

ISSN 2545-7179

Encuesta Nacional de Gastos de los Hogares 2017-2018

NOTA TÉCNICA

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

Mayo de 2020



NOTAS
TÉCNICAS
INDEC
N° 4



Ministerio de Economía
Argentina

Instituto Nacional de
Estadística y Censos
República Argentina



Encuesta Nacional de Gastos de los Hogares 2017-2018
Factores de expansión, estimación y cálculo de los errores de muestreo
Nota técnica n°4 - Mayo de 2020

Instituto Nacional de Estadística y Censos (INDEC)

Dirección: Marco Lavagna

Dirección Técnica: Pedro Ignacio Lines

Dirección Nacional de Difusión y Comunicación: María Silvina Viazzi

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

Coordinación de Producción Gráfica y Editorial: Marcelo Costanzo

Diseño y diagramación: Juan Garavaglia e Ignacio Pello

Revisión y corrección: Mariana Alonso, Horacio Barisani, Soledad Daffra y María Victoria Piñera

ISSN 2545-7179

ISBN 978-950-896-578-3

Instituto Nacional de Estadística y Censos - I.N.D.E.C.

Encuesta Nacional de Gastos de los Hogares 2017-2018 : Factores de expansión, estimación y cálculo de los errores de muestreo : nota técnica n°4. - 1a ed. - Ciudad Autónoma de Buenos Aires : Instituto Nacional de Estadística y Censos - INDEC, 2020.

Libro digital, PDF - (Notas técnicas ; 4)

Archivo Digital: descarga y online

ISBN 978-950-896-578-3

1. Metodología de la Investigación. 2. Encuestas. 3. Gastos. I. Título
CDD 310



Queda hecho el depósito que fija la ley n° 11.723

Libro de edición argentina

Buenos Aires, mayo de 2020

Publicaciones del INDEC

Las publicaciones editadas por el Instituto Nacional de Estadística y Censos están disponibles en www.indec.gob.ar y en el Centro Estadístico de Servicios, ubicado en Av. Presidente Julio A. Roca 609 C1067ABB, Ciudad Autónoma de Buenos Aires, Argentina. También pueden solicitarse al teléfono +54 11 51031-4632 en el horario de atención al público de 9:30 a 16:00. Correo electrónico: ces@indec.gob.ar

Calendario anual anticipado de informes: www.indec.gob.ar/indec/web/Calendario-Fecha-0

Índice

1. Introducción	4
2. Diseño muestral de la ENGHo.....	4
3. Factores de expansión de la encuesta.....	9
4. Estimación a partir de los datos de la encuesta	16
5. Indicadores de calidad asociados con el error de muestreo.....	18
6. Empleo de los pesos replicados con la base de datos para usuarios.....	21
7. Recomendaciones para el uso de los datos con fines estadísticos	31
Referencias.....	35
Anexo I. Total de UPM y USM por jurisdicción.....	37
Anexo II. Listado de localidades seleccionadas para la MMUVRA y la ENGHo	38
Anexo III. Distribución territorial de las UPM de la muestra de la ENGHo	40
Anexo IV. Cantidad de viviendas elegibles, no elegibles y de elegibilidad dudosa por jurisdicción.....	41
Anexo V. Cantidad de hogares en viviendas elegibles, con y sin respuesta por jurisdicción.....	42
Anexo VI. Cantidad de hogares por causa de no respuesta y jurisdicción.....	43
Anexo VII. Tasa de respuesta de los hogares.....	44
Glosario.....	46

1. Introducción

El Instituto Nacional de Estadística y Censos (INDEC) realizó la quinta Encuesta Nacional de Gastos de los Hogares (ENGHo) entre noviembre de 2017 y noviembre de 2018, con el objetivo de obtener información actualizada acerca de los gastos y los ingresos de los hogares y sus características sociodemográficas. La encuesta, de alcance nacional, permite caracterizar las condiciones de vida de los hogares, fundamentalmente en términos de su acceso a los bienes y servicios, y de sus ingresos monetarios o en especie.

Asimismo, la ENGHo proporciona información tanto para el cálculo de las ponderaciones del índice de precios al consumidor (IPC) como para la actualización de las estructuras de las canastas de bienes y servicios, que permite elaborar las líneas de pobreza e indigencia del Instituto. Además, se utiliza para las estimaciones de las cuentas nacionales y para el diseño de políticas públicas.

La presente publicación es una guía de referencia de la metodología adoptada para determinar los factores de expansión que se emplean en las estimaciones oficiales, y que permite calcular sus errores de muestreo para cualquiera de los resultados que surjan 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 la encuesta. A continuación, se describen los procesos de ajuste por elegibilidad, no respuesta y calibración que se llevaron a cabo sobre los factores de expansión o ponderadores surgidos del diseño muestral de la ENGHo.

Se exponen los motivos por los cuales se introduce una modalidad para el cálculo de los errores de muestreo que emplea replicaciones, y se detalla el proceso que da origen a los ponderadores asociados a las réplicas para estimar los errores de muestreo y cómo emplearlos en distintas herramientas de cálculo: R, Stata, SAS y Wesvar.

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

2. Diseño muestral de la ENGHo

El diseño muestral es relevante en toda operación estadística porque impacta en la calidad de las estimaciones, y en el costo y la organización de la encuesta. Dado que una porción significativa del presupuesto se destina a la recolección de los datos, el diseño muestral es un compromiso entre minimizar los costos de la colecta y maximizar la precisión en las estimaciones y la calidad de los datos.

En líneas generales, un diseño muestral está constituido por un marco de muestreo, en lo posible actualizado y que cubra lo mejor posible a la población objetivo; la cartografía asociada que permite identificar y alcanzar las unidades que la componen; información auxiliar que pueda ser empleada para determinar las probabilidades de inclusión de las unidades de muestreo; una regla probabilística que seleccione de manera aleatoria las unidades; un mecanismo de cálculo que brinde las estimaciones; y, finalmente, una estrategia que evalúe la precisión de los resultados a partir de la muestra a seleccionar.

Por lo general una muestra probabilística de viviendas para una encuesta a hogares se basa en un diseño muestral del tipo complejo; o sea, uno que emplea varias etapas para seleccionar la muestra y marcos de muestreo constituidos por unidades de áreas como unidades de muestreo. Involucra, asimismo, una estratificación y un muestreo probabilístico proporcional al tamaño, en una o más de todas sus etapas.

El empleo de diseños complejos surge principalmente por no contar con un marco de lista de todas las viviendas involucradas en el ámbito geográfico que abarca la encuesta y por restricciones de índole operativa. Cuando el estudio es de gran envergadura y 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, habría una alta probabilidad de que la muestra tenga una distribución geográfica muy dispersa.

Como resultado, los costos del operativo de campo de la encuesta serían excesivamente altos o prohibitivos para cualquier presupuesto. En particular, resultarían costosos los desplazamientos de los encuestadores para cubrir grandes distancias hasta alcanzar las viviendas, las posibles visitas para contactar a los informantes en distintos horarios, y las tareas de supervisión y control de la encuesta.

Para reducir los costos en la preparación de un diseño muestral específico para cada operativo, controlar los problemas que ocasiona la dispersión de las muestras señalada en los párrafos anteriores, e integrar y coordinar sus operaciones estadísticas, el INDEC emplea una modalidad bajo el esquema de muestra maestra. O sea, utiliza una única muestra probabilística, conocida como Muestra Maestra Urbana de Viviendas de la República Argentina (MMUVRA), que se caracteriza por mantener fijas las unidades de área que la conforman y su estructura probabilística asociada. Esta muestra maestra 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.

2.1 Unidades de muestreo y estructura probabilística

La MMUVRA es de alcance nacional y urbano. Para su diseño, inicialmente se desarrollaron dos etapas de selección probabilística. Cada unidad de primera etapa de muestreo (UPM) del diseño está definida por un aglomerado o localidad de al menos 2.000 habitantes según el Censo Nacional de Población, Hogares y Viviendas 2010 (CNPhyV 2010).

El conjunto de todas las UPM constituye el marco 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. A estas últimas se las denomina UPM “autorrepresentadas”.

Del resto de las UPM, se seleccionó un conjunto 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) con base en radios censales¹ y en la

¹ El radio censal es una de las unidades territoriales que emplea el Instituto para organizar la tarea en los censos; por lo general está constituido por aproximadamente 400 viviendas y, dependiendo de sus características, se lo clasifica en urbano, mixto o rural.

cartografía del CNPhyV 2010. En cada UPM, todas sus USM² en conjunto cubren territorialmente dicha unidad y determinan la envolvente o área de cobertura asociada, con lo que se conforma el marco de muestreo para la selección de segunda etapa. El diseño muestral se completa 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 que, con sus actualizaciones periódicas, da origen al marco de selección de viviendas de la MMUVRA. El listado de viviendas tiene un orden específico y una cartografía asociada, lo 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.³ Es sobre este listado que se realizan las subselecciones para las muestras de todas las encuestas a hogares del Instituto.⁴

Para determinar la muestra de viviendas de la ENGHo, se sumó al diseño de la MMUVRA una nueva etapa de selección probabilística de un tercer tipo de unidades de muestreo, denominados “segmentos”. Estos están constituidos por 3 viviendas particulares contiguas o próximas entre ellas dentro del listado. De esta forma se intenta concentrar los desplazamientos de los encuestadores y disminuir el costo de traslado en su recorrido.

Para fijar el tamaño del segmento se buscó balancear el impacto de una posible correlación interna, o “efecto conglomerado”, en las principales variables de la encuesta sobre las estimaciones, y la distribución de las cargas de trabajo durante el período en que se desarrolla el operativo de la encuesta. La ENGHo está en terreno durante un año y en todas las semanas del calendario hay una asignación de viviendas, que respeta en lo posible el balance entre la estructura y la estratificación de la muestra por jurisdicción.

La muestra definitiva de viviendas de la ENGHo se determina mediante una selección sistemática con igual probabilidad de segmentos sobre el listado de viviendas de la MMUVRA, con su última actualización a la fecha de la encuesta; la selección es independiente por jurisdicción y en cada unidad secundaria seleccionada para la MMUVRA.

En el siguiente cuadro se resumen las características básicas del diseño muestral de la ENGHo: las etapas de selección, las unidades definidas en cada etapa, cómo se establecieron las probabilidades de selección y los procedimientos aplicados.

² Estas unidades también son conocidas como “Áreas MMUVRA”, y en su conformación los radios censales por cuestiones operativas (extensión, densidad, inaccesibilidad, etc.) pueden sufrir recortes o agrupamientos (por ejemplo, para equilibrar la uniformidad de sus tamaños en términos de viviendas).

³ A la fecha de la ENGHo, la MMUVRA en su última actualización registraba un total de 2.053.958 viviendas particulares.

⁴ Esta propiedad de permitir definir submuestras de viviendas sobre la muestra maestra hace que se la identifique también como un marco secundario de muestreo de viviendas.

Cuadro 1. Características principales del diseño muestral

Etapas de selección	Definición de unidades	Probabilidades asignadas	Procedimientos de selección
Primera etapa	Aglomerados/localidades de 50.000 y más habitantes	1	Autorrepresentadas
	Aglomerados/localidades de 49.999 y menos habitantes	Proporcionales al total de población según datos del CNPVyH 2010	Estratificación y selección sistemática proporcional a tamaño
Segunda etapa	Áreas MMUVRA (área pequeña definida por radios censales dentro de cada aglomerado/localidad)	Proporcionales al total de viviendas según datos del CNPVyH 2010	Estratificación y selección sistemática proporcional a tamaño
Tercera etapa	Segmentos de 3 viviendas	Uniforme	Selección sistemática

Fuente: INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

En el Anexo I se presenta la composición de la MMUVRA en términos del total de UPM y USM involucradas en la ENGHo; en el Anexo II, el listado de las localidades que definen las UPM; y en el Anexo III, su distribución espacial dentro del territorio nacional.

2.3 Tamaño y distribución de la muestra

La muestra de viviendas seleccionada para la ENGHo es del orden de las 44.922 viviendas particulares. En el cálculo se tuvieron en consideración aspectos operativos, costos, precisión para las principales estimaciones en los dominios de interés, la distribución de la muestra en las 52 semanas de relevamiento y los niveles de no respuesta esperados.

Este último punto tuvo particular importancia en la determinación del tamaño de muestra, dado que la falta de respuesta para este tipo de encuestas suele ser importante. La pérdida total (por ausencias, rechazo, viviendas deshabitadas u otras causas) se estimó que oscilaba entre un 20 y 40% según el tamaño de la localidad o aglomerado. Para casos excepcionales, como la Ciudad Autónoma de Buenos Aires (CABA) y los partidos del GBA, se la fijó en 50%, que es un valor histórico alcanzado por la encuesta en esos territorios.

Como resultado, la muestra total seleccionada se distribuye por jurisdicción según el siguiente cuadro:

Cuadro 2. Distribución de la muestra de viviendas por jurisdicción

Jurisdicciones	Cantidad de viviendas
Total del país	44.922
CABA	4.320
Buenos Aires	10.038
Catamarca	1.230
Córdoba	2.286
Corrientes	1.224
Chaco	1.374
Chubut	1.338
Entre Ríos	1.554
Formosa	1.236
Jujuy	1.206
La Pampa	1.242
La Rioja	1.254
Mendoza	1.938
Misiones	1.248
Neuquén	1.296
Río Negro	1.332
Salta	1.296
San Juan	1.296
San Luis	1.266
Santa Cruz	1.254
Santa Fe	2.280
Santiago del Estero	1.314
Tucumán	1.236
Tierra del Fuego	864

Fuente: INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

2.4 Principales dominios de estimación

La población objetivo o de interés de la encuesta abarca a las personas residentes en los hogares de viviendas particulares de las localidades de la República Argentina con 2.000 o más habitantes. En el diseño muestral se definen como dominios geográficos de estimación las siguientes 6 regiones:

Gran Buenos Aires	Ciudad Autónoma de Buenos Aires y los 31 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, resto de partidos de la provincia de Buenos Aires no incluidos en región Gran Buenos Aires
Patagonia	Chubut, Neuquén, Río Negro, Santa Cruz y Tierra del Fuego

y a las 24 jurisdicciones que conforman el territorio nacional: 23 provincias y la CABA.

Uno de los objetivos principales de la encuesta es determinar la estructura de ponderaciones de un nuevo IPC, con lo cual se definieron otros niveles de agregación o dominios de análisis para los cuales se puede dar información confiable a partir de los resultados de la encuesta.

El siguiente cuadro resume la apertura definida como dominio de análisis para las estimaciones de gastos según la estructura de ponderaciones del IPC:

Cuadro 3. Dominios de análisis de la ENGHo 2017-2018

Dominio de análisis	Objetivo	
	Estructura de ponderaciones del IPC	Estructura de gastos e ingresos
Total nacional	División	Según distintos cortes de población hasta subclase de gasto
	Grupo	
	Clase	
	Subclase	
	Producto	
Región	División	Según distintos cortes de población hasta clase de gasto
	Grupo	
	Clase	
	Subclase	
	Producto	
Provincia	División	
	Grupo	

Fuente: INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

La muestra de viviendas de la ENGHo fue distribuida a lo largo de 48 semanas de relevamiento, de las 52 que tiene un año. Para cada región y para los grandes aglomerados, se determinaron 48 submuestras, una por semana, equivalentes en cuanto a sus tamaños y a su distribución por estratos.

En las UPM más pequeñas, en donde la cantidad de USM no alcanzaban para asegurar muestras semanales equivalentes, su distribución se realizó de modo de balancear las cantidades por mes y por tipo de semana (primera, segunda, tercera y cuarta del mes) a lo largo del año y por estrato de USM. Como consecuencia, en las UPM autorrepresentadas se realizaron encuestas a lo largo de las 48 semanas, mientras que en las localidades más pequeñas, solo durante 24 semanas.

3. Factores de expansión de la encuesta

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 expansión⁵ de las personas, hogares o viviendas, según corresponda, que cuentan con dicha característica.

⁵ Los términos “factores de expansión”, “ponderadores” o “pesos” en el contexto del documento hacen referencia siempre al mismo concepto.

Inicialmente una vivienda seleccionada para la ENGHo posee un factor de expansión atribuido por el diseño muestral y que es definido como:⁶

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

donde,

w_{1i} es la inversa de la probabilidad de inclusión de la i -ésima UPM;

w_{2ij} es la 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} es la inversa de la probabilidad de inclusión en la última etapa de muestreo de la k -ésima vivienda dentro de la j -ésima USM de la i -ésima UPM seleccionada.⁷

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. Lejos de la situación ideal, 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 por la encuesta.

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; 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 de la encuesta varios ajustes sobre el definido por diseño, lo que busca disminuir el impacto de estos inconvenientes sobre los estimadores y aumentar la calidad de los resultados.

3.1 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 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.

⁶ 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.

⁷ La probabilidad de inclusión de la k -ésima vivienda se corresponde con la probabilidad de selección sistemática de segmentos de 3 viviendas contiguas o próximas dentro de las USM seleccionadas.

El tratamiento de este ajuste lleva a clasificar las viviendas seleccionadas como “elegibles”, “no elegibles” y de “elegibilidad dudosa”. Para la ENGHo, y solo con el fin de ajustar los factores de expansión iniciales por no elegibilidad, se consideran:

- Viviendas elegibles (VEL) aquellas en donde se detecta una vivienda particular y en la que se responde la encuesta; o que presentan alguna de las siguientes categorías en la pregunta “Causa por la que no se realizó la entrevista”:
 - Ausencia: causas circunstanciales, viaje o vacaciones.
 - Rechazo: cualquiera de las razones expresadas.
 - Otras causas: duelo, alcoholismo, discapacidad, idioma extranjero.
- Viviendas no elegibles (VNE) son aquellas registradas como:
 - Deshabitada.
 - Demolida.
 - Fin de semana.
 - Construcción.
 - Vivienda usada como establecimiento.
 - Variaciones en el listado: no es vivienda, la dirección no existente.
- Viviendas de elegibilidad dudosa o elegibilidad desconocida (VED) son aquellas que se corresponden con alguna de las siguientes categorías:
 - Ausencia: no se pudo contactar en tres visitas o no se especificó ningún motivo de ausencia.
 - Variaciones en el listado: no existe lugar físico o no se especificó ningún motivo de variaciones en el listado.
 - Otras causas: problemas de seguridad, inaccesibilidad (problemas climáticos u otros) o cuando no se especificó ningún otro motivo.

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:⁸

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

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” disjuntos definidos exclusivamente para los ajustes. Estas clases surgen del cruce de la variable jurisdicción

⁸ En la simbología empleada en la guía, Σ representa la suma sobre todas las unidades que pertenecen al conjunto A .

(25), las divisiones “aglomerado EPH” y “resto de las UPM” (2), y los estratos de diseño de la MMUVRA para las USM⁹ (5).

En consecuencia, en cada clase c , con $c = 1, \dots, 250$, 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 encuesta¹⁰ 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 Anexo IV se presentan los resultados de la ENGHo en relación a la cantidad de VEL, VNE y VED, a nivel nacional y por jurisdicción, que intervienen con sus factores de expansión iniciales, w_{0ijk}^V , en los cálculos del factor a_{1c} .

3.2 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 respuesta¹¹ del hogar. Esto puede ocurrir debido a una serie de razones: que en el hogar ninguno de los que lo componen quiera responder, que haya ausencia temporal de sus miembros durante el período de la encuesta, o bien que hubo un primer contacto, pero por algún motivo o circunstancia fue imposible continuar con la entrevista.

En particular, en la ENGHo, se considera que un hogar no responde si se registra alguna de las siguientes categorías en “Causa por la que no se realizó la entrevista”, presente en el cuestionario:

- Ausencia: causas circunstanciales, viaje o vacaciones.
- Rechazo: cualquiera de las razones expresadas.
- No respuesta al cuestionario 2 o 3: el hogar no respondió a los cuestionarios de gastos diarios o gastos varios.
- Otras causas: duelo, alcoholismo, discapacidad, idioma extranjero.

Un objetivo ineludible en toda encuesta es tratar de disminuir la incidencia de la no respuesta, por lo que 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, visitar el hogar con ausentes en varias ocasiones antes de dar por concluida la encuesta.

Aun así, la no respuesta es un fenómeno siempre presente en una encuesta, y es una fuente potencial de sesgo en las estimaciones. La magnitud del sesgo debido a la falta de respuesta generalmente no se conoce, pero está directamente relacionada con las diferencias en las características bajo estudio entre los grupos de unidades que respondieron y los que no lo hicieron. También se ve afectada por un factor asociado a la correlación entre la característica que se indaga sobre la unidad y la probabilidad de que el encuestado de respuesta por ella.

⁹ En rigor, los estratos de USM por diseño varían entre dos y cinco dependiendo del estrato de UPM; por lo tanto, es posible que algunas de las celdas se encuentren vacías por definición.

¹⁰ En las fórmulas, $EL(c)$, $ED(c)$ y $NE(c)$ señalan los conjuntos EL, NE y ED restringidos a la clase de ajuste c .

¹¹ En ninguna circunstancia las viviendas seleccionadas para la encuesta son reemplazadas por otras viviendas por razones de no respuesta.

Por estos motivos, y en un intento de disminuir su efecto sobre las estimaciones, se ajustan los factores de expansión de los hogares que responden 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, 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 posible, 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 de la encuesta.

Para realizar las correcciones por no respuesta en la ENGHo, 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,¹² definido como:

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

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 a la ENGHo, y después de los dos ajustes realizados, viene dada por:

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

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} , en el Anexo V se presenta el total de los hogares con y sin respuesta registrados en la encuesta, y en el Anexo VI se incluye el total de hogares por causa de no respuesta a nivel nacional y por jurisdicción.

3.3 Ajuste por calibración

Los factores de expansión de cada hogar seleccionado y que responde \tilde{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 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

¹² 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, $w_{0ijkl}^H = w_{0ijk}^V$, dado que se incluyen en la muestra todos los hogares que forman parte de la vivienda seleccionada.

estimaciones de algunas variables de la encuesta 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 de la encuesta, w_{ijkl}^H , que surgen de la resolución del siguiente problema numérico de optimización:

$$\begin{aligned} & \text{Minimizar } \sum_{HR} G(\tilde{w}_{ijkl}^H, w_{ijkl}^H), \\ & \text{sujeto a: } \sum_{HR} w_{ijkl}^H \mathbf{x}_{ijkl}^H = \sum_U \mathbf{x}_q^H \end{aligned}$$

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, $\mathbf{x}_{ijkl}^H = (x_{ijkl1}^H, \dots, x_{ijklq}^H)^T$ medidas en la encuesta, a partir de los factores de expansión deseados, w_{ijkl}^H , reproduzcan sus totales poblacionales, $\sum_U \mathbf{x}_q^H = (t_{x1}^H, \dots, t_{xq}^H)$, 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. Para la ENGHo se emplearon 8 variables que reflejan la estructura demográfica por sexo y por grupos de edad, donde $\mathbf{x}_{ijkl}^H = (x_{ijkl1}^H, \dots, x_{ijkl8}^H)$ y cuyas componentes son:

- x_{ijkl1}^H = cantidad de mujeres en el hogar,
- x_{ijkl2}^H = cantidad de varones en el hogar,
- x_{ijkl3}^H = cantidad de personas en el hogar entre 0 y 14 años,
- x_{ijkl4}^H = cantidad de personas en el hogar entre 15 y 24 años,
- x_{ijkl5}^H = cantidad de personas en el hogar entre 25 y 34 años,
- x_{ijkl6}^H = cantidad de personas en el hogar entre 35 y 49 años,
- x_{ijkl7}^H = cantidad de personas en el hogar entre 50 y 64 años,
- x_{ijkl8}^H = cantidad de personas en el hogar de 65 años y más.

Los totales de población, involucrados como marginales para estas variables en el proceso iterativo, provienen de proyecciones poblacionales.¹³

En la calibración, se emplea la función de distancia “logit” (Deville y Särndal, 1992; Haziza y Beaumont, 2017) del paquete Survey de R,¹⁴ que permite controlar, en lo posible, el rango de los w_{ijkl}^H . Como la generación de pesos extremos impacta en la eficiencia del estimador y aumenta el riesgo de incrementar la variabilidad de las estimaciones, en el caso de la ENGHo, a la calibración se le sumó un proceso iterativo de recalibración, que consta de los siguientes pasos:

1. Calcular el percentil del 99% de los factores de expansión calibrados w_{ijkl}^H como punto de corte, α_{99} .

¹³ Los totales poblacionales proyectados fueron calculados a partir de datos censales de población según CNPhyV 2010 al 15 de agosto de 2018 y determinados por la Dirección Nacional de Estadísticas Sociales y Poblacionales del INDEC.

¹⁴ Versión 4.0 disponible en: <https://cran.r-project.org/web/packages/survey/index.html>.

2. Revertir los w_{ijkl}^H que superen el punto de corte al factor de expansión ajustado por no respuesta \tilde{w}_{ijkl}^H , o sea:

$$w_{ijkl}^H = \begin{cases} w_{ijkl}^H & \text{si } w_{ijkl}^H \leq \alpha_{99} \\ \tilde{w}_{ijkl}^H & \text{si } w_{ijkl}^H > \alpha_{99} \end{cases}$$

3. Volver a calibrar los factores de la calibración con las modificaciones según el paso 2.
4. Repetir los pasos 2 y 3 un máximo de cinco veces o hasta que ningún factor supere α_{99} definido en el paso 1.

El proceso de recalibración se efectúa en forma independiente por provincia o jurisdicción; 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; si el proceso no converge o continúa generando factores de expansión extremos, se lo reduce a nivel de jurisdicción.

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:

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

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 expresión es válida siempre que la vivienda y el hogar seleccionados pertenezcan a la clase de ajuste c , $c = 1, \dots, 250$.

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

4. Estimación a partir de los datos de la encuesta

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 a la mayoría se los 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, estadísticos de orden (medianas, quintiles, etc.) o más complejos (índice de Gini, índice de Atkinson, u otros de desigualdad).

Para alcanzar las estimaciones de esos parámetros en la ENGHo se emplean estimadores que recurren a los factores de expansión w_{ijkl}^H , que surgen de la última etapa de ajuste, y son del tipo de estimadores “calibrados” pertenecientes a la familia de los estimadores de regresión generalizada.¹⁵

El conjunto de pesos permite construir expresiones simples para los estimadores, aun cuando su estructura fuera compleja. En el caso de que Y y Z sean variables o características de interés medidas sobre los hogares, las formulaciones generales para los que más se emplean en la encuesta son:

Parámetro	Estimador ¹⁶
Total, T_y	$\hat{t}_y = \sum_R w_{ijkl}^H y_{ijkl}$
Promedio¹⁷, \bar{Y}	$\hat{y} = \frac{\sum_R w_{ijkl}^H y_{ijkl}}{\sum_R w_{ijkl}^H}$
Razón, $R_{yz} = \frac{T_y}{T_z}$	$\hat{R}_{yz} = \frac{\hat{t}_y}{\hat{t}_z} = \frac{\sum_R w_{ijkl}^H y_{ijkl}}{\sum_R w_{ijkl}^H z_{ijkl}}$

Los siguientes ejemplos se ajustan a los requerimientos principales de la encuesta en relación a parámetros asociados a las variables gastos, ingresos y cantidades:

Gastos de consumo totales: si es el “gasto de consumo de los hogares en educación”, una estimación del total, T_{GE} , se obtiene a partir de:

$$\hat{t}_{GE} = \sum_R w_{ijkl}^H g_{Eijkl}$$

donde w_{ijkl}^H y g_{Eijkl} son el factor de expansión y el gasto de consumo en educación del hogar l , de la vivienda k , en la USM j , de la UPM i a la cual pertenece la unidad, respectivamente.

¹⁵ Se recuerda que tanto una variable medida a nivel hogar como una medida a nivel individuo emplean el mismo factor de expansión w_{ijkl}^H , como se señala en el apartado 3.3.

¹⁶ En todos los casos, Σ en las fórmulas hace referencia a sumar sobre los hogares que responden a la encuesta..

¹⁷ La definición de los parámetros promedio y proporción coincide si Y es una variable binaria, que toma el valor de 1 cuando la unidad posee una característica dada y 0 en caso contrario.

Gastos de consumos relativos: en el caso del “gasto de consumo en educación relativo al gasto total”, $GR_E = \frac{T_{GE}}{T_{GT}}$ se lo estima a través del estimador por razón o cociente:

$$\widehat{GR}_E = \frac{\hat{t}_{GE}}{\hat{t}_{GT}} = \frac{\sum_R w_{ijkl}^H g_{Eijkl}}{\sum_R w_{ijkl}^H g_{Tijkl}}$$

donde g_{Tijkl} es el gasto en consumo total del hogar l , en la vivienda k , en la USM j , de la UPM i a la cual pertenece la unidad.

Gastos medios: para el caso del “gasto medio mensual por hogar”, \bar{G}_m , se emplea el estimador por cociente:

$$\hat{G}_m = \frac{\hat{t}_{G_m}}{\hat{N}_H} = \frac{\sum_R w_{ijkl}^H g_{Tijkl}}{\sum_R w_{ijkl}^H}$$

donde, \hat{N}_H es una estimación a partir de la encuesta del total de hogares.

Ingresos totales: para el caso del “ingreso total de los hogares”, T_{Ihog} , se lo estima como:

$$\hat{t}_{Ihog} = \sum_R w_{ijkl}^H I_{ijkl}$$

donde I_{ijkl} es el ingreso total del hogar l , de la vivienda k , en la USM j , de la UPM i a la cual pertenece la unidad.

Ingresos medios: para el caso del “ingreso medio mensual por hogar”, \bar{I}_{hog} , se emplea el estimador por cociente:

$$\hat{I}_{hog} = \frac{\hat{t}_{Ihog}}{\hat{N}_H} = \frac{\sum_R w_{ijkl}^H I_{ijkl}}{\sum_R w_{ijkl}^H}$$

Cantidad total adquirida: la cantidad adquirida del bien b , \hat{C}_b se calcula como:

$$\hat{C}_b = \sum_R w_{ijkl}^H C_{bijkl}$$

donde C_{bijkl} es la cantidad adquirida del bien b en el hogar l , de la vivienda k , en la USM j , de la UPM i a la cual pertenece la unidad, respectivamente.

Cantidad media adquirida: para el caso de la “cantidad media mensual adquirida por el hogar del bien b ”, \bar{C}_b , un estimador del parámetro se obtiene como cociente entre \hat{C}_b y el estimador de la cantidad de hogares a la fecha de la encuesta, \hat{N}_H :

$$\hat{C}_b = \frac{\hat{C}_b}{\hat{N}_H} = \frac{\sum_R w_{ijkl}^H C_{bijkl}}{\sum_R w_{ijkl}^H}$$

5. 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 permite a los usuarios cuantificar el grado de confianza y conocer las limitaciones que pueden llegar a tener los resultados, y así, restringir su uso cuando las estimaciones no alcanzan ciertos estándares definidos para la encuesta.

En una encuesta que emplea una muestra probabilística como la ENGHo, la inferencia estadística sobre la población objetivo se basa en los datos recopilados de solo una parte de esta población. Por tal motivo, los resultados probablemente diferirán de los que se pueden obtener a partir de un censo completo a la población objetivo de la encuesta.

El error que se genera al extraer conclusiones en términos estadísticos para toda la población basándose 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 de la encuesta depende de algunos aspectos del diseño muestral como el número de etapas y el método de selección; también influyen el tamaño de la muestra, el estimador que se emplea y 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.

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 con respecto al promedio de estas según el diseño muestral. A partir de la varianza muestral se pueden definir otras medidas más populares, como el error estándar (EE) y el coeficiente de variación (CV); o más complejas de interpretar, como el efecto de diseño (ED) o el intervalo de confianza (IC).

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; o sea, es una medida relativa que generalmente se expresa como un porcentaje. 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 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.

5.1 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 de la encuesta. 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 a 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 insumos para cuantificarlos;
- proteger y anonimizar la 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.

5.2 Determinación de las réplicas *Bootstrap*

Existen distintos métodos para conformar las réplicas (Wolter, 2007). El que se adopta para generar las submuestras en la ENGHo es el *bootstrap* propuesto en Rao y Wu (1998) 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 , con $b = 1, \dots, B$, el procedimiento consiste en seleccionar en cada estrato de diseño, h , una muestra simple al azar con reemplazo de $n_h - 1$ conglomerados a partir de la muestra original de n_h conglomerados. Luego, se define el peso *bootstrap* $w_{hml}^{*(b)}$ a partir de un peso inicial w_{hml} para la l -ésima unidad en el conglomerado m del estrato h en la réplica b según el siguiente ajuste:

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

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

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

donde:

$\hat{\theta}$ es el estimador¹⁸ 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:

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

y el del coeficiente de variación:

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

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 emplear este método 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 la ENGHo, 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 la ENGHo fueron determinadas en forma independiente en cada jurisdicción. Para ajustarse a los requerimientos del método, en las UPM autorrepresentadas de la encuesta, 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

¹⁸ Ver sección 5.

y limitar el tiempo de cálculo, entre otras cuestiones. Por estos motivos, en la ENGHo el total de réplicas es de 200 ($B=200$), cantidad que asegura la estabilidad del estimador de varianza para las principales estimaciones de la encuesta.

Todas las réplicas se obtienen de la muestra original, que incluye a todos los hogares y personas de las viviendas elegibles, cuyos factores de expansión vienen dados por $w_{0ijkl}^V a_{1c}$. 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 de la encuesta, se repiten dichos ajustes sobre los pesos replicados. Es decir, para cada una de las 200 réplicas, los pesos *bootstrap* son ajustados nuevamente por no respuesta y calibrados por sexo y edad de manera análoga a como se detalla en la sección 4. A diferencia de los pesos originales, los pesos *bootstrap* no son sometidos a un proceso de redondeo.

Como consecuencia, de todo el procedimiento detallado en la sección anterior, la ENGHo dispone de un conjunto de 200 réplicas, $\{w_{ijkl}^{*H(b)}, b = 1, \dots, 200\}$, que vinculados a la base con los microdatos o usuaria permiten calcular los errores muestrales para las estimaciones oficiales de la encuesta.

6. Empleo de los pesos replicados con la base de datos para usuarios

La presente sección constituye una guía de cómo deben ser empleadas las réplicas que acompañan la base usuario de la encuesta en distintas herramientas de cálculo: R,¹⁹ SAS,²⁰ Stata²¹ y Wesvar.²² En caso de no contar con ellas, se presenta un ejemplo que sugiere cómo efectuar el cálculo siguiendo la definición formulada en [1] de la sección 5.2, y que cualquier usuario puede poner en práctica con pocos recursos.²³

Se advierte que la guía no constituye un manual exhaustivo de cada una de las herramientas y sus opciones, 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 lograrlos.

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 5.

Para facilitar las indicaciones, la presentación adopta la notación empleada en (1) la base de usuarios de la ENGHo y en (2) la base con las réplicas, para las principales variables de interés para esta sección:²⁴

¹⁹ Versión 4.0 disponible en: www.r-project.org.

²⁰ Versión 9.4 M3 disponible en: www.sas.com.

²¹ Versión 15 disponible en: www.stata.com.

²² Versión 5.1 disponible en: www.westat.com/capability/information-systems-software/wesvar.

²³ No se incluye la herramienta de cálculo SPSS, ya que a la fecha no cuenta oficialmente con la posibilidad de emplear la metodología desarrollada sin recurrir a una programación *ad hoc*.

²⁴ Se sugiere al usuario leer el [Manual de uso de la base de datos usuarios](#) correspondiente a la encuesta para los detalles, así como los diccionarios vinculados a cada base.

- **id**: variable de identificación de registro, presente en (1) y (2);
- **pondera**: factor de expansión final de la encuesta,²⁵ presente en (1);
- **w_repb**: peso *bootstrap* replicado, donde *b* representa el número de réplica al cual corresponden los pesos, tomando los valores de 1 a 200, presente en (2).

Por otro lado, se asume que Y, Z son variables genéricas (continuas o categóricas), que hacen referencia a características indagadas por la encuesta para las cuales se requieren estimaciones de los parámetros poblacionales de interés (ver sección 5), y de sus respectivas estimaciones de los errores de muestreo.

Para poder seguir correctamente estas instrucciones, primero es necesario vincular la base de los microdatos de la encuesta 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_encuesta**, se puede construir por lo general a través de la sentencia u opción “**join**” o “**merge**” en la mayoría de las herramientas de cálculo propuestas.

Como resultado, cada unidad (persona u hogar) o registro de la base usuaria poseerá su factor de expansión asociado (**pondera**) y cada uno de los 200 valores de los pesos *bootstrap* replicados (**w_repb**).

6.1 Cálculo del error de muestreo a través de R

Una de las posibilidades disponibles en la comunidad *open source*, y que acepta la metodología propuesta en esta guía, es el paquete Survey de R, el mismo que se emplea para la etapa de calibración mencionado en los párrafos anteriores. Siguiendo las indicaciones del manual²⁶, y asumiendo que **base_encuesta** fue creada como se indica en la sección anterior, se define el objeto **disenio**²⁷, que incluye las componentes que se requieren para los cálculos a través de la opción **svrepdesign**.

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

```
library(survey)
disenio=svrepdesign(data=base_encuesta,
                   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:

²⁵ Etiqueta que hace referencia al factor de expansión w_{ijkl}^H , definido en la sección 3.

²⁶ Disponible en inglés en: <https://cran.r-project.org/web/packages/survey/survey.pdf>.

²⁷ El usuario puede optar por cualquier otro nombre para el objeto.

Estimador	Estimaciones por Survey
\hat{t}_y	svytotal(~Y,design= disenio) cv(svytotal(~Y,design= disenio))
\hat{y}	svymean(~Y,design= disenio) cv(svymean(~Y,design= disenio))
\hat{R}_{YZ}	svyratio(~Y,~Z,disenio) cv(svyratio(~Y,~Z,disenio))

6.2 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 **base_encuesta** 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 de la encuesta **pondera**, los pesos replicados **w_rep***, y el método para el cálculo de la varianza **bootstrap**. Asimismo, se debe incluir la opción **mse** e indicar que el estimador de varianza **bootstrap** es el considerado en la sección 5. Para preparar la herramienta para las estimaciones, el usuario debe invocar:

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

A continuación, y habiendo definido el **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, y una razón:

Estimador	Estimaciones por Stata
\hat{t}_y	svy bootstrap : total Y estat cv
\hat{y}	svy bootstrap : mean 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 intervalo de confianza 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, al no contar con la opción **bootstrap**, se debe proceder como se indicó en los párrafos anteriores, pero se tendrá que invocar al prefijo **svyset** con la opción **brweight**, y **brr** en la opción **vce**. De esta forma, se podrán obtener estimaciones válidas para el EE, CV o IC. 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).

6.3 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_encuesta** 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_rep200** vía **repweight** y el 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_encuesta sum cvsum varmethod= Bootstrap ; repweight w_rep1--w_rep200 ; weight pondera ; var Y; run;
\hat{y}	proc surveymeans data= base_encuesta mean cv varmethod= Bootstrap ; repweight w_rep1--w_rep200 ; weight pondera ; var Y; run;
\hat{R}_{YZ}	proc surveymeans data= base_encuesta varmethod= Bootstrap ; repweight w_rep1--w_rep200 ; 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).

6.4 Cálculo del error de muestreo a través de Wesvar

Wesvar²⁸ 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.

Debe advertirse que, a diferencia de las demás herramientas presentadas en esta sección, Wesvar no permite vincular la base de usuarios con la base de réplicas dentro de su entorno. 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

²⁸ Disponible en: www.westat.com/capability/information-systems-software/wesvar. Se puede acceder de forma gratuita a la documentación de Wesvar enviando un correo electrónico a: wesvar_tech_support@westat.com.

una base de datos Wesvar (.var) a partir de la base de la encuesta con las réplicas (**base_encuesta**), 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 el espacio de trabajo donde se encuentra.²⁹

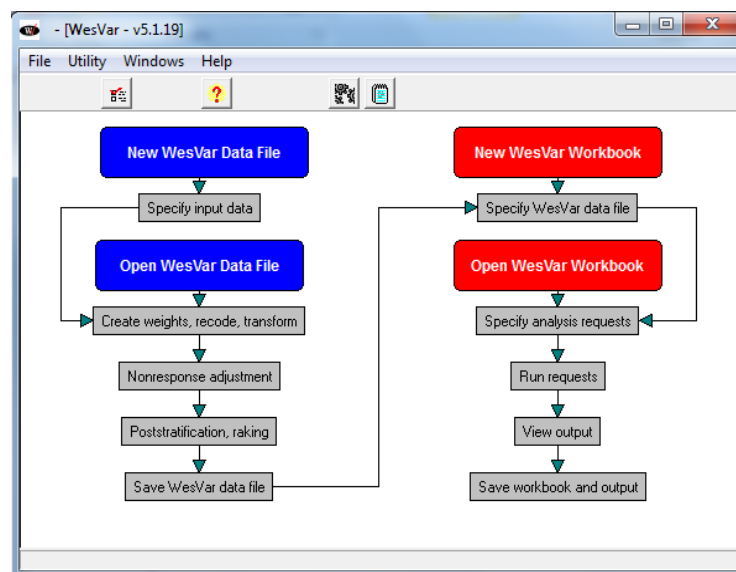


Figura 1

Al usuario le aparece una ventana como la que se ve en la figura 2, donde debe completar la información necesaria para comenzar 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* de la encuesta, **w_rep1,...,w_rep200**; y en el apartado **Full Sample**, el factor de expansión final de la encuesta, **pondera**. En **Method** se debe optar por **BRR**, que brinda resultados válidos para las estimaciones de los errores de muestreo empleando los pesos *bootstrap* de la encuesta (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 iniciar sus estimaciones.

²⁹ Se advierte que la herramienta tiene la posibilidad de importar datos en formato csv/txt con delimitadores, SAS o SPSS.

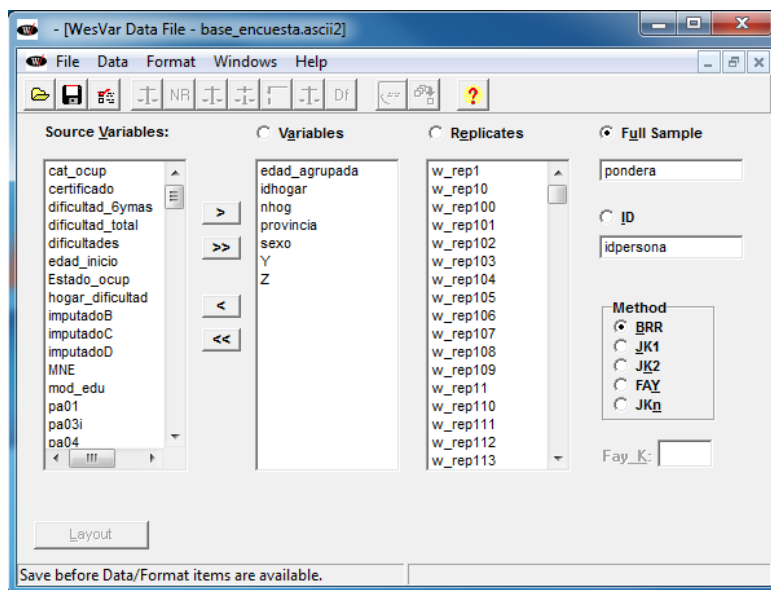


Figura 2

En el paso siguiente se debe crear un libro de trabajo haciendo clic sobre la etiqueta “New Wesvar Workbook” del diagrama de flujo (figura 1, en rojo), 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 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**).

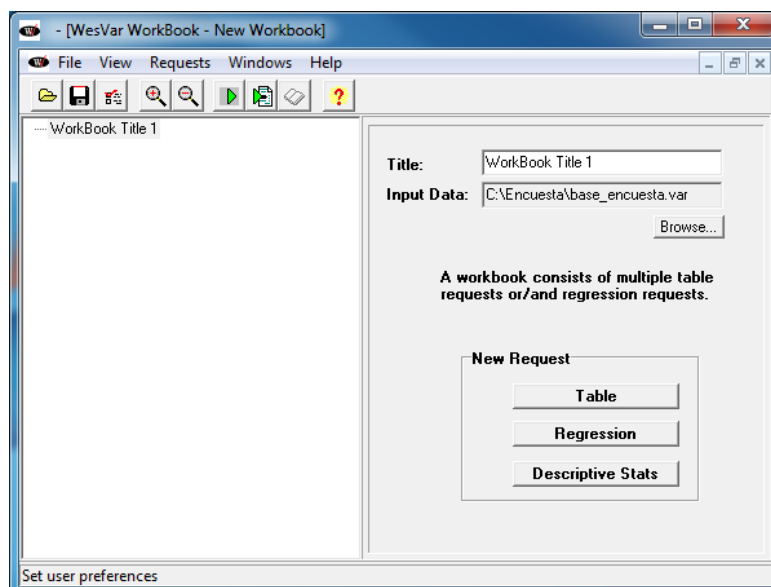


Figura 3

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.

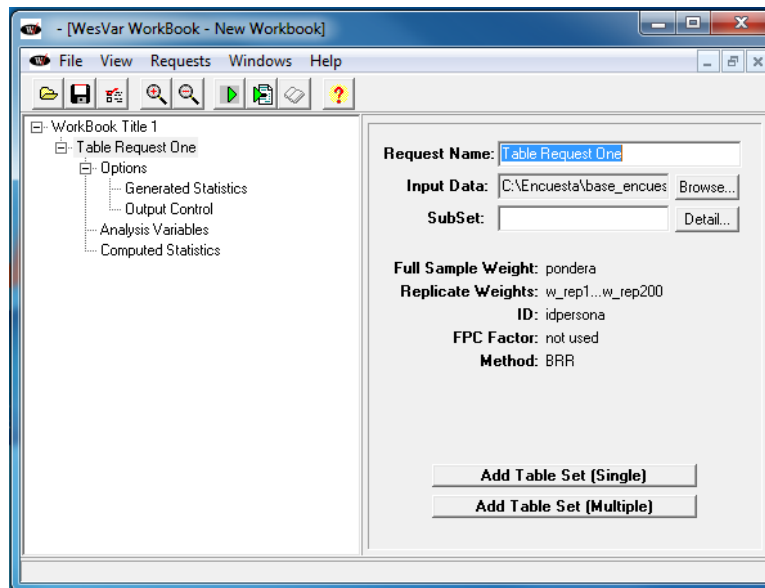


Figura 4

Haciendo clic en el nodo “Analysis Variables”, en el panel izquierdo, 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.

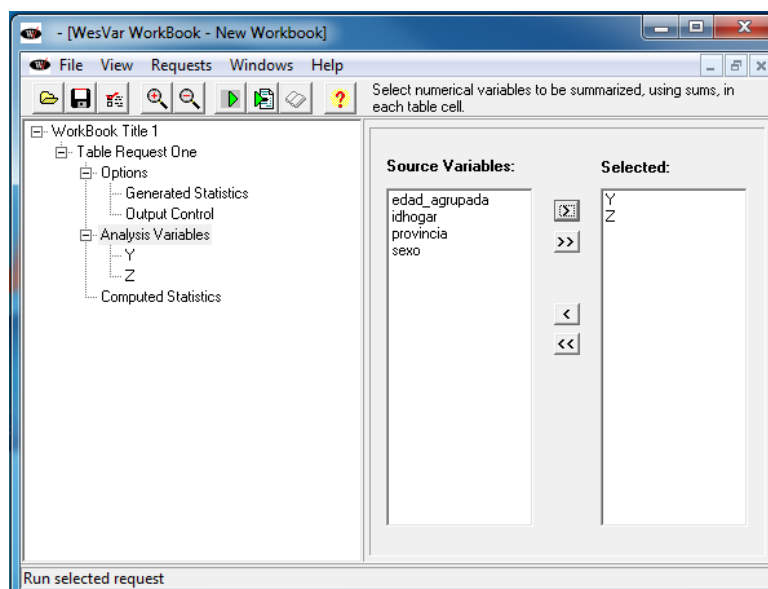


Figura 5

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).

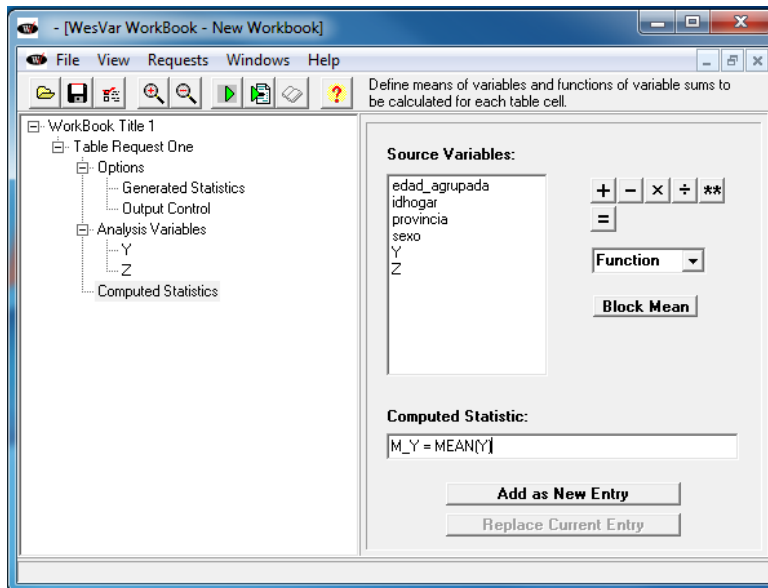


Figura 6

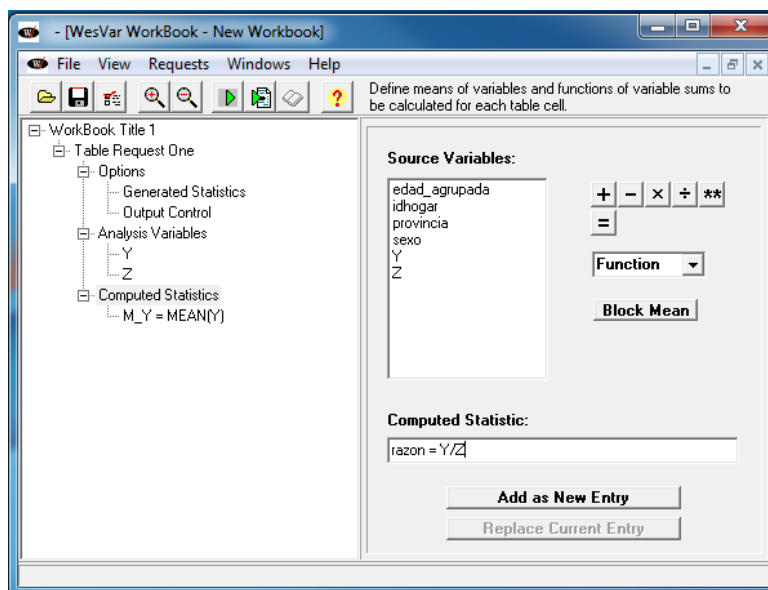


Figura 7

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

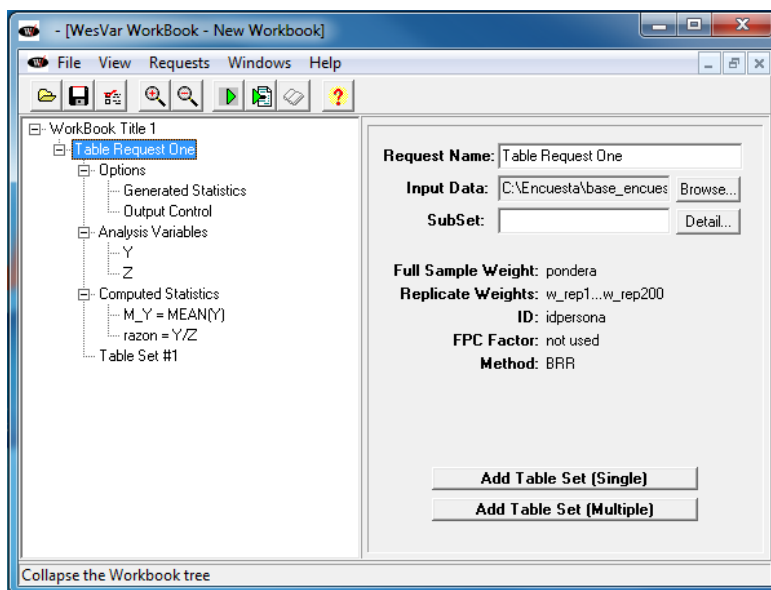


Figura 8



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  y seleccionando el nodo sobre el panel izquierdo “Overall”, 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.

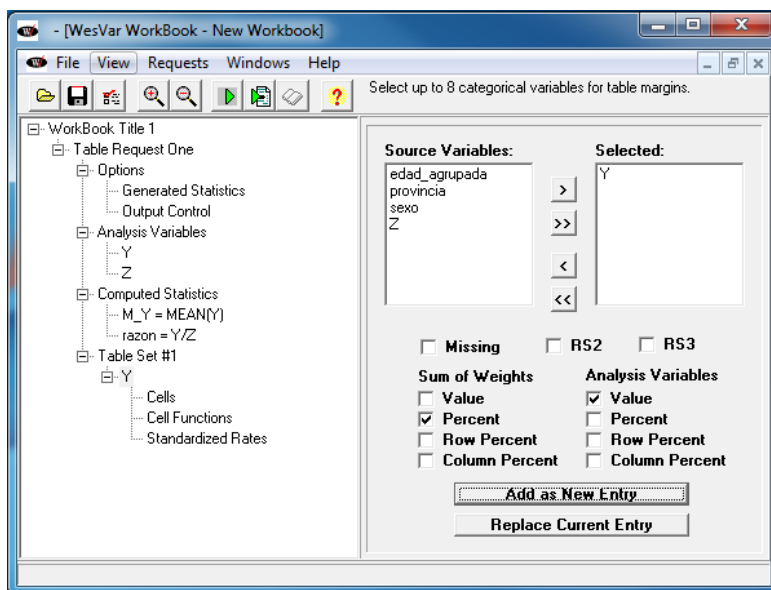


Figura 10

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.

6.5 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 5.2, empleando las fórmulas [1] a [3].

Por ejemplo, si se asume que la variable Y está medida sobre los hogares de la encuesta, la expresión que se debe emplear como estimador para un total t_y , según se lo definió en la sección 4, es:

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

Siguiendo lo señalado en la sección 5.2, la formulación para la varianza *bootstrap* [1] de un estimador es:

$$v_B(\hat{\theta}) = \frac{1}{200} \sum_{b=1}^{200} (\hat{\theta}_{(b)}^* - \hat{\theta})^2$$

Empleando el conjunto de réplicas $\{w_{ijkl}^{*H(b)}, b = 1, \dots, 200\}$ y reemplazando $\hat{\theta}$ por \hat{t}_y , y $\hat{\theta}_{(b)}^*$ por $\hat{t}_{y(b)}^*$, donde $\hat{t}_{y(b)}^* = \sum_R w_{ijkl}^{*H(b)} y_{ijkl}$ es la estimación del total a partir de los factores de expansión $w_{ijkl}^{*H(b)}$ para la l -ésima hogar en la b -ésima submuestra *bootstrap*, $b = 1, \dots, 200$, permite calcular estimaciones para la varianza *bootstrap* de \hat{t}_y , a través de:

$$v_B(\hat{t}_y) = \frac{1}{200} \sum_{b=1}^{200} (\hat{t}_{y(b)}^* - \hat{t}_y)^2 \quad [4]$$

para el error estándar, según

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

y para del coeficiente de variación con

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

De manera análoga se procede para los casos de un promedio, o un cociente o razón, reemplazando en [1] $\hat{\theta}$ por \hat{y} o por \hat{R}_{yz} , (ver sección 5) y las estimaciones *bootstrap* $\hat{\theta}_{(b)}^*$ que emplean las réplicas por:

$$\hat{y}_{(b)}^* = \frac{\sum_R w_{ijkl}^{*H(b)} y_{ijkl}}{\sum_R w_{ijkl}^{*H(b)}} \quad \text{o} \quad \hat{R}_{yz(b)}^* = \frac{\sum_R w_{ijkl}^{*H(b)} y_{ijkl}}{\sum_R w_{ijkl}^{*H(b)} z_{ijkl}}$$

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.

7. Recomendaciones para el uso de los datos con fines estadísticos

No es posible asumir en todos los resultados de la encuesta la misma confianza. 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 de la encuesta, o son marginales para la población o subpoblación en estudio;
- la cantidad de hogares o personas involucradas en la estimación es escasa como consecuencia de los niveles de desagregación deseados, o se trata de un fenómeno poco representado en la población;
- la estimación de un total involucrado en el denominador de un cociente o razón posee una variabilidad o coeficiente de variación muy alto.

En todas estas situaciones, el comportamiento del estimador empleado, tanto el del parámetro 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 las situaciones mencionadas.

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:

- en dominios de análisis donde participan pocas unidades de la muestra en los “últimos conglomerados”,
- si la característica que se estudia 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, como consecuencia de los procesos de ajuste.

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 de la encuesta para sus propias estimaciones que deberán poner atención y ser prudentes a la hora de sacar conclusiones en ciertas circunstancias.

7.1 Recomendaciones sobre las estimaciones

Para ayudar al usuario a interpretar los resultados de la encuesta, 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 evaluar su precisión o la confianza que se puede poner en ella. Cualquier lector de los resultados oficiales publicados de la encuesta, 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 de la encuesta.

Cuadro 4. 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, Encuesta Nacional de Gastos de los Hogares 2017-2018.

Se insiste con la recomendación de que, en el caso que algunas de las estimaciones sean consideradas no confiables o poco confiables para inferir al 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 4, definido por el Instituto como estándar para la encuesta.

7.2 Recomendaciones para estimaciones en dominios

Otro aspecto importante a tener en cuenta por los usuarios de la base de datos de la encuesta es la manera en que se calculan las estimaciones en subpoblaciones o dominios. 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 se emplea esa modalidad 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 subestimarla. Por lo general, la documentación que acompaña la herramienta contempla esta advertencia. En particular, en aquellas presentadas en las secciones 6.1 a 6.3, los usuarios que deseen calcular estimaciones en subpoblaciones o dominios pueden recurrir a las opciones **subset**³⁰ en R, **subpop** en Stata, y **domain** en SAS para obtener en forma adecuada la estimación del CV o del EE que esté calculando.³¹

7.3 Recomendaciones sobre el cálculo de intervalos de confianza

Los intervalos de confianza (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:

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

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 $\hat{\theta}$

³⁰ En el paquete Survey es posible utilizar también el comando **svyby** para obtener estimaciones en subpoblaciones.

³¹ 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.

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 de la encuesta no es un muestreo simple al azar (MSA), e involucra distintas etapas con probabilidades de selección proporcionales a tamaños y estratificaciones. Esta complejidad en el diseño, por lo general, lleva a que el conjunto de datos no siga la hipótesis *i. i. d.*, o sea, la de independientes e idénticamente distribuidos, requerida en este contexto para sostener el supuesto de normalidad (Heeringa, West y Berglund, 2017).

En virtud de lo expuesto, se sugiere a los usuarios 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

- 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. Proceedings American Statistical Associations. *Survey Research Methods Section - JSM*, 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. y Särndal, C.E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87. Recuperado de: [DOI:10.1080/01621459.1992.10475217](https://doi.org/10.1080/01621459.1992.10475217).
- Frankel, L. R. (1983). The Report of the CASRO Task Force on Response Rates. En F. 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*, 6(1). 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](https://doi.org/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](https://doi.org/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](https://doi.org/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, 231-241. Recuperado de: [DOI:10.1080/01621459.1988.10478591](https://doi.org/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, 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: 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. *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](https://doi.org/10.1007/978-0-387-35099-8).

Anexo I. Total de UPM y USM por jurisdicción

Cuadro 5. Total de UPM y USM de la MMUVRA presentes en la ENGHo

Jurisdicción	UPM			USM
	Autorrepresentadas	No autorrepresentadas	Total	Total
Total del país	205	187	392	5.351
CABA	1	-	1	179
Buenos Aires	14	26	40	1.190
Catamarca	14	-	14	152
Córdoba	5	19	24	344
Corrientes	9	13	22	204
Chaco	9	10	19	197
Chubut	12	-	12	188
Entre Ríos	11	10	21	259
Formosa	12	5	17	158
Jujuy	4	9	13	148
La Pampa	8	7	15	158
La Rioja	6	6	12	139
Mendoza	6	8	14	208
Misiones	9	12	21	194
Neuquén	11	3	14	156
Río Negro	11	8	19	221
Salta	9	12	21	190
San Juan	6	7	13	129
San Luis	15	-	15	167
Santa Cruz	14	-	14	131
Santa Fe	6	17	23	307
Santiago del Estero	5	7	12	116
Tucumán	5	8	13	154
Tierra del Fuego	3	-	3	62

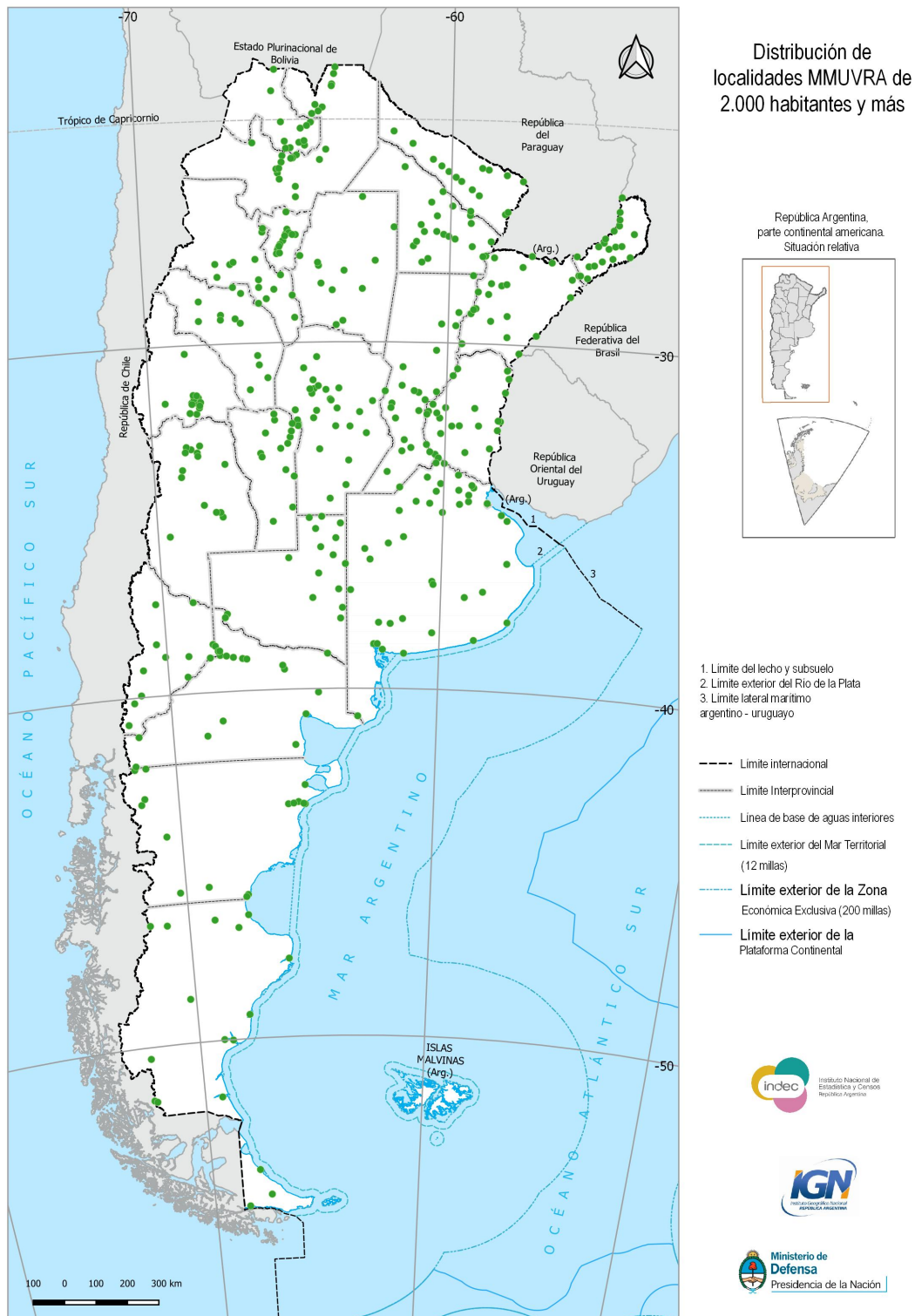
Fuente: INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

Anexo II. Listado de localidades seleccionadas para la MMUVRA y la ENGHo

Provincia	Localidades
	Ciudad Autónoma de Buenos Aires.
Buenos Aires	Partidos del Gran Buenos Aires 30 de Agosto, Arrecifes, Ayacucho, Bahía Blanca, Baradero, Berisso, Campana, Carlos Casares, Carmen de Patagones, Chacabuco, Chivilcoy, Coronel Pringles, Dolores, Ensenada, General Daniel Cerri, General Mansilla, General Rojo, Junín, La Plata, Lincoln, Luján, Mar del Plata, Mercedes, Monte Hermoso, Necochea-Quequén, Olavarría, Parada Robles-Pavón, Pehuajó, Pergamino, Punta Alta, Ramallo, Rivera, Ruta Sol, Salto, San Antonio de Areco, San Nicolás de los Arroyos, Sierra de la Ventana, Tandil, Tornquist, Trenque Lauquen, Tres Arroyos, Villa Alfredo Fortabat, Villa La Arcadia, Zárate.
Catamarca	Andalgalá, Belén, Chumbicha, Fiambalá, Huillapima, Londres, Los Altos, Pomán, Recreo, San Fernando del Valle de Catamarca, San Isidro, San José, Santa María, Saujil, Tinogasta
Córdoba	Bella Ville, Biale Massé, Córdoba, Cosquín, Cruz del Eje, Dean Funes, Estancia Vieja, Hernando, Huerta Grande, La Calera, La Carlota, La Falda, Laboulaye, Las Chacras, Las Higueras, Las Tapias, Malvinas Argentinas, Mendiolaza, Oncativo, Parque Norte-Ciudad de los Niños-Villa Pastora-Almirante Brown, Pilar, Pozo del Molle, Río Cuarto, Río Primero, Río Segundo, San Agustín, San Antonio de Arredondo, San Francisco, San Pedro, San Roque, Santa Magdalena, Saturnino María Laspiur, Tanti, Valle Hermoso, Villa Allende, Villa Carlos Paz, Villa de las Rosas, Villa del Dique, Villa Dolores, Villa María, Villa Nueva, Villa Río Icho Cruz, Villa Rumipal, Villa Sarmiento.
Corrientes	Bella Vista, Colonia Liebig's, Corrientes, Curuzú Cuatiá, Esquina, Gobernador Igr. Valentín Virasoro, Goya, Itá Ibaté, Ituzaingó, Lavalle, Mercedes, Monte Caseros, Nuestra Señora del Rosario de Caá Catí, Paso de los Libres, Perugorría, Saladas, San Luis del Palmar, San Roque, Santa Lucía, Santa Rosa, Santo Tomé, Tabay.
Chaco	Barranqueras, Campo Largo, Charata, Concepción del Bermejo, Coronel Du Graty, Fontana, General José de San Martín, Juan José Castelli, La Leonesa, Las Breñas, Las Palmas, Machagai, Makallé, Nueva Pompeya, Pampa del Indio, Presidencia de la Plaza, Presidencia Roque Sáenz Peña, Puerto Vilelas, Quitilipi, Resistencia, Tres Isletas, Villa Angela, Villa Río Bermejito.
Chubut	Comodoro Rivadavia, Dolavon, El Maitén, Esquel, Gaiman, Gobernador Costa, Lago Puelo, Playa Unión, Puerto Madryn, Rada Tilly, Rawson, Río Mayo, Sarmiento, Trelew, Trevelín.
Entre Ríos	Aldea Valle María, Basavilbaso, Chajarí, Colón, Colonia Avellaneda, Concepción del Uruguay, Concordia, Crespo, Diamante, Federación, General Ramírez, Gualeguay, Gualeguaychú, La Paz, Lucas González, Nogoyá, Paraná, San José, Santa Elena, Viale, Victoria, Villaguay.
Formosa	Clorinda, Comandante Fontana, El Colorado, Estanislao del Campo, Formosa, Ibarreta, Ingeniero Guillermo N. Juárez, Laguna Blanca, Laguna Yema, Las Lomitas, Misión Tacaaglé, Palo Santo, Pirané, Pozo del Tigre, Villa del Carmen, Villa General Manuel Belgrano, Villa Kilómetro 213.
Jujuy	Abra Pampa, Aguas Calientes, Caimancito, El Carmen, El Piquete, La Quiaca, Libertador General San Martín, Maimará, Palpalá, Perico, San Pedro, San Salvador de Jujuy, Santa Clara, Yuto.
La Pampa	25 de Mayo, Catrilo, Colonia Barón, Eduardo Castex, General Acha, General Pico, General San Martín, Guatraché, Ingeniero Luiggi, Intendente Alvear, Macachín, Rancul, Realicó, Santa Rosa, Toay, Victorica.
La Rioja	Aimogasta, Chamental, Chepes, Chilecito, La Rioja, Milagro, Nonogasta, Olta, Salicas-San Blas, Villa San José de Vinchina, Villa Sanagasta, Villa Unión.
Mendoza	Eugenio Bustos, General Alvear, Godoy Cruz, Guaymallén, La Paz, Las Heras, Luján de Cuyo, Maipú, Malargüe, Mendoza, Perdriel, Real del Padre, Rivadavia, Rodeo del Medio, San Martín, San Rafael, Tres Porteñas, Tunuyán, Villa Atuel.
Misiones	25 de Mayo, Apóstoles, Aristóbulo del Valle, Capioví, Cerro Azul, Concepción de la Sierra, Dos de Mayo, El Soberbio, Eldorado, Garupá, Jardín América, Leandro N. Alem, María Magdalena, Montecarlo, Oberá, Posadas, Puerto Esperanza, Puerto Iguazú, Puerto Rico, San Pedro, San Vicente.
Neuquén	Aluminé, Centenario, Chos Malal, Cutral Có, Junín de los Andes, Las Lajas, Neuquén, Picún Leufú, Plaza Huinca, Plottier, Rincón de los Sauces, San Martín de los Andes, San Patricio del Chañar, Senillosa, Villa La Angostura, Zapala.
Río Negro	Allen, Catriel, Cinco Saltos, Cipolletti, El Bolsón, General Conesa, General Roca, Ingeniero Luis A. Huergo, Lamarque, Los Menucos, Luis Beltrán, Maquinchao, Río Colorado, San Antonio Oeste, San Carlos de Bariloche, Sierra Grande, Viedma, Villa Manzano, Villa Regina.
Salta	Aguaray, Apolinario Saravia, Campo Santo, Cerrillos, Chicoana, Colonia Santa Rosa, Coronel Moldes, Embarcación, General Güemes, General Mosconi, La Caldera, La Merced, Las Lajitas, Misión El Cruce-El Milagro-El Jardín de San Martín, Pichanal, Profesor Salvador Mazza, Rosario de la Frontera, Rosario de Lerma, Salta, San Antonio de los Cobres, San José de Metán, San Ramón de la Nueva Orán, Tartagal, Vaqueros.
San Juan	9 de Julio, Alto de Sierra, Barreal-Villa Pituil, Caucete, Chimbass, Los Berros, Rawson, Rivadavia, San José de Jáchal, San Juan, Santa Lucía, Villa Aberastain-La Rinconada, Villa Barboza-Villa Nacusi, Villa Borjas-La Chimbera, Villa Centenario, Villa El Salvador-Villa Sefair, Villa General San Martín-Campo Afuera, Villa Media Agua, Villa San Martín, Villa Santa Rosa.

Provincia	Localidades
San Luis	Buena Esperanza, Candelaria, Carpintería, Concarán, Juana Koslay, Justo Daract, La Punta, La Toma, Merlo, Naschel, Quines, San Francisco del Monte de Oro, San Luis, Santa Rosa del Conlara, Tilisarao, Unión, Villa Mercedes.
Santa Cruz	28 de Noviembre, Caleta Olivia, Comandante Luis Piedrabuena, El Calafate, Gobernador Gregores, Las Heras, Los Antiguos, Perito Moreno, Pico Truncado, Puerto Deseado, Puerto San Julián, Puerto Santa Cruz, Río Gallegos, Yacimientos Río Turbio.
Santa Fe	Arequito, Arroyo Seco, Avellaneda, Barrio Arroyo del Medio, Cañada de Gómez, Casilda, Coronda, El Trébol, Esperanza, Florencia, Fray Luis Beltrán, Funes, Granadero Baigorria, Humboldt, La Criolla, Moisés Ville, Pérez, Puerto General San Martín, Rafaela, Reconquista, Roldán, Romang, Rosario, San Jorge, San José del Rincón, San Lorenzo, Santa Fe, Santa Rosa de Calchines, Santo Tomé, Sastre, Sauce Viejo, Teodelina, Venado Tuerto, Vera, Villa Constitución, Villa Gobernador Gálvez.
Santiago del Estero	Añatuya, El Zanjón, Frías, La Banda, La Dársena, Monte Quemado, Quimilí, Sachayoj, Santiago del Estero, Sumampa, Suncho Corral, Termas de Río Hondo, Villa Atamisqui, Villa Ojo de Agua, Villa San Martín (Est. Loreto).
Tucumán	Aguilares, Alderetes, Banda del Río Salí, Barrio San José III/020 Villa Carmela, Colombres, Concepción, Delfín Gallo, Diagonal Norte-Luz y Fuerza-Los Pocitos-Villa Nueva Italia, El Manantial, Famallá, Ingenio San Pablo, La Florida, La Trinidad, Los Ralos, Lules, Medina, Monteros, Pueblo Independencia, Río Seco, San Miguel de Tucumán, Tafí Viejo, Villa de Trancas, Villa Mariano Moreno-El Colmenar, Villa Quinteros, Yerba Buena-Marcos Paz.
Tierra del Fuego	Río Grande, Tolhuin, Ushuaia.

Anexo III. Distribución territorial de las UPM de la muestra de la ENGHo



Fuente: INDEC, Coordinación del Sistema Geoestadístico.

Anexo IV. Viviendas elegibles, no elegibles y de elegibilidad dudosa

Cuadro 6. Cantidad de viviendas elegibles, no elegibles y de elegibilidad dudosa, por jurisdicción

Jurisdicción	Viviendas en la muestra	Viviendas elegibles	Viviendas no elegibles	Viviendas de elegibilidad dudosa
Total del país	44.922	38.321	6.152	449
CABA	4.320	3.676	644	-
Buenos Aires	10.038	8.313	1.492	233
Catamarca	1.230	1.081	149	-
Córdoba	2.286	1.891	395	-
Corrientes	1.224	1.053	168	3
Chaco	1.374	1.186	179	9
Chubut	1.338	1.219	119	-
Entre Ríos	1.554	1.365	168	21
Formosa	1.236	989	245	2
Jujuy	1.206	1.046	154	6
La Pampa	1.242	1.073	169	-
La Rioja	1.254	1.033	221	-
Mendoza	1.938	1.558	242	138
Misiones	1.248	1.124	118	6
Neuquén	1.296	1.160	136	-
Río Negro	1.332	1.203	129	-
Salta	1.296	1.123	173	-
San Juan	1.296	1.109	186	1
San Luis	1.266	1.103	153	10
Santa Cruz	1.254	1.086	161	7
Santa Fe	2.280	1.949	321	10
Santiago del Estero	1.314	1.141	173	-
Tucumán	1.236	1.094	140	2
Tierra del Fuego	864	746	117	1

Fuente: INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

Anexo V. Hogares en viviendas elegibles, con y sin respuesta

Cuadro 7. Cantidad de hogares en viviendas elegibles, con y sin respuesta, por jurisdicción

Jurisdicción	Hogares elegibles	Hogares con respuesta	Hogares sin respuesta
Total del país	38.599	21.547	17.052
CABA	3.719	1.889	1.830
Buenos Aires	8.349	3.750	4.599
Catamarca	1.103	885	218
Córdoba	1.902	1.135	767
Corrientes	1.064	632	432
Chaco	1.186	611	575
Chubut	1.219	738	481
Entre Ríos	1.373	753	620
Formosa	1.007	925	82
Jujuy	1.078	665	413
La Pampa	1.073	573	500
La Rioja	1.034	840	194
Mendoza	1.573	819	754
Misiones	1.125	546	579
Neuquén	1.164	384	780
Río Negro	1.213	952	261
Salta	1.151	895	256
San Juan	1.111	758	353
San Luis	1.109	456	653
Santa Cruz	1.091	388	703
Santa Fe	1.959	1.122	837
Santiago del Estero	1.151	659	492
Tucumán	1.099	813	286
Tierra del Fuego	746	359	387

Fuente: INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

Anexo VI. Hogares no respondientes

Cuadro 8. Cantidad de hogares por causa de no respuesta y jurisdicción

Jurisdicción	Hogares no respondientes	Ausencia ⁽¹⁾	Rechazo ⁽²⁾	No respuesta C2 o C3 ⁽³⁾	Otras causas
Total del país	17.052	4.697	7.483	4.145	727
CABA	1.830	593	1.016	166	55
Buenos Aires	4.599	1.329	2.077	1.046	147
Catamarca	218	57	94	63	4
Córdoba	767	97	341	276	53
Corrientes	432	147	145	107	33
Chaco	575	226	193	135	21
Chubut	481	242	194	32	13
Entre Ríos	620	215	221	162	22
Formosa	82	30	36	9	7
Jujuy	413	88	132	170	23
La Pampa	500	207	165	111	17
La Rioja	194	29	82	80	3
Mendoza	754	103	457	178	16
Misiones	579	139	260	126	54
Neuquén	780	203	279	268	30
Río Negro	261	119	102	24	16
Salta	256	49	116	68	23
San Juan	353	85	126	119	23
San Luis	653	163	242	240	8
Santa Cruz	703	118	274	294	17
Santa Fe	837	136	341	293	67
Santiago del Estero	492	198	172	86	36
Tucumán	286	27	188	62	9
Tierra del Fuego	387	97	230	30	30

⁽¹⁾ Incluye ausencia momentánea (no se pudo contactar en las visitas realizadas) y ausencia temporal (viaje, vacaciones, etcétera).

⁽²⁾ Incluye rechazo, rechazo previo al análisis y rechazo posterior al análisis e ingreso.

⁽³⁾ Los cuestionarios 2 y 3 incluyen preguntas sobre los gastos diarios y gastos varios del hogar, respectivamente.

Fuente: INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

Anexo VII. 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 The 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 la encuesta. 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, visita por parte de supervisor, etc.). Existen situaciones donde el contacto es imposible por presencia de sistemas de seguridad, portones cerrados, unidades de vivienda múltiple de difícil acceso o 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,

$e = EL/(EL + NE)$: 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 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 jurisdicción y para el total del país.

Cuadro 9. Tasas de respuesta por jurisdicción

Jurisdicciones	RR3	RR5
	%	
Total del país	55,3	55,8
CABA	50,8	50,8
Partidos del Gran Buenos Aires	41,2	42,8
Buenos aires	47,5	47,7
Catamarca	80,2	80,2
Córdoba	59,7	59,7
Corrientes	59,3	59,4
Chaco	51,2	51,5
Chubut	60,5	60,5
Entre Ríos	54,1	54,8
Formosa	91,7	91,9
Jujuy	61,4	61,7
La Pampa	53,4	53,4
La Rioja	81,2	81,2
Mendoza	48,4	52,1
Misiones	48,3	48,5
Neuquén	33,0	33,0
Río Negro	78,5	78,5
Salta	77,8	77,8
San Juan	68,2	68,2
San Luis	40,8	41,1
Santa Cruz	35,4	35,6
Santa Fe	57,0	57,3
Santiago del Estero	57,3	57,3
Tucumán	73,9	74,0
Tierra del Fuego	48,1	48,1

Fuente: INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

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, Hogares 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 intenta 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.

Coefficiente 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 al que los elementos pertenecen (manzanas, radios censales, fracciones censales, localidades, departamentos), o a fracciones del tiempo (semanas, días, tramos horarios, etc.). Es utilizado generalmente en diseños multietápico, 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 a otros aspectos importantes de un estudio o 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, 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 y a los que, 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 (ED). Cociente entre la varianza de un estimador correspondiente al diseño muestral empleado para seleccionar la muestra (en general, complejo) y la varianza del estimador que se obtendría bajo un muestreo simple al azar (MSA) de igual tamaño. Empleado 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, y 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 varianza 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 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 con la componente aleatoria definida por la varianza muestral del estimador que da origen a la estimación.

Error estándar (EE). 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. También empleado 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 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 (IC). 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 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 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*.

Muestra Maestra Urbana de Viviendas de la República Argentina (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 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 de 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 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 con el mismo diseño sin considerar estratos.

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

Muestreo por conglomerados. Es una modalidad del muestreo probabilístico que emplea como unidad de muestreo el conglomerado. En encuestas a hogares, esta alternativa de muestreo permite disminuir los costos de la encuesta, generalmente, en perjuicio de la 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 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 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 carácter operativa y que posee límites conocidos y precisos, con un determinado número de viviendas, empleada por el INDEC en la organización de los censos de población. Por su clasificación, 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 de la MMUVRA. (Ver **Áreas 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. Sin embargo, 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 de 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.