Caldwell	34561	12301
53	Calhoun	21406	7766
54	Callahan	13311	5447
55	Cameron	402074	119628
56	Camp	12213	4678
57	Cass	30317	12429
58	Castro	7960	2744
59	Chambers	34831	11952
60	Cherokee	47552	17894
61	Clay	10889	4319
62	Coleman	8838	3857
63	Collin	778608	283754
64	Colorado	20540	8182
65	Comal	107739	41363
66	Comanche	13702	5580
67	Cooke	37605	14513
68	Coryell	64064	22533
69	Dallas	2333915	855917
70	Dawson	11829	4385
71	Deaf Smith	19169	6365
72	Delta	5186	2088
73	DeWitt	18215	7407
74	Dimmit	10004	3421
75	Duval	11112	4090

76	Eastland	17711	7465
77	Ector	134353	48688
78	El Paso	782168	256555
79	Ellis	147460	50503
80	Erath	36107	14569
81	Falls	15649	6302
82	Fannin	30462	12149
83	Fayette	24180	10078
84	Fort Bend	578726	187384
85	Franklin	10471	4159
86	Freestone	18095	7259
87	Frio	13919	4850
88	Gaines	17367	5606
89	Galveston	287224	108969
90	Gillespie	24536	10572
91	Gonzales	19341	7120
92	Gray	21184	8443
93	Grayson	118686	46905
94	Gregg	117063	45795
95	Grimes	23731	8902
96	Guadalupe	129870	45762
97	Hale	33293	11846
98	Hamilton	8122	3442
99	Hardin	54286	20462
100	Harris	4029032	1435117
101	Harrison	64306	24523
102	Hays	149766	55016
103	Henderson	77528	31020
104	Hidalgo	762416	216342
105	Hill	34165	13238
106	Hockley	22426	8242
107	Hood	50261	20795
108	Hopkins	34658	13308
109	Houston	20811	8645
110	Howard	29011	11333
111	Hunt	84344	32076
112	Hutchinson	22055	8812
113	Jack	7809	3136
114	Jackson	13874	5284
115	Jasper	34546	13770
116	Jefferson	235575	93439
117	Jim Wells	40391	13961
118	Johnson	148263	52193
119	Jones	14977	6033

120	Karnes	11477	4463
121	Kaufman	101911	34964
122	Kendall	32958	12617
123	Kleberg	29758	11097
124	Lamb	13711	5081
125	Lampasas	19467	7539
126	Lavaca	18782	7808
127	Lee	15884	6151
128	Leon	16753	6896
129	Liberty	70186	25073
130	Limestone	21493	8499
131	Llano	19122	9008
132	Lubbock	267634	105779
133	Lynn	5798	2246
134	Madison	11041	4187
135	Marion	10330	4595
136	Matagorda	36517	13894
137	Maverick	53166	15563
138	McCulloch	8155	3338
139	McLennan	224403	86891
140	Medina	43817	15530
141	Midland	135010	50845
142	Milam	24242	9408
143	Montague	19519	7989
144	Montgomery	451431	162530
145	Moore	21661	7197
146	Morris	12695	5226
147	Nacogdoches	59473	23861
148	Navarro	47106	17380
149	Newton	13651	5476
150	Nolan	14989	5999
151	Nueces	332866	124587
152	Ochiltree	10163	3617
153	Orange	81420	31031
154	Palo Pinto	27674	10871
155	Panola	23600	9271
156	Parker	113385	42069
157	Parmer	10257	3413
158	Polk	41071	16503
159	Potter	114475	42933
160	Presidio	7895	2906
161	Rains	10834	4377
162	Randall	119393	47975
163	Red River	12935	5469

164	Reeves	10685	3839
165	Robertson	16224	6541
166	Rockwall	77739	26448
167	Runnels	10204	4165
168	Rusk	48013	18476
169	Sabine	10628	4738
170	San Augustine	8703	3625
171	San Jacinto	25930	10096
172	San Patricio	64054	22637
173	Shelby	25090	9648
174	Smith	205387	79054
175	Somervell	8216	3078
176	Starr	60056	17001
177	Stephens	9080	3665
178	Tarrant	1785484	657129
179	Taylor	126050	50724
180	Terry	11441	4200
181	Titus	31734	10813
182	Tom Green	104834	42323
183	Travis	1000420	404461
184	Trinity	14551	6142
185	Tyler	19310	8007
186	Upshur	38841	14925
187	Uvalde	25921	9025
188	Val Verde	47119	15653
189	Van Zandt	51814	20047
190	Victoria	85076	32187
191	Walker	51250	20969
192	Waller	39503	14040
193	Washington	32502	13037
194	Webb	246960	67105
195	Wharton	40718	15132
196	Wichita	119408	49015
197	Wilbarger	12958	5289
198	Willacy	18877	5764
199	Williamson	416598	152606
200	Wilson	42458	15009
201	Wise	57985	21015
202	Wood	40899	17118
203	Yoakum	7885	2643
204	Young	18563	7343
205	Zapata	13922	4297
206	Zavala	11148	3573

Las Unidades Primarias de Muestreo 1, 4, 5, 6, 7, 10, 11, 12, 14, 16, 18, 19, 20, 23, 24, 30, 35, 61, 72, 75, 119, 120, 128, 133, 157, y 205 no pertenecen al marco muestral definitivo porque son eliminados en los recortes realizados posteriormente a la conformación de las UPM.

### Anexo 3: Descripción de estratos de UPM

Las unidades primarias de muestreo fueron estratificadas según la cantidad de viviendas.

Además, se agruparon en un estrato de inclusión forzosa a aquellas UPM que contienen a aglomerados urbanos de mayor relevancia.

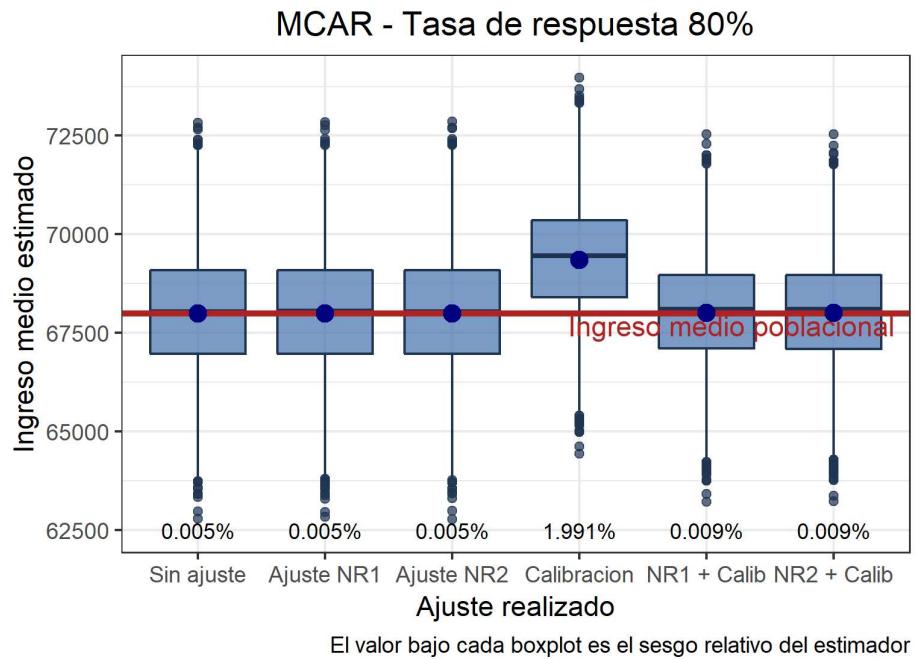
**Tabla A3.1: Composición de Estratos de UPM**

Estrato	Definición	Unidades Primarias de Muestreo	Cantidad de viviendas
1	Cantidad de viviendas menor o igual a 10,000	2, 3, 8, 9, 15, 21, 22, 25, 26, 27, 28, 29, 32, 34, 37, 38, 40, 43, 44, 48, 50, 52, 53, 54, 56, 57, 58, 59, 62, 64, 66, 67, 70, 71, 73, 74, 76, 81, 82, 83, 85, 86, 87, 88, 90, 91, 92, 95, 97, 98, 105, 106, 109, 110, 112, 113, 114, 117, 123, 124, 125, 126, 127, 130, 131, 134, 135, 136, 138, 142, 143, 145, 146, 148, 149, 150, 152, 154, 155, 160, 161, 163, 164, 165, 167, 169, 170, 171, 173, 175, 176, 177, 180, 181, 184, 185, 187, 192, 197, 198, 201, 203, 204, 206	499,295
2	Cantidad de viviendas mayor a 10,000 y menor o igual a 50,000	13, 31, 33, 36, 39, 45, 49, 51, 60, 65, 68, 77, 79, 80, 93, 94, 96, 99, 101, 102, 103, 107, 108, 111, 115, 118, 121, 122, 129, 137, 140, 141, 147, 153, 156, 158, 159, 162, 166, 168, 172, 179, 182, 186, 188, 189, 190, 191, 193, 195, 196, 200, 202	1,273,595
3	Cantidad de viviendas mayor a 50,000	17, 41, 46, 47, 55, 84, 89, 104, 116, 132, 139, 144, 151, 174, 194, 199	1,897,667
4	Aglomerados urbanos relevantes	42 = Bexar County 63 = Collin County 69 = Dallas County 78 = El Paso County 100 = Harris County 178 = Tarrant County 183 = Travis County	4,417,045

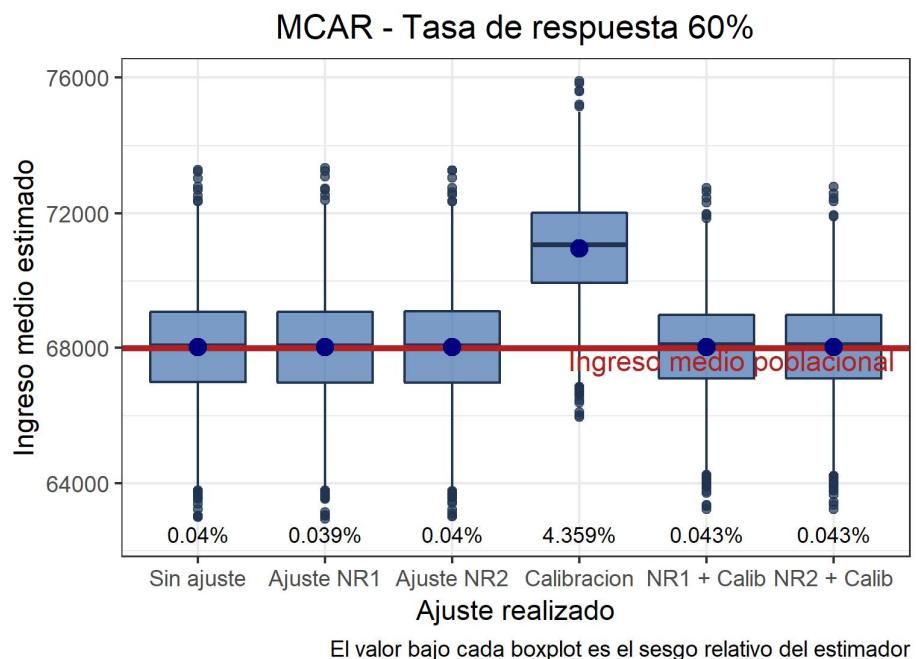
## **Anexo 4: Gráficos**

En este anexo se muestran diversos gráficos que resumen las distribuciones y propiedades de los estimadores bajo las diferentes condiciones de simulación.

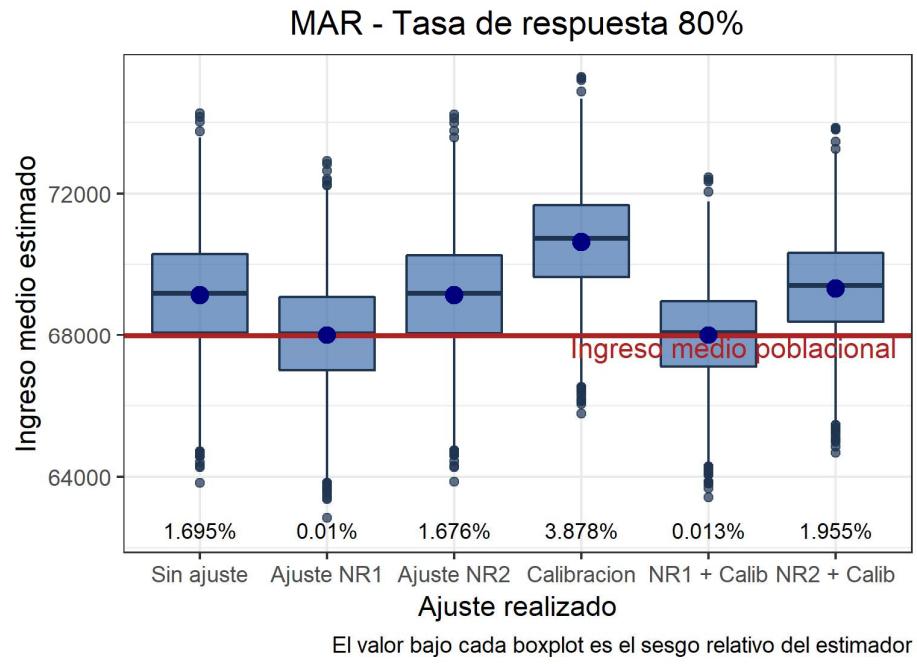
**Gráfico A4.1:** Distribución del estimador del ingreso medio. MCAR - Tasa de respuesta 80%.



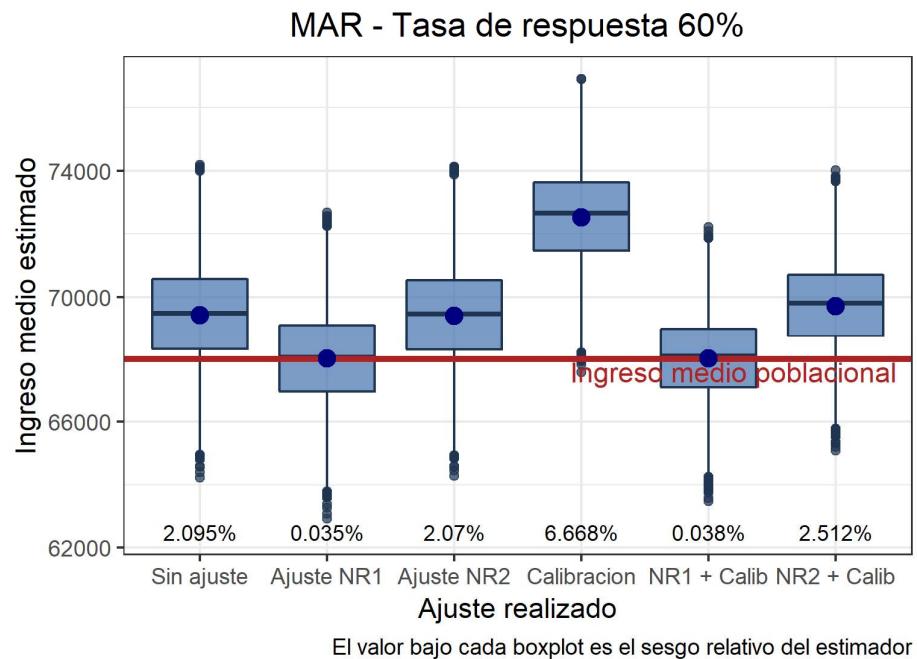
**Gráfico A4.2:** Distribución del estimador del ingreso medio. MCAR - Tasa de respuesta 60%



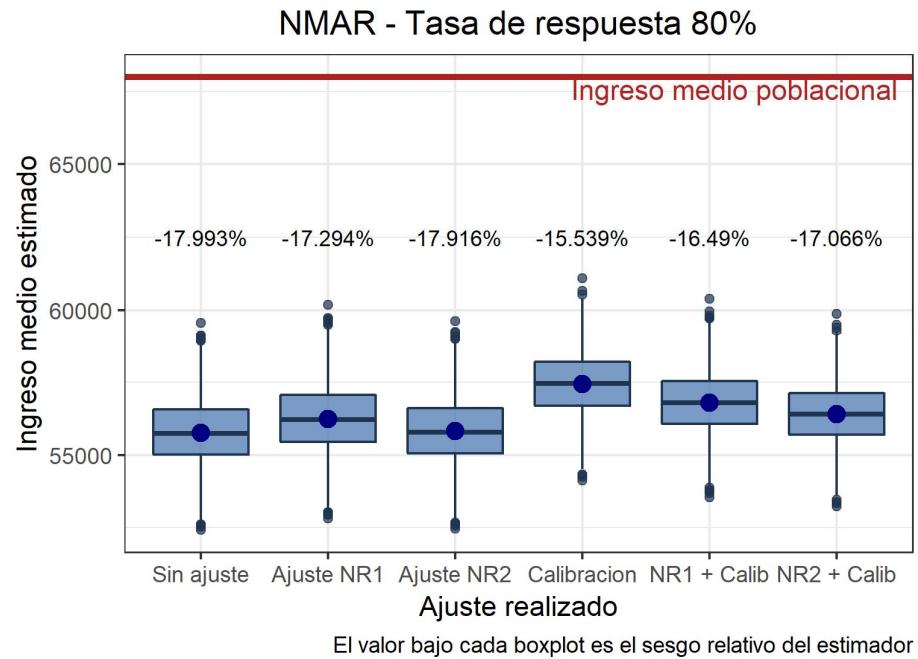
**Gráfico A4.3:** Distribución del estimador del ingreso medio. MAR - Tasa de respuesta 80%.



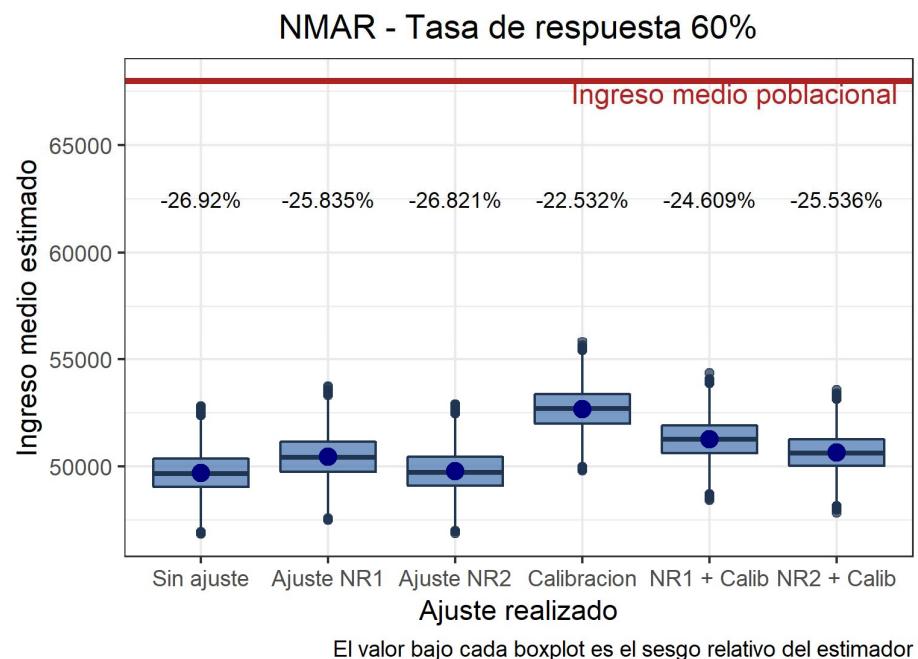
**Gráfico A4.4:** Distribución del estimador del ingreso medio. MAR - Tasa de respuesta 60%.



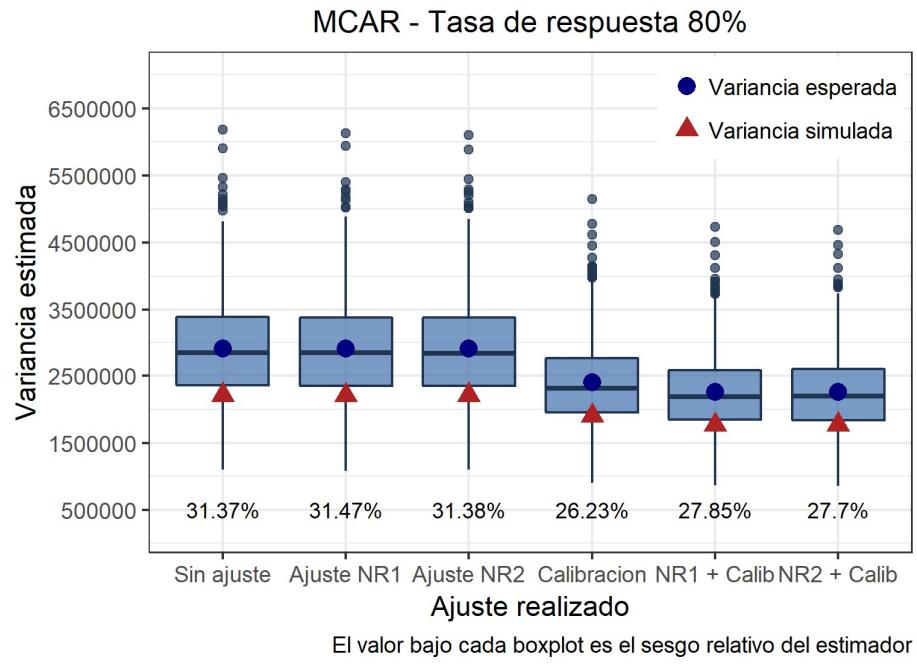
**Gráfico A4.5:** Distribución del estimador del ingreso medio. NMAR - Tasa de respuesta 80%.



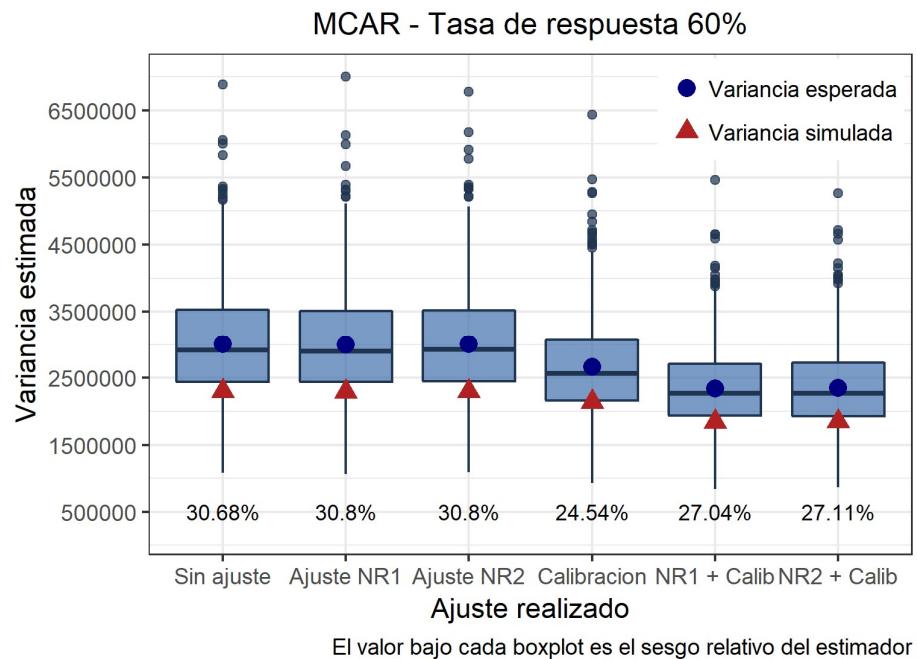
**Gráfico A4.6:** Distribución del estimador del ingreso medio. NMAR - Tasa de respuesta 60%.



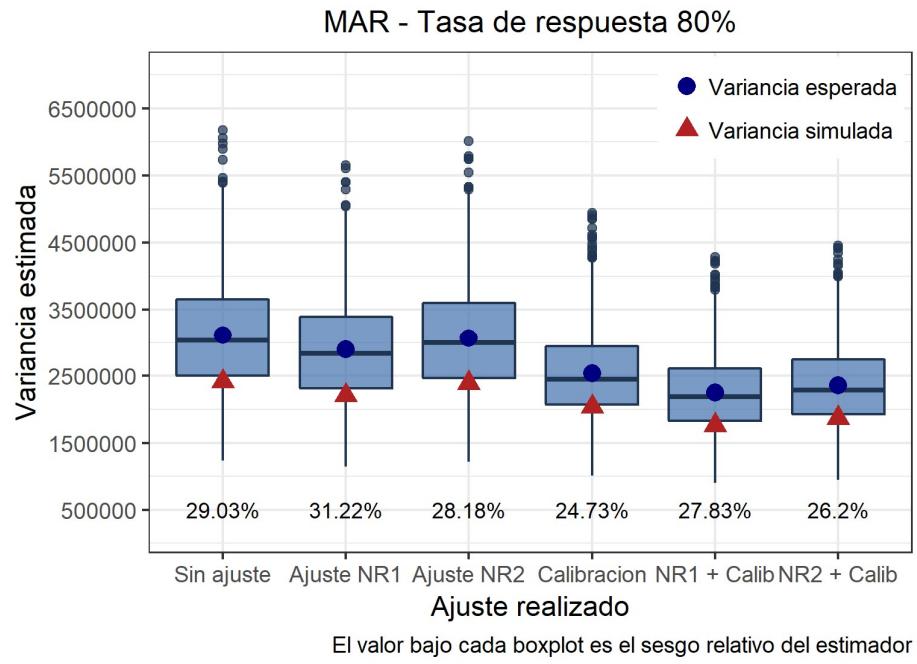
**Gráfico A4.7:** Distribución del estimador de variancia. MCAR - Tasa de respuesta 80%.



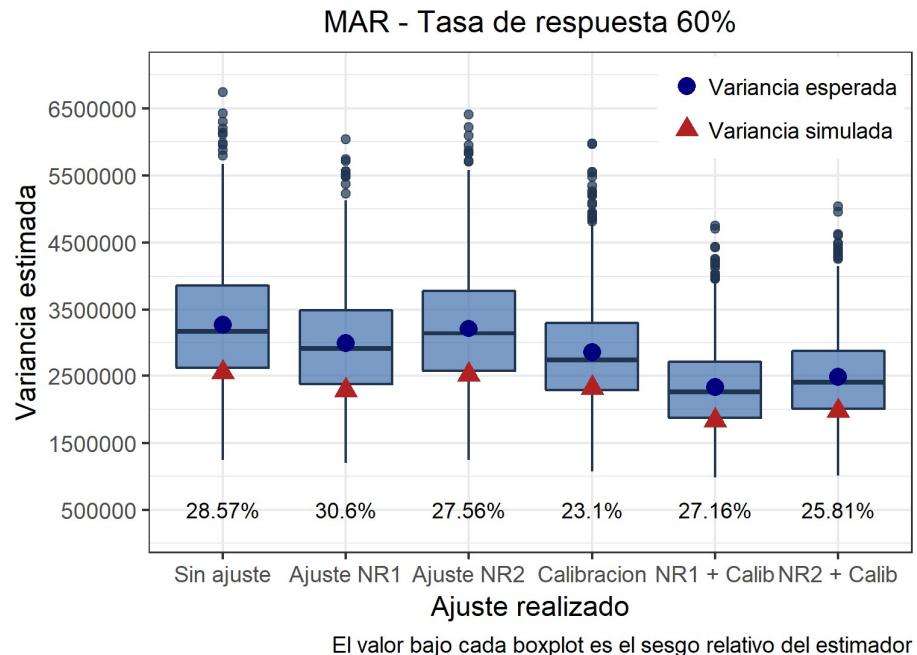
**Gráfico A4.8:** Distribución del estimador de variancia. MCAR - Tasa de respuesta 60%.



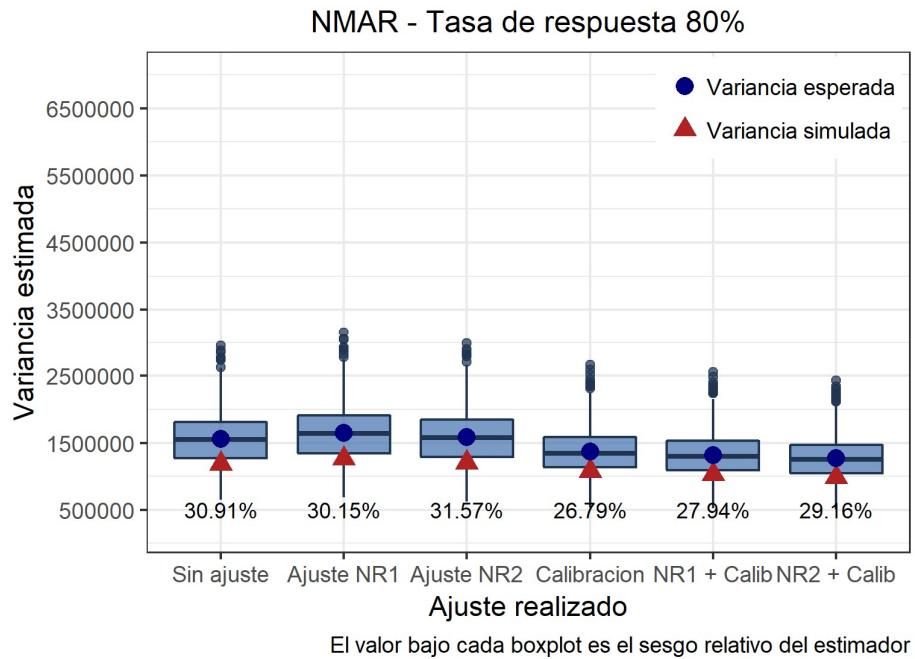
**Gráfico A4.9:** Distribución del estimador de variancia. MAR - Tasa de respuesta 80%.



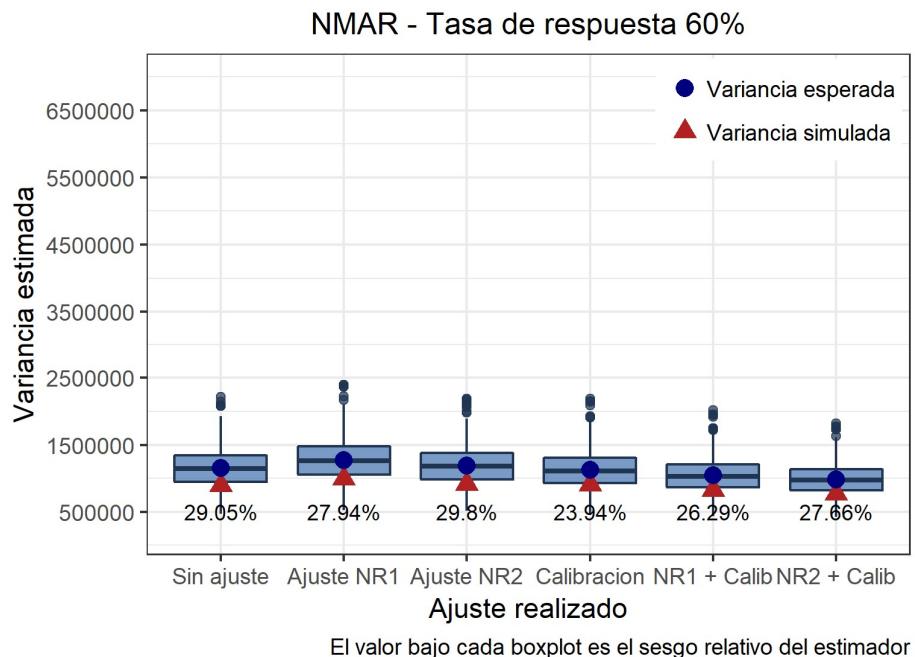
**Gráfico A4.10:** Distribución del estimador de variancia. MAR - Tasa de respuesta 60%.



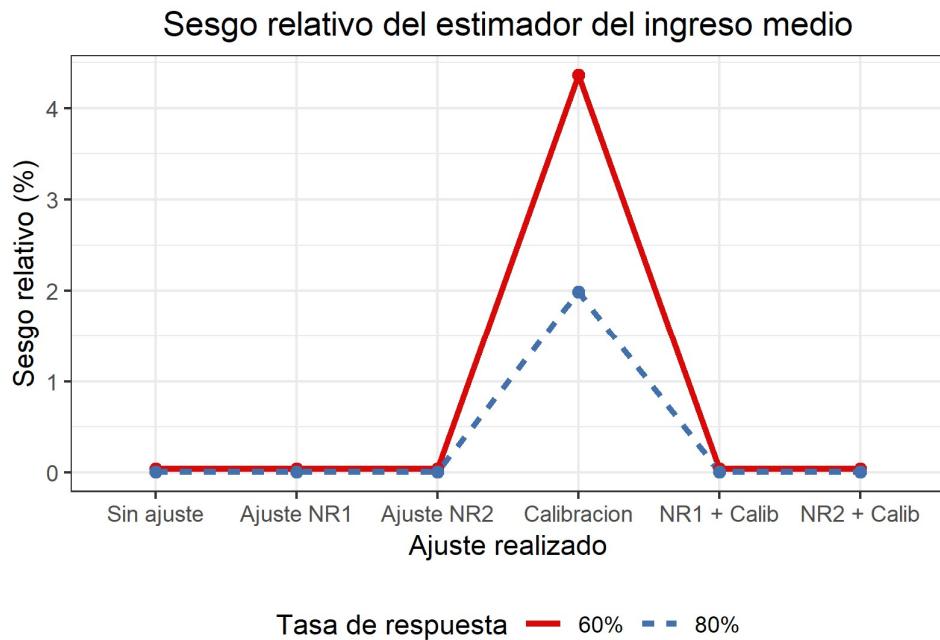
**Gráfico A4.11:** Distribución del estimador de variancia. NMAR - Tasa de respuesta 80%.



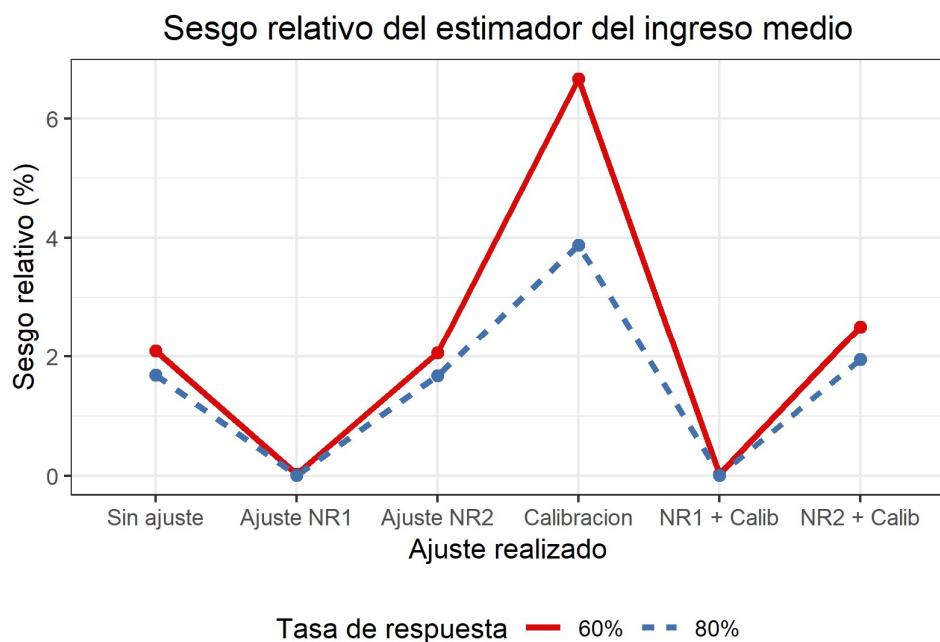
**Gráfico A4.12:** Distribución del estimador de variancia. NMAR - Tasa de respuesta 60%.



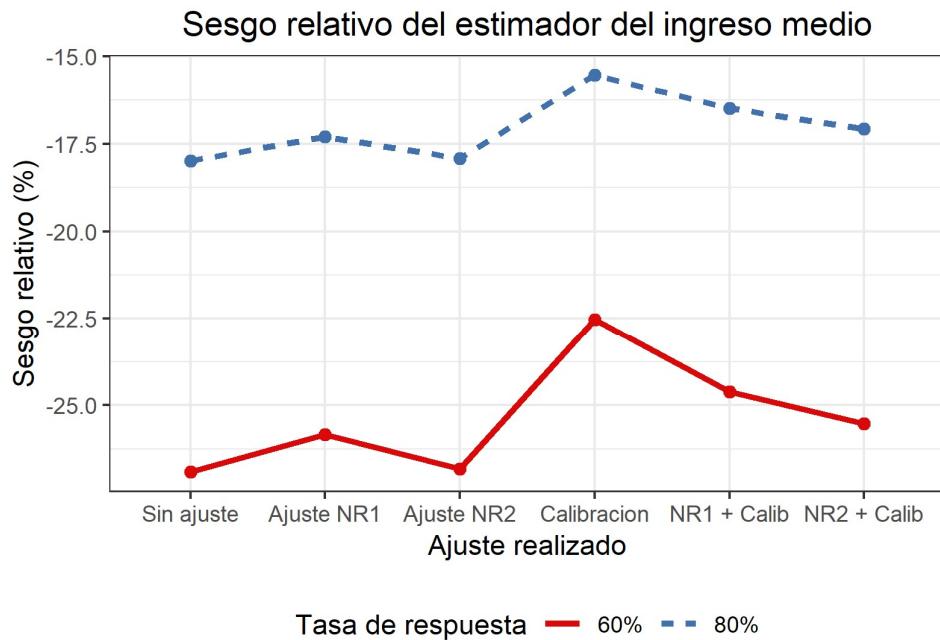
**Gráfico A4.13:** Sesgo relativo del estimador del ingreso medio según ajuste y tasa de respuesta – MCAR.



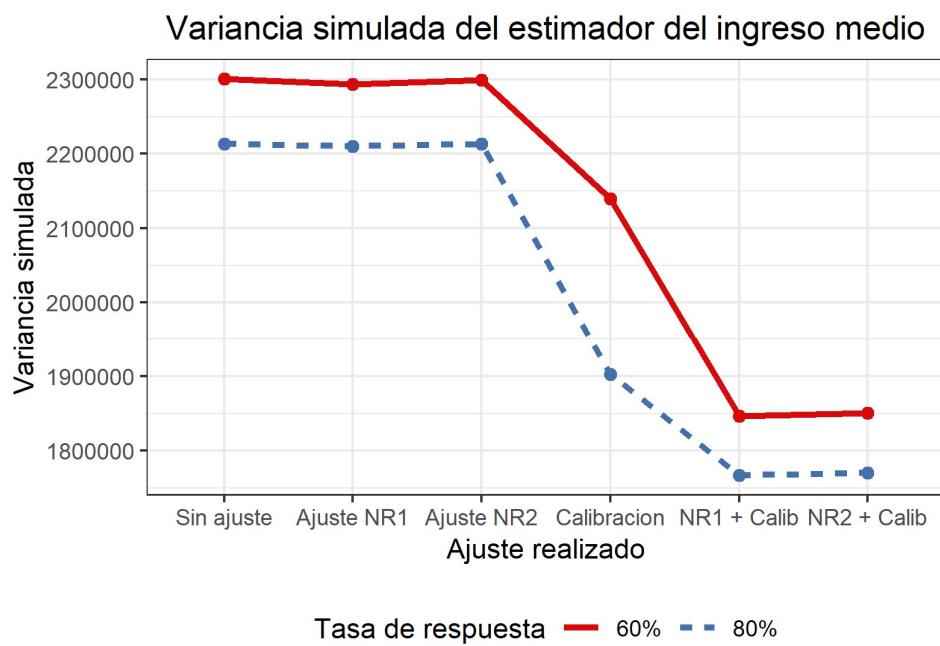
**Gráfico A4.14:** Sesgo relativo del estimador del ingreso medio según ajuste y tasa de respuesta – MAR.



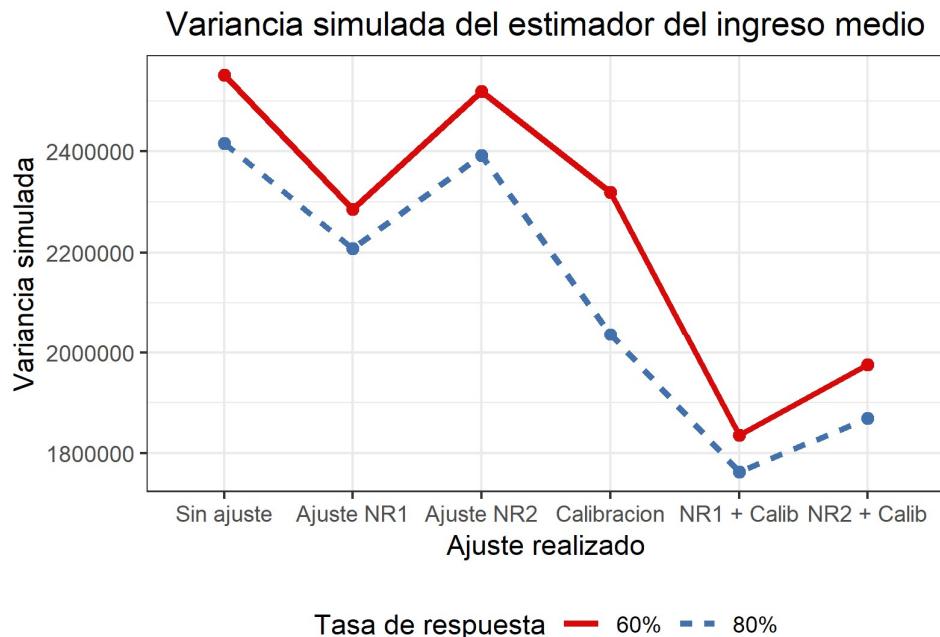
**Gráfico A4.15:** Sesgo relativo del estimador del ingreso medio según ajuste y tasa de respuesta – NMAR.



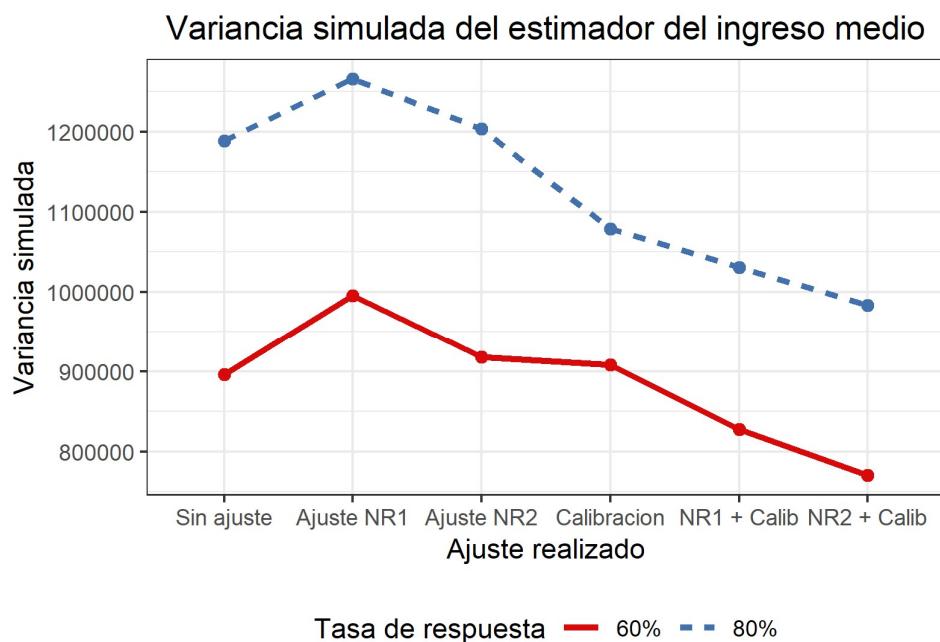
**Gráfico A4.16:** Variancia simulada del estimador del ingreso medio según ajuste y tasa de respuesta – MCAR.



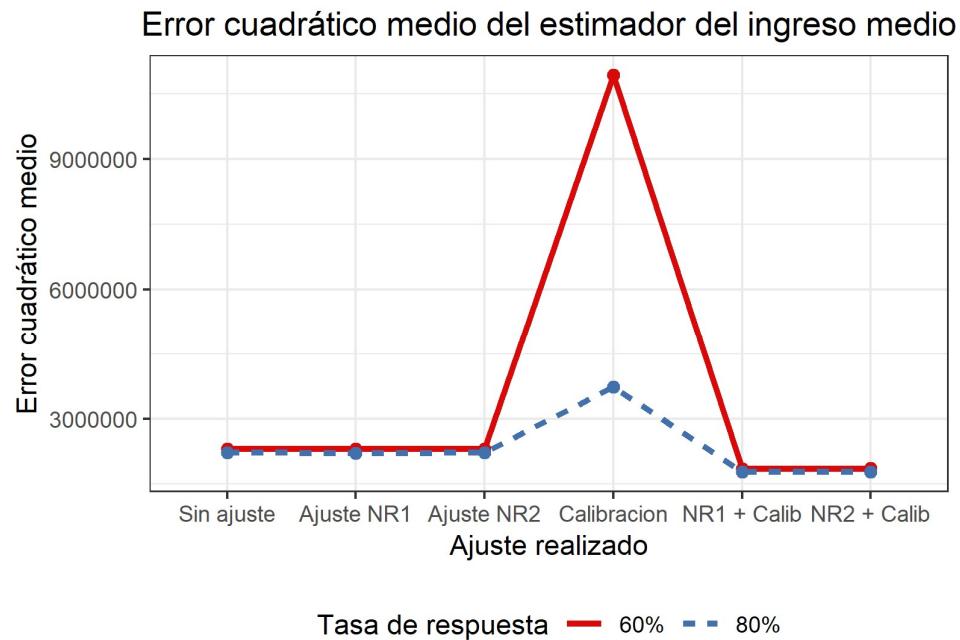
**Gráfico A4.17:** Variancia simulada del estimador del ingreso medio según ajuste y tasa de respuesta – MAR.



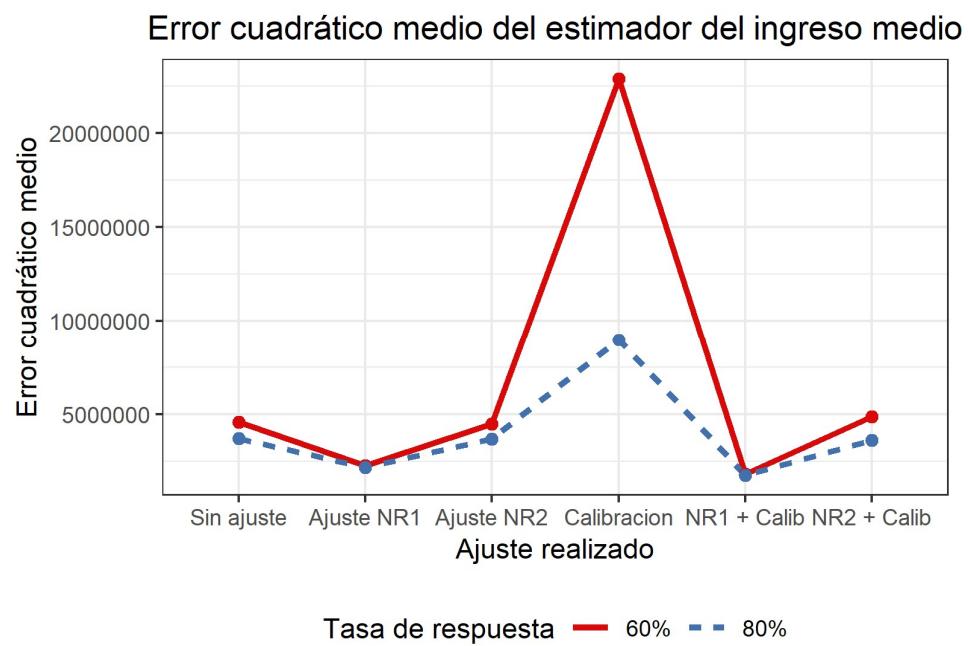
**Gráfico A4.18:** Variancia simulada del estimador del ingreso medio según ajuste y tasa de respuesta – NMAR.



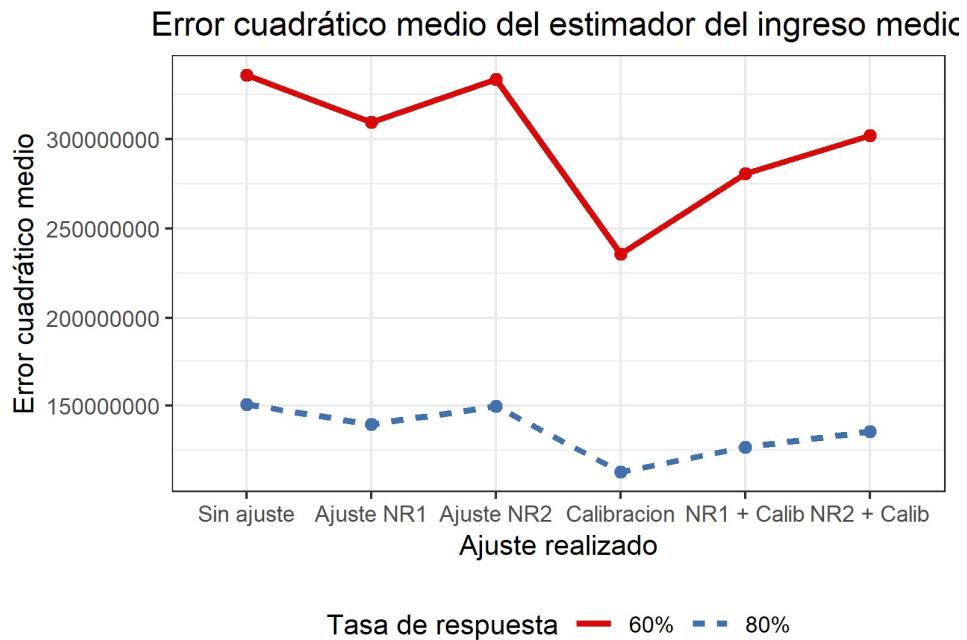
**Gráfico A4.19:** Error cuadrático medio del estimador del ingreso medio según ajuste y tasa de respuesta – MCAR.



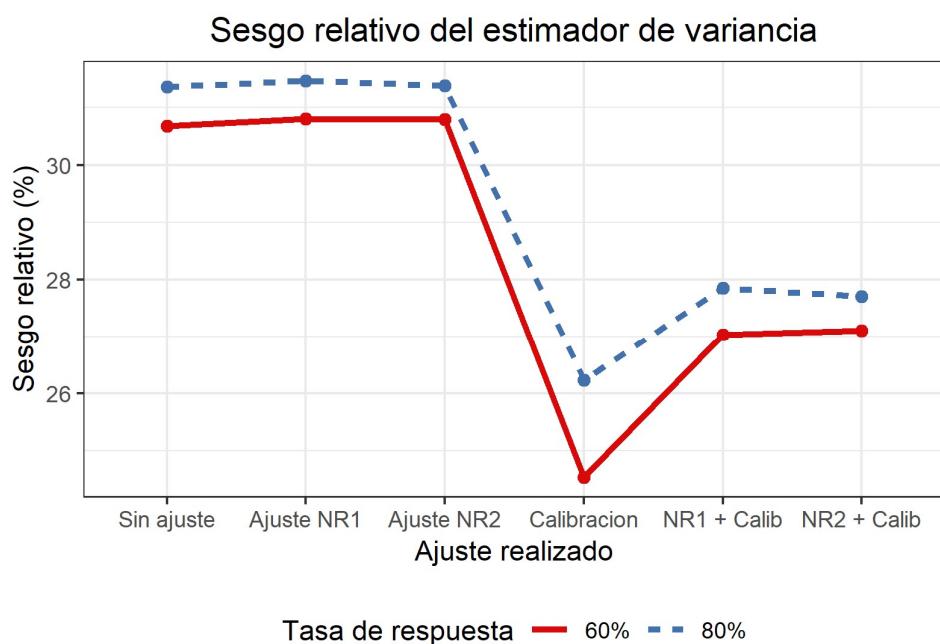
**Gráfico A4.20:** Error cuadrático medio del estimador del ingreso medio según ajuste y tasa de respuesta – MAR.



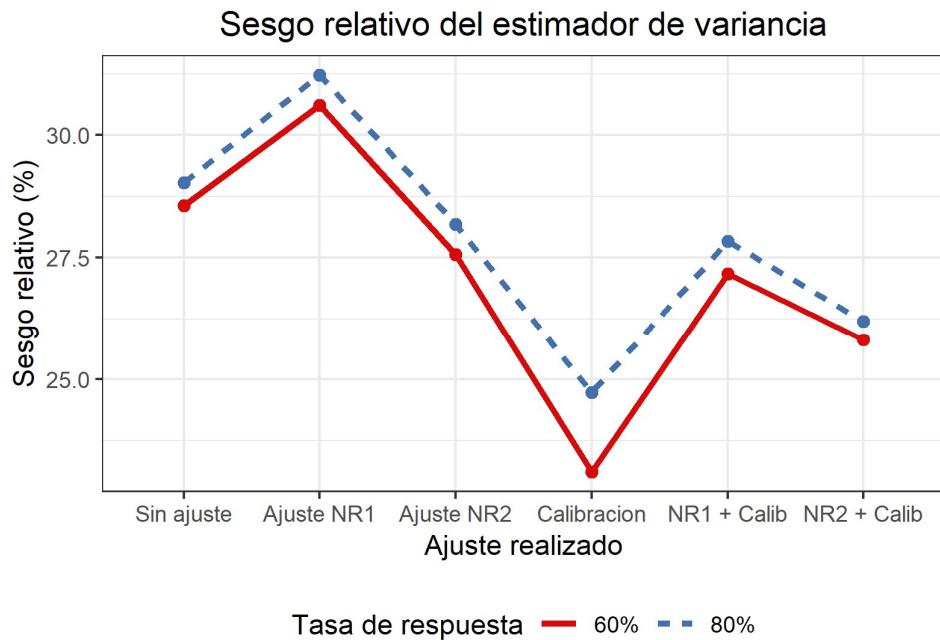
**Gráfico A4.21:** Error cuadrático medio del estimador del ingreso medio según ajuste y tasa de respuesta – NMAR.



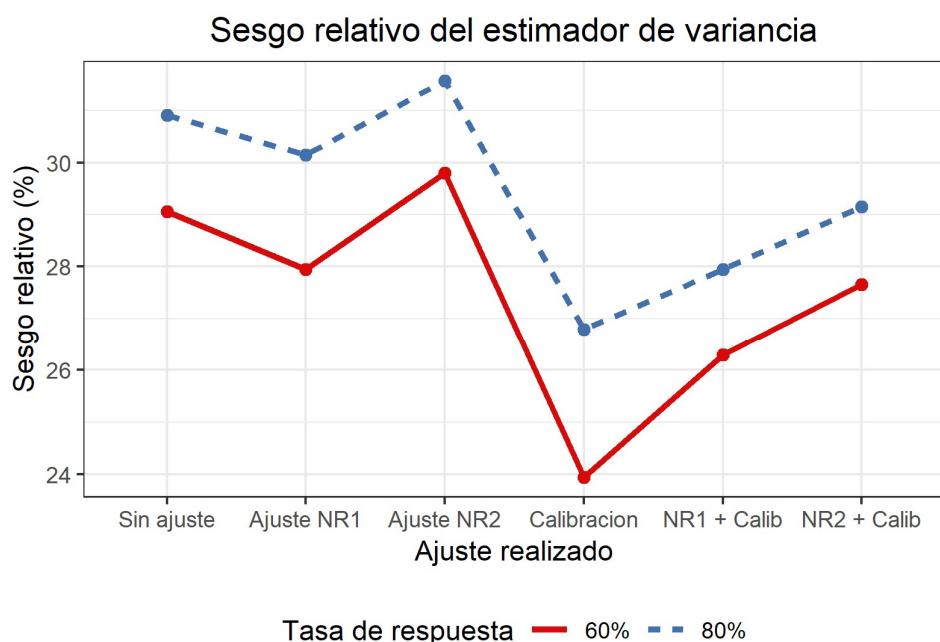
**Gráfico A4.22:** Sesgo relativo del estimador de variancia según ajuste y tasa de respuesta – MCAR.



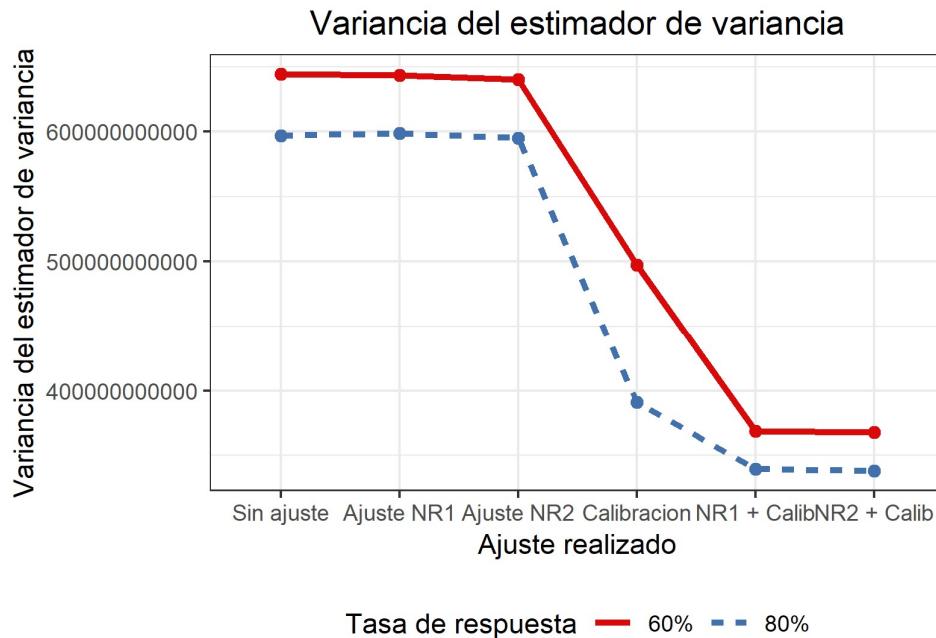
**Gráfico A4.23:** Sesgo relativo del estimador de variancia según ajuste y tasa de respuesta – MAR.



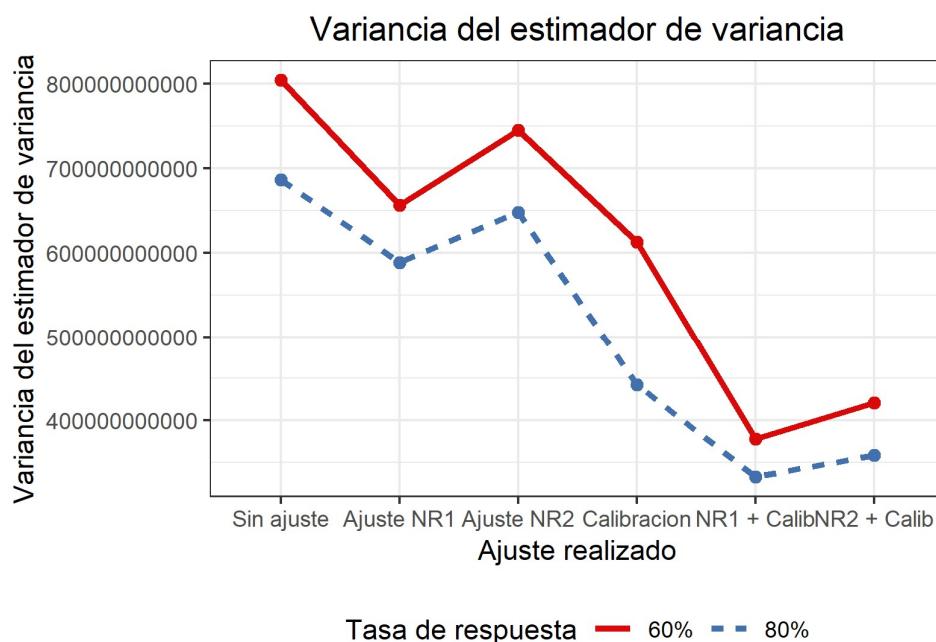
**Gráfico A4.24:** Sesgo relativo del estimador de variancia según ajuste y tasa de respuesta – NMAR.



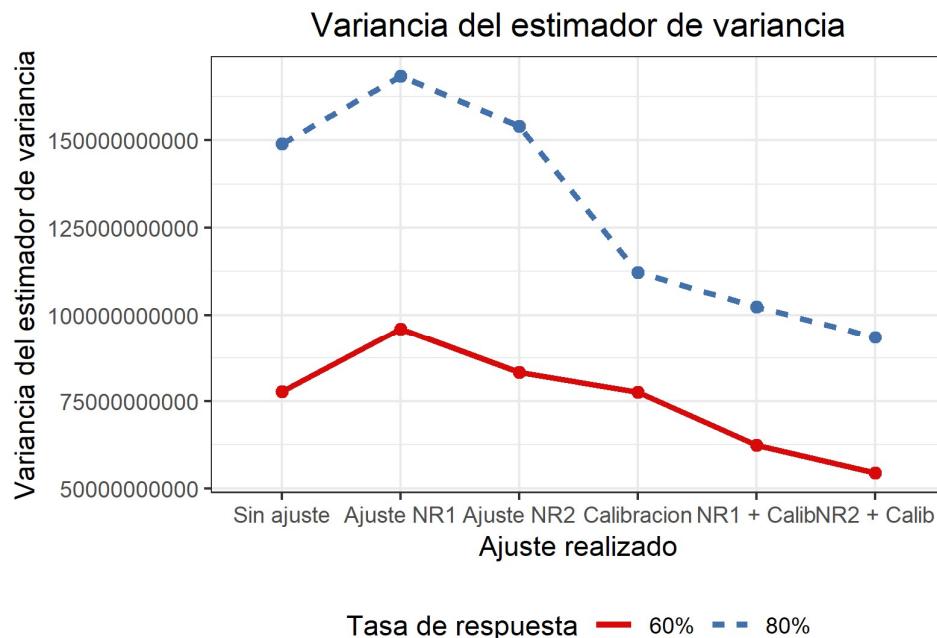
**Gráfico A4.25:** Variancia del estimador de variancia según ajuste y tasa de respuesta – MCAR.



**Gráfico A4.26:** Variancia del estimador de variancia según ajuste y tasa de respuesta – MAR.



**Gráfico A4.27:** Variancia del estimador de variancia según ajuste y tasa de respuesta – NMAR.



## **Anexo 5: Código utilizado**

El código utilizado en esta tesina se encuentra dividido en 10 bloques. Cada uno de ellos precede al siguiente. Cada bloque de código contiene una descripción al comienzo del mismo.

El enlace al repositorio de GitHub con el código es: <https://github.com/tomicapretto/tesina>