Encuesta de Actividades de Niños, Niñas y Adolescentes 2016-2017

NOTA TÉCNICA **EANNA**

Factores de expansión, estimación y cálculo de los errores por muestra para el dominio urbano

Septiembre de 2019

NOTAS TÉCNICAS INDEC N° 1

Ministerio de Hacienda
Presidencia de la Nación

Instituto Nacional de Estadística y Censos
República Argentina
indec

Encuesta de Actividad de Niñas, Niños y Adolescentes 2016-2017

Factores de expansión, estimación y cálculo de los errores por muestra para el dominio urbano

Nota técnica - Septiembre de 2019

Instituto Nacional de Estadística y Censos (INDEC)

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

ISBN en trámite

Libro de edición argentina

Responsable de la edición: Lic. Jorge Todesca

Director técnico: Mag. Pedro Lines

Directora de la publicación: Mag. Silvina Viazzi

Coordinación de producción editorial: Lic. Marcelo Costanzo

Buenos Aires, septiembre de 2019

Publicaciones del INDEC

Las publicaciones editadas por el Instituto Nacional de Estadística y Censos pueden ser consultadas 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. El horario de atención al público es de 9:30 a 16:00.

También pueden solicitarse al teléfono (54-11) 5031-4632

Correo electrónico: ces@indec.gob.ar

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

INDECArgentina

Índice

1. Introducción

El Instituto Nacional de Estadística y Censos (INDEC) conjuntamente con el Ministerio de Trabajo, Empleo y Seguridad Social (MTEySS) realizaron la Encuesta de Actividades de Niños, Niñas y Adolescentes (EANNA) 2016/2017, con el objetivo de medir y diagnosticar la situación del trabajo infantil en el país[1].

Se trata de la segunda encuesta de este tipo aplicada en el país y la primera de carácter nacional, ya que cubre a toda la población de la Argentina, tanto la residente en zonas urbanas como la que integra las áreas rurales.

Esta publicación es una guía de referencia básica de la metodología adoptada para determinar los factores de expansión que se emplean en las estimaciones oficiales, y que permite estimar sus errores de muestreo en el dominio urbano de la encuesta. Cabe mencionar que la metodología aplicada en el ámbito rural de la EANNA[2] en líneas generales se corresponde con la que se expone en este documento.

En primera instancia se describen 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 describe el proceso para la determinación y el ajuste de los factores de expansión o ponderadores de la encuesta, y se exponen los motivos por los cuales se introduce una metodología para el cálculo de los errores por muestra que emplea replicaciones. Se detalla el proceso que da origen a los ponderadores asociados a las réplicas y se incluyen indicaciones para estimar los principales indicadores del error de muestreo empleando la metodología 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 las publicaciones de la encuesta, y para aquellas que generen las propias a partir de la base provista a los usuarios.

2. Diseño muestral en el dominio urbano

El diseño muestral es relevante en toda encuesta a hogares que emplee el muestreo probabilístico, porque impacta en la calidad de las estimaciones y en el costo y la organización de la encuesta. Dado que una porción significante de su presupuesto es destinada a la recolección de los datos, el diseño muestral es un compromiso entre minimizar los costos de la colecta y maximizar la calidad de los datos.

En líneas generales debe estar constituido por un marco de muestreo, o sea, la lista de unidades que cubre a la población objetivo desde la cual se selecciona la muestra; la cartografía necesaria para definir, identificar y alcanzar a las unidades que lo componen; información sobre ellas que permita definir un diseño eficiente en términos del costo del operativo y de la precisión esperada para los resultados; una regla probabilística que seleccione de manera aleatoria a sus unidades; un mecanismo de cálculo que brinde las estimaciones; y finalmente una estrategia que evalúe la precisión alcanzada en los resultados a partir de la muestra.

[1] Ver https://www.indec.gob.ar/ftp/cuadros/sociedad/eanna_2018.pdf.
[2] El diseño muestral del dominio rural fue definido por el MTEySS, que también llevó adelante el operativo de la encuesta en ese ámbito.

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

Un diseño simple y eficiente en términos de precisión podría ser un muestreo simple al azar (MSA), donde las viviendas son seleccionadas aleatoriamente con igual probabilidad. Sin embargo, se requeriría de una lista de todas viviendas pertenecientes al ámbito geográfico que abarca la encuesta, lo cual es dificultoso o imposible de lograr en la práctica.

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

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

Para maximizar los recursos, integrar y coordinar sus operaciones estadísticas, el INDEC emplea una modalidad bajo el esquema de muestra maestra. O sea, una única gran muestra probabilística que mantiene fijas a las unidades de área que la conforman y a su estructura probabilística asociada, y que permite subseleccionar las muestras de viviendas para todas las encuestas a hogares del Instituto durante aproximadamente un decenio, o período intercensal.

De esta manera se busca mejorar la relación costo-beneficio, al reducir los costos en la preparación de un diseño muestral para cada operativo, y controlar los problemas que ocasiona la dispersión de las muestras señalada en los párrafos anteriores. A dicha muestra se la conoce como Muestra Maestra Urbana de Viviendas de la República Argentina (MMUVRA).

La MMUVRA es de alcance nacional y urbano, y permite subseleccionar muestras para las encuestas que tienen como principales dominios de estimación a las provincias y a los aglomerados que participan en la Encuesta Permanente de Hogares (EPH) que lleva a cabo el Instituto, u otros agregados territoriales.

Su diseño inicialmente emplea dos etapas de selección probabilística. Cada unidad de primera etapa de muestreo (UPM) del diseño está definida por un aglomerado o localidad de al menos 2.000 habitantes según el Censo Nacional de Población y Viviendas 2010 (CNPyV 2010). El conjunto de todas las UPM constituye el marco de muestreo o la lista de unidades de muestreo para la selección probabilística de primera etapa.

Estas son estratificadas según el total de población según CNPyV 2010 y aquellas UPM formadas por aglomerados o localidades de 50.000 habitantes o más son incluidas en la MMUVRA con probabilidad 1 por diseño, y se las denomina "UPM autorrepresentadas". Del resto de las UPM, un conjunto fue seleccionado por provincia mediante un muestreo sistemático con probabilidad proporcional a la cantidad total de habitantes[3]. Tanto las UPM autorrepresentadas como las seleccionadas conforman la muestra de aglomerados o localidades de la MMUVRA.

[3] Para el total y un listado de las UPM que componen la muestra maestra e involucradas en la EANNA, ver Anexo I.

Para la segunda etapa, en las UPM en seleccionadas para la MMUVRA, y solo para ellas, se definieron las “unidades de segunda etapa de muestreo” (USM) o “Áreas MMUVRA”[4] en base a radios censales[5], empleando la cartografía del CNPyV 2010. En cada UPM, todas sus USM en conjunto la cubren territorialmente y determinan la envolvente o área de cobertura asociada a dicha unidad, conformando el marco de muestreo para la selección de segunda etapa. Una muestra probabilística de USM conforma la segunda jerarquía de la MMUVRA, cuya selección fue realizada bajo un diseño estratificado definido a partir de variables sociodemográficas y mediante un muestreo sistemático proporcional a la cantidad total de viviendas, según el CNPyV 2010.

Por último, en cada una de las USM seleccionadas, se confeccionó inicialmente un listado exhaustivo de viviendas particulares, lo que dio origen al marco de selección de viviendas de la MMUVRA y sobre el cual se realizan las subselecciones para las muestras de todas las encuestas a hogares del Instituto[6]. Este marco se actualiza en forma periódica con los resultados de distintos operativos de campo que lleva adelante el INDEC. El listado de viviendas tiene un orden específico y una cartografía asociada, que facilita su actualización y ayuda a organizar la asignación de la carga de trabajo, y las tareas de campo y recorrido de los encuestadores.

En el caso de la EANNA, se realizó una nueva etapa de selección probabilística de un tercer tipo de unidades de muestreo, denominados "segmentos". Estos están constituidos por 5 viviendas particulares contiguas o próximas entre ellas dentro del listado de la MMUVRA. Una selección sistemática con igual probabilidad de estos segmentos permitió conformar la muestra definitiva de viviendas en el dominio urbano de la encuesta.

3. Dominios geográficos de estimación y tamaño de la muestra

La población objetivo o de interés de la EANNA abarca a niñas, niños y adolescentes de 5 a 17 años, residentes en viviendas particulares de las localidades de la República Argentina con 2.000 o más habitantes. La encuesta tiene como dominio de estimación el total del país, dividido en 6 regiones estadísticas:

- 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 y el resto de los partidos de Buenos Aires.
- Patagonia: Chubut, Neuquén, Río Negro, Santa Cruz y Tierra del Fuego.

La muestra seleccionada para la EANNA Urbana 2016/2017 está constituida por 38.165 viviendas, distribuidas por dominio de estimación según el siguiente cuadro[7]:

[4] En la conformación de las Áreas MMUVRA, los radios censales por cuestiones operativa (extensión, densidad, inaccesibilidad, etc.) pueden sufrir recortes o agrupamientos (por ejemplo, para equilibrar la uniformidad de sus tamaños en términos de viviendas).
[5] El radio censal es una de las unidades territoriales que emplea el Instituto para organizar la tarea en los censos; por lo general están constituidos por aproximadamente 400 viviendas y, dependiendo de sus características, se los clasifica en urbanos, mixtos o rurales.
[6] A la fecha de la EANNA, el total de viviendas particulares registradas en la MMUVRA era de 2.053.958.
[7] Para ver una distribución de la muestra de viviendas por jurisdicción, ver Anexo II.

Cuadro 1. Distribución de la muestra de viviendas por región

Regiones	Cantidad de viviendas
Gran Buenos Aires	9.750
Noroeste	4.975
Noreste	4.720
Cuyo	4.055
Pampeana	10.055
Patagonia	4.610
Total del país	38.165

Fuente: INDEC, *Encuesta de Actividades de Niños, Niñas y Adolescentes 2016-2017.*

4. Determinación y ajuste de los factores de expansión

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

Distintos factores, los señalados en relación a la complejidad del diseño muestral y la asignación de la muestra, y los ajustes por cobertura y no respuesta que se deben realizar, contribuyen a que las ponderaciones o factores de expansión[8] de la encuesta para todas las unidades de la muestra no sean uniformes.

El diseño muestral empleado en la EANNA a partir de la MMUVRA determina el factor de expansión inicial de cada vivienda. Surge de la multiplicación de las inversas de las probabilidades de inclusión de cada una de las tres etapas de selección definidas en el apartado 2. Por lo tanto, el factor de expansión de la k-ésima vivienda ubicada en la j-ésima USM dentro de la i-ésima UPM se define como[9]:

$$w_{ijk}^{(0)} = f_{1i} f_{2ij} f_{3ijk},$$

donde,

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

f_{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;

f_{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[10].

El proceso de cálculo de los factores de expansión finales de la encuesta que se emplean para las estimaciones oficiales involucra varias correcciones sobre los factores iniciales, a fin de compensar

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

distintos errores introducidos durante las complejas fases de una operación estadística de la envergadura de la EANNA.

Las principales fuentes de error de una encuesta son las de cobertura, a causa de la deficiencia o desactualización del marco de muestreo; la no respuesta; los errores de medición; y los que surgen en el procesamiento y edición de los datos.

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

Es por esto que un objetivo central de las encuestas del Instituto es minimizar el efecto de las distintas fuentes, 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 que no responde o a la persona que inicialmente no se encuentra para revertir su estado.

Aun tomando todos estos recaudos, los errores no desaparecen. Por eso, antes de la etapa de estimación, los factores de expansión deben ser ajustados o corregidos buscando, en lo posible, disminuir el potencial sesgo que pueden estar introduciendo en los resultados y así aumentar la calidad de las estimaciones de una encuesta.

4.1 Ajuste por elegibilidad dudosa

El primer ajuste que se realiza sobre los factores de expansión tiene como objetivo atender los problemas causados por las deficiencias en la elegibilidad de las unidades, ya sea por desactualización del listado o bien por las dificultades encontradas por los encuestadores en alcanzar las viviendas seleccionadas. El tratamiento de este ajuste lleva a clasificar a las viviendas como elegibles, no elegibles y de elegibilidad dudosa.

Para la EANNA, y solo con el fin de ajustar los factores de expansión iniciales por elegibilidad dudosa, se considera:

- Viviendas elegibles (VEL) a aquellas que respondieron la encuesta, o que presentan alguna de las siguientes “causa por la que no se realizó la entrevista” indicadas en el cuestionario:
    - 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

- 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, inaccesibles (problemas climáticos u otros) o no se especificó ningún motivo de otras causas.

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

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

donde,

$e = \sum_{EL} w_{ijk}^{(0)} / \left(\sum_{EL} w_{ijk}^{(0)} + \sum_{NE} w_{ijk}^{(0)}\right)$ corresponde a la tasa de elegibilidad,
EL es el conjunto de viviendas clasificadas como elegibles,
NE es el conjunto de viviendas clasificadas como no elegibles, y
ED es el conjunto de viviendas clasificadas como de elegibilidad dudosa.

Estos ajustes se realizan dentro de grupos o clases definidos exclusivamente a tal efecto, que surgen del cruce de la variable provincia, la división aglomerado EPH y resto de las UPM, y los estratos de diseño de la MMUVRA para las USM.

Por consiguiente, en cada grupo g se obtiene un primer factor de ajuste a_{1g}, definido por la proporción de viviendas que se estiman como elegibles sobre el total de viviendas de la encuesta[12]:

$$a_{1g} = \frac{\sum_{EL(g)} w_{ijk}^{(0)} + e \sum_{ED(g)} w_{ijk}^{(0)}}{\sum_{EL(g)} w_{ijk}^{(0)} + \sum_{NE(g)} w_{ijk}^{(0)} + \sum_{ED(g)} w_{ijk}^{(0)}}$$

que permite corregir el factor de expansión de la última etapa, f_{3ijk}, asociado a la k-ésima vivienda perteneciente al grupo g, como:

$$f'_{3ijk} = f_{3ijk} a_{1g}$$

y que origina el factor de expansión de la k-ésima vivienda elegible ubicada en la j-ésima USM dentro de la i-ésima UPM corregido por elegibilidad dudosa de las viviendas,

[11] En la simbología empleada en la guía, $\sum_A$ representa la suma sobre todas las unidades que pertenecen al conjunto A.
[12] En las fórmulas, *EL(g)*, *ED(g)* y *NE(g)* señalan a los conjuntos *EL, NE y ED* restringido al grupo *g*.

$$w_{ijk}^{(1)} = \frac{w_{ijk}^{(0)}}{f_{3ijk}} f'_{3ijk}$$

En el cuadro 2, se presentan los resultados de la encuesta en el ámbito urbano[13] en relación a la cantidad de VEL, VNE y VED, por dominio de estimación y a nivel nacional, que intervienen con sus factores de expansión iniciales, $w_{ijk}^{(0)}$, en los cálculos del factor a_1.

Cuadro 2. Cantidad de viviendas elegibles, no elegibles y de elegibilidad dudosa, por región

Regiones	Viviendas en la muestra	Viviendas elegibles	Viviendas no elegibles	Viviendas de elegibilidad dudosa
Gran Buenos	9.784	6.175	673	2.936
Noroeste	4.962	4.243	493	226
Noreste	4.732	3.911	494	327
Cuyo	4.058	3.395	440	223
Pampeana	10.119	7.722	1.183	1.214
Patagonia	4.620	3.735	476	409
Total país	38.275	29.181	3.759	5.335

Fuente: INDEC, *Encuesta de Actividades de Niños, Niñas y Adolescentes 2016-2017*.

Como, por diseño, la probabilidad de selección de un hogar dentro de una vivienda es la probabilidad de selección de dicha vivienda, en adelante, los factores de expansión corresponden a cada hogar identificado dentro de las viviendas elegibles. Es decir que el factor de expansión de cada vivienda es afectado a cada uno de los hogares que la componen, por lo tanto, el peso del l-ésimo hogar de la k-ésima vivienda en la j-ésima USM dentro de la i-ésima UPM es $w_{ijkl}^{(1)} = w_{ijk}^{(1)}$.

4.2 Ajuste por no respuesta

Cuando se identifica una vivienda como elegible para la encuesta, y por consiguiente los hogares que la componen, no siempre es posible hacer una entrevista a todos sus miembros originando una no respuesta[14] del hogar. Esto puede ocurrir debido a una serie de razones: que en el hogar ninguno quiera responder, que haya ausencia temporal de sus miembros, o bien que hubo un primer contacto, pero por algún motivo o circunstancia fue imposible continuar con la entrevista.

En particular, en la EANNA, 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.
- Otras causas: duelo, alcoholismo, discapacidad, idioma extranjero.

[13] En la EANNA se encuestó a todas las viviendas presentes en el domicilio, por esta razón, la cantidad de viviendas en la muestra es levemente mayor a las seleccionadas.
[14] Bajo ninguna circunstancia las viviendas seleccionadas para la encuesta son reemplazadas por otras viviendas por razones de no respuesta.

La no respuesta es un fenómeno siempre presente en una encuesta u operación estadística, y es una fuente de sesgo en las estimaciones. Su incidencia sobre ellas puede que aumente a medida que se incrementa la no respuesta, y es por esto que se hacen esfuerzos para mantener la tasa de respuesta lo más alta posible durante la recolección de los datos.

La magnitud del sesgo debido a la falta de respuesta generalmente no se conoce, pero está directamente relacionada con las diferencias en las variables que indaga la encuesta entre los grupos de unidades que respondieron y las que no lo hicieron. También, se ve afectada por un factor asociado a la correlación entre dichas variables y la probabilidad a la propensión a dar respuesta por las unidades. Ante la potencial presencia de este sesgo, y para atenuar su efecto sobre las estimaciones, los factores $w_{ijkl}^{(1)}$ de los respondentes se ajustan para compensar la no respuesta alcanzada en la encuesta.

Una de las claves para reducir este sesgo, y lograr el éxito del ajuste, es determinar clases o grupos de unidades que expliquen lo mejor posible el mecanismo de no respuesta que hay por detrás del fenómeno que se investiga en la encuesta. Para que el ajuste sea eficaz se busca:

- que los agrupamientos permitan sostener el supuesto de probabilidad de respuesta constante de las unidades dentro de ellos, y

- que estos agrupamientos sean lo más homogéneos posibles, para que valga en algún grado la hipótesis de que dentro de una clase dada los que responden sean similares a los que no lo hacen, en términos de las principales variables de interés.

Para realizar las correcciones por no respuesta en la EANNA, se emplean los mismos grupos o clases definidos para el ajuste por elegibilidad dudosa, que surgen de cruzar la provincia (25 categorías), la división aglomerado EPH y resto de las UPM (2 categorías) y el estrato de USM (5 categorías). O sea, en cada grupo g se obtiene un segundo factor de ajuste a_{2g}, definido por:

$$a_{2g} = \left(\sum_{R(g)} w_{ijkl}^{(1)} + \sum_{NR(g)} w_{ijkl}^{(1)} \right) \Big/ \left(\sum_{R(g)} w_{ijkl}^{(1)} \right)$$

donde R(g) y NR(g) representan a los conjuntos de hogares con respuesta y no respuesta en la encuesta en el grupo g, respectivamente. Este factor de ajuste multiplicado por el de expansión corregido por elegibilidad dudosa f'_{3ijk} permite obtener los ponderadores ajustados por no respuesta (f''_{3ijk}) para aquellos hogares que han respondido a la encuesta como:

$$f''_{3ijk} = f'_{3ijk} a_{2g} = f_{3ijk} a_{1g} a_{2g}$$

que lleva al factor de expansión corregido por no respuesta para el l-ésimo hogar que responde,

$$w_{ijkl}^{(2)} = \frac{w_{ijkl}^{(1)}}{f'_{3ijk}} f''_{3ijk}$$

Cabe mencionar que para obtener los factores de ajuste a_{2g} un reagrupamiento de clases es practicado cuando el factor es más grande que 2,5 dentro de una clase. Este reagrupamiento se realiza a nivel de la variable estrato, agrupando clases contiguas y recalculando el factor de ajuste en la clase redefinida

hasta lograr que no supere el umbral. La razón de esta estrategia es eliminar factores de ajustes muy grandes dado que ellos tienden a incrementar la variabilidad de las estimaciones.

Para ilustrar la cantidad de unidades de la muestra que con sus factores de expansión $w_{ijkl}^{(1)}$, se involucran en los cálculos de los factores de ajuste a_2, el siguiente cuadro totaliza los hogares con y sin respuesta registrados en la encuesta a nivel nacional[15] y por dominio de estimación.

Cuadro 3. Total de hogares en viviendas clasificadas como elegibles, hogares con y sin respuesta, por región

Regiones	Hogares elegibles	Hogares con respuesta	Hogares sin respuesta
Gran Buenos Aires	6.512	4.561	1.951
Noroeste	4.404	4.180	224
Noreste	3.981	3.536	445
Cuyo	3.465	3.299	166
Pampeana	7.880	7.077	803
Patagonia	3.820	3.462	358
Total país	30.062	26.115	3.947

Fuente: INDEC, *Encuesta de Actividades de Niños, Niñas y Adolescentes 2016-2017.*

4.3 Ajuste por calibración

Los factores de expansión de cada hogar ajustados por elegibilidad dudosa y no respuesta reciben una última modificación o ajuste, denominado "calibración", que emplea información auxiliar de alguna fuente externa disponible. Esta información puede ayudar a corregir ciertas deficiencias de cobertura, originadas cuando algunos grupos de la población no están bien representados por los que responden a la encuesta. Para disminuir estas discrepancias, la calibración busca la consistencia entre las estimaciones en algunas variables de la encuesta y totales poblacionales conocidos o *benchmarks*.

La información auxiliar también permite definir estimadores más eficientes, que aprovechan la correlación que pueda existir entre las características indagadas por la encuesta y la provista por la fuente externa. Generalmente, a mayor correlación, la precisión de los estimadores aumenta y disminuye el error de muestreo en las estimaciones.

El proceso de calibración que genera el sistema de ponderadores definitivos de la encuesta queda definido al fijar y minimizar una distancia entre los factores de expansión surgidos de la última corrección detallada en el apartado 4.2 y los calibrados, y el planteo de un conjunto de restricciones sobre estos (Valliant, Dever y Kreuter, 2013). Por lo tanto, se busca que:

a) $\sum_R G(w_{ijkl}^{(2)}, w_{ijkl}^{(3)})$ sea mínima, donde G es una función a determinar que define la proximidad entre los factores $w_{ijkl}^{(2)}$ y los calibrados $w_{ijkl}^{(3)}$,

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

b) los factores a determinar, $w_{ijkl}^{(3)}$, de los hogares con respuesta dentro de las viviendas elegibles de la muestra satisfagan q totales marginales conocidos de la población objetivo U, o sea:

$$\sum_R w_{ijkl}^{(3)} \mathbf{x}_{ijkl} = \sum_U \mathbf{x}_q,$$

donde $\mathbf{x}_{ijkl} = (x_{ijkl\,1}, \dots, x_{ijkl\,q})$ es un conjunto de q variables auxiliares disponibles para cada hogar de la muestra, y con totales de la población, $\sum_U \mathbf{x}_q = (t_{x1}, \dots, t_{xq})$, provistos por la fuente externa a la encuesta para cada variable.

Dada una función G, la resolución numérica que satisface a) y b) es un proceso iterativo, que bajo ciertas condiciones de regularidad converge y permite determinar constantes a_{3ijkl} para cada hogar con respuesta, y definir al sistema de pesos calibrados como:

$$w_{ijkl}^{(3)} = w_{ijkl}^{(2)} a_{3ijkl},$$

o sea, igual al último peso ajustado por elegibilidad dudosa y no respuesta por el factor a_{3ijkl}, que surge de la calibración.

Para la EANNA se emplearon 9 variables que reflejan la composición interna de cada hogar por sexo y grupos de edad, donde $\mathbf{x}_{ijkl} = (x_{ijkl\,1}, \dots, x_{ijkl\,9})$ y con:

$x_{ijkl\,1}$ cantidad de mujeres en el hogar,
$x_{ijkl\,2}$ cantidad de varones en el hogar,
$x_{ijkl\,3}$ cantidad de personas entre 0 y 4 años en el hogar,
$x_{ijkl\,4}$ cantidad de personas entre 5 y 13 años en el hogar,
$x_{ijkl\,5}$ cantidad de personas entre 14 y 15 años en el hogar,
$x_{ijkl\,6}$ cantidad de personas entre 16 y 17 años en el hogar,
$x_{ijkl\,7}$ cantidad de personas entre 18 y 44 años en el hogar,
$x_{ijkl\,8}$ cantidad de personas entre 45 y 64 años en el hogar,
$x_{ijkl\,9}$ cantidad de personas de 65 años y más en el hogar.

Por lo tanto, los factores $w_{ijkl}^{(2)}$ son calibrados de forma tal que satisfacen totales de población surgidos de proyecciones poblacionales para ese conjunto de variables sociodemográficas[16].

Para evitar producir dos conjuntos de pesos finales para la encuesta, uno para hogares y otro para personas, en la EANNA se emplea un método de calibración integrado que origina un peso único que permite estimaciones de parámetros tanto a nivel de personas como de hogares (Lemaître y Dufour, 1987). Es decir que, el factor de expansión final para la p-ésima persona residente del l-ésimo hogar que se aplica para todas las estimaciones de la encuesta es $w_{ijklp}^{(3)} = w_{ijkl}^{(3)}$, con $w_{ijkl}^{(3)}$ el surgido de la calibración.

El proceso de calibración se realiza en forma independiente por provincia o jurisdicción, y cuando es posible también se lo efectúa a los totales proyectados según la división aglomerado EPH y resto de las UPM dentro la provincia en cuestión.

[16] Los totales poblacionales proyectados fueron calculados a partir de datos censales de población según el CNPyV 2010 al 15 de noviembre de 2016 y determinados por la Dirección Nacional de Estadísticas Sociales y Poblacionales del INDEC.

Para resolver el problema numérico en la EANNA se emplea la función de distancia "logit" (Deville y Särndal, 1992; Haziza y Beaumont, 2017) del package Survey de R (Lumley, 2018), la elección permite controlar el rango de los factores $w_{ijkl}^{(3)}$ y así sus valores extremos y asegurar que sean positivos. La generación de pesos extremos impacta en la eficiencia del estimador y aumenta el riesgo de incrementar la variabilidad de las estimaciones.

Por último, los pesos que surgen del proceso iterativo de la calibración son tratados por un algoritmo de redondeo para eliminar la componente decimal dando origen a los $w_{ijkl}^{(3)}$ finales que se emplean para todas las estimaciones oficiales de la encuesta.

En el Anexo IV, y a manera ilustrativa de los efectos que tienen los distintos ajustes, se presentan las distribuciones de los factores de expansión iniciales de cada hogar $w_{ijkl}^{(0)}$, la de los factores ajustados por elegibilidad dudosa $w_{ijkl}^{(1)}$, los ajustados por no respuesta $w_{ijkl}^{(2)}$ y los factores de expansión calibrados $w_{ijkl}^{(3)}$ por los dominios geográficos de estimación de la encuesta.

5. 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 lo denomina estimación.

Los parámetros poblacionales que resultan de interés para estimar a partir de los datos de una encuesta son por lo general descriptivos y la mayoría puede definirse 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 o más complejos.

Para alcanzar las estimaciones de esos parámetros en la EANNA se emplean estimadores que recurren a los factores de expansión finales $w_{ijkl}^{(3)}$, o $w_{ijklp}^{(3)}$ según sea el caso, que surgen de la última etapa de ajuste.

A modo de ejemplo, y en el caso de que Y y Z sean variables o características de interés medidas a nivel de persona o individuo, la expresión de los estimadores más empleados son:

Parámetro	Estimador[17]
Total, $\boldsymbol{t_y}$	$\widehat{t_y} = \sum_R w_{ijklp}^{(3)} y_{ijklp},$
Promedio, $\overline{\boldsymbol{y}}$	$\hat{\bar{y}} = \frac{\sum_R w_{ijklp}^{(3)} y_{ijklp}}{\sum_R w_{ijklp}^{(3)}}$

[17] En todos los casos, $\sum_R$ en las fórmulas hace referencia a sumar sobre las personas que responden a la encuesta.

Proporción[18], $\boldsymbol{p}$	$\hat{p} = \frac{\sum_R w_{ijklp}^{(3)} y_{ijklp}}{\sum_R w_{ijklp}^{(3)}}$
Razón, $\boldsymbol{R_{yz}} = \frac{\boldsymbol{t_y}}{\boldsymbol{t_z}}$	$\hat{R}_{yz} = \frac{\widehat{t_y}}{\widehat{t_z}} = \frac{\sum_R w_{ijklp}^{(3)} y_{ijklp}}{\sum_R w_{ijklp}^{(3)} z_{ijklp}}$

6. 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 en una encuesta permite a los usuarios cuantificar el grado de confianza o bondad 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 un estudio que emplea una muestra probabilística, como la EANNA, la inferencia estadística sobre la población objetivo se basa en los datos recopilados de solo una parte de esta población. Es así como los resultados probablemente diferirán de los que se pueden obtener a partir de un censo completo.

El error que se genera al extraer conclusiones en términos estadísticos para toda la población 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 tamaño de la muestra, el número de etapas y el método de selección, el estimador empleado 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 estas no se distribuyen con cierto grado de uniformidad en la población.

Una medida del error de muestreo es la varianza muestral del estimador. Representa la variabilidad de las estimaciones que se obtienen a partir de todas las muestras posibles según el diseño muestral, con respecto al valor poblacional de la característica bajo estudio.

A partir de la varianza muestral se pueden definir otras medidas más populares como son 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). Cuanto más pequeño es el EE, el CV o el ED, o la amplitud del IC, más precisa es la estimación.

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

[18] La definición de los parámetros promedio y proporción coinciden si *Y* es una variable binaria, que toma el valor de 1 cuando el individuo posee una característica dada y 0, en caso contrario.

importancia que tiene en cualquier estudio por muestreo, es central su estimación como indicador de la calidad de las estimaciones en una encuesta.

7. 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 puedan emplearse para estimar el error muestral.

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 (elegibilidad dudosa, no respuesta y calibración), para incorporar la variabilidad propia de estas correcciones en los cálculos del error por muestra, 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, quintiles, deciles, etc.) o los de desigualdad (índice de Gini, curva de Lorentz, etc.) que en otros métodos son difíciles de implementar;
- habilitar a los usuarios a calcular por sus propios medios los errores de muestreo para sus estimaciones, con transparencia y de la misma manera en que los obtiene el Instituto, sin tener que depender de tablas u otros elementos para cuantificarlos;
- proteger y anonimizar cierta información que puede vulnerar el secreto estadístico que pesa sobre el microdato, por ejemplo, al no involucrar al usuario con las variables que definen el diseño muestral (estratos, UPM, USM), y que son necesarias para determinar el error de muestreo en una estimación.

Existen distintos métodos para conformar las réplicas (Wolter, 2007), y el que se adopta para generar las submuestras en la EANNA 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 $b, b = 1, \dots, B$, el procedimiento lleva a que en cada estrato de diseño, h, se seleccione una muestra simple al azar con reemplazo de $n_h - 1$ conglomerados a partir

de la muestra original de n_h conglomerado. Se define el peso *bootstrap* $w_{hcl}^{*(b)}$ a partir de un peso inicial w_{hcl} para la *l*-ésima unidad en el conglomerado c del estrato h en la réplica b según el siguiente ajuste:

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

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

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

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

donde:

$\hat{\theta}$ es el estimador[19] de θ calculado a partir de los ponderadores w_{hcl} 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 $w_{hcl}^{*(b)}$ de la réplica $b, b = 1, \ldots, B$.

De $[1]$ se puede obtener el estimador del error estándar,

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

y el del coeficiente de variación,

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

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

Sin embargo, la adopción de estos supuestos habilita emplearlo como un estimador de varianza para un diseño PPT sin reemplazo, si la selección de las UPM sin reemplazo es más eficiente que la selección de UPM con reemplazo (West, 2012; Särndal, Swensson, y Wretman,1992), como es el caso de la EANNA, lo que convierte al 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 EANNA 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

[19] Ver apartado 5.

las USM; en cambio, en las UPM no autorrepresentadas, los estratos *bootstrap* se corresponden con los estratos 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. Pero se debe alcanzar un compromiso entre garantizar la estabilidad, controlar el tamaño de la base de réplicas y limitar el tiempo de cálculo. Por estos motivos, en la EANNA el total de réplicas es de 300 ($B = 300$). Esta cantidad asegura la estabilidad de las estimaciones de varianza para las principales estimaciones de la encuesta.

Todas las réplicas son obtenidas de la muestra inicial, incluyendo todos los hogares encuestados o no en las viviendas elegibles, cuyo factor de expansión viene dado por $w_{ijkl}^{(1)}$. Este pasa a ser corregido según el estrato h y el "último conglomerado" c al cual pertenece el hogar, como lo requiere el procedimiento *bootstrap* descripto y que origina los pesos replicados $w_{ijkl}^{*(1,b)}$, b=1,…,300, para cada una de las submuestras.

Con el fin de incorporar la variabilidad en las estimaciones que introducen los ajustes efectuados en los factores de expansión, se repiten los mismos ajustes sobre los pesos replicados $w_{ijkl}^{*(1,b)}$. Es decir, para cada una de las 300 réplicas, los pesos *bootstrap* son ajustados nuevamente por no respuesta y calibrados por sexo y edad de manera análoga a como lo fueron los pesos originales $w_{ijkl}^{(1)}$, como se detalla en el apartado 4.

Por ejemplo, para obtener el peso ajustado por no respuesta, $w_{ijkl}^{*(2,b)}$, se multiplica el peso de hogar en la réplica b o submuestra *bootstrap*, $w_{ijkl}^{*(1,b)}$, por el factor de ajuste $a_{2g}^{(b)}$ calculado a partir de la réplica en cuestión. De la misma manera, se obtiene el ajuste $a_{3ijkl}^{(b)}$ correspondiente a la calibración de $w_{ijkl}^{*(2,b)}$ de cada réplica b, empleado para determinar el factor de expansión $w_{ijkl}^{*(3,b)}$. A diferencia de los pesos originales, los pesos *bootstrap* no son sometidos a un proceso de redondeo.

Finalmente, para el cálculo de la varianza *bootstrap* $v_B(\hat{\theta})$ expresada en [1], se emplean los factores de expansión $w_{ijkl}^{(3)}$ originales para estimar $\hat{\theta}$ y los factores $w_{ijkl}^{*(3,b)}$ de cada una de las réplicas para generar $\hat{\theta}_{(1)}^{*}, \dots, \hat{\theta}_{(300)}^{*}$, donde $\hat{\theta}_{(b)}^{*}$ es el estimador de θ definido a partir de la réplica $b, b = 1, \dots, 300$.

8. Modo de empleo de los pesos replicados

Para poder calcular las estimaciones de los errores de muestreo correspondientes a los resultados de la EANNA, se incluye en la base para usuarios de la encuesta un conjunto de 300 columnas que corresponden a los ponderadores o factores de expansión de cada réplica *bootstrap* para cada hogar/persona de la base[20], que surgieron del proceso detallado en el apartado anterior.

La presente sección constituye una guía de cómo deben ser empleadas las réplicas en distintas herramientas de cálculo: R[21], SAS[22], Stata[23] y Wesvar[24]. En caso de no contar con ellas, se presenta

[20] Ver INDEC (2019). *Manual de uso de las bases usuarios de la EANNA urbana 2016-2017.*
[21] www.r-project.org. Versión 3.6.
[22] www.sas.com. Versión 9.4 M3.
[23] www.stata.com. Versión 15.
[24] www.westat.com/capability/information-systems-software/wesvar. Versión 5.1.

un ejemplo que sugiere cómo efectuar el cálculo siguiendo la definición formulada en [1] del apartado 7, y que cualquier usuario puede poner en práctica con pocos recursos[25].

Se advierte que la guía no constituye un manual exhaustivo de cada una de las herramientas y sus opciones. Se asume que el usuario tiene 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.

Solo se incluyen los códigos que brindan las estimaciones puntuales, y el que permite alcanzar una medida del error a través del error estándar o el coeficiente de variación. En los ejemplos se consideran la estimación de un total, de un promedio, de una proporción, y de una razón o cociente entre dos totales, asumiendo que son los parámetros que más se requieren estimar a partir de los datos de la encuesta.

Para facilitar las indicaciones se asume que el usuario cuenta con la siguiente información, incluida en la base para usuarios:

- $\boldsymbol{w}$: factor de expansión final de la encuesta[26].
- $\boldsymbol{w\_repb}$: peso *bootstrap* replicado, donde b representa el número de réplica al cual corresponden los pesos, tomando los valores de 1 a 300[27].
- $\boldsymbol{Y, Z}$: variables genéricas (continuas, categóricas o binarias), hacen referencia a características para las cuales se requieren estimaciones de los parámetros poblacionales de interés (ver apartado 5), y las respectivas estimaciones de los errores de muestreo.

8.1 Cálculo del error de muestreo a través de R

Una de las posibilidades disponibles, y que acepta la metodología propuesta en esta herramienta, es el paquete Survey[28] (Lumley, 2018). Siguiendo las indicaciones del manual de Survey[29], y asumiendo que la base para el usuario con los datos de la encuesta importada a R se denomina ***base_encuesta***, se define el objeto **disenio**[30] 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 (***w***), el método que generó las réplicas (*bootstrap*), el conjunto de replicaciones (***w_rep[1-9]+***) que se encuentran en la base, y la opción ***mse=T***. Estas indicaciones preparan la herramienta para obtener las estimaciones y las estimaciones del error de muestreo, bajo las siguientes sentencias:

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

[25] No se incluye a la herramienta de cálculo SPSS, ya que no cuenta oficialmente a la fecha con la posibilidad de emplear la metodología desarrollada sin recurrir a una programación *adhoc*.
[26] Los valores corresponden a los $w_{ijkl}^{(3)}$ del apartado 4; en la base usuario se la etiqueta como PONDERA.
[27] Los valores corresponden a los $w_{ijkl}^{*(3b)}$ del apartado 7; en la base de réplicas poseen el mismo etiquetado.
[28] https://cran.r-project.org/web/packages/survey/index.html. Versión 3.36.
[29] https://cran.r-project.org/web/packages/survey/survey.pdf.
[30] El usuario puede optar por cualquier otro nombre para el objeto.

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

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

8.2 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 forma es la que se adopta para la presentación.

El comando ***svyset*** es el que se emplea para gestionar los cálculos. En él se deben identificar: el factor de expansión de la encuesta **w**, los pesos replicados ***w_rep****, el método para el cálculo de la varianza ***bootstrap***. Asimismo, se debe incluir la opción ***mse*** para obtener el estimador de varianza *bootstrap* considerado en el punto 7. Para preparar la herramienta para las estimaciones, el usuario debe invocar:

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

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

Estimador	Estimaciones por Stata
$\widehat{t_y}$	svy bootstrap : total Y estat cv
$\hat{\bar{y}}$	svy bootstrap : mean Y estat cv
$\hat{p}$	svy bootstrap : proportion Y estat cv
$\widehat{R_{YZ}}$	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, se debe proceder como se indicó en los párrafos anteriores, pero se tendrá que invocar al prefijo ***svyset*** con la opción ***brrweight***, y ***brr*** en la opción ***vce***. De esta forma, se podrán obtener estimaciones válidas para el EE, el CV o el IC, al no contar en esa versión con la opción ***bootstrap***. En la versión 9 o anteriores, la herramienta no cuenta con el prefijo ***svy*** para invocar estimaciones con pesos replicados, y obliga a cambiar el procedimiento para obtener estimaciones de varianzas (Chowhan y Buckley, 2005).

8.3 Cálculo del error de muestreo a través de SAS

El sistema 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 Inc., 2017), incluye los procedimientos ***surveymeans*** y ***surveyfreq*** que permiten brindar estimaciones de parámetros descriptivos de una población.

La opción que se debe emplear en cualquiera de ellos es *varmethod=**Bootstrap***, invocando los pesos replicados **w_rep1--w_rep300** vía *repweight* y al factor de expansión de la encuesta **w** en *weight*. En particular, para la estimación de los parámetros señalados se presentan los siguientes códigos orientativos:

Estimador	Estimaciones por SAS
$\widehat{t_y}$	proc surveymeans data=base_encuesta sum cvsum varmethod=Bootstrap; repweight w_rep1--w_rep300; weight w; var Y; run;
$\hat{\bar{y}}$	proc surveymeans data=base_encuesta mean cv varmethod= Bootstrap; repweight w_rep1--w_rep300; weight w; var Y; run;
$\hat{p}$	proc surveyfreq data= base_encuesta varmethod=Bootstrap; repweight w_rep1--w_rep300; weight w; table Y; run;
$\widehat{R_{YZ}}$	proc surveymeans data=base_encuesta varmethod=Bootstrap; repweight w_rep1--w_rep300; weight w; ratio Y/Z; run;

Se advierte que el método *bootstrap* para el cálculo de errores por muestra para diseños complejos está disponible para 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).

8.4 Cálculo del error de muestreo a través de Wesvar

Wesvar[31] 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 en base a replicaciones (Brick, Morganstein y Valliant, 2000). Solo tiene versión para plataforma Windows, y emplea un conjunto de menús desplegables e interactivos a través de su interfaz visual. A continuación, se brinda una descripción sencilla de su empleo y de las opciones básicas que hay que invocar para operar con ella, empleando la versión 5.1.19.

En la figura 1 se observa la ventana de inicio donde aparece el árbol de actividades y opciones que guían al usuario dentro de la herramienta. En primera instancia se debe crear una base de datos *Wesvar* (.var) a partir de la base de la encuesta, con el objetivo de utilizarla para realizar los análisis o estimaciones. Para esto el usuario deberá hacer clic en *New Wesvar Data File*, y elegir la base con las réplicas en la carpeta o espacio de trabajo donde se encuentra[32].

[31] www.westat.com/capability/information-systems-software/wesvar. Se puede acceder de forma gratuita a la documentación de WesVar enviando un mail a: wesvar_tech_support@westat.com.
[32] Se advierte que la herramienta tiene la posibilidad de importar datos en formato csv/txt con delimitadores, SAS o SPSS.

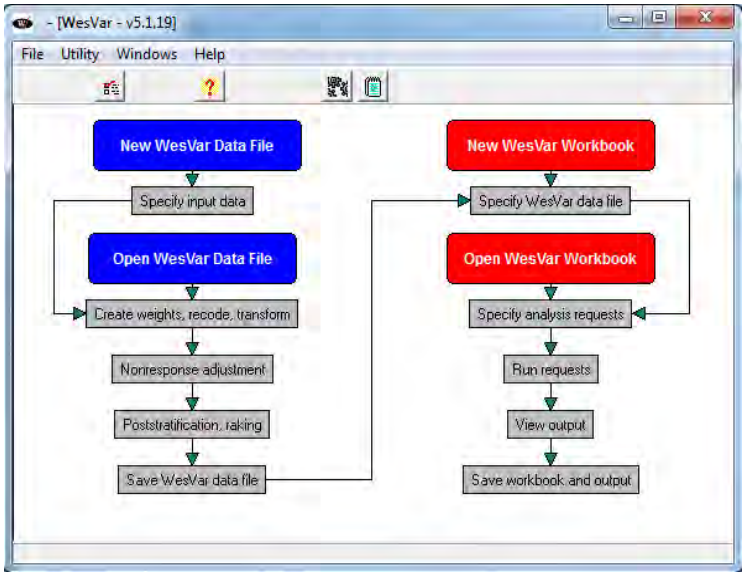


Figura 1

Al usuario le aparece una ventana como la que se ve en la figura 2, donde debe completar la información necesaria para iniciar 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_rep300**; y en el apartado **Full Sample**, el factor de expansión final de la encuesta, **w**. 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 continuar con las estimaciones.

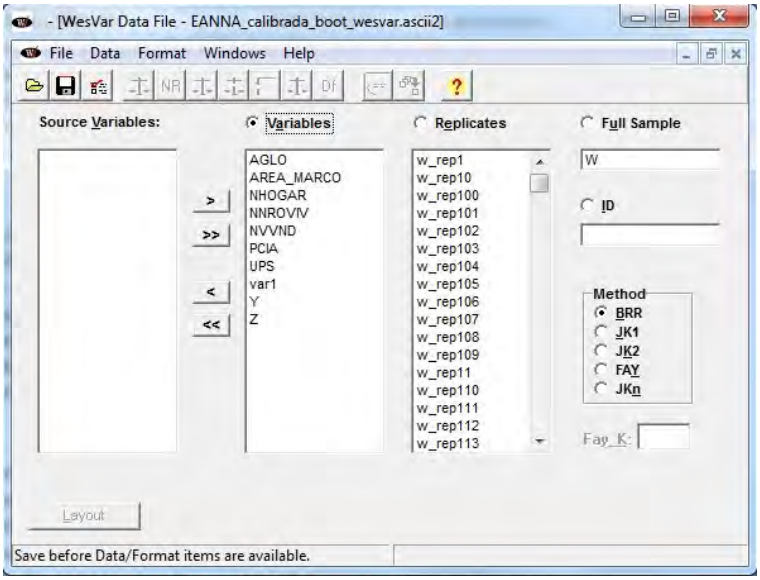


Figura 2

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

En la figura 3, se presenta la ventana a partir de la cual Wesvar permite gestionar los distintos análisis o 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 lo emplea para definir y cambiar los análisis o los tipos de estimaciones que ofrece la herramienta: tablas con totales o frecuencias, modelos de regresión o estadísticos descriptivos (***Table***, ***Regression***, ***Descriptive Stats***), respectivamente.

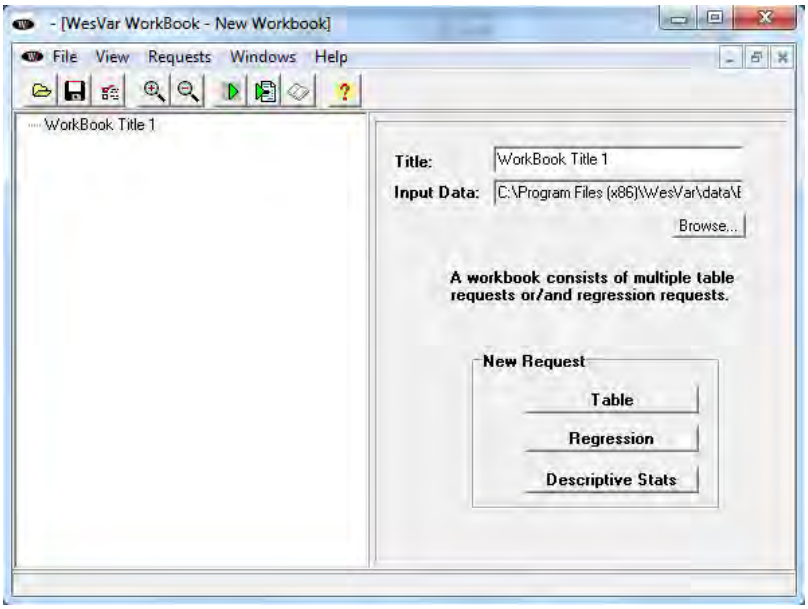


Figura 3

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

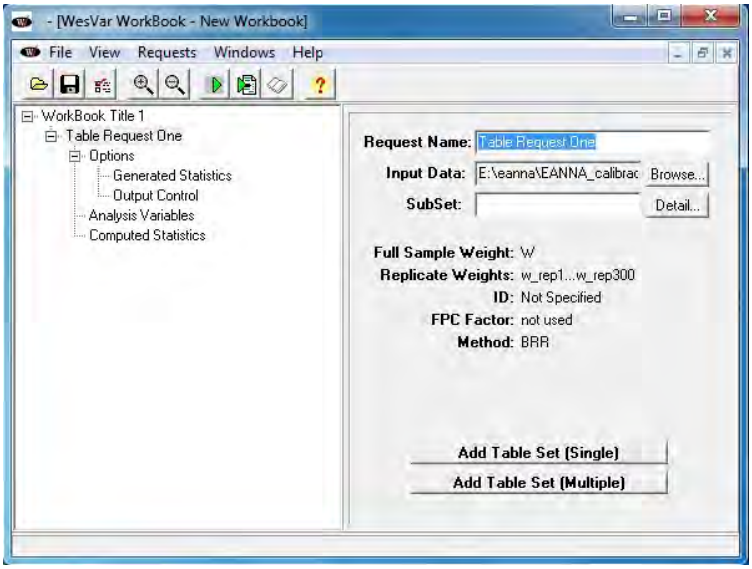


Figura 4

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

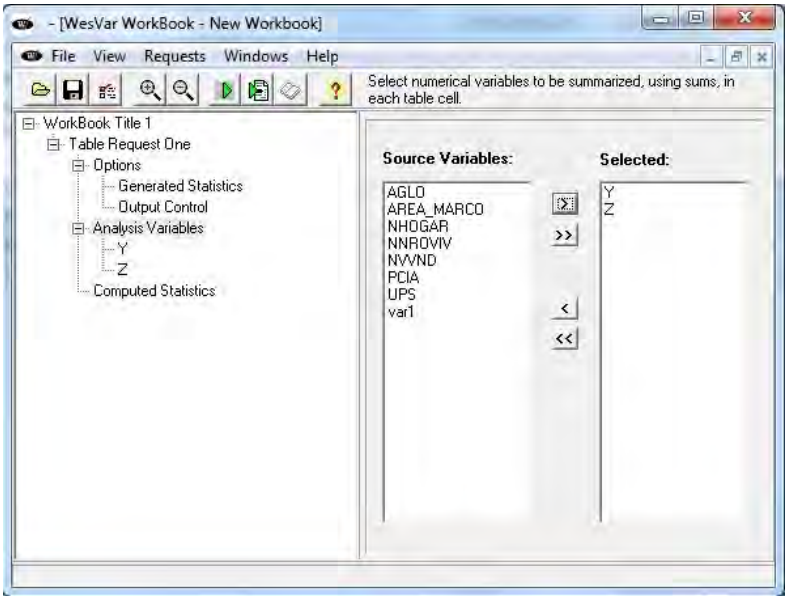


Figura 5

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

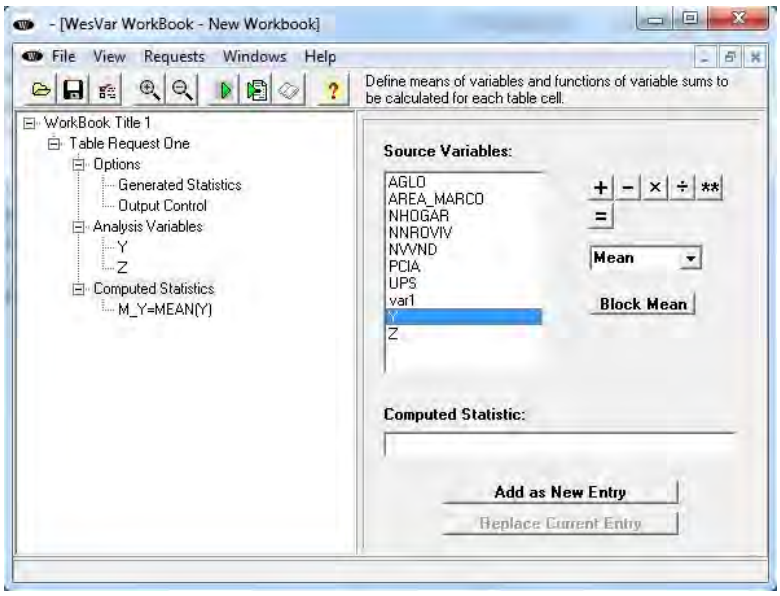


Figura 6

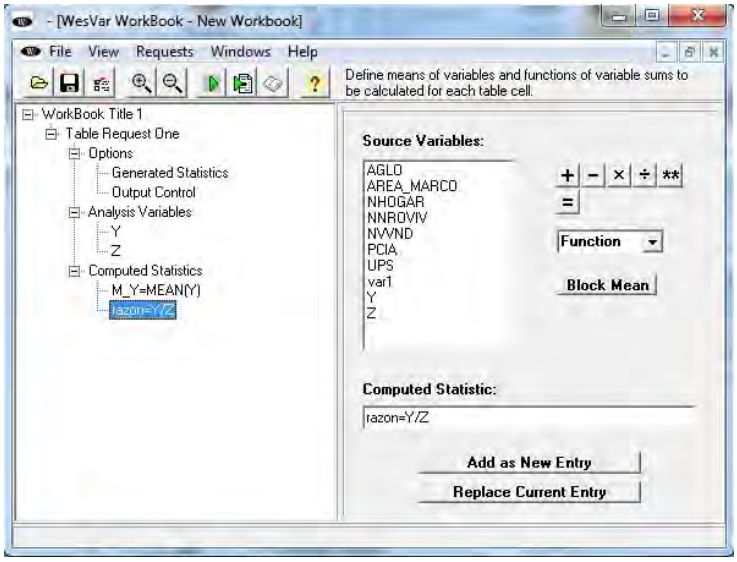


Figura 7

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

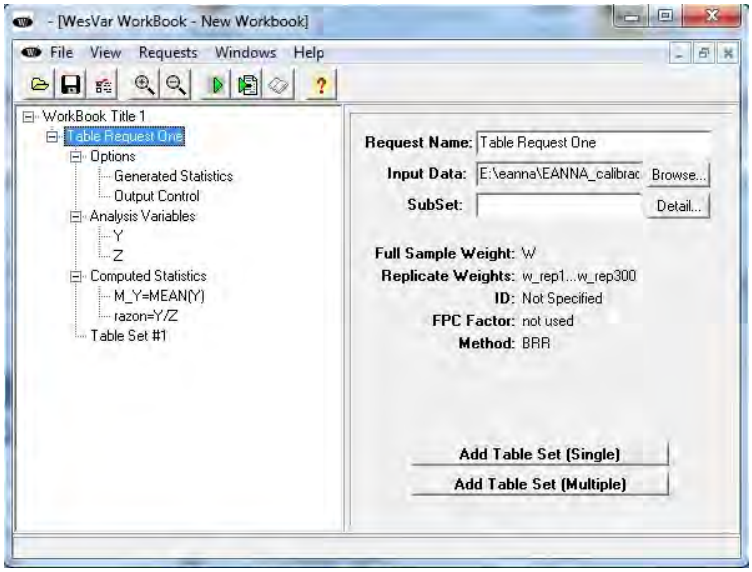


Figura 8

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

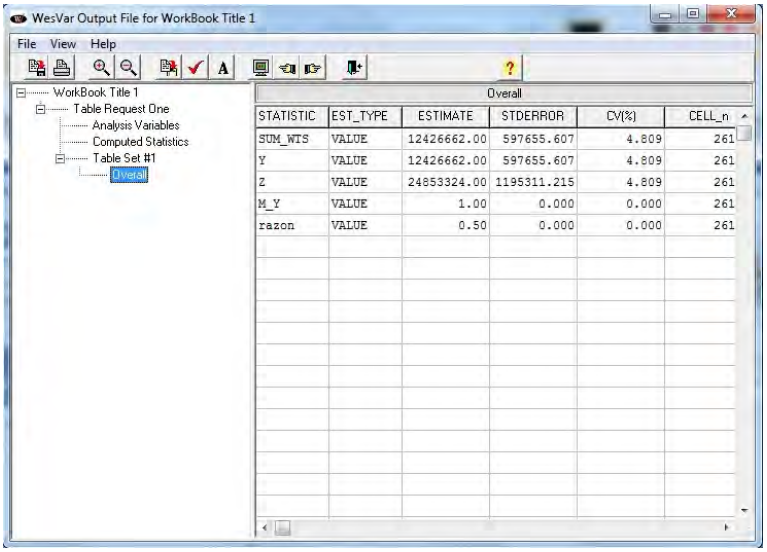


Figura 9

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 obtener la estimación de proporciones, asumiendo que Y es del tipo categórica, se debe generar una tabla 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.

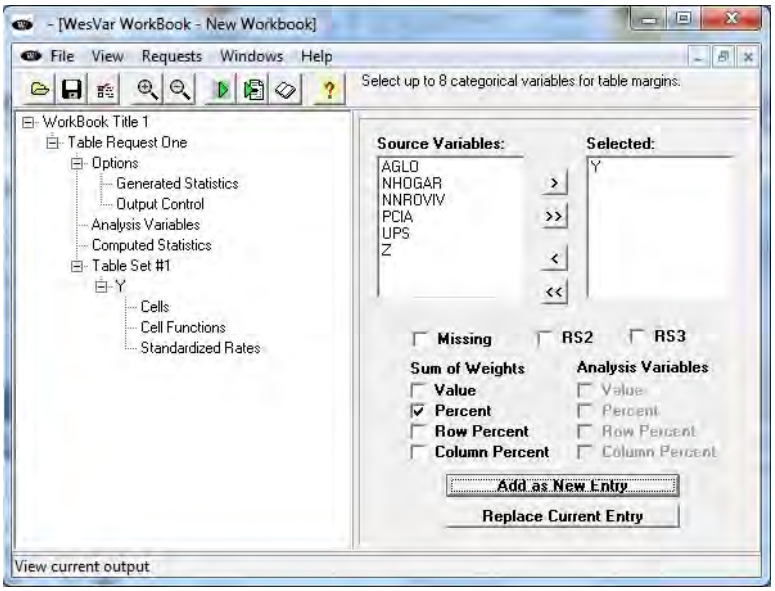


Figura 10

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.

8.5 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 el apartado 7 empleando las fórmulas [1] a [3].

Por ejemplo, si se asume que la variable Y está medida sobre las personas de la encuesta, la fórmula que se debe emplear como estimador para un total t_y , según se definió en el apartado 5, es:

$$\widehat{t_y} = \sum_R w_{ijklp}^{(3)} * y_{ijklp}$$

Siguiendo lo señalado en el apartado 7, la formulación para la varianza *bootstrap* [1] de un estimador es:

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

Reemplazando en ella $\hat{\theta}$ por $\widehat{t_y}$, y $\hat{\theta}_{(b)}^{*}$ por $\widehat{t_{y(b)}^{*}}$, donde $\widehat{t_{y(b)}^{*}} = \sum w_{ijklp}^{*(3,b)} * y_{ijklp}$, es la estimación del total a partir de los factores de expansión $w_{ijklp}^{*(3,b)}$ para la p-ésima persona en la b-ésima submuestra *bootstrap*, $b = 1 \ldots, 300$, permite calcular estimaciones para la varianza *bootstrap* de $\widehat{t_y}$, a través de:

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

para el error estándar, según

$$ee_B(\widehat{t_y}) = \sqrt{v_B(\widehat{t_y})}$$

y para el coeficiente de variación con

$$cv_B(\widehat{t_y}) = \frac{ee_B(\widehat{t_y})}{\widehat{t_y}}$$

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

$$\hat{\bar{y}}^*_{(b)} = \frac{\sum w^{*(3,b)}_{ijklp} * y_{ijklp}}{\sum w^{*(3,b)}_{ijklp}},$$

$$\hat{p}^*_{A(b)} = \frac{\sum w^{(3,b)}_{ijklp} * y_{ijklp}}{\sum w^{(3,b)}_{ijklp}},$$

o,

$$\hat{R}^*_{(b)} = \frac{\sum w^{*(3,b)}_{ijklp} * y_{ijklp}}{\sum w^{*(3,b)}_{ijklp} * z_{ijklp}}$$

según sea el caso, para obtener las respectivas varianzas estimadas por *bootstrap*, como también para los estimadores de ee_B y cv_B de la estimación en cuestión.

9. Recomendaciones para el uso con fines estadísticos de los datos de la encuesta

No en todos los resultados de la encuesta se puede poner la misma confianza. Inclusive, en algunas situaciones no es aconsejable tomarlos como válidos para hacer inferencia estadística. Distintos motivos, algunos señalados a través de la guía, 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 los calculan en dominios de estimación 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;
- la estimación de un total involucrado en el denominador de un cociente posee una variabilidad o coeficiente de variación muy alto.

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

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

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

En los casos mencionados en los párrafos anteriores 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 de las estimaciones publicadas, y en particular a los que empleen la base con los datos de la encuesta para alcanzarlas, que deberán poner atención y ser prudentes a la hora de sacar conclusiones en ciertas circunstancias.

9.1 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 tener que 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 deber 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 50. b) La estimación de una razón es menor a 0.05. c) La estimación de una proporción es menor al 5%. d) El denominador de un cociente, razón, o proporción, tiene un CV > 15%. 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 ≤ 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 ≤ 16,6%	La estimación puede ser considerada sin restricciones. No se requiere una notación especial.

Fuente: INDEC, *Encuesta de Actividades de Niños, Niñas y Adolescentes 2016-2017.*

Se insiste con la recomendación de que, en el caso de que algunas de las estimaciones sean consideradas no confiables o poco confiables para inferir el total de la población o las subpoblaciones y el usuario aun así desee incorporarlas en una publicación, se incluya una advertencia y se haga 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.

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

La herramienta que se emplee para las estimaciones del error de muestreo debe hacer uso de todas las observaciones de la muestra, para obtener una medida confiable y no estar subestimándola. Por lo general la documentación que acompaña a la herramienta contempla esta advertencia. En particular, en aquellas presentadas en los apartados 8.1 a 8.3, los usuarios que deseen obtener estimaciones en subpoblaciones o dominios, pueden recurrir a las opciones **subset**[33] en R, **subpop** en Stata, y **DOMAIN** en SAS para obtener en forma adecuada la estimación del CV o del EE que esté calculando[34].

9.3 Recomendaciones sobre el cálculo de intervalos de confianza

Los intervalos de confianza IC proveen otro camino para evaluar la variabilidad inherente a 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 con un nivel de confianza de 0,95 significa que, si un gran número de muestras 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 (IC) 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 un rol importante la distribución probabilística del estimador y las propiedades asintóticas del estimador empleado para la varianza. A diferencia del EE y el CV, el IC obliga a adoptar algunos supuestos sobre el estimador $\hat{\theta}$ empleado para estimar el parámetro de interés. Entre ellos, que de manera aproximada siga en distribución una ley normal, de difícil verificación en la práctica.

Como se advierte en distintos apartados, el diseño muestral de la encuesta no es un 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 independencia y distribución idéntica requeridas en este contexto para sostener el supuesto de normalidad (Heeringa, West, Berglung, 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 los apartados 9.1 y 9.2.

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

Referencias

American Association for Public Opinion Research (2016). Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. (9° ed.). Outbrook Terrace: AAPOR. Recuperado: https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf

Brick M, Morganstein D., Valliant R. (2000). Analysis of Complex Sample Data Using Replication, Westat.
Recuperado:
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. (2013). Response Rates Revisited. Proceedings American Statistical Associations. Survey Research Methods Section, JSM 2013, pp. 1200-1208.
Recuperado:
http://www.asasrms.org/Proceedings/y2013/files/308173_80404.pdf

Chowhan J., Buckley N. (2005). Using Mean Bootstrap Weights in Stata: A BSWREG Revision. The Research Data Centres Information and Technical Bulletin, 2(1), pp. 23-37. Statistics Canada.
Recuperado:
http://www.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=12-002-X20040016890&lang=eng

Deville J., Särndal C.E. (1992). Calibration Estimators in Survey Sampling. Journal of the American Statistical Association, 87, pp. 376-382.
DOI:10.1080/01621459.1992.10475217

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

Gagné C., Roberts G., 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) Statistics Canada.
Recuperado:
https://www150.statcan.gc.ca/n1/pub/12-002-x/2014001/article/11901-eng.htm

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

Heeringa S., West B., Berglund P. (2017). Applied Survey Data Analysis. (2^0 ed.) Chapman & Hall/CRC.
DOI:10.1201/9781315153278

Lemaître G., Dufour J. (1987), An Integrated Method for Weighting Persons and Families. Survey Methodology, 13, pp. 199-207.
Recuperado:
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.
DOI:10.1002/9780470580066

Lumley T. (2018). Survey: Analysis of Complex Survey Samples. R package version 3.33-2. Recuperado:
https://cran.r-project.org/package=survey

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

Rao J.N.K., Wu C.F.J., Yue K. (1992). Some Recent Work on Resampling Methods for Complex Surveys. Survey Methodology, 18, pp. 209-217.
Recuperado:
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), pp. 6-15. Recuperado:
http://www.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=12-002- X20040027032&lang=eng

Sarndall C., Swensson B., 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., Kreuter F. (2013). Practical Tools for Designing and Weighting Survey Samples, Nueva York: Springer.
DOI: 10.1007/978-1-4614-6449-5_14.

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

Wolter, K.M. (2007). Introduction to Variance Estimation (2° ed.). Nueva York: Springer-Verlag.
DOI: 10.1007/978-0-387-35099-8

Anexo I.A. Total de UPM y USM

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

Regiones	UPM	USM
Gran Buenos Aires	2	657
Noroeste	85	892
Noreste	79	745
Cuyo	42	486
Pampeana	122	1.738
Patagonia	62	754
Total del país	392	5.272

Fuente: INDEC, *Encuesta de Actividades de Niños, Niñas y Adolescentes 2016-2017*.

Anexo I.B. Listado de localidades seleccionadas para la MMUVRA y la EANNA

Provincia	Localidad
	Ciudad Autónoma de Buenos Aires
Buenos Aires	Almirante Brown, Avellaneda, Bahía Blanca, Baradero, Berazategui, Berisso, Campana, Carmen de Patagones, Chacabuco, Chivilcoy, Coronel Pringles, Dolores, Ensenada, Escobar, Esteban Echeverría, Ezeiza, Florencio Varela, General Daniel Cerri, General Rodríguez, General San Martín, Hurlingham, Ituzaingó, José C. Paz, Junín, La Matanza, La Plata, Lanús, Lincoln, Lomas de Zamora, Luján, Malvinas Argentinas, Mar del Plata, Marcos Paz, Máximo Paz ,Mercedes, Merlo, Monte Hermoso, Moreno, Morón, Necochea, Quequén, Olavarría, Pehuajó, Pergamino, Pilar, Presidente Perón, Punta Alta, Quilmes, Ramallo, Rivera, Ruta Sol, Salto, San Antonio de Areco, San Fernando, San Isidro, San Miguel, San Nicolás de los Arroyos, San Vicente, Tandil, Tigre, Trenque Lauquen, Tres Arroyos, Tres de Febrero, Vicente López, Villa Alfredo Fortabat, Zárate
Catamarca	Andalgalá, Belén, Londres, Los Altos, Pomán, Recreo, San Fernando del Valle de Catamarca, San Isidro, San José, Santa María, Saujil, Tinogasta
Córdoba	Córdoba, Cosquín, Dean Funes, La Calera, La Carlota, La Falda, Las Higueras, Mendiolaza, Oncativo, Parque Norte Ciudad de los Niños, Villa Pastora, Almirante Brown, Pilar, Río Cuarto, Río Primero, Río Segundo, San Agustín, San Francisco, Saturnino María Laspiur, Valle Hermoso, Villa Allende, Villa Carlos Paz, Villa de las Rosas, Villa Dolores, Villa María, Villa Nueva, Villa Río Icho Cruz, Villa Sarmiento
Corrientes	Bella Vista, Corrientes, Goya, Itá Ibaté, Ituzaingó, Monte Caseros, Nuestra Señora del Rosario de Caá Catí, Paso de los Libres, Perugorría, Saladas, San Roque, Santa Lucía, Santo Tomé
Chaco	Barranqueras, Campo Largo, Charata, Fontana, La Leonesa, Las Breñas, Las Palmas, Machagai, Pampa del Indio, Presidencia de la Plaza, Presidencia Roque Sáenz Peña, Puerto Vilelas, Resistencia, Tres Isletas, Villa Angela, Villa Río Bermejito
Chubut	Comodoro Rivadavia, Dolavon, Esquel, Gaiman, Gobernador Costa, Lago Puelo, Playa Unión, Puerto Madryn, Rada Tilly, Rawson, Sarmiento, Trelew

Provincia	Localidad
Entre Ríos	Aldea Valle María, Basavilbaso, Chajarí, Colonia Avellaneda, Concepción del Uruguay, Concordia, Crespo, Federación, General Ramírez, Gualeguaychú, La Paz, Lucas González, Paraná, San José, Santa Elena
Formosa	Clorinda, El Colorado, Estanislao del Campo, Formosa, Ibarreta, Laguna Yema, Las Lomitas, Misión Tacaaglé, Palo Santo, Pirané, Pozo del Tigre, Villa Kilómetro 213
Jujuy	Abra Pampa, Aguas Calientes, Caimancito, El Piquete, Libertador General San Martín, Palpalá, Perico, San Pedro, San Salvador de Jujuy, Santa Clara
La Pampa	25 de Mayo, Catriló, Colonia Barón, Eduardo Castex, General Acha, General Pico, Ingeniero Luiggi, Macachín, Rancul, Realicó, Santa Rosa, Toay
La Rioja	Chamical, Chepes, Chilecito, La Rioja, Milagro, Salicas-San Blas, Villa San José de Vinchina, Villa Unión
Mendoza	Eugenio Bustos, Godoy Cruz, Guaymallén, Las Heras, Luján de Cuyo, Maipú, Malargüe, Mendoza, Perdriel, Real del Padre, Rivadavia, San Martín, San Rafael, Tres Porteñas, Villa Atuel
Misiones	25 de Mayo, Cerro Azul, Concepción de la Sierra, Dos de Mayo, Eldorado, Garupá, Oberá, Posadas, Posadas (Expansión), Puerto Esperanza, Puerto Iguazú, Puerto Rico, San Vicente
Neuquén	Aluminé, Centenario, Chos Malal, Cutral Có, Las Lajas, Neuquén, Plaza Huincul, Plottier, San Patricio del Chañar, 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, General Güemes, General Mosconi, Misión El Cruce-El Milagro-El Jardín de San Martín, Rosario de la Frontera, Rosario de Lerma, Salta, San Antonio de los Cobres, San Ramón de la Nueva Orán, Tartagal, Vaqueros
San Juan	Barreal-Villa Pituil, Caucete, Chimbas, Los Berros, Rawson, Rivadavia, San José de Jáchal, San Juan, Santa Lucía, Villa Aberastain-La Rinconada, Villa Barboza-Villa Nacusi, Villa El Salvador-Villa Sefair, Villa General San Martín-Campo Afuera, Villa Media Agua, Villa Santa Rosa
San Luis	Buena Esperanza, Juana Koslay, Justo Daract, La Punta, La Toma, Merlo, Quines, San Francisco del Monte de Oro, San Luis, Santa Rosa del Conlara, Villa Mercedes
Santa Cruz	28 de Noviembre, Caleta Olivia, El Calafate, Gobernador Gregores, Las Heras, Los Antiguos, Pico Truncado, Puerto San Julián, Puerto Santa Cruz, Río Gallegos
Santa Fe	Arequito, Arroyo Seco, Avellaneda, Cañada de Gómez, El Trébol, Florencia, Fray Luis Beltrán, Funes, Granadero Baigorria, La Criolla, Moisés Ville, Pérez, Puerto General San Martín, Rafaela, Reconquista, Roldán, Rosario, San José del Rincón, San Lorenzo, Santa Fe, Santo Tomé, Sastre, Sauce Viejo, Venado Tuerto, Villa Constitución, Villa Gobernador Gálvez
Santiago del Estero	Frías, La Banda, La Dársena, Monte Quemado, Quimilí, Sachayoj, Santiago del Estero, Termas de Río Hondo, Villa Atamisqui, Villa Ojo de Agua, Villa San Martín (Est. Loreto)
Tucumán	Alderetes, Banda del Río Salí, Concepción, Diagonal Norte-Luz y Fuerza-Los Pocitos-Villa Nueva Italia, El Manantial, Ingenio San Pablo, Los Ralos, Lules, 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, Ushuaia

Anexo II. Distribución de la muestra de viviendas seleccionadas por jurisdicción

Cuadro 6. Distribución de la muestra de viviendas seleccionadas por jurisdicción

Jurisdicción	Cantidad de viviendas
CABA	3.145
Partidos del GBA	6.605
Resto de Buenos aires	3.685
Catamarca	630
Córdoba	2.205
Corrientes	1.110
Chaco	1.155
Chubut	1.275
Entre Ríos	1.305
Formosa	1.265
Jujuy	875
La Pampa	850
La Rioja	895
Mendoza	1.505
Misiones	1.190
Neuquén	785
Rio Negro	1.285
Salta	1.000
San Juan	1.210
San Luis	1.340
Santa Cruz	810
Santa Fe	2.010
Santiago del Estero	695
Tucumán	880
Tierra del Fuego	455
Total del país	38.165

Fuente: INDEC, *Encuesta de Actividades de Niños, Niñas y Adolescentes 2016-2017*.

Anexo III. 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 de calidad importante y permite evaluar en forma general el desempeño en la operación de captura de datos en una encuesta. Los estándares o protocolos adoptados por la comunidad estadística, por ejemplo, el de la American Association for Public Opinion Research (AAPOR, 2016) o por el Council of American Survey Research Organizations (Frankel, 1983) sugieren realizar los cálculos a partir de considerar no solo las unidades elegibles y con respuesta, sino también las de ilegibilidad dudosa o desconocida.

Esta modalidad permite tener en cuenta explícitamente la incertidumbre que a menudo supone la elegibilidad de una dirección, vivienda u otra unidad para una encuesta. Por ejemplo, los casos no contactados incluyen aquellos donde no se sabe si existe una vivienda particular en la dirección asignada a un encuestador y se desconoce si es elegible para el estudio. Ante la falta de contacto, la elegibilidad será desconocida, a menos que pueda ser determinada de alguna otra forma (información adicional del marco muestral, afirmación de un vecino, inspección ocular de la unidad seleccionada, revisita por parte de supervisor, etcétera). Existen situaciones en las que el contacto es imposible por presencia de sistemas de seguridad, portones cerrados, unidades de vivienda múltiple de difícil acceso, o por tratarse de áreas inaccesibles ya sea por inclemencias climáticas o cuestiones de inseguridad. También es posible que la dirección brindada sea errónea, 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 contemplan 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 todos 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 identificados 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: $RR3 = \frac{R}{EL+e*ED}$.

El siguiente cuadro presenta la tasa de respuesta con la cota superior o valor máximo estimado a partir de $RR5 = \frac{R}{EL}$, cuando se asume $e = 0$, por región y total país[35].

Cuadro 7. Tasas de respuesta por regiones y total del país

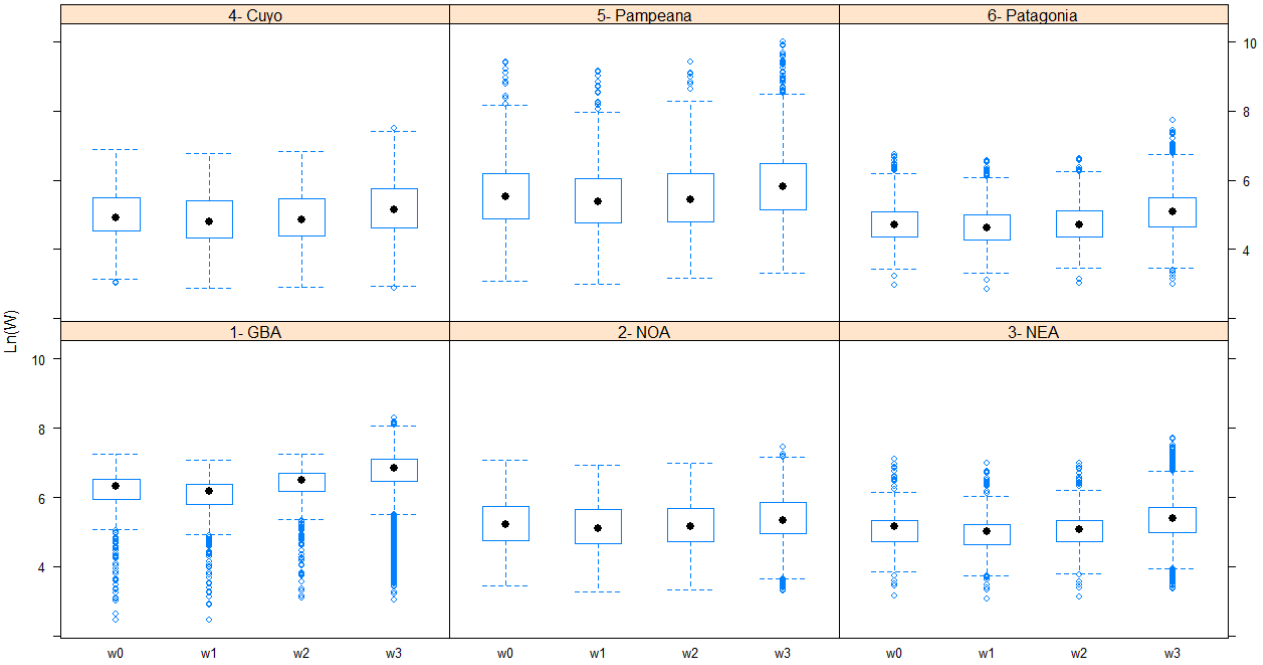
Regiones	RR3	RR5
Gran Buenos Aires	49,7%	70,0%
Noroeste	90,7%	94,9%
Noreste	82,8%	88,8%
Cuyo	90,1%	95,2%
Pampeana	79,2%	89,8%
Patagonia	82,8%	90,6%
Total del país	75,0%	86,9%

Fuente: INDEC, *Encuesta de Actividades de Niños, Niñas y Adolescentes 2016/2017*.

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

Anexo IV. Distribución de los factores de expansión resultantes de cada ajuste

Gráfico 1. Distribución de las ponderaciones de los hogares, en escala logarítmica, según los distintos ajustes, por regiones



Fuente: INDEC, *Encuesta de Actividades de Niños, Niñas y Adolescentes 2016-2017*.

Donde,

w0 es la ponderación de diseño de los hogares $w_{ijkl}^{(0)}$,
w1 es la ponderación de los hogares ajustada por elegibilidad $w_{ijkl}^{(1)}$,
w2 es la ponderación de los hogares ajustada por no respuesta $w_{ijkl}^{(2)}$,
w3 es la ponderación de los hogares calibrada a los totales poblacionales $w_{ijkl}^{(3)}$

$w_{ijkl}^{(t)}$ es el ponderador del l-ésimo hogar, de la k-ésima vivienda ubicada en la j-ésima USM dentro de la i-ésima UPM en el paso de ajuste t-ésimo, $t = 0{,}1{,}2\ y\ 3$.

Glosario

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

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

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

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

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

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

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

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

Coeficiente de variación (CV). Dentro del ámbito del muestreo en poblaciones finitas, constituye otra forma de presentar el error de muestreo. Se lo obtiene a partir del cociente entre el error estándar del estimador y el estimador. En general, se lo calcula en términos porcentuales, siendo esto un beneficio, dado que es una cantidad libre de unidad de medición, lo que permite la comparabilidad.

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

Diseño muestral. Marco metodológico y de trabajo que sirve de base para la selección de la muestra, y que afecta a otros aspectos importantes de un estudio o encuesta. Se 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; 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, o porque no fueron previstos, o 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 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. Cociente entre la variancia de un estimador correspondiente al diseño muestral empleado para seleccionar la muestra (en general, complejo) y la variancia del estimador que se obtendría bajo un muestreo simple al azar (MSA) de igual tamaño. 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. Referida a si una unidad de la muestra es parte de la población objetivo o no. Errores en la determinación de la elegibilidad afectan directamente a 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 las asume 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.

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 que promediadas sobre todas las realizaciones de la muestra, hace que difiera consistentemente de su verdadero valor poblacional o parámetro. A diferencia de la varianza muestral del estimador que se puede estimar desde la propia muestra, el sesgo necesita de valores poblacionales, desconocidos a menos que se realice un censo, para poder ser cuantificado. Aun así, el ECM es una medida importante que se emplea para estudiar el comportamiento teórico de un estimador, y su formulación analítica corresponde a la suma de la variancia muestral del estimador y el sesgo al cuadrado.

Error de cobertura. Diferencias entre la población objetivo y la población que cubre el marco muestral producen errores de esta índole en un estimador. Pueden deberse a problemas de subcobertura y sobrecobertura del marco (ver **Cobertura**). En el primer caso, algunos elementos de la población objetivo tienen una probabilidad nula de ser seleccionados para una muestra. En el segundo, por incluir erróneamente o duplicar algunos de los elementos, estos poseen una probabilidad de ser seleccionados cuando no la deben tener, o 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 o error muestral o error por muestra. Error asociado con la no observación, es decir, ocurre porque no todos los miembros de la población se incluyen en la muestra. Se refiere a la diferencia entre la estimación derivada de la muestra y el valor "verdadero" que resultaría si se realizara un censo de toda la población bajo las mismas condiciones en las que se llevó adelante la muestra. Tiene la particularidad de ir disminuyendo a medida que aumenta el tamaño de la muestra, y a través del muestreo probabilístico es posible estimarlo a partir de la propia muestra. En ausencia de sesgo, este error se corresponde a la componente aleatoria definida por la varianza muestral del estimador que da origen a la estimación.

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

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

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

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

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

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

Localidad. Unidad geoestadística urbana, determinada por criterios físicos y territoriales. Por su clasificación, puede ser simple, si se extiende sobre una sola jurisdicción y no está atravesada por ningún límite de provincia, departamento o partido, ni de gobierno local; o compuesta (también “aglomerado”), cuando se extiende sobre más de una jurisdicción. Para la MMUVRA, todas las localidades de 2.000 o más habitantes, según el Censo Nacional de Población y Viviendas 2010, conforman las UPM del marco de muestreo adoptado para el diseño muestral.

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

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

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

MMUVRA. Muestra maestra urbana empleada por el INDEC con alcance nacional restringido a las localidades de 2.000 o más habitantes, que se utiliza como marco secundario de selección de viviendas particulares para todas sus encuestas a hogares entre dos censos de población y viviendas. Posee un diseño muestral complejo, y se le 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, pero que suele ser elegida por motivos asociados a costos, eficiencia u oportunidad.

Muestra aleatoria. Ver **Muestra probabilística**.

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

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

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

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

Muestreo con probabilidad proporcional al tamaño. Modalidad del muestreo probabilístico que puede llevarse a cabo cuando las unidades del marco de muestreo tienen una medida de tamaño asignada. La probabilidad de inclusión de una unidad en una muestra queda definida por la relación entre su tamaño y la suma de tamaños de todas las unidades de la población, o una función de ellas. Bajo esta estrategia, las unidades de mayor tamaño tienen una probabilidad más alta de participar en

una muestra. En encuestas a hogares, conjuntamente con el muestreo por conglomerados, es la estrategia más adoptada por las oficinas nacionales de estadísticas (ONE) para seleccionar las muestras de viviendas de sus principales operativos estadísticos.

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

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 al conglomerado. En encuestas a hogares, esta alternativa de muestreo permite disminuir los costos de la encuesta, en perjuicio de perder, generalmente, precisión en las estimaciones al depender de la homogeneidad interna entre las unidades con respecto a las características que se están estudiando.

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

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

Nivel de confianza. Probabilidad, fijada *a priori*, de que una afirmación sobre el valor de un parámetro poblacional sea correcta. Generalmente, 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 muestral. Peso o ponderación de cada unidad muestral que se emplea en todo el proceso inferencial y que genera los resultados de la encuesta. A causa del muestreo, cada unidad representa a otras de la población objetivo por un factor o peso definido por la inversa de la probabilidad de inclusión de la unidad, según el diseño muestral empleado. Estos pesos iniciales pueden sufrir ajustes de distinta índole, y que dan origen a los factores de expansión finales de la encuesta. (Ver **Factor de expansión**).

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 que posee límites conocidos y precisos, con un determinado número de viviendas, y de carácter operativa 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 (USM) 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. Aunque existen otros factores más determinantes que impactan en la magnitud del sesgo, en particular, el grado de asociación que existe entre la probabilidad a dar respuesta de los individuos de la población y las características que están siendo estudiadas.

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

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