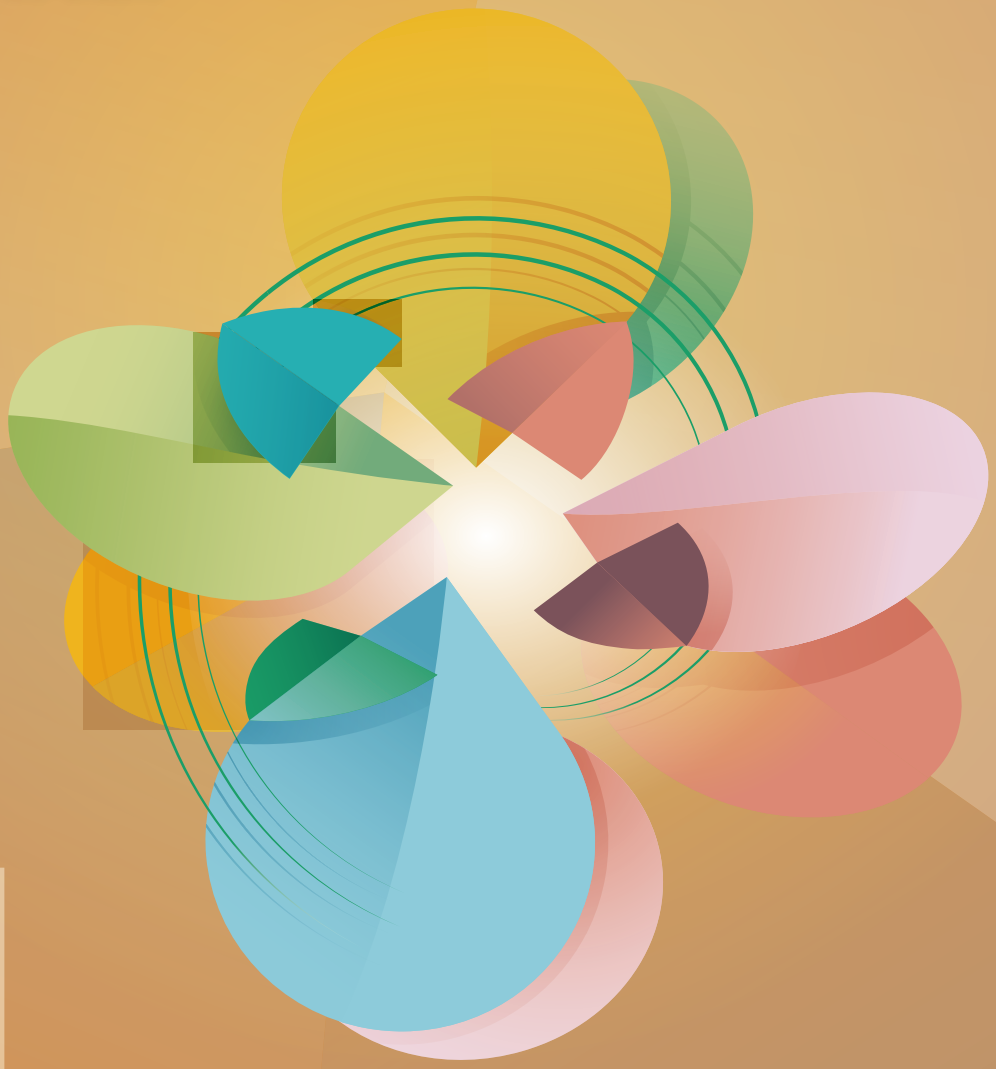
ISSN 2683-8478

Estudio Nacional sobre el Perfil de las Personas con Discapacidad

NOTA TÉCNICA

Factores de expansión, estimación, imputación y cálculo de los errores de muestreo

Enero de 2020



NOTAS TÉCNICAS INDEC N°3

Ministerio de Economía
Presidencia de la Nación

Instituto Nacional de Estadística y Censos
República Argentina

Estudio Nacional sobre el Perfil de las Personas con Discapacidad

Factores de expansión, estimación, imputación y cálculo de los errores de muestreo

Nota técnica - Enero de 2020

Instituto Nacional de Estadística y Censos (INDEC)

Esta publicación fue realizada por el equipo técnico de la Dirección Nacional de Metodología Estadística, a cargo del Lic. Gerardo Antonio Mitas, y de la Coordinación de Muestreo, a cargo de la Lic. María de los Ángeles Barbará, y el equipo de trabajo integrado por el Mag. Gonzalo Marí, el Lic. Gregorio García y la Lic. Fernanda Bonifazzi.

ISSN 2683-8478
ISBN 978-950-896-570-7

Instituto Nacional de Estadística y Censos - I.N.D.E.C.
Estudio nacional sobre el perfil de las personas con discapacidad : factores de expansión, estimación, imputación y cálculo de los errores de muestreo : nota técnica / 1a ed . - Ciudad Autónoma de Buenos Aires : Instituto Nacional de Estadística y Censos - INDEC, 2020.
Libro digital, PDF - (Notas técnicas ; 3)

Archivo Digital: descarga y online
ISBN 978-950-896-570-7

1. Estadísticas. 2. Personas Con Discapacidad. 3. Discapacidad. I. Título.
CDD 318

Libro de edición argentina

Responsable de la edición: Lic. Marco Lavagna

Director técnico: Mag. Pedro Lines

Directora de la publicación: Mag. Silvina Viazzi

Coordinación de producción editorial: Lic. Marcelo Costanzo

Buenos Aires, enero de 2020

Índice

1. Introducción

El Instituto Nacional de Estadística y Censos (INDEC) conjuntamente con la Agencia Nacional de Discapacidad (ANDIS) y el Consejo Federal de Discapacidad (COFEDIS) realizaron el Estudio Nacional sobre el Perfil de las Personas con Discapacidad 2018 (en adelante Estudio), con el objetivo de cuantificar a la población con dificultades y describir su perfil según diversas variables sociodemográficas.[1]

Esta publicación es una guía de referencia básica de la metodología adoptada para determinar los factores de expansión que se emplean en las estimaciones oficiales, y que permite estimar sus errores de muestreo para cualquiera de las estimaciones que surgen de ella.

En primera instancia se presentan las características principales del diseño muestral, el tamaño de la muestra, su asignación territorial y los dominios de estimación definidos para el Estudio.

A continuación, se detalla el proceso de ajuste por elegibilidad y no respuesta de los factores de expansión o ponderadores del Estudio. Se exponen los motivos por los cuales se introduce una modalidad para el cálculo de los errores por muestra que emplea replicaciones y se incluyen indicaciones para estimarlos en distintas herramientas de cálculo: R, Stata, SAS y Wesvar.

A su vez, se describe el proceso de imputación de datos faltantes que se aplica en las principales variables del Estudio.

Finalmente se explicita una serie de recomendaciones y advertencias sobre la confiabilidad y las limitaciones de las estimaciones que aparecen en los cuadros del Estudio publicados, o para aquellas que generen con fines estadísticos a partir de la base usuario del Estudio.

2. Diseño muestral del estudio

2.1 Antecedentes y limitaciones

El diseño muestral determinado para el Estudio no contó con las ventajas y los beneficios que tuvo la Encuesta Nacional de Personas con Discapacidad 2002-2003 (ENDI 2002-2003) al apoyarse en el Censo Nacional de Población, Hogares y Viviendas 2001 (ENPHyV 2001) dentro del plan de encuestas complementarias a dicho censo.

En el CNPHyV 2001 se empleó una pregunta que buscó relevar los hogares con al menos una persona con discapacidad. Esto permitió definir un diseño muestral para la ENDI 2002-2003 que posibilitó la combinación de dos muestras probabilísticas: una de viviendas con casos de discapacidad detectados por el censo y otra como complemento de la población sin discapacidad. Esta modalidad es altamente eficiente en estudios de prevalencia, ya que lleva a reducir costos, maximizar la precisión de las estimaciones, y disminuir considerablemente el tamaño de la muestra final.

[1] Ver https://www.indec.gob.ar/ftp/cuadros/poblacion/estudio_discapacidad_12_18.pdf.

El Censo Nacional de Población, Hogares y Viviendas 2010 (CNPHyV 2010) también incluyó una pregunta sobre la temática, pero el tiempo transcurrido a la fecha del Estudio (8 años) y el hecho de que dicha pregunta solo formó parte del cuestionario ampliado del censo,[2] y no del cuestionario básico, hizo inviable definir una estrategia similar a la aplicada en la ENDI 2002-2003 para el diseño del Estudio.

Como consecuencia, las aspiraciones de lograr estimaciones con desagregados equivalentes y con precisión comparables a los alcanzados con la ENDI 2002-2003 estarán limitadas a lo que puede arrojar una muestra con diseño en una fase, es decir, sin una previa identificación aproximada de la población objetivo a través de otra fuente.

2.2 Diseño y selección de la muestra

El diseño muestral del Estudio se apoya en el de la Muestra Maestra Urbana de Viviendas de la República Argentina (MMUVRA), cuyo diseño muestral es del tipo complejo. Un diseño simple y eficiente en términos de precisión podría ser un muestreo simple al azar (MSA), en el que las viviendas son seleccionadas aleatoriamente con igual probabilidad. Sin embargo, se requeriría de una lista de todas las viviendas pertenecientes al ámbito geográfico que abarca el Estudio, lo cual es dificultoso o imposible de lograr en la práctica.

También existen restricciones de índole operativa que pueden llevar a requerir un diseño complejo. Cuando el estudio es de gran envergadura y se aspira a alcanzar estimaciones con representatividad a nivel nacional u otros dominios territoriales de gran extensión, aun si se dispone de una lista completa de viviendas, bajo un MSA habría una alta probabilidad de que la muestra tenga una distribución geográfica muy dispersa.

Como resultado, los costos del operativo de campo del Estudio serían excesivamente altos o prohibitivos para cualquier presupuesto, en particular, los costos asociados a los desplazamientos de los encuestadores para cubrir grandes distancias hasta alcanzar las viviendas, a las posibles revisitas para contactar a los informantes en distintos horarios, y a las tareas de supervisión y control por parte de los supervisores.

Para reducir los costos en la preparación de un diseño muestral para cada operativo, controlar los problemas que ocasiona la dispersión de las muestras e integrar y coordinar las operaciones estadísticas, el INDEC emplea una modalidad bajo el esquema de muestra maestra. O sea, utiliza una única gran muestra probabilística, conocida como MMUVRA, que mantiene fijas las unidades de área que la conforman y su estructura probabilística asociada, y que permite subseleccionar las muestras de viviendas en el ámbito urbano para todas las encuestas a hogares del Instituto durante aproximadamente un decenio, o período intercensal.

La MMUVRA es de alcance nacional y urbano, y tiene como principales dominios de estimación las provincias y los aglomerados que participan en la Encuesta Permanente de Hogares (EPH) que lleva a cabo el INDEC. Su diseño inicialmente emplea dos etapas de selección probabilística. Cada unidad de primera etapa de muestreo (UPM) del diseño está definida por un aglomerado o una localidad de al menos 2.000 habitantes según el CNPHyV 2010. El conjunto de todas las UPM constituye el marco de muestreo o la lista de unidades de muestreo para la selección probabilística de primera etapa. Estas son estratificadas de acuerdo al total de población según CNPHyV 2010, y aquellas UPM formadas por aglomerados o localidades de 50.000 habitantes o más son incluidas en la MMUVRA con probabilidad 1 por diseño, y se las denomina “UPM autorrepresentadas”.

[2] El cuestionario ampliado fue aplicado en toda la población, salvo en las localidades de 50.000 o más habitantes donde solo lo respondía por muestra una parte de la población.

Del resto de las UPM, un conjunto fue seleccionado por provincia mediante un muestreo sistemático con probabilidad proporcional a la cantidad total de habitantes. Tanto las UPM autorrepresentadas como las seleccionadas conforman la muestra de aglomerados o localidades de la MMUVRA.

Para la segunda etapa, en las UPM seleccionadas, y solo para ellas, se definieron las "unidades de segunda etapa de muestreo" (USM) o "Áreas MMUVRA"[3] con base en radios censales y en la cartografía del CNPHyV 2010. En cada UPM, todas sus USM en conjunto la cubren territorialmente, determinan su envolvente y conforman el marco de muestreo para la selección de segunda etapa. Esta se completó con la selección de una muestra probabilística de USM, que emplea un diseño estratificado definido a partir de variables sociodemográficas y mediante un muestreo sistemático proporcional a la cantidad total de viviendas particulares ocupadas, según el CNPHyV 2010.

Por último, en cada una de las USM seleccionadas, se confeccionó inicialmente un listado exhaustivo de viviendas particulares, lo que dio origen al marco de selección de viviendas de la MMUVRA y sobre el cual se realizan las subselecciones para las muestras de todas las encuestas a hogares del INDEC.[4] El listado de viviendas tiene un orden específico y una cartografía asociada, que facilita su actualización y ayuda a organizar la asignación de la carga de trabajo, y las tareas de campo y recorrido de los encuestadores.[5]

Para el Estudio se definió una nueva etapa de selección sobre la MMUVRA, constituida por un tercer tipo de unidades de muestreo denominados "segmentos". Estos están conformados por 5 viviendas particulares contiguas o próximas dentro del listado de la MMUVRA, con el objetivo de concentrar los desplazamientos en terreno de los encuestadores y así reducir el costo del operativo. Una selección sistemática con igual probabilidad de estos segmentos permitió conformar la muestra definitiva de viviendas del Estudio.

3. Dominios geográficos de estimación y tamaño de la muestra

La población objetivo o de interés del Estudio abarca a las personas con dificultad residentes en viviendas particulares de las localidades de la República Argentina con 5.000 o más habitantes.[6] En el diseño muestral se definen como dominios geográficos de estimación del Estudio el total del país y su desagregación a nivel de 6 regiones:

- Gran Buenos Aires: Ciudad Autónoma de Buenos Aires y los partidos del Gran Buenos Aires.
- Noroeste: Catamarca, Jujuy, Salta, Tucumán, La Rioja y Santiago del Estero.
- Noreste: Chaco, Corrientes, Formosa y Misiones.
- Cuyo: Mendoza, San Juan y San Luis.
- Pampeana: Córdoba, Santa Fe, Entre Ríos, La Pampa y el resto de los partidos de Buenos Aires.

[3] En la conformación de las áreas MMUVRA, por cuestiones operativas (extensión, densidad, inaccesibilidad, etc.), los radios censales pueden sufrir recortes o agrupamientos (por ejemplo, para equilibrar la uniformidad de sus tamaños en términos de viviendas).
[4] Esta propiedad de permitir submuestrear viviendas sobre la muestra maestra hace que se la identifique también como un marco secundario de muestreo de viviendas.
[5] A la fecha del Estudio, la MMUVRA en su última actualización registraba un total de 2.053.958 viviendas particulares.
[6] Por este motivo, para la selección de las viviendas se la restringe la MMUVRA a localidades de 5.000 o más habitantes.

- Patagonia: Chubut, Neuquén, Río Negro, Santa Cruz y Tierra del Fuego.

Como consecuencia, la muestra del Estudio está constituida por 40.885 viviendas. Los siguientes cuadros, a manera de resumen, dan cuenta de la distribución de la muestra por región y jurisdicción.

Cuadro 1. Distribución de la muestra de viviendas por región

Regiones	Cantidad de viviendas
Gran Buenos Aires	10.548
Noroeste	6.154
Noreste	5.518
Cuyo	5.360
Pampeana	9.187
Patagonia	4.118
Total del país	**40.885**

Fuente: INDEC, Estudio Nacional sobre el Perfil de Personas con Discapacidad 2018.

Cuadro 2. Distribución de la muestra de viviendas por jurisdicción

Jurisdicción	Cantidad de viviendas
CABA	4.824
Partidos del Gran Buenos Aires	5.724
Resto de Buenos aires	2.995
Catamarca	970
Córdoba	2.136
Corrientes	1.572
Chaco	1.596
Chubut	840
Entre Ríos	1.040
Formosa	1.070
Jujuy	992
La Pampa	888
La Rioja	980
Mendoza	2.730
Misiones	1.280
Neuquén	928
Rio Negro	950
Salta	1.168
San Juan	1.590
San Luis	1.040[7]
Santa Cruz	800
Santa Fe	2.128
Santiago del Estero	1.068
Tucumán	976
Tierra del Fuego	600
Total del país	**40.885**

Fuente: INDEC, Estudio Nacional sobre el Perfil de Personas con Discapacidad 2018.

[7] Por cuestiones operativas ajenas al diseño de la muestra, la cantidad de viviendas en San Luis se redujo a 520.

4. Determinación y ajuste de los factores de expansión

La estimación de parámetros poblacionales a partir de una encuesta por muestreo probabilístico se basa en la premisa de que cada unidad de la muestra representa un cierto número de otras unidades en la población, además de sí misma. Por ejemplo, el total de unidades que poseen una característica dada se estima sumando los factores de expansión[8] de las personas, los hogares o las viviendas que tienen la característica buscada dentro de la muestra.

Inicialmente una vivienda seleccionada para el Estudio posee un factor de expansión atribuido por el diseño muestral del Estudio y que es definido como:[9]

$$w_{0ijk}^{V} = w_{1i}w_{2ij}w_{3ijk}$$

donde:

w_{1i} = inversa de la probabilidad de inclusión de la i-ésima UPM,

w_{2ij} = inversa de la probabilidad de inclusión en la segunda etapa de muestreo de la j-ésima USM dentro de la i-ésima UPM seleccionada,

w_{3ijk} = inversa de la probabilidad de inclusión de la k-ésima vivienda dentro de la j-ésima USM de la i-ésima UPM seleccionada.[10]

Sin embargo, en la práctica estos factores de expansión iniciales suelen ser modificados por diversos motivos y no terminan siendo los que se emplean para obtener las estimaciones de una encuesta. Durante el desarrollo de cualquier operativo estadístico se presentan una serie de problemas, algunos vinculados a errores de cobertura por desactualización del marco de muestreo, a la no respuesta de las unidades, o a la falta de eficacia en la captura de ciertos grupos de la población.

Todos estos errores forman parte de los denominados errores “no muestrales”, y que, sumados a otros, contribuyen a la componente del “error total” en una estimación. Son difíciles de cuantificar y afectan la calidad del dato en dos direcciones. Si son introducidos de manera aleatoria, la probabilidad de incrementar la variabilidad de la estimación es alta; pero si no son aleatorios, el principal impacto es introducir sesgo en los resultados.

Un objetivo central de las encuestas es minimizar el efecto de las distintas fuentes de error sobre los resultados; por ejemplo, manteniendo actualizados los marcos de muestreo, evaluando la estrategia de captura del dato en pruebas piloto, capacitando y entrenando a los encuestadores, o visitando en varias ocasiones y en distintos horarios el hogar o a la persona que no responde, para revertir su estado.

Pero aun tomando todos estos recaudos, los errores no desaparecen y llevan a que en la etapa previa a la estimación se incorporen en la determinación de los factores de expansión finales del Estudio varios ajustes

[8] Los términos “factores de expansión”, “ponderadores” o “pesos” en el contexto del documento hacen referencia siempre al mismo concepto.
[9] Para facilitar la lectura en la notación se omiten los subíndices correspondientes a los estratos definidos por el diseño muestral de las UPM y las USM, por lo que queda implícita la pertenencia a estos cada vez que se refiera al subíndice i de las UPM y al j de las USM.
[10] La probabilidad de inclusión de la k-ésima vivienda se corresponde con la probabilidad de selección sistemática de segmentos de 5 viviendas contiguas o próximas dentro de las USM seleccionadas.

sobre el factor de expansión definido por diseño, lo que busca disminuir el impacto de estos inconvenientes sobre los estimadores y aumentar la calidad de los resultados.

4.1 Ajuste por viviendas no elegibles

El primer ajuste que se realiza sobre los factores de expansión iniciales tiene como objetivo atender los problemas causados por deficiencias en la elegibilidad de la vivienda. Estas ocurren, ya sea por desactualización del listado de viviendas de la MMUVRA, o por la imposibilidad de los encuestadores de alcanzar o detectar las viviendas seleccionadas para la encuesta.

El tratamiento de este ajuste lleva a clasificar las viviendas seleccionadas como “elegibles”, “no elegibles” y de “elegibilidad dudosa”. Para el Estudio, y solo con el fin de ajustar los factores de expansión iniciales por no elegibilidad, se definen los siguientes tipos de vivienda.

- Viviendas elegibles (VEL), aquellas en donde se detecta una vivienda particular y se realiza una entrevista; o que presentan alguna de las siguientes categorías en la pregunta “Estado de la vivienda”:
    - con personas presentes,
    - con todas las personas ausentes temporalmente (causas circunstanciales, viaje o vacaciones),
    - rechazo,
    - otras causas.
- Viviendas no elegibles (VNE), aquellas registradas como:
    - deshabitada,
    - demolida, en demolición,
    - fin de semana o temporada,
    - en construcción o refacción,
    - usada como establecimiento,
    - local o comercio sin vivienda,
    - variaciones en el listado: no es vivienda.
- Viviendas de elegibilidad dudosa o elegibilidad desconocida (VED), aquellas que se corresponden con alguna de las siguientes categorías:
    - con todas las personas ausentes momentáneamente,
    - área insegura,
    - vivienda no identificada o dirección no existente,
    - variaciones en el listado: no se especificó ningún motivo de variaciones en el listado.

Teniendo en cuenta la clasificación, se estima la cantidad total de viviendas elegibles ajustada por elegibilidad dudosa como la suma de VEL más la proporción de VED que se asumen elegibles, mediante la siguiente expresión:[11]

[11] En la simbología empleada en la guía, $\sum_A$ representa la suma sobre todas las unidades que pertenecen al conjunto A.

$$\sum_{EL} w_{0ijk}^{V} + e \sum_{ED} w_{0ijk}^{V}$$

donde,

$e = \frac{\sum_{EL} w_{0ijk}^{V}}{\sum_{EL} w_{0ijk}^{V} + \sum_{NE} w_{0ijk}^{V}}$ es la tasa de elegibilidad,
EL = conjunto de viviendas clasificadas como elegibles,
NE = conjunto de viviendas clasificadas como no elegibles, y
ED = conjunto de viviendas clasificadas como de elegibilidad dudosa.

Los cálculos se realizan dentro de grupos o "clases de ajuste" disjuntas definidas exclusivamente para los cálculos. Estas clases surgen del cruce de la variable provincia o jurisdicción (24) y los estratos de diseño de la MMUVRA para las USM[12] (5). En consecuencia, en cada clase c, $c = 1, \ldots, 120$, el primer factor de ajuste, a_{1c}, es definido por la proporción de viviendas que se estiman como elegibles sobre el total de viviendas estimadas por la encuesta[13] empleando la tasa de elegibilidad e_c dentro de la clase c:

$$a_{1c} = \frac{\sum_{EL(c)} w_{0ijk}^{V} + e_c \sum_{ED(c)} w_{0ijk}^{V}}{\sum_{EL(c)} w_{0ijk}^{V} + \sum_{NE(c)} w_{0ijk}^{V} + \sum_{ED(c)} w_{0ijk}^{V}}$$

En el cuadro 3, se presentan los resultados del Estudio en relación con la cantidad de VEL, VNE y VED, por dominio de estimación y a nivel nacional, que intervienen con sus factores de expansión iniciales, w_{0ijk}^{V}, en los cálculos del factor a_{1c}.

Cuadro 3. Cantidad de viviendas elegibles, no elegibles y de elegibilidad dudosa, por región

Regiones	Viviendas en la muestra	Viviendas elegibles	Viviendas no elegibles	Viviendas de elegibilidad dudosa
Gran Buenos Aires	10.548	6.632	1.029	2.887
Noroeste	6.154	4.897	762	495
Noreste	5.518	4.515	543	460
Cuyo[14]	4.840	3.589	486	765
Pampeana	9.187	6.717	1.151	1.319
Patagonia	4.118	3.307	381	430
Total del país	**40.365**	**29.657**	**4.352**	**6.356**

Fuente: INDEC, Estudio Nacional sobre el Perfil de Personas con Discapacidad 2018.

[12] En rigor, los estratos de USM varían entre dos y cinco, dependiendo del estrato de UPM. Por lo tanto, puede haber 48 o 120 clases de ajuste dependiendo de la UPM en la que se realiza.
[13] En las fórmulas, *EL(c)*, *ED(c)* y *NE(c)* señalan los conjuntos EL, NE y ED restringidos a la clase de ajuste c.
[14] Por cuestiones operativas ajenas al diseño de la muestra, la cantidad de viviendas en San Luis se redujo a 520.

4.2 Ajuste por no respuesta

Cuando se identifica una vivienda como elegible para la encuesta, y por consiguiente los hogares que la componen, no siempre es posible hacer una entrevista, lo cual origina una no respuesta[15] del hogar. Esto puede ocurrir debido a una serie de razones: que en el hogar ninguno quiera responder, que haya ausencia temporal de sus miembros durante el período del Estudio, o bien que hubo un primer contacto, pero por algún motivo o alguna circunstancia fue imposible continuar con la entrevista. En particular, en el Estudio, se considera que un hogar no responde si se registra alguna de las siguientes categorías en "Estado de la vivienda", presente en el cuestionario:

- con todas las personas ausentes (causas circunstanciales, viaje o vacaciones),
- rechazo,
- otras causas.[16]

La no respuesta es un fenómeno siempre presente en una encuesta u operación estadística y es una fuente de sesgo en las estimaciones. En las etapas previas al cálculo de las estimaciones, y para disminuir la incidencia de la no respuesta sobre ellas, se hacen distintos esfuerzos para mantener la tasa de respuesta lo más alta posible. Algunas prácticas habituales son capacitar a los encuestadores con técnicas especiales de abordaje para lograr un cambio de actitud en el entrevistado que rechaza participar y, durante la recolección de los datos, revisitar el hogar con ausentes en varias ocasiones antes de dar concluida la encuesta.

La magnitud del sesgo debido a la falta de respuesta generalmente no se conoce, pero está directamente relacionada con las diferencias entre los grupos de unidades que respondieron y los que no lo hicieron, en las características bajo estudio. También, se ve afectada por un factor asociado a la correlación entre la característica que se indaga sobre la unidad y la probabilidad de propensión a dar respuesta por ella. Por estos motivos, y en un intento de disminuir el efecto del sesgo sobre las estimaciones, se ajustan los factores de expansión de los hogares para compensar la no respuesta alcanzada en la encuesta.

Una de las claves para lograr el éxito del ajuste es poder definir un modelo que explique lo mejor posible el mecanismo de no respuesta que hay por detrás del fenómeno. Habitualmente, y con la ayuda de información disponible tanto para los que responden como para los que no, se emplean clases o grupos de unidades en la población con la ayuda de variables o información auxiliar disponible para todas las unidades.

Desde el punto de vista de la bondad del modelo subyacente y de la eficiencia de los estimadores a emplear para las estimaciones, se busca que las clases:

- permitan sostener en lo posible el supuesto de probabilidad de respuesta constante de las unidades dentro de ellas, y

- sean lo más homogéneas posibles, para que valga en algún grado la hipótesis de que, en una clase dada, los encuestados sean similares a los no encuestados en términos de las principales variables de interés.

[15] Bajo ninguna circunstancia las viviendas seleccionadas para la encuesta son reemplazadas por otras viviendas por razones de no respuesta.
[16] En otras causas si incluyen, por ejemplo, las encuestas que se invalidan en la etapa de edición por inconsistencias severas en la información brindada o por incompletitud insalvable a causa de un error de sincronización del dispositivo móvil empleado para la captura.

Para realizar las correcciones por no respuesta en el Estudio, se emplean las mismas clases conformadas para el ajuste por no elegibilidad, definidas en la sección anterior. A partir de ellas, en cada clase c se obtiene un segundo factor de ajuste, a_{2c}, por "no respuesta" del hogar para los factores de expansión de los hogares[17], definido como:

$$a_{2c} = \frac{\sum_{HR(c)} w^{H}_{0ijkl} a_{1c} + \sum_{HNR(c)} w^{H}_{0ijkl} a_{1c}}{\sum_{HR(c)} w^{H}_{0ijkl} a_{1c}}$$

donde $HR(c)$ y $HNR(c)$ representan los conjuntos de hogares que responden o no a la encuesta en la clase c, respectivamente.

En consecuencia, la expresión del factor de expansión de un hogar seleccionado que responde al Estudio, y después de los dos ajustes realizados viene dado por:

$$\widetilde{w}^{H}_{ijkl} = w^{H}_{0ijkl} a_{1c} a_{2c}$$

Para ilustrar la cantidad de unidades de la muestra que, con sus factores de expansión, se involucran en los cálculos de los factores de ajuste a_{2c}, el siguiente cuadro presenta el total de los hogares con y sin respuesta registrados en el Estudio a nivel nacional[18] y por región.

Cuadro 4. Total de hogares en viviendas clasificadas como elegibles, hogares con y sin respuesta, por región

Regiones	Hogares elegibles	Hogares con respuesta	Hogares sin respuesta
Gran Buenos Aires	6.792	5.555	1.237
Noroeste	5.008	4.691	317
Noreste	4.613	4.346	267
Cuyo	3.645	3.345	300
Pampeana	6.827	6.189	638
Patagonia	3.319	3.021	298
Total del país	**30.204**	**27.147**	**3.057**

Fuente: INDEC, Estudio Nacional sobre el Perfil de Personas con Discapacidad 2018.

[17] Cabe destacar que el factor de expansión inicial correspondiente a un hogar coincide con el de la vivienda de la cual forma parte, o sea, $w^{H}_{0ijkl} = w^{V}_{0ijk}$, dado que se incluyen en la muestra todos los hogares que forman parte de la vivienda seleccionada.

[18] En el Anexo se presenta la tasa de respuesta a nivel nacional y por dominio calculados con el estándar habitual de la AAPOR (2016).

4.3 Ajuste por calibración

Los factores de expansión de cada hogar seleccionado y que responde hasta esta instancia, $\widetilde{w}_{ijkl}^{H}$, reciben una última modificación o ajuste, denominado "calibración". Este procedimiento emplea información auxiliar de una fuente externa disponible, y tiene por objetivo contribuir a una mejora en los ajustes ya realizados, y a corregir posibles sub o sobrerrepresentaciones en algunos grupos de la población, originadas cuando no están bien captados por la encuesta. Para disminuir estas discrepancias, la calibración busca la consistencia entre las estimaciones de algunas variables del Estudio y totales poblacionales conocidos, o *benchmarks,* para esas variables.

La información auxiliar incorporada en la calibración permite definir estimadores más eficientes que el habitual estimador de expansión simple en términos del error muestral, dado que aprovechan la correlación que pueda existir entre las características indagadas por la encuesta y la información provista por la fuente externa.

El proceso de calibración que opera sobre el conjunto de hogares que responden genera el sistema de ponderadores definitivos del Estudio, w_{ijkl}^{H}, que surgen de la resolución del siguiente problema numérico de optimización:

$$\text{Minimizar } \sum_{HR} G(\widetilde{w}_{ijkl}^{H}, w_{ijkl}^{H}),$$
$$\text{sujeto a: } \sum_{HR} w_{ijkl}^{H}\, \mathbf{x}_{\text{ijkl}}^{H} = \sum_{U} \mathbf{x}_{\text{q}}^{H}$$

donde G es una función que define la proximidad entre los factores deseados y los surgidos del último ajuste, y la igualdad propone que las estimaciones para un conjunto de q variables auxiliares, $\mathbf{x}_{\text{ijkl}}^{H} = \left(x_{ijkl\,1}^{H}, \dots, x_{ijklq}^{H}\right)^{T}$ medidas en la encuesta, a partir de los factores de expansión deseados, w_{ijkl}^{H}, reproduzcan sus totales poblacionales, $\sum_{U} \mathbf{x}_{\text{q}}^{H} = (t_{x1}^{H}, \dots, t_{xq}^{H})$, provistos por una fuente externa a la encuesta (Valliant, Dever y Kreuter, 2013).

Dada G, la resolución numérica es un proceso iterativo, que bajo ciertas condiciones de regularidad converge y permite obtener factores de ajuste por calibración, λ_{ijkl}, para cada hogar con respuesta. En el Estudio se emplearon 8 variables que reflejan la estructura demográfica por sexo y por grupos de edad, donde $\mathbf{x}_{\text{ijkl}}^{H} = (x_{ijkl\,1}^{H}, \dots, x_{ijkl\,8}^{H})$ y cuyas componentes son:

$x_{ijkl\,1}^{H}$ = cantidad de mujeres en el hogar,
$x_{ijkl\,2}^{H}$ = cantidad de varones en el hogar,
$x_{ijkl\,3}^{H}$ = cantidad de personas en el hogar entre 0 y 4 años,
$x_{ijkl\,4}^{H}$ = cantidad de personas en el hogar entre 5 y 14 años,
$x_{ijkl\,5}^{H}$ = cantidad de personas en el hogar entre 15 y 39 años,
$x_{ijkl\,6}^{H}$ = cantidad de personas en el hogar entre 40 y 64 años,
$x_{ijkl\,7}^{H}$ = cantidad de personas en el hogar entre 65 y 79 años,
$x_{ijkl\,8}^{H}$ = cantidad de personas en el hogar de 80 años y más,

y los totales de población, involucrados como marginales para estas variables en el proceso iterativo, provienen de proyecciones poblacionales.[19]

Para la calibración en el Estudio, se emplea la función de distancia "logit" (Deville y Särndal, 1992; Haziza y Beaumont, 2017) del paquete *survey* de R (Lumley, 2010), que permite controlar el rango de los w_{ijkl}^{H}, y así sus valores extremos. De esta forma se limita el riesgo de incrementar el error de muestreo en las estimaciones del Estudio a causa del ajuste.

El proceso de calibración se efectúa en forma independiente por provincia o jurisdicción y, en lo posible, el ajuste involucra los totales proyectados por sexo y grupos de edad según la división aglomerado EPH y resto de las UPM dentro de la provincia en cuestión. La expresión definitiva del factor de expansión de un hogar seleccionado, que responde a la encuesta y que incluye todos los ajustes, viene dada por:

$$w_{ijkl}^{H} = \widetilde{w}_{ijkl}^{H}\lambda_{ijkl} = w_{0ijkl}^{H}a_{1c}a_{2c}\lambda_{ijkl}$$

donde:

w_{0ijkl}^{H} es el factor de expansión inicial del l-ésimo hogar, de la k-ésima vivienda ubicada en la j-ésima USM dentro de la i-ésima UPM,

a_{1c} es el factor de corrección por viviendas no elegibles perteneciente a la clase c de ajuste,

a_{2c} es el factor de corrección por no respuesta del hogar perteneciente a la clase c de ajuste,

λ_{ijkl}*es el* factor de ajuste que surge de la calibración correspondiente al hogar l-ésimo, de la k-ésima vivienda ubicada en la j-ésima USM dentro de la i-ésima UPM.

Esta formulación es válida siempre que la vivienda y el hogar seleccionado pertenezca a la clase de ajuste c, $c = 1, \ldots, 120$.

Para evitar producir dos conjuntos de pesos finales para la encuesta, uno para hogares y otro para personas, en el Estudio se emplea un método de calibración integrado que origina un peso único, que permite estimaciones de parámetros tanto a nivel de personas como de hogares (Lemaître y Dufour, 1987). Es decir que el factor de expansión final para la persona residente del l-ésimo hogar que se aplica para todas las estimaciones del Estudio es w_{ijkl}^{H}.

Por último, los pesos que surgen del proceso iterativo de la calibración son tratados por un algoritmo de redondeo para eliminar la componente decimal, lo que da origen a los w_{ijkl}^{H} finales que se emplean para todas las estimaciones oficiales del Estudio.

[19] Los totales poblacionales proyectados fueron calculados a partir de datos censales de población según el CNPHyV 2010 al 15 de mayo de 2018 y determinados por la Dirección Nacional de Estadísticas Sociales y Poblacionales del INDEC.

5. Imputación de datos faltantes

Como se mencionó en secciones anteriores, la no respuesta es un problema común en casi todas las encuestas, pero indeseable, porque influye en la calidad de la inferencia. Es habitual distinguir la no respuesta total de la no respuesta parcial.

En el primer caso, la unidad seleccionada no responde ninguna pregunta del Estudio, lo que se traduce en una pérdida total de información. El rechazo a participar o la inhabilidad para establecer el contacto con el informante son las causas más comunes para este tipo de no respuesta.

En cambio, la no respuesta parcial ocurre cuando la ausencia de información se limita a algunas variables o algunos ítems del Estudio, ya sea porque el individuo no contesta algunas preguntas o porque la información que brinda para ellas es rechazada por una inconsistencia detectada durante la etapa de revisión y edición y es, por lo tanto, considerada inválida.

La reponderación y la imputación son dos técnicas de corrección de la no respuesta una vez concluida la encuesta. Por lo general, la reponderación es empleada para compensar la no respuesta total, y consiste en incrementar los factores de expansión de los que responden la encuesta para tener en cuenta a los que no lo hacen (ver sección 4.2).

Cuando la no respuesta es parcial, la imputación es el método más habitual para su tratamiento y consiste en reemplazar los datos faltantes o inconsistentes por uno determinado para completarlo. El principal objetivo de cualquier técnica de imputación es disminuir en lo posible el sesgo por no respuesta, y al mismo tiempo proveer un conjunto de datos completos que permita obtener resultados consistentes para los distintos tipos de análisis que surgen a partir del Estudio, siendo esta una propiedad atractiva desde el punto de vista de los usuarios de los datos.

Sin embargo, para alcanzarlo se deben asumir algunos riesgos, dado que por lo general las técnicas se sostienen empleando una serie de supuestos de difícil verificación en la práctica. Entre ellos, está el que asume que el mecanismo o patrón de respuesta desconocido que origina los datos faltantes en una variable depende de los datos observados y puede ser explicado por variables propias del Estudio.

5.1 Procedimientos de imputación aplicados en la encuesta

En la etapa de imputación de una encuesta, la tarea central reside en seleccionar uno o varios procedimientos de imputación que se consideren apropiados para completar valores faltantes. Estos, en lo posible, deben poseer un grado de automatización que favorezca el flujo de procesamiento del Estudio para alcanzar las estimaciones, deben poder reproducirse bajo cualquier circunstancia, y deben emplear en forma eficiente la información válida y disponible de las unidades que responden y de las que no responden.

Los procedimientos se clasifican en dos grandes grupos. Los determinísticos son aquellos que, al repetir el procedimiento de imputación basado en el mismo conjunto de unidades que responden, llevan siempre al mismo conjunto de datos completos. Ejemplos de este tipo son: la imputación por el promedio, por el cociente, o por el vecino más cercano; también se incluyen en este grupo los deductivos, o sea aquellos que, por información brindada por la unidad en otros ítems o variables del cuestionario, permiten deducir el valor faltante.

En contraste, los procedimientos de imputación aleatorios (o estocásticos) son aquellos que llevan a un conjunto de datos completos distinto cada vez que el proceso de imputación es repetido. En general estos métodos pueden ser vistos como una imputación determinística más la suma de una componente aleatoria. Una de sus características es que tienden a preservar la distribución de la variable que requiere imputación, pero incorporan una componente adicional de error en la estimación debido al empleo de un mecanismo de imputación aleatorio.

Para completar los datos faltantes, el método adoptado para el Estudio, salvo en situaciones marginales, se basa en la imputación por Hot Deck. Su versión más general consiste en reemplazar el valor faltante de una o más variables de una unidad que no respondió (receptor) por valores observados de otra que tiene la información (donante), y que es similar al receptor con respecto a características observadas en ambos casos (Andridge y Little, 2010).

Este método posee una serie de ventajas: es simple, existe un gran número de herramientas informáticas para ponerlo en práctica y no obliga a asumir un modelo paramétrico para obtener el valor imputado.

Para poner en práctica distintas variantes del Hot Deck, se debe clasificar a los donantes y receptores en clases o celdas de imputación constituidas a partir de variables auxiliares propias del Estudio. Una manera habitual para definirlas es a partir de la clasificación cruzada de variables discretas, lo que obliga a categorizar las variables continuas (si se desea considerarlas como variables de clasificación) y controlar el número de variables para no generar celdas con pocos o ningún donante.

Si la estrategia es incorporar la mayor cantidad de variables auxiliares disponibles al procedimiento, asumiendo que beneficia la reducción del sesgo por no respuesta, una posibilidad para remediar la falta de donantes en las celdas es agrupar categorías o eliminar variables hasta obtener donantes. Esta modalidad, conocida como **"**Hot Deck jerárquico", lleva a determinar *a priori* la manera de priorizar las variables a resignar en el proceso de eliminación, y cómo proceder ante la necesidad de realizar los agrupamientos entre celdas cuando no hay donantes.

En el Estudio se realizaron imputaciones a través de Hot Deck jerárquico en los módulos C y D[20] del cuestionario.[21] Para el módulo C, la imputación busca resolver un problema de no respuesta total al módulo, es decir, cuando el módulo debía contar con información y por distintas causas no la tenía: en cambio, en el módulo D, solo afecta a un conjunto de preguntas del módulo con no respuesta parcial.

5.2 Imputación del módulo C

En este módulo, se imputaron por no respuesta total a 449 personas que debían tener un módulo C sobre un total de 8127 (5,5%), y que, por razones de edición, consistencia, o pérdida de datos en la sincronización entre el dispositivo móvil y el sistema de recepción de los datos, no fueron completados. Se emplearon 4 variables de clasificación: provincia o jurisdicción (24 categorías), sexo (2 categorías), grupos de edad (6 categorías), y cantidad y tipo de dificultad (10 categorías) para definir las celdas de imputación; y una jerarquía inducida por 3 niveles en el orden de resignar variables luego de agotar el proceso de reagrupamiento de las celdas en una misma variable:

[20] Para el módulo A no se realizaron imputaciones y para el Módulo B se imputaron 637 personas sobre un total de 82.327 (0,77%).
[21] Ver cuestionario de la encuesta https://www.indec.gob.ar/ftp/cuadros/poblacion/cuestionario_enpd_2018.pdf.

- nivel 1: provincia, sexo, grupo de edad, y cantidad y tipo de dificultad
- nivel 2: sexo, grupo de edad, y cantidad y tipo de dificultad
- nivel 3: sexo y grupo de edad

con la pauta general de que, si en la celda de un receptor dado hay menos de 10 donantes, se pasa a generar un reagrupamiento de la celda con la contigua hasta conseguir 10 donantes o más.

5.3 Imputación del módulo D

En este módulo se imputan hogares en lugar de personas, por tratarse de un módulo de características del hogar. En total, se imputaron 205 hogares sobre un total de 6649 (3,1%). Las variables del Estudio que determinaron las celdas iniciales de imputación fueron: provincia o jurisdicción (24 categorías), UPM (de 2 a 21 categorías, dependiendo de la provincia o jurisdicción) y cantidad de miembros del hogar (3 categorías); la jerarquía de resignación de variables es definida por los siguientes 3 niveles:

- nivel 1: UPM, provincia, y cantidad de miembros del hogar
- nivel 2: provincia y cantidad de miembros del hogar
- nivel 3: cantidad de miembros del hogar.

Si en la celda de un receptor dado hay menos de 10 donantes, se aplica la misma pauta señalada para la imputación del módulo C.

6. Estimación a partir de los datos del Estudio

El proceso inferencial por el cual se obtienen aproximaciones a los parámetros desconocidos de la población bajo estudio a partir de los datos de una muestra se denomina estimación. Los parámetros poblacionales que resultan de interés a estimar son, por lo general, descriptivos, y la mayoría se puede definir a partir de totales: los promedios, las proporciones y las razones o tasas. No obstante, puede haber interés en otros que involucran, por ejemplo, a estadísticos de orden o más complejos.

Para alcanzar las estimaciones de esos parámetros en el Estudio se emplean estimadores que recurren a los factores de expansión w_{ijkl}^{H}, que surgen de la última etapa de ajuste y que pertenecen al tipo de estimadores “calibrados”.

En el caso de que Y y Z sean variables o características de interés del Estudio medidas a nivel de personas y teniendo presente que w_{ijkl}^{H} es el factor de expansión para todas aquellas que componen un hogar, como se señala en la sección 4, la expresión de los estimadores más empleados en el Estudio toman la siguiente forma:

Parámetro	Estimador [22]
Total, $\boldsymbol{t_y}$	$\hat{t}_y = \sum_R w_{ijkl}^H y_{ijklp},$
Promedio[23], $\bar{\boldsymbol{y}}$	$\hat{\bar{y}} = \frac{\sum_R w_{ijkl}^H y_{ijklp}}{\sum_R w_{ijkl}^H}$
Proporción, $\boldsymbol{p}$	$\hat{p} = \frac{\sum_R w_{ijkl}^H y_{ijklp}}{\sum_R w_{ijkl}^H}$
Razón, $\boldsymbol{R_{yz}} = \frac{\boldsymbol{t_y}}{\boldsymbol{t_z}}$	$\hat{R}_{yz} = \frac{\hat{t}_y}{\hat{t}_z} = \frac{\sum_R w_{ijkl}^H y_{ijklp}}{\sum_R w_{ijkl}^H z_{ijklp}}$

7. Indicadores de calidad asociados con el error de muestreo

Una de las etapas centrales de toda encuesta es la que evalúa la calidad de los datos, o sea, el proceso de analizar el producto final en términos de precisión y confiabilidad. Contar con indicadores de calidad de una encuesta permite a los usuarios cuantificar el grado de confianza y conocer las limitaciones que pueden llegar a tener los resultados, y restringir así su uso cuando las estimaciones no alcanzan ciertos estándares definidos para la encuesta.

En un estudio que emplea una muestra probabilística como el Estudio, la inferencia estadística sobre la población objetivo se basa en los datos recopilados de solo una parte de esta población. Es así que los resultados probablemente diferirán de los que se pueden obtener a partir de un censo completo.

El error que se genera al extraer conclusiones en términos estadísticos para toda la población basadas solo en una muestra se denomina “error de muestreo”, y es necesario tenerlo en cuenta en todo el proceso inferencial. El efecto que tiene en las estimaciones del Estudio depende de algunos aspectos del diseño muestral tales como el número de etapas y el método de selección, también del tamaño de la muestra, del estimador que se emplea y de la variabilidad propia de la característica de interés que se mide.

Por lo general, a medida que aumenta la muestra, y el resto de los factores intervinientes se mantienen constantes, se espera que su magnitud disminuya. Esto es consistente con el hecho de que debería ser cero una vez que se censa a toda la población. Difiere de una variable a otra, siendo en general mayor para características relativamente raras o cuando no se distribuye con cierto grado de uniformidad en la población.

[22] En todos los casos, $\sum_R$ en las fórmulas hace referencia a sumar sobre las personas que responden a la encuesta.
[23] La definición de los parámetros promedio y proporción coincide si Y es una variable binaria, que toma el valor de 1 cuando el individuo posee una característica dada y 0, en caso contrario.

Una medida del error de muestreo es la varianza muestral del estimador. Esta representa la variabilidad de las estimaciones que se obtienen a partir de todas las muestras posibles según el diseño muestral, con respecto al promedio de las estimaciones.

A partir de la varianza muestral se pueden definir otras medidas más populares tales como el error estándar (EE) y el coeficiente de variación (CV), o más complejas de interpretar, por ejemplo, el efecto de diseño (ED) o el intervalo de confianza (IC). Cuanto más pequeños son el EE, el CV, el ED o la amplitud del IC, más precisa es la estimación.

El EE se define como la raíz cuadrada de la varianza muestral del estimador. A diferencia de la varianza, el EE es medido en las mismas unidades de escala de la característica, lo cual facilita su interpretación. En cambio, el CV se define como el cociente entre el EE y el estimador. No depende de las unidades en que se mide la estimación, en virtud de que es una medida relativa a esta. Generalmente se lo expresa como un porcentaje y, en la práctica, una estimación del CV es una de las más empleadas para informar el error de muestreo de las estimaciones de una encuesta.

Aunque el concepto de varianza se basa en la idea de seleccionar todas las muestras posibles según el diseño muestral, en la práctica solo se extrae una, a partir de la cual la varianza puede ser estimada. Dada la importancia que tiene en cualquier estudio por muestreo, es central su estimación como indicador de la calidad de las estimaciones en una encuesta.

8. Estimación de los errores de muestro mediante replicaciones

La complejidad del diseño de la muestra y del método de estimación empleados para la encuesta presenta un desafío particular a la hora de estimar la varianza, debido a la dificultad para obtener su expresión analítica. Sin embargo, el aumento de la eficiencia informática ha hecho posible el uso de técnicas que emplean réplicas para resolver el problema.

Estos métodos son fáciles de implementar porque siempre utilizan el mismo proceso de estimación repitiéndolo muchas veces y no requieren de una fórmula analítica del estimador de la varianza muestral.

Por eso, para los cálculos que cuantifican el error por muestra en la encuesta se ha implementado una metodología con base en replicaciones. La idea básica de esta estrategia es tratar el conjunto de datos de la muestra como si esta fuera la población y generar de una manera sistemática un conjunto de submuestras que pueden emplearse para estimar el error muestral en las estimaciones.

El proceso de cálculo puede ser implementado de manera eficiente, aun por usuarios con pocos conocimientos en muestreo, sumando una serie de pesos replicados al conjunto de datos que se emplea para obtener los resultados del Estudio. Además de las razones señaladas, existen otras por las cuales se opta por emplear esta metodología, entre ellas:

- incluir en la etapa de la conformación de las réplicas el conjunto de ajustes que sufren los factores de expansión iniciales (no elegibilidad, no respuesta y calibración), para incorporar la variabilidad propia de estas correcciones en los cálculos del error de muestreo y que resultan dificultosas con otros métodos;
- brindar una solución al problema de obtener estimaciones del error por muestra para un número diverso de estimadores, incluyendo los de orden (mediana, deciles, percentiles,

etc.) o los de desigualdad (índice de Gini, curva de Lorentz, etc.), que en otros métodos son complejos para implementar;

- habilitar a los usuarios a calcular por sus propios medios los errores de muestreo para sus estimaciones, con transparencia y de la misma manera en que los obtiene el Instituto, sin tener que depender de tablas u otros elementos para cuantificarlos;
- proteger y anonimizar cierta información que puede vulnerar el secreto estadístico que pesa sobre el microdato, por ejemplo, al no involucrar al usuario con las variables que definen el diseño muestral (estratos, UPM, USM), y que son necesarias para determinar el error de muestreo en una estimación.

Existen distintos métodos para conformar las réplicas (Wolter, 2007), y el que se adopta para generar las submuestras en el Estudio es el *bootstrap* propuesto en Rao y Wu (1988) y en Rao, Wu y Yue (1992). Su formulación más general consiste en definir B submuestras *bootstrap* independientes de la muestra original. Para cada submuestra $b, b = 1, \dots, B$, el procedimiento lleva a que en cada estrato de diseño, h, se seleccione una muestra simple al azar con reemplazo de $n_h - 1$ conglomerados a partir de la muestra original de n_h conglomerados. Se define el peso *bootstrap* $w_{hml}^{*(b)}$ a partir de un peso inicial w_{hml} para la g-ésima unidad en el conglomerado m del estrato h en la réplica b según el siguiente ajuste:

$$w_{hmg}^{*(b)} = \frac{n_h}{n_h - 1} m_{hm}^{*(b)} \, w_{hmg}$$

donde $m_{hm}^{*(b)}$ es el número de veces que el conglomerado m del estrato h fue seleccionado en la réplica b.

Estos pesos replicados *bootstrap* permiten calcular la estimación de interés en cada una de las B submuestras, y con la variabilidad de los resultados obtenidos se calcula una medida del error muestral para la estimación en cuestión. A tal efecto, se define la varianza *bootstrap* de $\hat{\theta}$ a partir de las réplicas como:

$$v_B(\hat{\theta}) = \frac{1}{B} \sum_{b=1}^{B} \left(\hat{\theta}_{(b)}^{*} - \hat{\theta} \right)^2, \qquad [1]$$

donde:

$\hat{\theta}$ es el estimador[24] de θ calculado a partir de los ponderadores w_{hmg} definidos para la muestra; y θ, un parámetro poblacional de interés para una característica dada,

y

$\hat{\theta}_{(b)}^{*}$ es el estimador de θ a partir de los ponderadores $w_{hmg}^{*(b)}$ de la réplica $b, b = 1, \dots, B$.

De $[1]$ es inmediato obtener el del error estándar:

$$ee_B(\hat{\theta}) = \sqrt{v_B(\hat{\theta})} \qquad [2]$$

[24] Ver apartado 6.

y el del coeficiente de variación:

$$cv_B(\hat{\theta}) = \frac{ee_B(\hat{\theta})}{\hat{\theta}} \qquad [3]$$

El método en su formulación teórica es propuesto para diseños estratificados multietápicos, con UPM seleccionadas mediante probabilidad proporcional a un tamaño (PPT) con reemplazo, y asumiendo una expresión para la varianza bajo un diseño con reposición con el supuesto de "último conglomerado". Este último sostiene que la primera etapa de muestreo (UPM) brinda la información necesaria para alcanzar una estimación del error por muestra, ignorando las restantes etapas definidas en el diseño.

Sin embargo, la adopción de estos supuestos habilita emplearlo como un estimador de varianza para un diseño PPT sin reemplazo, si la selección de las UPM sin reemplazo es más eficiente que la selección de UPM con reemplazo (West, 2012; Särndal, Swensson y Wretman,1992), como es el caso del Estudio, lo que convierte el proceso inferencial en conservador y válido para la encuesta.

Las réplicas para calcular la estimación de la varianza o del error por muestra en el Estudio fueron determinadas en forma independiente en cada jurisdicción. Para ajustarse a los requerimientos del método, en las UPM autorrepresentadas del Estudio, los estratos para el procedimiento *bootstrap* quedaron definidos por el estrato de la segunda etapa de muestreo y los últimos conglomerados por las USM; en cambio, en las UPM no autorrepresentadas, los estratos *bootstrap* se corresponden con el estrato de las UPM y los últimos conglomerados, con las UPM.

Para obtener estimaciones de varianza estables para varios tipos de análisis, deberían estar disponibles tantas réplicas como sea posible. Sin embargo, se debe alcanzar un compromiso entre garantizar la estabilidad, controlar el tamaño de la base con las réplicas y limitar el tiempo de cálculo, entre otras cuestiones. Por estos motivos, en el Estudio, el total de réplicas es de 300 (B=300), cantidad que asegura la estabilidad del estimador de varianza para las principales estimaciones.

Todas las réplicas se obtienen de la muestra original, que incluye a todos los hogares y las personas de las viviendas elegibles, cuyos factores iniciales vienen dados por $w_{0ijkl}^{H} a_{1c}$. Estos pasan a ser corregidos o reescalados según el estrato h y el último conglomerado m al cual pertenece el hogar, como lo requiere el procedimiento *bootstrap* descripto.

Con el fin de incorporar en la variabilidad que introducen los ajustes efectuados en los factores de expansión del Estudio, se repiten los mismos ajustes sobre los pesos replicados. Es decir, para cada una de las 300 réplicas, los pesos *bootstrap* son ajustados nuevamente por no respuesta y calibrados por sexo y edad de manera análoga a como lo fueron los pesos originales, como se detalla en la sección 4, correspondientes a cada uno de los pasos del Estudio. A diferencia de los pesos originales, los pesos *bootstrap* no son sometidos a un proceso de redondeo.

9. Modo de empleo de los pesos replicados

Como consecuencia de todo el procedimiento detallado en la sección anterior, el Estudio dispone de un conjunto de 300 réplicas,

$$\left\{w_{ijkl}^{*H(b)}, b = 1, \dots, 300\right\},$$

que, vinculadas a la base con los microdatos, permiten calcular los errores muestrales para las estimaciones oficiales del Estudio.

La presente sección constituye una guía de cómo deben ser empleadas las réplicas en distintas herramientas de cálculo: R[25], SAS[26], Stata[27] y Wesvar.[28] En caso de no contar con ellas, se presenta un ejemplo que sugiere cómo efectuar el cálculo siguiendo la definición formulada en [1] del apartado 8, y que cualquier usuario puede poner en práctica con pocos recursos.[29]

Se advierte que la guía no constituye un manual exhaustivo de cada una de las herramientas y sus opciones, y que es aconsejable que el usuario tenga una mínima experiencia en aquella que va a emplear. En resumen, se trata de cubrir los aspectos que hacen a la estimación de los errores muestrales bajo la metodología adoptada con el objetivo de orientar al usuario para lograr esta estimación.

Por otro lado, solo se incluyen los códigos que brindan las estimaciones puntuales, y el que permite alcanzar una medida del error vía el error estándar o el coeficiente de variación. Se consideran en los ejemplos, la estimación de los parámetros definidos en la sección 6.

Para facilitar las indicaciones, la presentación adopta la notación empleada en (1) la base usuario del Estudio y (2) en la base con las réplicas, para las principales variables de interés para este apartado:[30]

- $\boldsymbol{id}$: variable de identificación de registro, presente en (1) y (2);
- $\boldsymbol{pondera}$: factor de expansión final del Estudio,[31] presente en (1);
- $\boldsymbol{w\_repb}$: peso *bootstrap* replicado, donde b representa el número de réplica al cual corresponden los pesos, tomando los valores de 1 a 300, presente en (2).

Por otro lado, se asume que $\boldsymbol{Y}, \boldsymbol{Z}$ son variables genéricas (continuas o categóricas), que hacen referencia a características indagadas por el Estudio para las cuales se requieren estimaciones de los parámetros poblacionales de interés (ver sección 6), y de sus respectivas estimaciones de los errores de muestreo.

[25] www.r-project.org. Versión 3.6.
[26] www.sas.com. Versión 9.4 M3.
[27] www.stata.com. Versión 15.
[28] www.westat.com/capability/information-systems-software/wesvar. Versión 5.1.
[29] No se incluye a la herramienta de cálculo SPSS, ya que no cuenta oficialmente a la fecha con la posibilidad de emplear la metodología desarrollada sin recurrir a una programación *ad hoc*.

[30] Se sugiere al usuario leer el *Manual de uso de la base de datos usuario* correspondiente al Estudio para los detalles, así como los diccionarios vinculados a cada base.

[31] Etiqueta que hace referencia al factor de expansión w_{ijkl}^{H}, definido en el apartado 4.

Para poder seguir correctamente estas instrucciones, primero es necesario vincular la base de los microdatos del Estudio con la de las réplicas de manera unívoca, a través de la variable identificadora ***id*** presente en ambas bases, y componer una nueva base para los cálculos. Esta base, de ahora en más ***base_Estudio***, se la puede construir por lo general a través de la sentencia u opción **merge** en la mayoría de las herramientas de cálculo propuestas.

Como resultado, cada unidad (persona u hogar) o registro de la base usuario poseerá su factor de expansión asociado (***pondera***) y cada uno de los 300 valores de los pesos *bootstrap* replicados (***w_repb***).

9.1 Cálculo del error de muestreo a través de R

Una de las posibilidades disponibles, y que acepta la metodología propuesta en esta guía, es el paquete *survey*[32]. Siguiendo las indicaciones del manual,[33] y asumiendo que ***base_Estudio*** fue creada como se indica en el paso anterior, se define el objeto **disenio**,[34] que incluye las componentes que se requieren para los cálculos a través de la opción ***svrepdesign***.

En ***svrepdesign*** se invocan el factor de expansión del Estudio (***pondera***), el método que generó las réplicas (***bootstrap***), el conjunto de replicaciones (***w_rep[1-9]+***) que se encuentran en la base, y la opción ***mse=T***. Estas indicaciones preparan a la herramienta para obtener las estimaciones, también las del error de muestreo, bajo las siguientes sentencias:

```
library(survey)
disenio=svrepdesign(data=base_Estudio,
                    weights=~pondera,
                    repweights="w_rep[1-9]+",
                    type="bootstrap", mse=T)
```

A manera de ejemplo, se detallan los códigos que brindan la estimación puntual y la del error estándar a partir de los pesos *bootstrap*, respetando la metodología adoptada. Se suma también la función que permite la estimación del CV correspondiente a la estimación en cuestión:

Estimador	Estimaciones por Survey
$\hat{t}_y$	svytotal(~Y,design=***disenio***) cv(svytotal(~Y,design=***disenio***))
$\hat{\bar{y}}$	svymean(~Y,design=***disenio***) cv(svymean(~Y,design=***disenio***))
$\hat{p}$	svymean(~as.factor(Y),design=***disenio***) cv(svymean(~as.factor(Y),design=***disenio***))
$\hat{R}_{YZ}$	svyratio(~Y,~Z,disenio) cv(svyratio(~Y,~Z,disenio))

[32] https://cran.r-project.org/web/packages/survey/index.html. Versión 3.36.
[33] https://cran.r-project.org/web/packages/survey/survey.pdf.
[34] El usuario puede optar por cualquier otro nombre para el objeto.

9.2 Cálculo del error de muestreo a través de Stata

Esta herramienta estadística presenta un módulo específico para efectuar estimaciones y análisis de datos provenientes de encuestas con diseños complejos. Las indicaciones que se brindan están habilitadas a partir de la versión 12 o superior (StataCorp, 2017). Stata permite operar con menús desplegables o bien vía sentencias o comandos; esta última es la forma que se adopta para la presentación.

Asumiendo que el usuario incorporó a ***base_Estudio*** en el entorno de Stata, el comando ***svyset*** es el que se emplea para gestionar los cálculos para las estimaciones. En él se deben identificar el factor de expansión del Estudio ***pondera***, los pesos replicados ***w_rep****, y el método para el cálculo de la varianza ***bootstrap***. Asimismo, se debe incluir la opción ***mse*** para obtener el estimador de varianza *bootstrap* considerado en la sección 8. Para preparar a la herramienta para las estimaciones el usuario debe invocar:

svyset [pw=***pondera***], bsrweight(***w_rep****) vce(***bootstrap***) ***mse***

A continuación, y habiendo definido a ***svyset***, se debe emplear el prefijo ***svy*** para las estimaciones de los parámetros y de los errores de muestreo asociados. A manera de ejemplo, se muestran los códigos correspondientes para la estimación de un total, una media, una proporción y una razón:

Estimador	Estimaciones por Stata
$\hat{t}_y$	***svy*** bootstrap : total Y estat cv
$\hat{\bar{y}}$	***svy*** bootstrap : mean Y estat cv
$\hat{p}$	***svy*** bootstrap : proportion Y estat cv
$\hat{R}_{YZ}$	***svy*** bootstrap : ratio (Y/Z) estat cv

En respuesta a la primera línea del código, y para cada caso, la herramienta brinda el resultado de la estimación del parámetro, la estimación de su error estándar a través del método *bootstrap*, y los límites para el IC del 95% para la estimación. La segunda línea de código (*estat cv*) permite obtener una aproximación al CV de la estimación.

En el caso de que se disponga de la versión 10 de Stata, se debe proceder como se indicó en los párrafos anteriores, pero se tendrá que invocar al prefijo ***svyset*** con la opción ***brrweight***, y ***brr*** en la opción ***vce***. De esta forma, se podrán obtener estimaciones válidas para el EE, CV o IC, al no contar en esa versión con la opción ***bootstrap***. En la versión 9 o anteriores, la herramienta no cuenta con el prefijo ***svy*** para invocar estimaciones con pesos replicados, y obliga a cambiar el procedimiento para obtener estimaciones de varianzas (Chowhan y Buckley, 2005).

9.3 Cálculo del error de muestreo a través de SAS

La herramienta para el análisis estadístico, SAS, emplea procedimientos específicos para el tratamiento de datos provenientes de muestras con diseños complejos. La componente SAS/STAT (SAS Institute, 2017) incluye los procedimientos ***surveymeans*** y ***surveyfreq*** que permiten brindar estimaciones de parámetros descriptivos de una población.

Habiendo incorporado ***base_Estudio*** al entorno SAS, la opción a emplear en cualquiera de los procedimientos para alcanzar los errores muestrales es *varmethod*=***Bootstrap***, invocando los pesos replicados ***w_rep1--w_rep300*** vía *repweight* y al factor de expansión para las estimaciones ***pondera*** en *weight*. En particular, para la estimación de los parámetros señalados se presentan los siguientes códigos orientativos:

Estimador	Estimaciones por SAS
$\hat{t}_y$	proc surveymeans data=***base_Estudio*** sum cvsum varmethod=***Bootstrap***; repweight ***w_rep1--w_rep300***; weight ***pondera***; var Y; run;
$\hat{\bar{y}}$	proc surveymeans data=***base_Estudio*** mean cv varmethod= ***Bootstrap***; repweight ***w_rep1--w_rep300***; weight ***pondera***; var Y; run;
$\hat{p}$	proc surveyfreq data= ***base_Estudio*** varmethod=***Bootstrap***; repweight ***w_rep1--w_rep300***; weight ***pondera***; table Y; run;
$\hat{R}_{YZ}$	proc surveymeans data=***base_Estudio*** varmethod=***Bootstrap***; repweight ***w_rep1--w_rep300***; weight ***pondera***; ratio Y/Z; run;

Se advierte que el método *bootstrap* para el cálculo de errores por muestra para diseños complejos está disponible desde la versión 14.3 del componente SAS/STAT (SAS v.9.4 M3). En versiones anteriores, los usuarios podrán indicar ***BRR*** en *varmethod* como método de estimación de varianza, ya que esta opción permite obtener resultados válidos para hacer inferencia con los pesos *bootstrap* (Gagné, Roberts y Keown, 2014).

9.4 Cálculo del error de muestreo a través de WESVAR

Wesvar[35] es una herramienta estadística con una opción de descarga libre al igual que R. Fue desarrollada por la empresa Westat y permite emplear la metodología de cálculo de errores por muestra con base en replicaciones (Brick, Morganstein y Valliant, 2000). A continuación, se brinda una descripción sencilla de cómo operar con ella y de las opciones básicas que hay que invocar, empleando la versión 5.1.19.

[35] www.westat.com/capability/information-systems-software/wesvar. Se puede acceder de forma gratuita a la documentación de Wesvar enviando un correo electrónico a: wesvartechsupport@westat.com.

Debe advertirse que, a diferencia de las demás herramientas presentadas en esta sección, Wesvar no dispone de medios en su entorno para vincular la base usuario con la base de réplicas. Es por eso que se sugiere al usuario vincular estas bases por otros medios, por ejemplo, mediante alguna otra de las herramientas presentadas, antes de proceder con las instrucciones de esta sección.

En la figura 1, se observa la ventana de inicio donde aparece el árbol de actividades y opciones que guían al usuario dentro de la herramienta. En primera instancia, se debe crear una base de datos Wesvar (.var) a partir de la base del Estudio con las réplicas (***base_Estudio***), con el objetivo de utilizarla para realizar los análisis o las estimaciones. Para esto el usuario deberá hacer clic en "New Wesvar Data File", y elegir la base en la carpeta o espacio de trabajo donde se encuentra.[36]

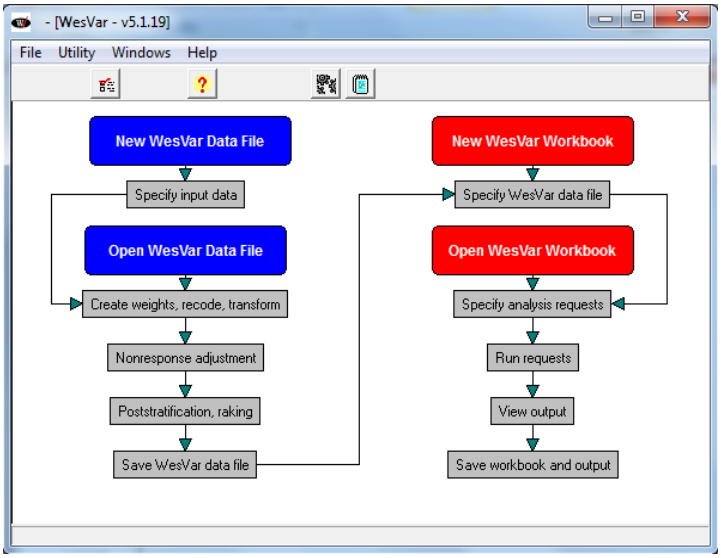


Figura 1

Al usuario le aparece una ventana como la que se ve en la figura 2, donde debe completar la información necesaria para iniciar a operar con las estimaciones. En el apartado **Variables** se deben indicar aquellas del panel **Source Variables** para las cuales se requieren estimaciones de parámetros. En **Replicates** se deben incluir las variables correspondientes a los pesos replicados de las muestras *bootstrap* del Estudio, **w_rep1,..,w_rep300**; y en el apartado **Full Sample**, el factor de expansión final del Estudio, ***pondera***. En **Method** se debe optar por **BRR**, que brinda resultados válidos para las estimaciones de los errores de muestreo empleando los pesos *bootstrap* del Estudio (Phillips, 2004).

Una vez hecha la asignación, se procede a guardar la base Wesvar generada en la carpeta de trabajo que emplea el usuario, quien ya queda en condiciones de continuar con las estimaciones.

[36] Se advierte que la herramienta tiene la posibilidad de importar datos en formato csv/txt con delimitadores, SAS o SPSS.

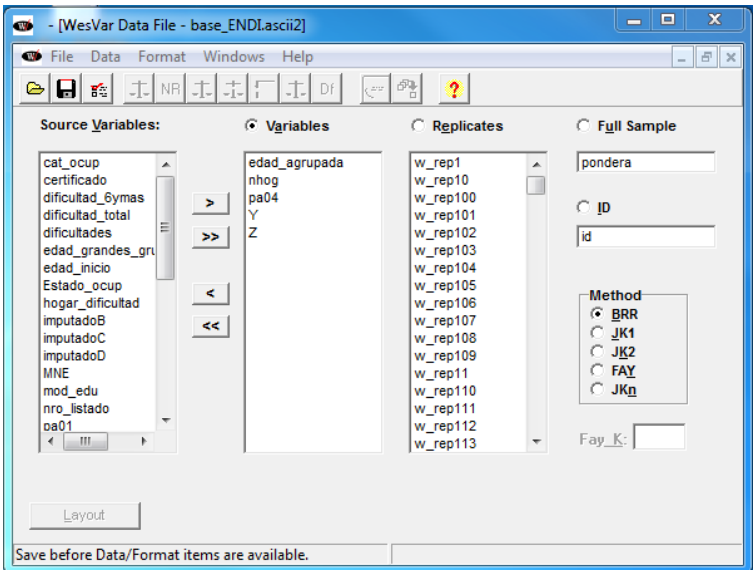


Figura 2

En el paso siguiente se debe crear un libro de trabajo haciendo clic sobre la etiqueta *New Wesvar Workbook* (figura 1), que obliga al usuario a seleccionar la base Wesvar constituida según lo detallado en los párrafos anteriores.

En la figura 3, se presenta la ventana a partir de la cual Wesvar permite gestionar los distintos análisis o las distintas estimaciones que el usuario desea llevar a cabo. Dicha ventana está dividida en dos paneles. El de la izquierda permite visualizar el árbol de trabajo que progresa a medida que se van introduciendo requerimientos de estimaciones o cálculos. En cambio, el panel derecho se emplea para definir y cambiar los análisis o los tipos de estimaciones que ofrece la herramienta: tablas con totales o frecuencias, modelos de regresión o estadísticos descriptivos (***Table, Regression, Descriptive Stats***), respectivamente.

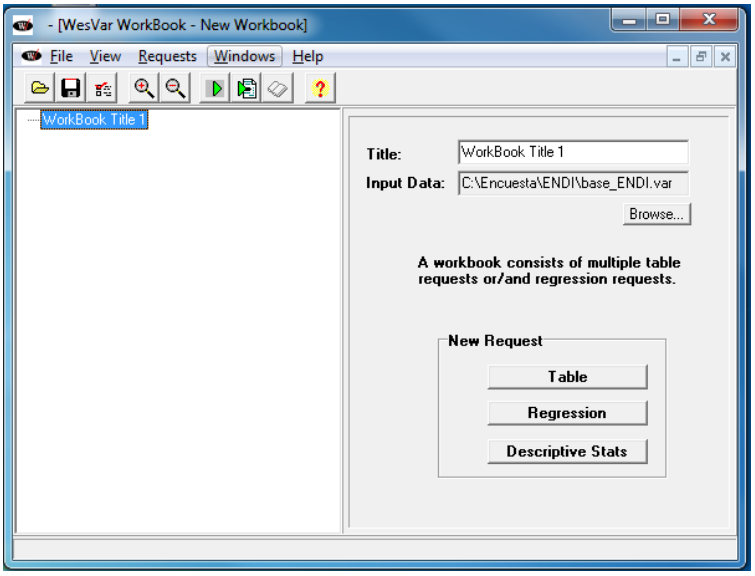


Figura 3

Una alternativa para obtener las estimaciones de los parámetros considerados en esta guía es a partir de la generación de una tabla (**Table**) del apartado **New Request** (figura 3). Al hacer clic en **Table**, se habilita una ventana similar a la que presenta la figura 4.

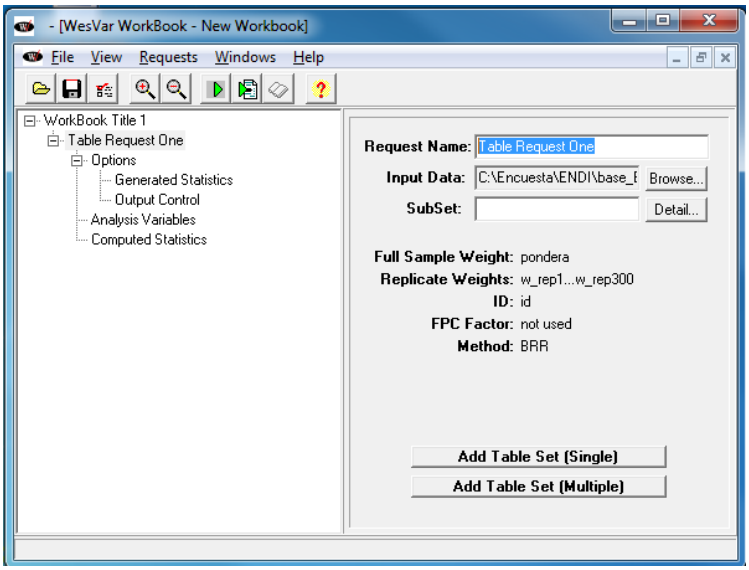


Figura 4

Sobre el panel izquierdo y haciendo clic en el nodo “Analysis Variables”, la herramienta habilita a definir las variables que requieren estimaciones de totales, por ejemplo, *Y* y *Z*. Como se muestra en la figura 5, las variables deben ser seleccionadas en **Source Variables** e incorporadas al apartado **Selected** del panel derecho.

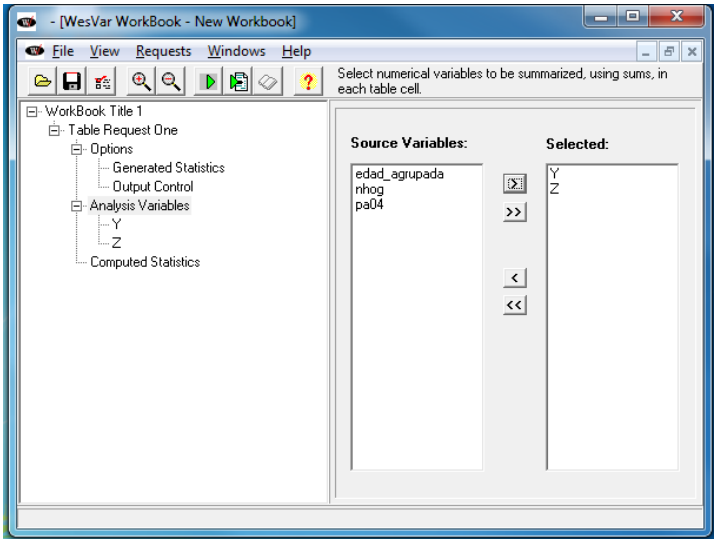


Figura 5

En forma adicional, haciendo clic sobre el nodo “Computed Statistics” del panel izquierdo sobre el árbol, se pueden definir otros estimadores alternativos como funciones de totales; por ejemplo, al promedio de la variable Y se lo define en **Computed Statistics** del panel derecho como *M_Y=MEAN(Y)* (figura 6) y la razón entre los totales de las variables Y y Z, como *razon=Y/Z* en el mismo apartado (figura 7).

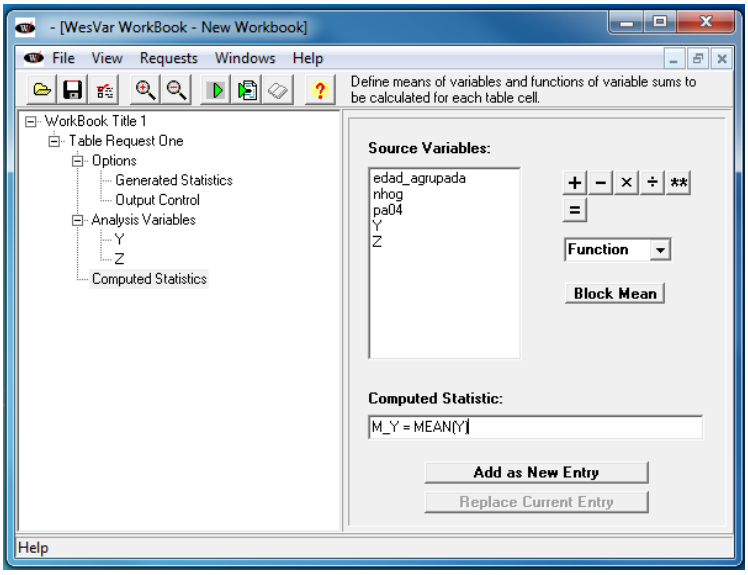


Figura 6

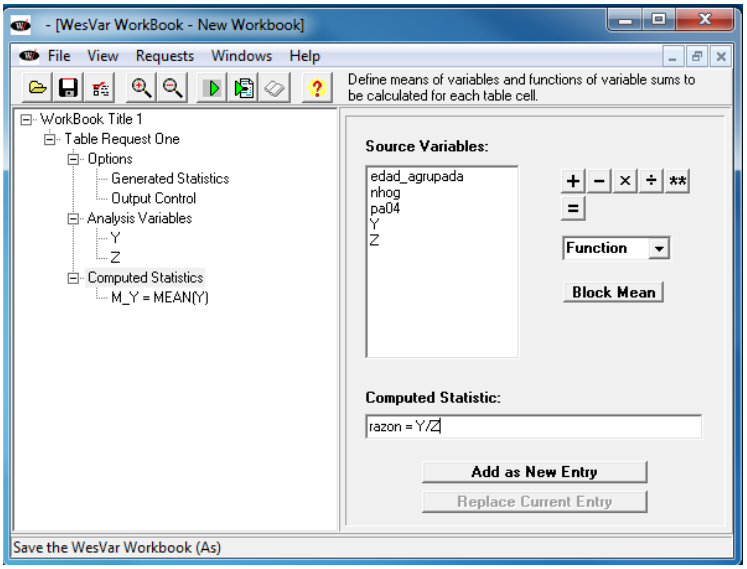


Figura 7

Por último, en el panel izquierdo y sobre el nodo *Table Request One*, la herramienta habilita a seleccionar la opción *Add Table Set* (*Single*) sobre el panel derecho para visualizar los resultados de los cálculos (figura 8).

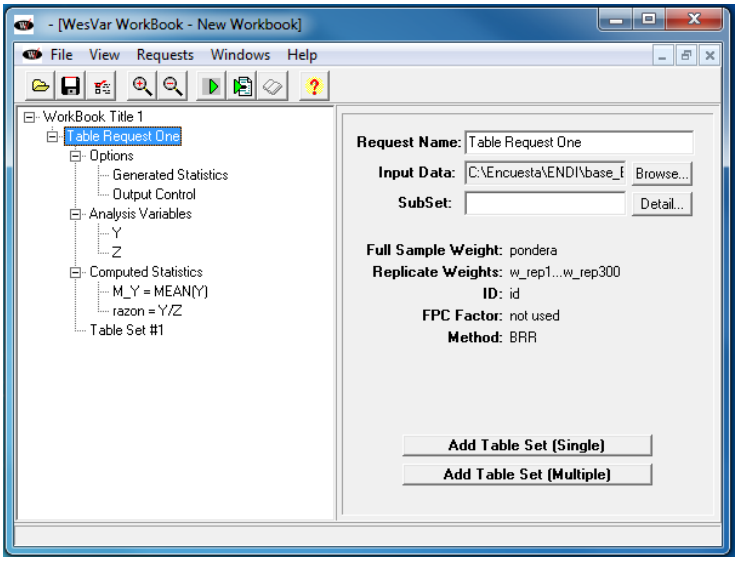


Figura 8

Aplicando sobre el ícono  del menú de la herramienta, se ejecutan los requerimientos o análisis definidos por el usuario; los resultados aparecen al hacer clic sobre  *Overall* y seleccionando el nodo sobre el panel izquierdo, como muestra la figura 9.

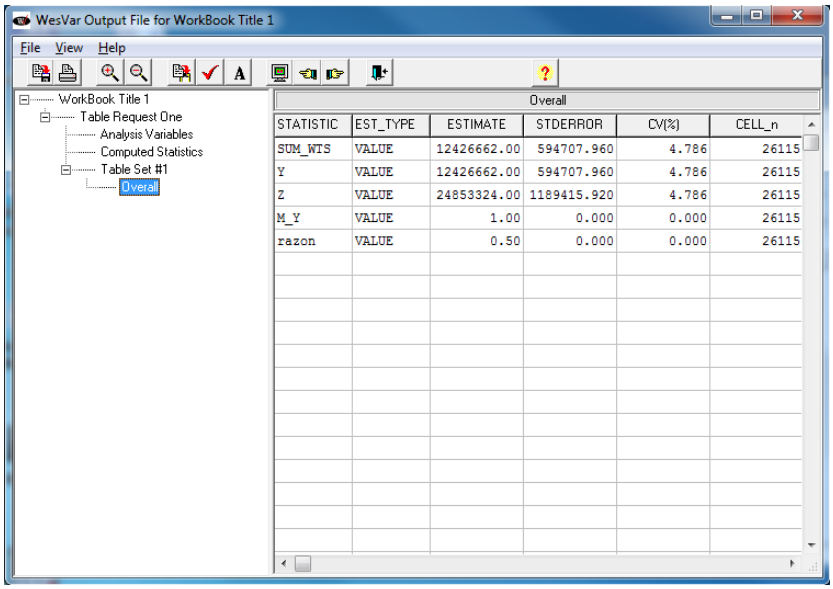


Figura 9

El usuario podrá advertir que, por defecto, Wesvar calcula para las estimaciones requeridas (ESTIMATE) una estimación del error estándar (STDERROR) y del coeficiente de variación (CV(%)).

En el caso de que se desee la estimación de proporciones, asumiendo que Y es del tipo categórica, se debe generar una tabla (**Table**) en la ventana de la figura 3, agregar una tabla con la opción **Add Table Set (Single)** (figura 4), e indicar cuál es la variable para la que se desean las estimaciones, como muestra la figura 10.

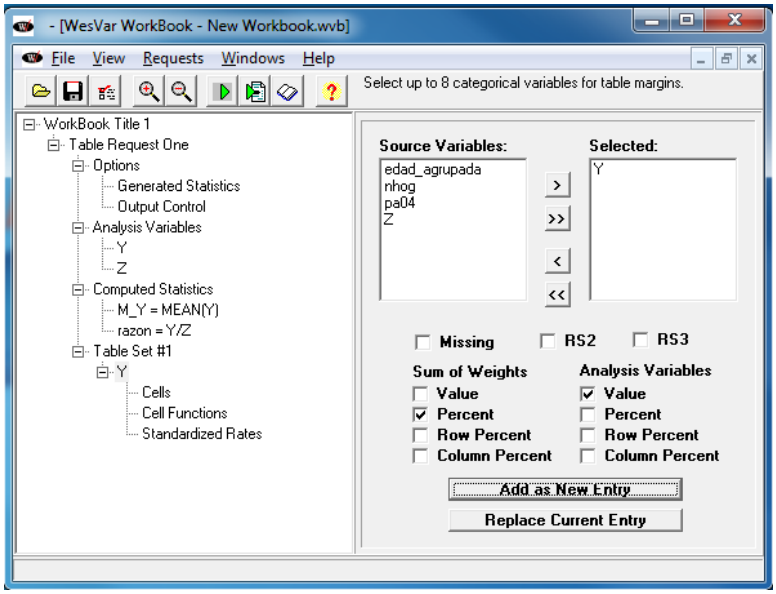


Figura 10

Para los usuarios que deseen emplear esta herramienta, el manual brinda un tratamiento detallado de las distintas opciones con las que cuenta y en el que se amplía lo presentado en esta guía.

9.5 Alternativa para el cálculo del error de muestreo

Si no se cuenta con las herramientas que se presentaron para efectuar los cálculos de los errores de muestreo, y dependiendo del volumen de estimaciones que desea el usuario, existe la posibilidad de recurrir a la operatoria que se presentó en la sección 8 empleando las fórmulas [1] a [3].

Por ejemplo, si se asume que la variable Y está medida sobre las personas del Estudio, la expresión que se debe emplear como estimador para un total t_y , según se lo definió en la sección 6, es:

$$\hat{t}_y = \sum_R w_{ijkl}^H \, y_{ijkl}$$

Siguiendo lo señalado en la sección 8, la formulación para la varianza *bootstrap* [1] de un estimador es:

$$v_B(\hat{\theta}) = \frac{1}{300}\sum_{b=1}^{300}\left(\hat{\theta}^*_{(b)} - \hat{\theta}\right)^2$$

Empleando al conjunto de réplicas $\left\{w^{*H(b)}_{ijkl}, b = 1, \ldots, 300\right\}$ y reemplazando a $\hat{\theta}$ por $\hat{t}_y$, y a $\hat{\theta}^*_{(b)}$ por $\hat{t}_{y^*_{(b)}}$, donde $\hat{t}_{y^*_{(b)}} = \sum_R w^{*H(b)}_{ijkl} y_{ijkl}$ es la estimación del total a partir de los factores de expansión $w^{*P(b)}_{ijkl}$ para la p-ésima persona en la b-ésima submuestra *bootstrap*, $b = 1, \ldots, 300$, permite calcular estimaciones para la varianza *bootstrap* de $\hat{t}_y$, a través de:

$$v_B\left(\hat{t}_y\right) = \frac{1}{300}\sum_{b=1}^{300}\left(\hat{t}_{y^*_{(b)}} - \hat{t}_y\right)^2 \quad [4]$$

para el error estándar, según

$$ee_B\left(\hat{t}_y\right) = \sqrt{v_B\left(\hat{t}_y\right)}$$

y para del coeficiente de variación con

$$cv_B\left(\hat{t}_y\right) = \frac{ee_B\left(\hat{t}_y\right)}{\hat{t}_y}$$

De manera análoga se procede para los casos de un promedio, una proporción, o un cociente o una razón, reemplazando en [1] a $\hat{\theta}$ por $\hat{\bar{y}}, \hat{p}_A$, o $\hat{R}_{yz}$, respectivamente (ver sección 6) y a las estimaciones *bootstrap* $\hat{\theta}^*_{(b)}$ que emplean a las réplicas por:

$$\hat{\bar{y}}^*_{(b)} = \frac{\sum w^{*H(b)}_{ijkl} y_{ijkl}}{\sum w^{*H(b)}_{ijkl}},$$

$$\hat{p}^*_{A(b)} = \frac{\sum w^{*H(b)}_{ijkl} y_{ijkl}}{\sum w^{*H(b)}_{ijkl}},$$

o,

$$\hat{R}^*_{\mathrm{yz}(b)} = \frac{\sum w^{*H(b)}_{ijkl} y_{ijkl}}{\sum w^{*H(b)}_{ijkl} z_{ijkl}}$$

según sea el caso, para obtener las respectivas varianzas estimadas por *bootstrap*, como también ee_B y cv_B, para cualquiera de las estimaciones en cuestión.

10. Recomendaciones para el uso con fines estadísticos de los datos del Estudio

No es posible asumir la misma confianza en todos los resultados del Estudio. Incluso en algunas situaciones no es aconsejable tomarlos como válidos para hacer inferencia estadística. Distintos motivos pueden

afectar las estimaciones y, en consecuencia, la inferencia que se haga a partir de ellas. Por ejemplo, las estimaciones pueden no representar a la población objetivo de interés, cuando:

- los parámetros de interés se estiman en dominios no previstos en el diseño del Estudio, o son marginales para la población o subpoblación en estudio;
- la cantidad de hogares o personas involucradas en la estimación es escasa;
- la estimación de un total involucrado en el denominador de un cociente posee una variabilidad o coeficiente de variación muy alto.

En todas estas situaciones, el comportamiento del estimador empleado, tanto el del parámetro o como el de la varianza, puede sufrir un deterioro importante en términos de precisión. Si bien se realizaron ajustes para disminuir el impacto del sesgo que introducen algunos de los errores no muestrales, este puede persistir y acentuarse si se está en presencia de algunas de estas situaciones.

A su vez, algunos de los supuestos en los que se sostiene la metodología para el cálculo de los errores de muestreo pueden no cumplirse o verse afectados. Por ejemplo:

- si se calculan estimaciones a niveles de desagregación muy alta,
- en dominios de análisis donde participan pocas unidades en los “últimos conglomerados”,
- si la característica no está presente en la mayoría de los últimos conglomerados, y
- si en las estimaciones participan factores de expansión con alta variabilidad, o con algunos valores extremos.

En los casos mencionados, la estimación del parámetro puede tener un nivel de error muy alto, o bien la estimación del error de muestra puede ser inestable como para suponerlo confiable. Por lo tanto, se advierte a todos los usuarios que empleen la base con los datos del Estudio para sus propias estimaciones que deberán poner atención y ser prudentes a la hora de sacar conclusiones en ciertas circunstancias.

10.1 Recomendaciones sobre las estimaciones

Para ayudar al usuario a interpretar los resultados del Estudio, se presentan algunas recomendaciones y sugerencias para identificar estimaciones en las que se debe poner poca o ninguna confianza.

El siguiente cuadro cubre algunas de las situaciones más generales por las que puede atravesar una estimación a la hora de tener que evaluar su precisión o la confianza que se puede poner en ella. Cualquier lector de los resultados oficiales publicados del Estudio, o los usuarios que generen sus propias estimaciones a partir de la base que entrega el Instituto, las deben tener presentes a la hora de sacar sus conclusiones del fenómeno que están estudiando a partir del Estudio.

Cuadro 5. Recomendaciones para interpretar las estimaciones

Calidad de la estimación	Condición	Recomendaciones
No confiable	Si se cumple alguna de las siguientes: a) El total de unidades involucradas en el cálculo de la estimación es menor a 100. b) La estimación de una razón es menor a 0,03. c) La estimación de una proporción es menor al 3%. d) El denominador de un cociente, razón, o proporción, tiene un $CV > 10\%$. e) La estimación posee un $CV > 33{,}3\%$.	Se recomienda no emplear la estimación en este caso. Si existe la necesidad de publicarla, se debe advertir que las conclusiones basadas en ella no son confiables o válidas.
Poco confiable	La estimación posee un CV en el rango: $16{,}6\% < CV \leq 33{,}3\%$	La estimación debe ser considerada con precaución. Hay una alta probabilidad de que la inferencia resultante presente un nivel de error elevado. Se recomienda presentarla con alguna notación en la que se advierta de esta situación.
Confiable	La estimación posee un CV en el rango: $CV \leq 16{,}6\%$	La estimación puede ser considerada sin restricciones. No se requiere una notación especial.

Fuente: INDEC, Estudio Nacional sobre el Perfil de Personas con Discapacidad 2018.

Se insiste con la recomendación de que, en el caso que algunas de las estimaciones sean consideradas no confiables o poco confiables para inferir el total de la población o en subpoblaciones y el usuario desee incorporarlas en una publicación, se incluya una advertencia y se haga una referencia a las limitaciones del caso citando la presente guía metodológica, en particular el cuadro 5, definido por el INDEC como estándar para la encuesta.

10.2 Recomendaciones para estimaciones en dominios

Otro aspecto importante a tener en cuenta por los usuarios de la base de datos del Estudio es la manera en que se calculan las estimaciones en subpoblaciones. Una práctica habitual es filtrar o seleccionar los casos que componen el dominio o la subpoblación, y a partir de ellos obtener una estimación del parámetro de interés para ese subconjunto de la población. Si esa modalidad se la emplea para el cálculo del error muestral, es importante señalar que generalmente puede llevar a subestimarlo y en algunas circunstancias de manera grosera.

La herramienta que se emplee para la estimación del error de muestreo debe hacer uso de todas las observaciones de la muestra, para obtener una medida confiable y no estar subestimándola. Por lo general, la documentación que acompaña la herramienta contempla esta advertencia. En particular, en aquellas presentadas en los apartados 9.1 a 9.3, los usuarios que deseen calcular estimaciones en subpoblaciones o dominios pueden recurrir a las opciones **subset**[37] en *R*, **subpop** en *Stata*, y **domain** en *SAS* para obtener en forma adecuada la estimación del CV o del EE que esté calculando.[38]

10.3 Recomendaciones sobre el cálculo de intervalos de confianza

Los IC brindan otro camino para evaluar la variabilidad inherente en las estimaciones provenientes de una muestra probabilística. Un intervalo de confianza es un rango de valores que tiene una probabilidad, conocida como "nivel de confianza", de contener el valor poblacional del parámetro. En otras palabras, un intervalo de confianza al 0,95 significa que, si todas las muestras posibles son seleccionadas y un IC es calculado para cada una de ellas, el 95% de los IC construidos deberían contener al valor verdadero del parámetro.

Para aquellos usuarios que deseen acompañar sus estimaciones con un intervalo de confianza y cuenten con la estimación de su varianza o de su error estándar, un IC con un nivel de confianza del 95% se puede calcular en forma aproximada de la siguiente manera:

$$IC_{\theta,95\%}: \left(\hat{\theta} - 1.96 * \sqrt{v_B(\hat{\theta})};\ \hat{\theta} + 1.96 * \sqrt{v_B(\hat{\theta})}\right),$$

donde $v_B(\hat{\theta})$ es la varianza *bootstrap*; o a partir de $cv_B(\hat{\theta})$, como:

$$IC_{\theta,95\%}: (\hat{\theta} - 1.96 * cv_B(\hat{\theta}) * \hat{\theta};\quad \hat{\theta} + 1.96 * cv_B(\hat{\theta}) * \hat{\theta})$$

En la determinación de un IC juegan roles importantes la distribución probabilística del estimador y las propiedades asintóticas del estimador empleado para la varianza. A diferencia del EE y el CV, el IC obliga a adoptar algunos supuestos sobre el estimador $\hat{\theta}$ empleado para estimar el parámetro de interés. Entre ellos, que de manera aproximada siga en distribución una ley normal, de difícil verificación en la práctica.

Como se advierte en distintos apartados, el diseño muestral del Estudio no es un MSA, e involucra distintas etapas con probabilidades de selección proporcionales a tamaños y estratificaciones. Esta complejidad en

[37] En el paquete Survey es posible utilizar también el comando **svyby** para obtener estimaciones en subpoblaciones.
[38] En Wesvar no es necesario emplear una opción para advertirle que se van a realizar estimaciones en dominios o subpoblaciones; al crear una tabla donde se involucre una variable que defina a la subpoblación (dominio) la herramienta procede correctamente al efectuar los cálculos del error por muestra.

el diseño por lo general lleva a que el conjunto de datos no siga la hipótesis $i.i.d,$ o sea, la de independencia e idénticamente distribuidos, requerida en este contexto para sostener el supuesto de normalidad (Heeringa, West y Berglung, 2017).

En virtud de lo expuesto, se sugiere a los usuarios a tener precaución al construir un IC para las estimaciones y no abusar de los supuestos cuando algunos pueden no cumplirse, en particular en las situaciones señaladas en párrafos anteriores de esta sección.

Referencias bibliográficas

Andridge, R. y Little, R. J. A. (2010). A Review of Hot deck Imputation for Survey Non-response. *International Statistical Review*, 78(1), 40-64.

American Association for Public Opinion Research (2016). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. Recuperado de https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf.

Brick, M., Morganstein, D. y Valliant, R. (2000). *Analysis of Complex Sample Data Using Replication*. Recuperado de https://www.researchgate.net/profile/David_Morganstein/publication/252297575_Analysis_of_Complex_Sample_Data_Using_Replication/links/55562a2e08ae6fd2d8235fbf/Analysis-of-Complex-Sample-Data-Using-Replication.pdf.

Carlson, B. L. (2013). Response Rates Revisited. *JSM*, *Survey Research Methods Section,* 1200-1208. Recuperado de http://www.asasrms.org/Proceedings/y2013/files/308173_80404.pdf.

Chowhan, J. y Buckley, N. (2005). Using Mean Bootstrap Weights in Stata: A BSWREG Revision. *The Research Data Centres Information and Technical Bulletin*, 2(1), 23-37. Recuperado de https://www150.statcan.gc.ca/n1/en/pub/12-002-x/12-002-x2005001-eng.pdf?st=LJqB8hAc

Deville J., Särndal C.E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87. Recuperado de DOI:10.1080/01621459.1992.10475217.

Frankel, L. R. (1983). The Report of the CASRO Task Force on Response Rates. En Frederick Wiseman (Ed.), *Improving Data Quality in a Sample Survey*. Cambridge: Marketing Science Institute.

Gagné, C., Roberts, G. y Keown, L. (2014). Weighted Estimation and Bootstrap Variance Estimation for Analyzing Survey Data: How to Implement in Selected Software. *The Research Data Centres Information and Technical Bulletin.* Recuperado de https://www150.statcan.gc.ca/n1/pub/12-002-x/2014001/article/11901-eng.htm.

Haziza, D. y Beaumont, J. F. (2017). Construction of Weights in Surveys: A Review. *Statistical Science*, 32(2), 206-226. Recuperado de DOI:10.1214/16-STS608.

Heeringa, S., West, B. y Berglund, P. (2017). *Applied Survey Data Analysis*. Nueva York: Chapman & Hall/CRC. Recuperado de DOI:10.1201/9781315153278.

Lemaître, G. y Dufour J. (1987). An Integrated Method for Weighting Persons and Families. *Survey Methodology*, 13(2), 199-207. Recuperado de https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X198700214607.

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R.* Nueva Jersey: J. Wiley & Sons. Recuperado de DOI:10.1002/9780470580066.

Rao, J. N. K. y Wu, C. F. J. (1988). Resampling Inference with Complex Surveys Data. *Journal of American Statistical Association*, 83(401), 231-241. Recuperado de DOI: 10.1080/01621459.1988.10478591.

Rao, J. N. K., Wu, C. F. J. y Yue K. (1992). Some Recent Work on Resampling Methods for Complex Surveys. *Survey Methodology*, 18(2), 209-217. Recuperado de https://www150.statcan.gc.ca/n1/pub/12-001-x/1992002/article/14486-eng.pdf.

Phillips, O. (2004). Using Bootstrap Weights with WesVar and SUDAAN. *Research Data Centres, Information and Technical Bulletin*, 1(2), 6-15. Recuperado de https://www150.statcan.gc.ca/n1/en/pub/12-002-x/12-002-x2004002-eng.pdf?st=JeakLQDY.

Särndall, C., Swensson, B. y Wretman, J. (1992). *Model Assisted Survey Sampling*. Nueva York: Springer-Verlag Publishing.

SAS Institute Inc. (2017). *SAS/STAT® 14.3 User's Guide*. Cary: SAS Institute Inc.

StataCorp (2017). *Stata Survey Data Reference Manual. Release 15*. College Station, Texas: StataCorp LLC.

Valliant, R., Dever J. A. y Kreuter F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. Nueva York: Springer. Recuperado de https://www.springer.com/gp/book/9783319936314.

West, B. (2012). Accounting for Multi-stage Sample Designs in Complex Sample Variance Estimation. Michigan Program in Survey Methodology. Recuperado de http://www.isr.umich.edu/src/smp/asda/first_stage_ve_new.pdf.

Wolter, K. M. (2007). *Introduction to Variance Estimation*. Nueva York: Springer-Verlag. Recuperado de DOI: 10.1007/978-0-387-35099-8.

Anexo. Tasa de respuesta de los hogares

La tasa de respuesta de los hogares es la proporción de hogares en viviendas elegibles que completó la encuesta. Es una medida importante de calidad y permite evaluar en forma general el desempeño en la operación de captura de datos en una encuesta. Los estándares o protocolos adoptados por la comunidad estadística, por ejemplo, el de la American Association for Public Opinion Research (AAPOR, 2016) o del Council of American Survey Research Organizations (CASRO) (Frankel, 1983) sugieren realizar los cálculos a partir de considerar no solo las unidades elegibles y con respuesta, sino también las de elegibilidad dudosa o desconocida.

Esta modalidad permite tener en cuenta explícitamente la incertidumbre que a menudo rodea la elegibilidad de una dirección, vivienda u otra unidad seleccionada para una encuesta. Por ejemplo, los casos no contactados incluyen aquellos en los que no se sabe si existe una vivienda particular en la dirección asignada a un encuestador y se desconoce si es elegible para el Estudio. Ante la falta de contacto, la elegibilidad será desconocida a menos que pueda ser determinada de alguna otra forma (información adicional del marco muestral, afirmación de un vecino, inspección ocular de la unidad seleccionada, revisita por parte de supervisor, etc.). Existen situaciones en las que el contacto es imposible por presencia de sistemas de seguridad, portones cerrados, por tratarse de unidades de vivienda múltiple de difícil acceso o encontrarse en áreas inaccesibles, ya sea por inclemencias climáticas o cuestiones de inseguridad. También es posible que la dirección brindada sea errónea, que se cuente con información insuficiente para ubicarla o sea inexistente para el encuestador o supervisor de la encuesta.

Todas las alternativas propuestas para el cálculo de la tasa de respuesta realizan algún supuesto sobre las unidades cuya elegibilidad está en duda o es desconocida, e involucran en su expresión la tasa de elegibilidad e ($0 \leq e \leq 1$), o sea, la proporción estimada de casos con elegibilidad desconocida o dudosa que son elegibles (Carlson, 2013).

El valor máximo, $e = 1$, es el que se corresponde con asumir que todos los casos con elegibilidad desconocida o dudosa son elegibles. El supuesto origina la mayor subestimación de la tasa de respuesta ($RR1$, en la notación de la AAPOR). La propuesta mínima asume que la proporción de unidades con elegibilidad desconocida son no elegibles, o sea $e = 0$, maximizando el valor de la tasa de respuesta ($RR5$, en la notación de la AAPOR).

Un valor intermedio, adoptado para el cálculo de la tasa de respuesta de la encuesta, es el que emplea el método de asignación proporcional o método de CASRO. Se asume que la proporción de unidades elegibles para el conjunto de unidades con elegibilidad determinada es igual que para el conjunto de unidades cuya elegibilidad es desconocida o dudosa. En otras palabras, la proporción de unidades inelegibles es igual para unidades con elegibilidad conocida y para unidades con elegibilidad desconocida o dudosa. Este supuesto tiene la ventaja de facilitar los cálculos y de proveer estimaciones conservadoras para la tasa de respuesta ($RR3$, en la notación de la AAPOR). Si,

R: cantidad de hogares con respuesta dentro de cada vivienda elegible,

EL: cantidad total de hogares dentro de cada vivienda elegible,

NE: cantidad de hogares o viviendas no elegibles,

ED: cantidad de hogares o viviendas con elegibilidad dudosa o desconocida

$\mathrm{e} = EL/(EL + NE)$: tasa de elegibilidad, o proporción estimada de hogares con elegibilidad desconocida,

la variante $RR3$ para la tasa de respuesta queda definida como: $RR3 = \frac{R}{EL+e*ED}$.

Los siguientes cuadros presentan las tasas de respuesta con la versión $RR3$, y una cota superior o un valor máximo estimado cuando se asume $e = 0$, $RR5 = \frac{R}{EL}$, para hogares; y la $RR3$ para la tasa de respuesta a nivel de personas por región y para el total del país.[39]

Cuadro 6. Tasas de respuesta por regiones y total del país

Regiones	RR3	RR5
	%	%
Gran Buenos Aires	59,7	81,8
Noroeste	86,3	93,7
Noreste	86,5	94,2
Cuyo	77,4	91,8
Pampeana	77,8	90,7
Patagonia	81,5	91,0
Total del país	**75,9**	**89,9**

Fuente: INDEC Estudio Nacional sobre el Perfil de Personas con Discapacidad 2018.

[39] Para los cálculos no se emplearon los factores de expansión, dado que se busca poner de manifiesto el éxito del esfuerzo en la captura de los datos de la encuesta, independientemente de cuánto representa en la población una unidad.

Glosario

Aglomerado o localidad compuesta. Unidad geoestadística urbana, determinada por criterios físicos y territoriales, que se extiende sobre dos o más áreas político-administrativas, sean ellas jurisdicciones de primer orden (provincia), segundo orden (departamento o partido) o áreas de gobierno local. Es una unidad de área y es la unidad de muestreo de primera etapa (UPM) del marco de muestreo de la Muestra Maestra Urbana de Viviendas de la República Argentina (MMUVRA). (Ver **Localidad**).

Aleatorio. Concepto que permite calificar un evento vinculado a un resultado posible entre otros y desconocido antes de ser ejecutado. Dentro del muestreo probabilístico es el propio mecanismo el que asegura que la muestra resultante no pueda ser predicha de antemano. En ese contexto, las respuestas a las variables indagadas por la encuesta son tratadas como valores fijos, y la componente aleatoria es solo atribuida al proceso de selección que origina la muestra.

Área MMUVRA. Unidad de área que coincide en general con el radio censal definido sobre la base cartográfica del Censo Nacional de Población y Viviendas 2010. Sin embargo, también puede estar determinada por un agrupamiento de radios contiguos para ajustarse a requerimientos de tamaño en términos de viviendas; o por recortes operativos en algunos radios por baja densidad de viviendas, o economía de recursos, o de costos. Estas áreas son las unidades de segunda etapa de muestreo (USM) de la MMUVRA y, en cada UPM seleccionada, el conjunto compone el marco de muestreo para la selección de segunda etapa del diseño muestral.

Autorrepresentada. Dentro del muestreo de poblaciones finitas, se considera que una unidad muestral está autorrepresentada cuando se la incluye sin pasar por el proceso de selección aleatorio de una muestra; equivale a que la unidad tenga probabilidad 1 de ser seleccionada y siempre forme parte de cualquiera de las muestras surgida del diseño muestral. Como consecuencia, en el proceso inferencial, los valores de las características observadas en dicha unidad participan sin ponderarse o expandirse, y sin sumar al error muestral del estimador.

Bootstrap. Método no paramétrico que utiliza en forma intensiva recursos computacionales para realizar inferencias estadísticas. En líneas generales, emplea un remuestreo aleatorio intensivo, desde la muestra original, para generar un conjunto de réplicas o muestras *bootstrap*. A partir de ellas, se determina una aproximación empírica de la función de distribución muestral del estimador, que permite construir las medidas usuales del error: varianza, desvío estándar, intervalos de confianza, etcétera.

Calibración. Conjunto de procedimientos o técnicas de corrección de los factores de expansión que se utiliza en las encuestas por muestreo. Emplea la información agregada (totales), disponible para un conjunto de variables (de calibración) indagadas, que proviene de fuentes externas a la encuesta para el total de la población. Permite ajustar los factores o ponderadores, de manera tal que las estimaciones de totales para ese conjunto de variables coincidan con sus totales poblacionales. Esta práctica por lo general propicia la precisión en las estimaciones o la corrección de problemas de cobertura del marco de muestreo.

Censo. Operativo que intentar enumerar el total de elementos que conforma una población y medir una o más características sobre ellos. Puede brindar información con un nivel de desagregación geográfico y detalle muy alto. Se lo puede considerar como una muestra al 100% de la población. Debido a esta característica, los resultados que se obtienen están libres de error muestral; no así de errores ajenos al muestreo (tales como no respuesta, cobertura, medición, procesamiento, u otras fuentes siempre presentes en una operación estadística).

Cobertura. Grado de inclusión de los elementos de la población objetivo en el marco muestral. Si el marco no contiene todos los elementos de la población objetivo, se está en presencia de una subcobertura de la población; por el contrario, habrá sobrecobertura si existe la duplicación de elementos o la inclusión en el marco de unidades que no forman parte de la población objetivo.

Coeficiente de variación (CV). Dentro del ámbito del muestreo en poblaciones finitas, constituye otra forma de presentar el error de muestreo. Se lo obtiene a partir del cociente entre el error estándar del

estimador y el estimador. En general, se lo calcula en términos porcentuales, siendo esto un beneficio, dado que es una cantidad libre de unidad de medición, lo que permite la comparabilidad.

Conglomerado. Conjunto de unidades o elementos de la población agrupados por naturaleza propia o sobre la base de un criterio de proximidad. El conglomerado puede ser un agrupamiento ya existente de la población (vivienda u hogar, hospital, escuela); o bien, estar definido por divisiones administrativas, operativas o geográficas del territorio en donde los elementos pertenecen (manzanas, radios censales, fracciones censales, localidades, departamentos), o a fracciones del tiempo (semanas, días, tramos horarios, etc.). Utilizado generalmente en diseños multietápicos, en los que la selección de elementos o miembros de la población en forma directa resulta impracticable, por ausencia de listados o por motivos relacionados a los costos operativos.

Diseño muestral. Marco metodológico y de trabajo que sirve de base para la selección de la muestra, y que afecta otros aspectos importantes de un estudio o una encuesta. Define la población objetivo de la encuesta; el marco de muestreo que se emplea y que la representa, y el tipo de vínculo que tienen sus unidades con las de la población; las distintas etapas y el/los método/s involucrado/s en la selección de la muestra; las probabilidades asociadas a esas etapas y unidades; el tamaño de la muestra; los principales dominios de estimación; y las fórmulas de cálculo o los estimadores a emplear para obtener los resultados a partir de los datos obtenidos por la encuesta.

Diseño muestral complejo. Diseño que emplea una o varias etapas de selección y distintos tipos de estratificación y de conglomeración de las unidades, y que involucra probabilidades no uniformes en los procesos de selección de la muestra. Se adopta generalmente para las encuestas a hogares, ya que presenta la mejor opción cuando no se cuenta con un marco de lista de viviendas o cuando confeccionar uno es costoso.

Dominios de análisis. Subconjuntos de respondentes de una encuesta, determinados, por lo general, por características sociodemográficas, sobre los cuales se desea realizar el análisis de la información que provee la encuesta. A diferencia de los dominios de estimación, estos dominios no fueron contemplados por el diseño muestral, porque no fueron previstos, o porque no fue posible determinar la pertenencia de los elementos de la muestra a cada dominio *a priori*. Por lo tanto, no existió un control sobre la precisión para las estimaciones para estos dominios ni sobre sus tamaños de muestra que pasan a ser aleatorios para el diseño muestral.

Dominios de estimación. Subconjuntos de la población objetivo cuyos elementos pueden ser identificados en el marco muestral sin ambigüedad. A estos elementos, en la etapa de diseño de la encuesta, se les determina un tamaño de muestra y un nivel de precisión predefinido para obtener estimaciones de interés en ellos. Por lo general, son los dominios de publicación en los que el diseño muestral permite desagregar los resultados de la encuesta. En una encuesta a hogares, suelen ser agregados geográficos, o agrupamientos geopolíticos o administrativos del territorio (región, provincia, aglomerado o localidad principal, etcétera).

Efecto de diseño. Cociente entre la variancia de un estimador correspondiente al diseño muestral empleado para seleccionar la muestra (en general, complejo) y la variancia del estimador que se obtendría bajo un muestreo simple al azar (MSA) de igual tamaño. Se lo emplea para evaluar la precisión en las estimaciones; por lo general, se lo vincula a diseños muestrales que involucran conglomerados por la relación que tiene este indicador con la medida de homogeneidad interna en este tipo de unidades. Tiene otros potenciales usos, en particular a la hora de determinar tamaños de muestra en diseños complejos. Se debe tener en cuenta que es el cociente de dos cantidades poblacionales desconocidas y, por lo tanto, debe ser estimado a partir de la muestra.

Elegibilidad. Refiere a si una unidad de la muestra es parte de la población objetivo o no. Errores en la determinación de la elegibilidad afectan directamente dos aspectos importantes de la calidad de una encuesta. En primer lugar, si las reglas que determinan la condición de elegible o no de una unidad no son

claras y precisas, puede generarse un sesgo o error de cobertura. En segundo lugar, la tasa de respuesta de una encuesta puede estar subestimada si muchas unidades inelegibles se asumen como elegibles en los cálculos.

Encuesta Permanente de Hogares (EPH). Uno de los principales operativos con fines estadísticos del INDEC. Dicho relevamiento indaga sobre las características de la población en términos de mercado de trabajo, ocupación e ingresos, entre otras. Tiene una periodicidad trimestral, con un alcance geográfico sobre 31 entidades geográficas denominadas "aglomerados EPH". En el tercer trimestre del año calendario se amplía la cobertura a nivel nacional y provincial, para la población urbana, que se denomina "Encuesta Anual de Hogares Urbanos" (EAHU).

Error aleatorio. Error causado por cambios desconocidos e impredecibles en un proceso de medición.

Error cuadrático medio (ECM). Forma más general que toma el error muestral de un estimador en presencia de sesgo. Esta última componente resulta de una fuente de error que sistemáticamente distorsiona las estimaciones en una dirección, y cuyo promedio sobre todas las realizaciones de la muestra hace que difiera consistentemente de su verdadero valor poblacional o parámetro. A diferencia de la varianza muestral del estimador que se puede estimar desde la propia muestra, el sesgo necesita de valores poblacionales, desconocidos a menos que se realice un censo, para poder ser cuantificado. Aun así, el ECM es una medida importante que se emplea para estudiar el comportamiento teórico de un estimador, y su formulación analítica corresponde a la suma de la variancia muestral del estimador y el sesgo al cuadrado.

Error de cobertura. Diferencias entre la población objetivo y la población que cubre el marco muestral producen errores de esta índole en un estimador. Pueden deberse a problemas de subcobertura y sobrecobertura del marco (ver **Cobertura**). En el primer caso, algunos elementos de la población objetivo tienen una probabilidad nula de ser seleccionados para una muestra. En el segundo, por incluir erróneamente o duplicar algunos de los elementos, estos poseen una probabilidad de ser seleccionados cuando no la deben tener, o la probabilidad es más alta de la que le corresponde, respectivamente. El error neto de cobertura es la diferencia entre la subcobertura y la sobrecobertura.

Error de medición. Cualquier desviación aleatoria o sistemática entre el verdadero valor de la medición y el valor obtenido a partir del proceso o instrumento que origina la medida.

Error de muestreo, error muestral o error por muestra. Error asociado con la no observación, es decir, ocurre porque no todos los miembros de la población se incluyen en la muestra. Se refiere a la diferencia entre la estimación derivada de la muestra y el valor "verdadero" que resultaría si se realizara un censo de toda la población bajo las mismas condiciones en las que se llevó adelante la muestra. Tiene la particularidad de ir disminuyendo a medida que aumenta el tamaño de la muestra, y a través del muestreo probabilístico es posible estimarlo a partir de la propia muestra. En ausencia de sesgo, este error se corresponde a la componente aleatoria definida por la varianza muestral del estimador que da origen a la estimación.

Error estándar. Medida de la variabilidad de una estimación debida al muestreo. Se obtiene a partir de la raíz cuadrada de la varianza del estimador. Posee las mismas unidades de medición que la estimación y se calcula a partir de la muestra.

Error de no respuesta. Sesgo sobre el estimador que produce la diferencia entre las unidades muestrales que responden y las que no responden. Su magnitud depende de la tasa de no respuesta, y de la asociación entre la probabilidad de respuesta de las unidades y la característica que está siendo estudiada. (Ver **No respuesta**).

Error de respuesta. Error que ocurre cuando se obtienen respuestas incorrectas, de manera deliberada o no, a las preguntas del cuestionario. Diversos motivos llevan a los encuestados a brindar información errónea: de forma intencional, por temor a que se descubra su información, vergüenza, desconfianza; o de manera no intencional, por falta de comprensión de las preguntas, falta de memoria, entre otras. La existencia de estos errores limita la validez de los resultados que se extraen de los datos y, por ende, afecta la calidad de una encuesta.

Error no muestral. Conjunto de todos los tipos y las fuentes de error que potencialmente pueden afectar una encuesta, con la excepción de aquel asociado al muestreo (ver **Error de muestreo**). Forman parte de este conjunto los errores de cobertura del marco muestral, los del instrumento de medición o la modalidad empleada en la captura de la información, los que surgen de la interacción entre el entrevistador y el respondente, los que ocasionan la no respuesta, los que aparecen en la etapa de procesamiento de los datos, y los inducidos por modelización, entre otros. A diferencia del error de muestreo, los no muestrales no disminuyen al aumentar el tamaño de muestra, son difíciles de controlar y cuantificar, y la mayoría se traducen en sesgo para el estimador.

Error sistemático. Tendencia, en un proceso de medición, a generar resultados diferentes al verdadero de manera consistente en una dirección.

Estimación. Proceso por el cual se obtiene un valor numérico o un rango de valores para un parámetro desconocido de la población a partir de los datos de una muestra. Se lo emplea también para denominar el resultado del proceso.

Estimador. Expresión analítica de una función que, utilizada con los datos de una muestra, permite estimar un parámetro de interés desconocido.

Estimador consistente. Estimador que, al incrementar el tamaño de muestra, se acerca cada vez más al parámetro poblacional. En el contexto de poblaciones finitas, un estimador es consistente si coincide con el parámetro cuando la muestra coincide con la población (censo).

Estimador insesgado. Estimador en el que el valor central de su distribución probabilística o muestral coincide con el parámetro poblacional que intenta estimar.

Estratificación. Proceso de dividir las unidades del marco de muestreo, basado en un criterio, en grupos homogéneos y mutuamente excluyentes llamados “estratos”. Su principal objetivo en un diseño muestral es reducir el error de muestreo en una estimación. En ocasiones, los estratos pueden ser dominios de estimación de una encuesta, en cuyo caso el tamaño de la muestra deberá contemplar la precisión preestablecida para las estimaciones en los estratos.

Factor de expansión. Valor asociado a cada unidad elegible y que responde a la muestra, que se construye a partir de la inversa de la probabilidad de inclusión de cada unidad o peso muestral inicial. Puede incluir distintos tipos de ajustes, para disminuir en lo posible los errores de cobertura y de no respuesta que afectan a la encuesta, y ser tratados por un proceso de calibración que lleva en general a ganar eficiencia y precisión en las estimaciones. Los factores de expansión finales son los que se emplean tanto para generar todas las estimaciones de una encuesta, como en los cálculos del error muestral al determinar la precisión alcanzada.

Inferencia estadística. Conjunto de métodos y técnicas que permiten inducir o extraer conclusiones de características objetivas (parámetros) de una determinada población, con un riesgo de error medible en términos de probabilidad. Se realiza a partir de la información empírica proporcionada por una muestra y la teoría de probabilidades. Incluye la estimación puntual, la estimación por intervalos y la prueba de hipótesis estadísticas.

Intervalo de confianza. Declaración sobre el nivel de confianza de que el valor verdadero para la población se encuentra dentro de un rango específico de valores. La probabilidad, es decir, el nivel de confianza, de que el intervalo contenga al parámetro se determina *a priori* y de ella depende la longitud del intervalo. El intervalo de confianza es otra forma de presentar el error muestral de un estimador.

Localidad. Unidad geoestadística urbana, determinada por criterios físicos y territoriales. Por su clasificación, puede ser simple, si se extiende sobre una sola jurisdicción y no está atravesada por ningún límite de provincia, departamento o partido, ni de gobierno local; o compuesta (también “aglomerado”), cuando se extiende sobre más de una jurisdicción. Para la MMUVRA, todas las localidades de 2.000 o más

habitantes, según el Censo Nacional de Población, Hogares y Viviendas 2010, conforman las UPM del marco de muestreo adoptado para el diseño muestral.

Marco de muestreo. Cualquier lista o recurso que delimita, identifica y permite acceso a las unidades de muestreo de un diseño muestral con el objetivo de seleccionar un subconjunto de ellas. En los diseños muestrales para encuestas a hogares, cobran relevancia los marcos de muestreo de áreas. Estos son una colección de unidades territoriales o espaciales con definiciones cartográficas precisas, que pueden involucrar mapas, fotografías aéreas o imágenes satelitales sobre el territorio. Las unidades más usuales en un marco de área pueden involucrar a provincias, departamentos, aglomerados, localidades, radios censales, manzanas, entre otras. Este tipo de marcos juegan un papel importante en los diseños muestrales que emplean varias etapas de selección y conglomerados, o en los que utilizan marcos múltiples. A menudo, se usan cuando una lista de unidades de muestreo finales no existe, o cuando otros marcos tienen problemas de cobertura.

Medida de tamaño. Cantidad que refleja el tamaño de una unidad de muestreo; por lo general, en encuestas a hogares es el número de viviendas o el total de población. Se la emplea para definir probabilidades para las unidades de muestreo en métodos que seleccionan las unidades para la muestra con probabilidad proporcional al tamaño.

Métodos por replicaciones. Métodos empleados para la estimación de varianza en diseños muestrales complejos, especialmente útiles cuando no se cuenta con una formulación analítica de la varianza del estimador. La parte central de estos métodos consiste en la selección de submuestras o remuestreo, que se realiza a partir de la muestra original respetando, en lo posible, el diseño muestral en cuestión. Con el cálculo del estimador en cada una de ellas, y a partir de la variabilidad de las estimaciones obtenidas respecto al estimador para la muestra original, los métodos permiten calcular una estimación para la varianza del estimador y así del error muestral para una estimación. Los más divulgados e implementados en las principales herramientas estadísticas de cálculo son el método *jackknife*, el de replicaciones repetidas balanceadas y el *bootstrap*.

MMUVRA. Muestra maestra urbana empleada por el INDEC con alcance nacional restringido a las localidades de 2.000 o más habitantes, que se utiliza como marco secundario de selección de viviendas particulares para todas sus encuestas a hogares entre dos censos de población y viviendas. Posee un diseño muestral complejo y se realiza actualizaciones periódicas de sus listados de viviendas y de su cartografía asociada.

Muestra. Subconjunto de unidades de una población, que es seleccionado bajo condiciones preestablecidas para ser incluido en el estudio o la encuesta. Alternativa a un censo, en donde toda la población es objeto de estudio, que suele ser elegida por motivos asociados a costos, eficiencia u oportunidad.

Muestra aleatoria. Ver **Muestra probabilística**.

Muestra maestra. Muestra aleatoria de gran tamaño donde permanecen invariantes las probabilidades determinadas por el diseño muestral. Empleada como un único marco de muestreo para subseleccionar muestras para distintas encuestas. (Ver **MMUVRA**).

Muestra no probabilística. Muestra en la que la selección de las unidades se determina por conveniencia, por cuotas, de acuerdo a la experiencia o el juicio del investigador; es decir, no involucra un proceso de selección aleatorio.

Muestra probabilística. Subconjunto de la población seleccionado mediante un método basado en la teoría de la probabilidad, y que emplea el conocimiento *a priori* de las posibilidades que tienen las unidades a ser incluidas en una muestra.

Muestreo. Proceso o conjunto de procesos que permiten seleccionar un número no nulo de elementos de todos los que componen un marco de muestreo, para observar y facilitar la estimación de parámetros de la población bajo estudio sin tener que recurrir a un censo.

Muestreo con probabilidad proporcional al tamaño. Modalidad del muestreo probabilístico que puede llevarse a cabo cuando las unidades del marco de muestreo tienen una medida de tamaño asignada. La probabilidad de inclusión de una unidad en una muestra queda definida por la relación entre su tamaño y la suma de tamaños de todas las unidades de la población, o una función de ellas. Bajo esta estrategia, las unidades de mayor tamaño tienen una probabilidad más alta de participar en una muestra. En encuestas a hogares, conjuntamente con el muestreo por conglomerados, es la estrategia más adoptada por las oficinas nacionales de estadísticas (ONE) para seleccionar las muestras de viviendas de sus principales operativos estadísticos.

Muestreo estratificado. Modalidad del muestreo probabilístico que se basa en una estratificación de las unidades del marco de muestreo, definida *a priori* por el diseño muestral. El proceso de selección de las unidades es independiente en cada estrato y no necesita ser el mismo. Si la estratificación es eficiente, es decir, si los estratos son homogéneos internamente y heterogéneos entre ellos respecto a las principales características a estudiar en la población, con este tipo de muestreo las estimaciones ganan en precisión comparadas al mismo diseño sin considerar estratos.

Muestreo multietápico. Método de muestreo que selecciona una muestra en dos o más etapas.

Muestreo por conglomerados. Modalidad del muestreo probabilístico que emplea el conglomerado como unidad de muestreo. En encuestas a hogares, esta alternativa de muestreo permite disminuir los costos de la encuesta, en perjuicio de perder, generalmente, precisión en las estimaciones al depender de la homogeneidad interna entre las unidades con respecto a las características que se están estudiando.

Muestreo simple al azar (MSA). Método de muestreo probabilístico que asigna a todas las muestras posibles de igual tamaño la misma probabilidad de ser seleccionadas; como consecuencia, cada elemento de la población tiene la misma probabilidad de estar incluido en una muestra. Es simple de seleccionar si se cuenta con un marco de muestreo de las unidades que conforman la población objetivo, pero no es la más adecuada para las encuestas a hogares. Entre los motivos está el poco o nulo control sobre la dispersión geográfica de las unidades a seleccionar que impacta sobremanera en los costos y en la organización de una encuesta.

Muestreo sistemático. Familia de métodos de muestreo probabilístico que se caracteriza por la elección aleatoria de la primera unidad de la muestra de la población (arranque aleatorio); mientras que el resto queda determinado por un intervalo de selección fijado *a priori* por el diseño muestral.

Nivel de confianza. Probabilidad, fijada *a priori*, de que una afirmación sobre el valor de un parámetro poblacional sea correcta. Generalmente, es empleado en la determinación de un intervalo de confianza.

No respuesta. Imposibilidad de obtener datos sobre las unidades elegibles de la población objetivo, en un censo o una encuesta. Son diversos los motivos que generan una no respuesta, entre los cuales sobresalen dos: el rechazo y el no contacto con la unidad. Puede ser total, o sea, cuando para la unidad no se logra la información requerida por el cuestionario; o parcial, cuando solo para algunos de los ítems incluidos en el cuestionario se falla en obtener información.

Parámetros. Medidas cuantitativas de interés desconocidas de la población objetivo o de cualquier dominio de estimación específico, que son factibles de ser estimadas a partir de una muestra. Algunos, usualmente considerados en las encuestas por muestreo, son del tipo descriptivo (como totales, medias, proporciones, varianzas, etcétera).

Peso replicado. Peso asignado a las unidades que aparecen en cada una de las muestras replicadas, el cual es generado por el propio método de replicaciones empleado para el cálculo de la varianza. Este peso, por lo general, sufre los mismos ajustes aplicados al peso muestral inicial por diseño (elegibilidad, no respuesta y calibración) para capturar la incidencia y variabilidad atribuida a este en la estimación de la varianza o el error muestral.

Población objetivo. Población de interés sobre la cual se desea obtener información estadística.

Ponderador. Ver **Factor de expansión**.

Precisión. Consistencia con la que se obtienen los resultados o las mediciones a partir de la muestra aplicando el mismo diseño muestral con respecto al valor verdadero o parámetro poblacional de interés. (Ver **Error de muestreo**).

Probabilidad. Cuantificación de la posibilidad de ocurrencia de un evento aleatorio. Toma valores entre 0 y 1, y es el pilar fundamental en el que sostiene el proceso de inferencia estadística.

Probabilidad de selección. Medida de la posibilidad que tiene cada unidad de la población del marco de muestreo de ser incluida en una muestra según el diseño muestral. Con cierto grado de generalidad, en el muestreo probabilístico también hace referencia a la probabilidad de inclusión de una unidad.

Radio censal. Unidad de área de límites conocidos y precisos, con un determinado número de viviendas, y de carácter operativo empleada por el INDEC en la organización de los censos de población. Por su clasificación, el radio censal puede ser urbano, rural o mixto, de acuerdo a pautas que involucran la distribución espacial y la densidad en términos de viviendas. Es la unidad empleada como base para definir las unidades de segunda etapa de muestreo (USM) de la MMUVRA. (Ver **Área MMUVRA**).

Rechazo. Ver **No respuesta**.

Segmento. Conglomerado compuesto por un número fijo de viviendas contiguas con límites conocidos y de fácil identificación en terreno, empleado como unidad de muestreo en algunas encuestas. En los censos de población y viviendas que conduce el INDEC, es la carga de trabajo de un censista.

Sesgo. Diferencia entre el valor esperado de un estimador y el valor del parámetro poblacional.

Sesgo por no respuesta. Sesgo que ocurre cuando el valor observado se desvía del parámetro poblacional debido a diferencias entre quienes responden la encuesta y los que no lo hacen. Es probable que ocurra cuando no se obtiene el 100% de respuesta de los casos elegibles para la encuesta. Aunque existen otros factores más determinantes que impactan en la magnitud del sesgo, en particular, el grado de asociación que existe entre la probabilidad a dar respuesta de los individuos de la población y las características que están siendo estudiadas.

Tasa de respuesta. Proporción de unidades de la muestra elegibles que respondieron al operativo. Se puede calcular la tasa de respuesta total y parcial de acuerdo a la ocurrencia de respuesta total (todo el cuestionario) o parcial (ítems con no respuesta), respectivamente.

Unidad de muestreo. Componente básico de un marco muestral. Unidad sobre la que el diseño muestral asigna una probabilidad positiva a ser seleccionada o incluida en una muestra. Pueden definirse distintas unidades de muestreo si el diseño involucra varias etapas; en cuyo caso, su denominación contiene una referencia que indica la etapa a la cual pertenece, por ejemplo, unidad de primera etapa de muestreo, UPM; unidad de segunda etapa de muestreo, USM; etcétera.

Varianza muestral. Grado por el cual las estimaciones de un parámetro poblacional, obtenidas a partir de todas las muestras posibles seleccionadas bajo un mismo diseño muestral, difieren unas de otras. Es calculada como el promedio del cuadrado de las diferencias entre el estimador y su valor esperado. Dentro del muestreo en poblaciones finitas, es el principal insumo para determinar el error muestral de una estimación y expresar sus distintas variantes.