

4° Encuesta Nacional de Factores de Riesgo

NOTA TÉCNICA ENFR

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

Octubre de 2019



NOTAS
TÉCNICAS
INDEC
N° 2

4° Encuesta Nacional de Factores de Riesgo

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

Nota técnica - Octubre 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 Marí y el Lic. Gregorio García.

ISSN 2683-8478

ISBN 978-950-896-555-4

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

4° Encuesta Nacional de Factores de Riesgo: factores de expansión, estimación y cálculo de los errores de muestreo. - 1a ed. - Ciudad Autónoma de Buenos Aires: Instituto Nacional de Estadística y Censos - INDEC, 2019.

Libro digital, PDF - (Notas técnicas; 2)

Archivo Digital: descarga y online

ISBN 978-950-896-555-4

1. Estadísticas. 2. Matemática Estadística. 3. Factores de Riesgo. I. Título
CDD 510

Libro de edición argentina



Esta publicación utiliza una licencia Creative Commons. Se permite su reproducción con atribución de la fuente.

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, octubre de 2019

Publicaciones del INDEC

Las publicaciones editadas por el Instituto Nacional de Estadística y Censos pueden ser consultadas en www.indec.gov.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.gov.ar

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



Índice

1. Introducción	4
2. Diseño muestral de la encuesta	4
3. Pasos de la encuesta	6
4. Dominios de estimación, tamaño y asignación de la muestra	8
5. Determinación de los factores de expansión para el Paso 1	9
6. Determinación de los factores de expansión para el Paso 2	16
7. Determinación de los factores de expansión para el Paso 3	17
8. Determinación de los factores de expansión para hogar	18
9. Estimación a partir de los datos de la encuesta	19
10. Indicadores de calidad asociados al error de muestreo	20
11. Estimación de los errores de muestro mediante replicaciones	21
12. Modo de empleo de los pesos replicados	23
13. Recomendaciones para el uso con fines estadísticos de los datos de la encuesta	33
Referencias	37
Anexo I.A Total de UPM y USM de la MMUVRA presentes en la ENFR	39
Anexo I.B Localidades de 5.000 habitantes y más de la MMUVRA involucradas en la ENFR	40
Anexo I.C Distribución territorial de los aglomerados y localidades que participan en la ENFR	42
Anexo II. Dominio de estimación del Paso 3 - Aglomerados de 150.000 habitantes y más	43
Anexo III. Distribución del total de la muestra de viviendas seleccionadas por jurisdicción y paso	44
Anexo IV.A Total de viviendas elegibles, no elegibles y de elegibilidad dudosa por jurisdicción - Paso 1	45
Anexo IV.B Causas de no elegibilidad o dudosa de las viviendas por jurisdicción - Paso 1	46
Anexo IV.C Total de hogares con y sin respuesta por jurisdicción - Paso 1	47
Anexo IV.D Total de personas seleccionadas con y sin respuesta por jurisdicción - Paso 1	48
Anexo IV.E Total de personas por causa de no respuesta por jurisdicción - Paso 1	49
Anexo V.A Total de viviendas elegibles o no y de elegibilidad dudosa por jurisdicción - Paso 2	50
Anexo V.B Total de hogares con y sin respuesta por jurisdicción - Paso 2	51
Anexo V.C Total de personas elegibles, con y sin respuesta por jurisdicción - Paso 2	52
Anexo V.D Total de personas que no responden por causa de no respuesta y jurisdicción - Paso 2	53
Anexo VI.A Total de personas con y sin respuesta por aglomerado - Paso 3	54
Anexo VI.B Total de personas sin respuesta por causa de no respuesta y aglomerado - Paso 3	55
Anexo VII. Tasa de respuesta de los hogares y las personas	56
Glosario	60

1. Introducción

El Instituto Nacional de Estadística y Censos (INDEC) conjuntamente con el Ministerio de Salud y Desarrollo Social (MSyDS) realizaron la Encuesta Nacional de Factores de Riesgo (ENFR) 2018, con el objetivo de proporcionar información válida, confiable y oportuna sobre factores de riesgo (como consumo de tabaco, alcohol, alimentación, actividad física, entre otros), procesos de atención en el sistema de salud y principales enfermedades no transmisibles en la población (hipertensión, diabetes, obesidad y otras). Se trata de la cuarta encuesta de este tipo realizada en el país,¹ las anteriores se realizaron en los años 2005, 2009 y 2013.

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

En primera instancia, se presentan las características principales del diseño muestral, el tamaño de la muestra, su asignación territorial y los dominios de estimación definidos para 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 los cuadros de la encuesta publicados, o para aquellas que se generen con fines estadísticos a partir de la base para los usuarios de la encuesta.

2. Diseño muestral de la encuesta

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

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

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 su selección, marcos

¹ INDEC (2019). *4° Encuesta Nacional de Factores de Riesgo*. Disponible en: www.indec.gob.ar/ftp/cuadros/publicaciones/enfr_2018_resultados_definitivos.pdf.

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), en el que las viviendas son seleccionadas aleatoriamente con igual probabilidad. Sin embargo, se requeriría de una lista de todas las viviendas pertenecientes al ámbito geográfico que abarca 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 visitas para contactar a los informantes en distintos horarios, y de otros equipos que realizan la tarea 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 las unidades de área que la conforman y 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.

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, Hogares y Viviendas 2010 (CNPHV 2010); el conjunto de todas las UPM constituyen el marco de muestreo o la lista de unidades de muestreo para la selección probabilística de primera etapa. Estas son estratificadas de acuerdo al total de población según CNPHV 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. Tanto las UPM autorrepresentadas como las seleccionadas conforman la muestra de aglomerados o localidades de la MMUVRA.

Para la segunda etapa, en las UPM seleccionadas, y solo para ellas, se definieron las “unidades de segunda etapa de muestreo” (USM) o “Áreas MMUVRA”² con base en radios censales, y a la cartografía del CNPHV 2010. En cada UPM, todas sus USM en conjunto cubren territorialmente y determinan la envolvente o área de cobertura asociada a dicha unidad, conformando el marco de

² En la conformación de las Áreas MMUVRA, los radios censales por cuestiones operativas (extensión, densidad, inaccesibilidad, etc.) pueden sufrir recortes o agrupamientos (por ejemplo, para equilibrar la uniformidad de sus tamaños en términos de viviendas).

muestreo para la selección de segunda etapa. Esta se completa con la selección de una muestra probabilística de USM, que emplea un diseño estratificado definido a partir de variables sociodemográficas y mediante un muestreo sistemático proporcional a la cantidad total de viviendas particulares ocupadas, según el CNPHV 2010.

Por último, en cada una de las USM seleccionadas, se confeccionó inicialmente un listado exhaustivo de viviendas particulares, lo que da 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.³ 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.⁴

Por los motivos señalados, el diseño muestral de la ENFR se apoya en el diseño de la MMUVRA ajustado a los requerimientos de la encuesta, que como población objetivo incluye a las localidades y aglomerados de 5.000 o más habitantes.⁵ Para la muestra definitiva de viviendas de la encuesta se realiza 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, y cuyo principal objetivo es concentrar los desplazamientos en terreno de los encuestadores y del personal de salud, para reducir el costo del operativo. Una selección sistemática con igual probabilidad de estos segmentos permitió conformar la muestra definitiva de viviendas de la encuesta.

3. Pasos de la encuesta

La ENFR emplea un cuestionario incorporado a un dispositivo móvil de captura, que se aplica en tres pasos con intervención de un encuestador para la entrevista y un representante de salud para realizar las mediciones físicas y bioquímicas. Cada paso define hitos importantes que afectan a la organización del operativo de campo de la encuesta, el diseño y asignación de la muestra, la definición de los dominios de estimación, y a los cálculos y ajustes de los distintos factores de expansión que se emplean para lograr los resultados.

3.1 Paso 1: características de los hogares y selección del miembro para las mediciones

La ENFR indaga distintas características de los hogares y de todos sus miembros en la vivienda seleccionada a través de un módulo inicial para el hogar⁶ y sus componentes. Este es muy similar a los aplicados en las ediciones anteriores de la ENFR, e incluye preguntas asociadas a las características de la vivienda, el ingreso del hogar, la situación laboral del jefe del hogar, educación, cobertura de salud y situación conyugal de todos los miembros del hogar.

En esta instancia, el encuestador selecciona con igual probabilidad a una persona de 18 años o más, asistido por un algoritmo que emplea la lista de miembros que introdujo en el aplicativo móvil al contactar el hogar, para completar lo que se denomina “Paso 1” de la encuesta.

³ Esta propiedad de permitir submuestrear viviendas sobre la muestra maestra hace que se la identifique también como un marco secundario de muestreo de viviendas.

⁴ A la fecha de la ENFR, la MMUVRA en su última actualización registraba un total de 2.053.958 viviendas particulares.

⁵ Para el total de UPM y USM ver Anexo I.A; para un detalle y la distribución territorial de las localidades de la MMUVRA involucradas en la ENFR, ver Anexo I.B y I.C, respectivamente.

⁶ Ver bloque Hogar (BH) en el cuestionario de la encuesta.

El miembro seleccionado responde a preguntas del cuestionario de carácter individual,⁷ donde declara sobre su situación laboral, salud general, actividad física, consumo de tabaco, hipertensión arterial, peso corporal, alimentación, colesterol, consumo de alcohol, diabetes, lesiones, prácticas preventivas, cáncer colorrectal, y, si es mujer, sobre embarazo.

3.2 Paso 2: mediciones antropométricas

Finalizado el Paso 1, el miembro seleccionado, con su consentimiento, habilita al personal de salud a iniciar el segundo paso de la encuesta (Paso 2). Este incluye una serie de mediciones antropométricas (MA): presión arterial, peso, talla y perímetro de la cintura.

Por razones presupuestarias, disponibilidad de instrumentos para la toma de las mediciones, y la complejidad de la coordinación del trabajo entre los encuestadores y el personal de salud, el Paso 2 por diseño se aplica a una submuestra probabilística del 75% de las viviendas de la encuesta.

La submuestra es una selección de USM que componen el Paso 1, bajo un muestreo simple al azar (MSA) y con una fracción de muestreo de 75% en cada estrato de las USM definida para la MMUVRA⁸, que determina una muestra para el Paso 2 de aproximadamente el 75% de los segmentos y de viviendas de la original afectada al Paso 1.

La selección se realizó al momento de la extracción de la muestra de viviendas de la encuesta, lo que permitió, en forma anticipada, la distribución y asignación de las cargas de trabajo a los encuestadores, en particular la del personal de salud que comienza a participar de la encuesta en este paso.

3.3 Paso 3: mediciones bioquímicas

La encuesta finaliza (Paso 3) con mediciones bioquímicas (MQ). Al miembro del hogar seleccionado y, nuevamente con su consentimiento, se le realizan mediciones de glucemia capilar y colesterol total. Para cumplimentar correctamente con ellas se requiere como condición necesaria que el miembro esté en ayunas, incorporando una complejidad adicional al operativo de campo. Es importante advertir que no todos los miembros afectados al Paso 2 de la encuesta continúan con el Paso 3, solo aquellos que habitan en un aglomerado o localidad de 150.000 habitantes y más lo completan por decisión operativa.

⁷ Ver bloque individual (BI) en el cuestionario de la encuesta.

⁸ Un algoritmo de balanceo aleatorio se aplica en la selección cuando el total de USM por estrato no es múltiplo de 4, para asegurar una redistribución entre los estratos y lograr aproximadamente el 75% de la muestra de viviendas para Paso 2 respetando la estructura por diseño de la estratificación de la MMUVRA.

4. Dominios de estimación, tamaño y asignación de la muestra

Dadas las características señaladas de los pasos, la encuesta por diseño tiene dominios de estimación diferentes para cada uno de ellos. Desde el punto de vista de la muestra, el 100% de las viviendas seleccionadas para el operativo forman parte del Paso 1, lo cual permite, además de resultados a nivel total del país, desagregarlos a nivel de 6 regiones:

- Gran Buenos Aires Ciudad Autónoma de Buenos Aires y los partidos del Gran Buenos Aires
- Noroeste Catamarca, Jujuy, Salta, Tucumán, La Rioja y Santiago del Estero
- Noreste Chaco, Corrientes, Formosa y Misiones
- Cuyo Mendoza, San Juan y San Luis
- Pampeana Córdoba, Santa Fe, Entre Ríos, La Pampa y resto de partidos de la provincia de Buenos Aires
- Patagonia Chubut, Neuquén, Río Negro, Santa Cruz y Tierra del Fuego

Permite, asimismo, desagregarlos para las 24 jurisdicciones de primer orden (23 provincias y CABA), y para cada uno de los ocho aglomerados urbanos de más de 500.000 habitantes:⁹ Gran Buenos Aires, Gran Córdoba, Gran La Plata, Gran Mendoza, Gran Rosario, Gran Salta, Gran Tucumán y Mar del Plata.

Para el Paso 2 de mediciones físicas y antropomórficas, y como consecuencia de la reducción de la muestra de viviendas en un 25% señalada en la sección 3, las estimaciones responden al total del país; en cambio para el Paso 3, al ser un recorte geográfico de la muestra del Paso 2, se define un único dominio compuesto por el conjunto de aglomerados o localidades de 150.000 habitantes y más.¹⁰

El tamaño de la muestra se ajustó a las restricciones presupuestarias y a los requerimientos de precisión para las principales estimaciones en los dominios de estimación previstos para cada paso. La muestra para el Paso 1 es de 49.170 viviendas, de las cuales, por diseño, se seleccionaron 36.870 para continuar con el Paso 2, y 17.390, para el Paso 3. El siguiente cuadro, a manera de resumen, da cuenta de la distribución de la muestra por región y paso.¹¹

Cuadro 1. Distribución de la muestra de viviendas por región y por paso

Regiones	Viviendas		
	Paso 1	Paso 2	Paso 3
Gran Buenos Aires	7.550	5.650	5.650
Noroeste	7.810	5.850	2.880
Noreste	6.690	5.020	1.700
Cuyo	4.120	3.090	1.480
Pampeana	16.260	12.210	4.880
Patagonia	6.740	5.050	800
Total del país	49.170	36.870	17.390

Fuente: INDEC, 4° Encuesta Nacional de Factores de Riesgo.

⁹ Estos aglomerados fueron dominios de estimación en ENFR anteriores y se los mantuvieron en la presente a efectos de permitir la comparabilidad entre los resultados.

¹⁰ Para un detalle de las localidades que componen el dominio de estimación del Paso 3, ver Anexo II.

¹¹ Para un detalle de la distribución de la muestra por jurisdicción, ver Anexo III.

5. Determinación de los factores de expansión para el Paso 1

La estimación de parámetros poblacionales de interés 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 en la población que posee una característica dada se estima sumando las ponderaciones de las personas, hogares o viviendas que tienen la característica en cuestión en la muestra.

La ENFR emplea cuatro juegos de factores de expansión¹² para alcanzar las estimaciones oficiales de la encuesta. Tres están vinculados a las personas que participan de los pasos señalados en las secciones previas y el restante, a los hogares en el caso de que se deseen estimaciones para el módulo inicial del cuestionario para el hogar y sus componentes. Cabe destacar que, para el desarrollo de los ponderadores, será necesario definir el correspondiente a la vivienda que contiene los hogares y personas que participan de la encuesta, si bien no forma parte del conjunto de factores de expansión de interés.

Hipotéticamente, las estimaciones del Paso 1 se obtendrían empleando el factor de expansión del miembro seleccionado que surge del diseño muestral, como producto de las inversas de las probabilidades de inclusión de la vivienda¹³ y la del miembro seleccionado de manera aleatoria en el hogar. Su expresión es:

$$w_{1i}w_{2ij}w_{3ijk}m_{ijkl}$$

Donde:

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

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

w_{3ijk} es la inversa de la probabilidad de inclusión de la k -ésima vivienda dentro de la j -ésima USM de la i -ésima UPM seleccionada¹⁴,

m_{ijkl} es el total de miembros de la población objetivo en el l -ésimo hogar de la k -ésima vivienda, en la j -ésima USM de la i -ésima UPM.

Sin embargo, en la práctica los factores de expansión iniciales suelen ser modificados por diversos motivos y no terminan siendo los que se emplean para obtener las estimaciones de una encuesta. Durante el desarrollo de este operativo estadístico se presentan una serie de problemas, algunos vinculados a errores de cobertura por desactualización del marco de muestreo, a la no respuesta de las unidades, o a la falta de eficacia en la captura de ciertos grupos de la población por la encuesta. Todos estos errores forman parte de los denominados errores “no muestrales”, y que, sumados a otros, contribuyen a la componente del “error total” en una estimación. Son difíciles de cuantificar y afectan a la calidad del dato en dos direcciones. Si son introducidos de manera aleatoria, la probabilidad de incrementar la variabilidad de la estimación es alta; pero si ellos no son aleatorios el principal impacto es introducir sesgo en los resultados.

¹² Los términos “factores de expansión”, “ponderadores” o “pesos” en el contexto del documento hacen referencia siempre al mismo concepto.

¹³ Se asume que todos los hogares de la vivienda seleccionada forman parte de la muestra, por lo tanto la probabilidad de inclusión de un hogar es la de la vivienda a la cual pertenece; o sea, $w_{3ijk} = w_{3ijkl}$ si el hogar l pertenece a la vivienda k .

¹⁴ 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.

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

Pero aun tomando todos estos recaudos, estos errores no desaparecen y llevan a que en la etapa previa a la estimación se incorporen en la determinación de los factores de expansión finales de la encuesta varios ajustes sobre el definido por diseño, lo que busca disminuir el impacto de estos inconvenientes sobre los estimadores y aumentar la calidad de los resultados.

En las secciones 5.1 y 5.2 se describen los ajustes, por “no elegibilidad” y “no respuesta”, que afectan a las ponderaciones de las viviendas y los hogares, respectivamente. A continuación, se detallan los que llevan a determinar el factor de ponderación definitivo para las personas que responden al Paso 1 de la encuesta.

5.1 Ajuste por viviendas no elegibles

Inicialmente, el factor de expansión de una vivienda seleccionada para la encuesta surge de la multiplicación de las inversas de las probabilidades de inclusión de cada una de las etapas de selección definidas en la sección 2. Por lo tanto, para la k -ésima vivienda ubicada en la j -ésima USM del Paso 1 dentro de la i -ésima UPM, se lo define como:¹⁵

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

Donde:

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

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

w_{3ijk} es la inversa de la probabilidad de inclusión de la k -ésima vivienda dentro de la j -ésima USM del Paso 1 de la i -ésima UPM seleccionada.¹⁶

Cabe destacar que el factor de expansión inicial correspondiente a un hogar coincide con el de la vivienda de la cual forma parte, o sea, $w_{0ijkl}^{H(1)} = w_{0ijk}^{V(1)}$, dado que se incluyen en la muestra a todos los hogares que forman parte de la vivienda seleccionada.

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

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

¹⁵ Para facilitar la lectura en la notación se omiten los subíndices correspondientes a los estratos definidos por el diseño muestral de las UPM y las USM, por lo que queda implícita la pertenencia a los mismos cada vez que se refiera al subíndice i de las UPM y al j de las USM.

¹⁶ 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.

- Viviendas elegibles (VEL) a aquellas en donde se detecta una vivienda particular y se realiza una entrevista; 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
 - de fin de semana
 - en construcción
 - vivienda usada como establecimiento
 - variaciones en el listado: no es vivienda

- Viviendas de elegibilidad dudosa o 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 por problemas climáticos u otros, o no se especificó ninguna.

Teniendo en cuenta esta 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 asume que son elegibles, mediante la siguiente expresión:¹⁷

$$\sum_{EL} w_{0ijk}^{V(1)} + e_1 \sum_{ED} w_{0ijk}^{V(1)}$$

Donde:

$$e_1 = \frac{\sum_{EL} w_{0ijk}^{V(1)}}{\sum_{EL} w_{0ijk}^{V(1)} + \sum_{NE} w_{0ijk}^{V(1)}}$$

es la tasa de elegibilidad en el Paso 1,

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.

Los cálculos se realizan dentro de grupos o “clases de ajuste” disjuntas definidas exclusivamente para los cálculos, y que surgen del cruce de la variable provincia o jurisdicción (25), con la división “aglomerado EPH” y “resto de las UPM” (2), y los estratos de diseño de la MMUVRA para las USM (5). En consecuencia, en cada clase *c*, *c* = 1, ..., 250, el primer factor de ajuste, $a_{1c}^{(1)}$, es definido por la

¹⁷ En la simbología empleada en la guía, \sum_A representa la suma sobre todas las unidades que pertenecen al conjunto *A*.

proporción de viviendas que se estiman como elegibles sobre el total de viviendas estimadas por la encuesta para el Paso 1¹⁸ empleando la tasa de elegibilidad e_{1c} dentro de la clase c :

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

En el Anexo IV.A, se presentan los resultados de la encuesta en relación a la cantidad de VEL, VNE y VED por jurisdicción, que intervienen con sus factores de expansión iniciales, $w_{0ijk}^{V(1)}$, en los cálculos de los factores $a_{1c}^{(1)}$. Para un detalle de los motivos o causas de no elegibilidad o elegibilidad dudosa detectados en el operativo de campo en la ENFR, ver Anexo IV.B.

5.2 Ajuste por no respuesta de los hogares

Cuando se identifica una vivienda como elegible para la encuesta, y por consiguiente los hogares que la componen, no siempre es posible hacer una entrevista, lo cual origina una no respuesta¹⁹ del hogar. Esto puede ocurrir debido a una serie de razones: que en el hogar contactado ningún integrante quiera responder; que haya ausencia temporal de todos sus miembros durante el período de la encuesta; o bien si existió un contacto, por algún motivo o alguna circunstancia fue imposible continuar con la entrevista. En particular, en la ENFR 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: por causas circunstanciales, viaje o vacaciones.
- Rechazo: por cualquiera de las razones expresadas.
- Otras causas: duelo, alcoholismo, discapacidad, idioma extranjero.

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

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

Una de las claves para lograr el éxito del ajuste es determinar clases o grupos de unidades en la población que expliquen lo mejor posible el mecanismo de no respuesta que hay por detrás del

¹⁸ En las fórmulas, $EL(c)$, $ED(c)$ y $NE(c)$ señalan a los conjuntos EL, NE y ED restringidos a la clase c .

¹⁹ Bajo ninguna circunstancia las viviendas seleccionadas para la encuesta son reemplazadas por otras viviendas por razones de no respuesta.

fenómeno. Desde el punto de vista de la eficiencia en las estimaciones, se busca que los agrupamientos:

- permitan sostener en lo posible el supuesto de probabilidad de respuesta constante de las unidades dentro de ellos, y
- sean lo más homogéneos posibles, para que valga en algún grado la hipótesis de que, en una clase dada, los encuestados sean similares a los no encuestados en términos de las principales variables de interés.

Para realizar las correcciones por no respuesta en la ENFR, se emplean las mismas clases conformadas para el ajuste por no elegibilidad, definidas en la sección anterior. A partir de ellas, en cada clase c se obtiene un segundo factor de ajuste, $a_{2c}^{(1)}$, por “no respuesta” del hogar para los factores de expansión de los hogares del Paso 1, definido como:

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

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

Cabe mencionar que en esta instancia se realiza un reagrupamiento de clases cuando el valor del ajuste es más grande que 2,5 dentro de una clase. Este se efectúa a nivel de la variable estrato de USM, colapsando 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 muy grandes que tienden a incrementar el error de muestreo en las estimaciones.

Para ilustrar la cantidad de hogares de la muestra que, con sus factores de expansión ajustados por viviendas no elegibles, se involucran en los cálculos de este nuevo factor, $a_{2c}^{(1)}$, en el Anexo IV.C se presenta el total de los hogares con y sin respuesta del Paso 1 de la encuesta por jurisdicción.

5.3 Ajuste por no respuesta de las personas seleccionadas

Como se describe en las primeras secciones, el encuestador selecciona al azar a un miembro de 18 años o más de todos los que habitan el hogar para continuar con el Paso 1 (y el resto de los pasos si correspondiera). El factor inicial de expansión de esta persona se determina de manera condicional, a partir del conjunto de hogares con respuesta dentro de las viviendas elegibles. Este factor inicial, para un miembro seleccionado del hogar l , viene dado por:

$$w_{0ijkl}^{P(1)} = w_{0ijkl}^{H(1)} a_{1c}^{(1)} a_{2c}^{(1)} m_{ijkl}$$

Es decir, el factor de expansión del hogar al que pertenece con los ajustes que se le aplicaron hasta esta instancia, multiplicado por la inversa de la probabilidad de selección del individuo dentro del hogar.

Los $w_{0ijkl}^{P(1)}$ también son ajustados por la no respuesta de los miembros seleccionados del Paso 1. Las causas, consignadas por el encuestador, que determinan una no respuesta del individuo que se pretende entrevistar son:

- Ausencias reiteradas.
- Rechazo:
 - negativa rotunda,
 - rechazo por portero eléctrico,
 - se acordaron entrevistas que no se concretaron,
 - la encuesta demanda mucho tiempo,
 - no quiere hablar del tema,
 - desconfía de qué van a hacer con los datos.
- No brinda el consentimiento a continuar con la encuesta.
- Otra causa.

Una vez más, se emplean las mismas clases “de ajuste” introducidas en las secciones anteriores para llevar adelante los cálculos; o sea, en cada clase c , $c = 1, \dots, 250$, se obtiene un factor de ajuste por no respuesta a nivel de persona, $a_{3c}^{(1)}$, definido por:

$$a_{3c}^{(1)} = \frac{\sum_{PR(c)} w_{0ijkl}^{P(1)} + \sum_{PNR(c)} w_{0ijkl}^{P(1)}}{\sum_{PR(c)} w_{0ijkl}^{P(1)}}$$

Donde $PR(c)$ y $PNR(c)$ son el conjunto de personas que responden o no al Paso 1 en la clase c , respectivamente; también para este factor se procede a reagrupamientos de clases, con los mismos criterios de adoptados en el ajuste de la sección 5.2. En consecuencia, la expresión del factor de expansión de un miembro seleccionado y que responde al Paso 1 de la encuesta viene dada por:

$$\tilde{w}_{ijkl}^{P(1)} = w_{0ijkl}^{H(1)} a_{1c}^{(1)} a_{2c}^{(1)} m_{ijkl} a_{3c}^{(1)}$$

En el Anexo IV.D, un cuadro ilustra la cantidad total de personas seleccionadas con y sin respuesta del Paso 1 de la encuesta, que con sus ponderaciones iniciales intervienen en los cálculos de este ajuste; en el Anexo IV.E se presenta las causas de no respuesta de los miembros elegidos consignadas por el encuestador en este paso.

5.4 Ajuste por calibración de los factores de expansión de personas

Los factores de expansión de cada persona seleccionada y que responde a la encuesta hasta esta instancia, $\tilde{w}_{ijkl}^{P(1)}$, reciben una última modificación o ajuste, denominado “calibración”. Este procedimiento emplea información auxiliar de una fuente externa disponible, y tiene por objetivo contribuir a una mejora en los ajustes ya realizados, y a corregir posibles sub o sobre representaciones en algunos grupos de la población, originadas cuando estos no están bien captados por la encuesta. Para disminuir estas discrepancias, la calibración busca la consistencia entre las estimaciones de algunas variables de la encuesta y totales poblacionales conocidos, o *benchmarks*, para esas variables.

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

Al proceso de calibración que opera sobre el conjunto de personas que responden al paso y genera el sistema de ponderadores definitivos de la encuesta para el Paso 1, $w_{ijkl}^{P(1)}$, se lo puede traducir en el siguiente problema numérico de optimización:

$$\begin{aligned} & \text{Minimizar } \sum_R G(\tilde{w}_{ijkl}^{P(1)}, w_{ijkl}^{P(1)}), \\ & \text{sujeto a: } \sum_R w_{ijkl}^{P(1)} \mathbf{x}_{ijkl}^P = \sum_U \mathbf{x}_q^P \end{aligned}$$

En donde G es una función que define la proximidad entre los factores deseados y los surgidos del último ajuste, y la igualdad propone que las estimaciones para un conjunto de q variables auxiliares, $\mathbf{x}_{ijkl}^P = (x_{ijkl1}^P, \dots, x_{ijklq}^P)^T$ medidas en la encuesta, a partir de los factores de expansión deseados, $w_{ijkl}^{P(1)}$, reproduzcan sus totales poblacionales, $\sum_U \mathbf{x}_q^P = (t_{x1}^P, \dots, t_{xq}^P)$, provistos por una fuente externa a la encuesta (Valliant, Dever y Kreuter, 2013).

Dada G , la resolución numérica es un proceso iterativo, que bajo ciertas condiciones de regularidad converge y permite obtener factores de ajuste por calibración, $\lambda_{ijkl}^{(1)}$, para cada persona con respuesta del Paso 1. En la ENFR se emplearon 7 variables que reflejan la estructura demográfica por sexo y grupos de edad, donde $\mathbf{x}_{ijkl}^P = (x_{ijkl1}^P, \dots, x_{ijkl7}^P)$ y cuyas componentes son:

$$\begin{aligned} x_{ijkl1}^P &= 1 \text{ si la persona es mujer de 18 años o más, y 0 en otro caso,} \\ x_{ijkl2}^P &= 1 \text{ si la persona es varón de 18 años o más, y 0 en otro caso,} \\ x_{ijkl3}^P &= 1 \text{ si la persona tiene entre 18 y 24 años, y 0 en otro caso,} \\ x_{ijkl4}^P &= 1 \text{ si la persona tiene entre 25 y 34 años, y 0 en otro caso,} \\ x_{ijkl5}^P &= 1 \text{ si la persona tiene entre 35 y 49 años, y 0 en otro caso,} \\ x_{ijkl6}^P &= 1 \text{ si la persona tiene entre 50 y 64 años, y 0 en otro caso,} \\ x_{ijkl7}^P &= 1 \text{ si la persona tiene 65 años y más, y 0 en otro caso,} \end{aligned}$$

y los totales de población, involucrados como marginales para estas variables en el proceso iterativo, provienen de proyecciones poblacionales.²⁰

Para la calibración en la ENFR, se emplea la función de distancia “logit” (Deville y Särndal, 1992; Haziza y Beaumont, 2017) del *package* Survey de R (Lumley, 2018), que permite controlar el rango de los $w_{ijkl}^{P(1)}$, y así sus valores extremos. De esta forma se limita el riesgo de incrementar el error de muestreo en las estimaciones de la encuesta.

El proceso de calibración se efectúa en forma independiente por provincia o jurisdicción, y en lo posible el ajuste involucra los totales proyectados por sexo y grupos de edad según la división aglomerado EPH y resto de las UPM dentro de la provincia en cuestión. La expresión definitiva del factor de expansión de una persona seleccionada para el Paso 1, que responde a la encuesta y que incluye todos los ajustes, viene dada por:

$$w_{ijkl}^{P(1)} = w_{0ijkl}^{H(1)} a_{1c}^{(1)} a_{2c}^{(1)} m_{ijkl} a_{3c}^{(1)} \lambda_{ijkl}^{(1)}$$

²⁰ Los totales poblacionales proyectados fueron calculados a partir de datos censales de población según CNPHV 2010 al 15 de noviembre de 2018 y determinados por la Dirección Nacional de Estadísticas Sociales y Poblacionales del INDEC.

Donde:

$w_{0ijkl}^{H(1)}$ es el factor de expansión inicial del l -ésimo hogar, de la k -ésima vivienda del Paso 1 ubicada en la j -ésima USM dentro de la i -ésima UPM,

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

$a_{2c}^{(1)}$ es el factor de corrección por no respuesta del hogar perteneciente a la clase c de ajuste,

m_{ijkl} es el total de miembros de la población objetivo en el l -ésimo hogar, en la k -ésima vivienda, en la j -ésima USM de la i -ésima UPM,

$a_{3c}^{(1)}$ es el factor de corrección por no respuesta de personas perteneciente a la clase c de ajuste, y

$\lambda_{ijkl}^{(1)}$ es el factor de ajuste que surge de la calibración correspondiente a la persona seleccionada del hogar l -ésimo, de la k -ésima vivienda del Paso 1 ubicada en la j -ésima USM dentro de la i -ésima UPM.

Esta formulación es válida siempre que la vivienda, el hogar y el miembro seleccionado pertenezcan a la clase de ajuste c , $c = 1, \dots, 250$.

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 $w_{ijkl}^{P(1)}$ finales que se emplean para todas las estimaciones oficiales del Paso 1 de la encuesta.

6. Determinación de los factores de expansión para el Paso 2

A diferencia del Paso 1, las viviendas que participan en el Paso 2 provienen de una subselección de USM y, como se señala en la sección 3.2, esta subselección origina una muestra de viviendas de aproximadamente el 75% de las seleccionadas para el Paso 1. Como consecuencia, solo las probabilidades de inclusión correspondientes a las USM de la submuestra del Paso 2 se ven afectadas por esta circunstancia y la expresión correspondiente al factor de expansión de una vivienda que continúa en el Paso 2 es:

$$w_{0ijk}^{V(2)} = w_{1i} w'_{2ij} w_{3ijk}$$

Donde:

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

w'_{2ij} es la inversa de la probabilidad de inclusión en la segunda etapa de muestreo de la j -ésima USM de Paso 2 dentro de la i -ésima UPM seleccionada;

w_{3ijk} es la inversa de la probabilidad de inclusión de la k -ésima vivienda dentro de la j -ésima USM de Paso 1 de la i -ésima UPM seleccionada.

El valor de w'_{2ij} es calculado multiplicando a w_{2ij} por la inversa de la probabilidad de inclusión debida a la subselección de USM para el Paso 2, a_{sub2} , que se determina por diseño a nivel de provincia, UPM y estrato de USM, y se define como:

$$a_{sub2} = \frac{n_{USM1}}{n_{USM2}}$$

Donde:

n_{USM1} es el total de USM seleccionadas en el Paso 1 para la ENFR de la MMUVRA en una provincia, UPM, y estrato de USM, y

n_{USM2} es el total de USM seleccionadas de las anteriores a través de un muestreo simple al azar con fracción de muestreo de 75% para el Paso 2, en la provincia, UPM, y estrato de USM en cuestión.

A partir de $w_{0ijk}^{V(2)}$ y con el mismo procedimiento para los ajustes detallado en la sección 5, se inicia con el de vivienda “no elegible”, para continuar con el de “no respuesta” de hogar y dar origen al factor de ponderación inicial de la persona del Paso 2,

$$w_{0ijkl}^{P(2)} = w_{0ijkl}^{H(2)} a_{1c}^{(2)} a_{2c}^{(2)} m_{ijkl}$$

Por último, $w_{0ijkl}^{P(2)}$ es ajustado por no respuesta a nivel de persona y por calibración para obtener el factor de expansión final para una persona que responde al Paso 2. Su expresión viene dada por:

$$w_{ijkl}^{P(2)} = w_{0ijkl}^{H(2)} a_{1c}^{(2)} a_{2c}^{(2)} m_{ijkl} a_{3c}^{(2)} \lambda_{ijkl}^{(2)}$$

Es conveniente insistir en que en el Paso 2 solo intervienen las unidades que conforman la submuestra del 75% de las viviendas del Paso 1. Este subconjunto contribuye con totales de viviendas elegibles, no elegibles o de elegibilidad dudosa²¹ y no respuesta de hogares²² y de personas²³, diferentes a los del Paso 1 en los cálculos para determinar los factores $a_{1c}^{(2)}$, $a_{2c}^{(2)}$, $a_{3c}^{(2)}$ y $\lambda_{ijkl}^{(2)}$.

7. Determinación de los factores de expansión para el Paso 3

Como se señala en la sección 3.3, el Paso 3 involucra a las personas que responden al Paso 2 de los aglomerados y localidades de 150.000 habitantes y más. Por este motivo, el factor de expansión inicial para las personas del Paso 3, $w_{0ijkl}^{P(3)}$, es $w_{0ijkl}^{P(2)}$, el definido para el Paso 2.²⁴

Los únicos dos ajustes que recibe el ponderador inicial en este paso son por no respuesta de persona y calibración. Estos ajustes sufren ligeras modificaciones con respecto a los determinados para los otros dos pasos. En el primero, $a_{3c}^{(3)}$, las clases de ajuste son definidas solo a nivel de los “aglomerados EPH”, por lo tanto $c' = 1, \dots, 26$. En el segundo, el proceso de calibración mediante el cual se determinan factores de ajuste $\lambda_{ijkl}^{(3)}$ es realizado utilizando las mismas variables y totales

²¹ Ver Anexo V.A.

²² Ver Anexo V.B.

²³ Ver Anexo V.C.

²⁴ Como las viviendas y los hogares del dominio geográfico del Paso 3 son las del Paso 2, los factores de ajuste $a_{1c}^{(3)}$ y $a_{2c}^{(3)}$ son $a_{1c}^{(2)}$, $a_{2c}^{(2)}$, respectivamente.

proyectados de los pasos anteriores a nivel de cada aglomerado EPH involucrado en el dominio de estimación del Paso 3.

La expresión del factor de expansión final y que se emplea para los resultados del Paso 3 es:

$$w_{ijkl}^{P(3)} = w_{0ijkl}^{P(2)} a_{3c}^{(3)} \lambda_{ijkl}^{(3)}$$

En el Anexo VI.A y VI.B se detalla el total de personas con y sin respuesta en el Paso 3 por aglomerado EPH, y el total de personas clasificadas por causa de no respuesta por aglomerado EPH, respectivamente.

8. Determinación de los factores de expansión para el bloque Hogar

En la sección 3.1 se menciona que la encuesta se inicia con un módulo o bloque Hogar (BH) que permite indagar por características de sus componentes (CH), de la vivienda (CV), el ingreso del hogar (IH) y la situación laboral del jefe del hogar (SLJ). Este conjunto de variables es un intento por describir a grandes rasgos el entorno sociodemográfico y económico del miembro seleccionado que va a continuar con las preguntas y mediciones. Por este motivo, se suma a la encuesta un factor de expansión para el hogar que permite obtener estimaciones para cualquiera de las características estudiadas en el bloque.

La determinación de los factores de expansión para los hogares comparte también los ajustes que se presentan en la sección 5. Teniendo en cuenta los correspondientes a viviendas no elegibles y no respuesta de hogar de los apartados 5.1 y 5.2, $a_{1c}^{(1)}$ y $a_{2c}^{(1)}$, respectivamente, el peso muestral inicial del hogar, $w_{0ijkl}^{H(1)}$, toma la expresión:

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

Los pesos $\tilde{w}_{ijkl}^{H(1)}$ son calibrados para reflejar la estructura demográfica de la población por sexo y grupos de edad siguiendo la metodología expuesta en el punto 5.4. En este caso, se emplean 8 variables que reflejan la composición interna de cada hogar:

$$\mathbf{x}_{ijkl}^H = (x_{ijkl}^H 1, \dots, x_{ijkl}^H 8)$$

Donde:

- $x_{ijkl}^H 1$ = cantidad de mujeres en el hogar,
- $x_{ijkl}^H 2$ = cantidad de varones en el hogar,
- $x_{ijkl}^H 3$ = cantidad de personas entre 0 y 17 años en el hogar,
- $x_{ijkl}^H 4$ = cantidad de personas entre 18 y 24 años en el hogar,
- $x_{ijkl}^H 5$ = cantidad de personas entre 25 y 34 años en el hogar,
- $x_{ijkl}^H 6$ = cantidad de personas entre 35 y 49 años en el hogar,
- $x_{ijkl}^H 7$ = cantidad de personas entre 50 y 64 años en el hogar,
- $x_{ijkl}^H 8$ = cantidad de personas de 65 años y más en el hogar.

Como resultado, el factor de expansión final correspondiente al hogar l -ésimo de la vivienda k -ésima en el Paso 1 en la j -ésima USM de la i -ésima UPM, $w_{ijkl}^{H(1)}$, viene dado por:

$$w_{ijkl}^{H(1)} = w_{0ijk}^{V(1)} a_{1c}^{(1)} a_{2c}^{(1)} \lambda_{Hijkl}^{(1)}$$

Donde:

$w_{0ijk}^{V(1)}$ = factor de expansión inicial de la k -ésima vivienda del Paso 1 ubicada en la j -ésima USM dentro de la i -ésima UPM,

$a_{1c}^{(1)}$ = factor de ajuste por vivienda no elegible en la clase c de ajuste,

$a_{2g}^{(1)}$ es el factor de ajuste por no respuesta del hogar en la clase c de ajuste,

$\lambda_{Hijkl}^{(1)}$ es el factor de ajuste que surge de la calibración correspondiente al hogar l -ésimo de la k -ésima vivienda en el Paso 1 ubicada en la j -ésima USM dentro de la i -ésima UPM.

9. Estimación a partir de los datos de la encuesta

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

Para alcanzar las estimaciones de esos parámetros en la ENFR se emplean estimadores que recurren a los factores de expansión $w_{ijkl}^{P(1)}$, $w_{ijkl}^{P(2)}$, $w_{ijkl}^{P(3)}$ y $w_{ijkl}^{H(1)}$, según sea el caso, que surgen de la última etapa de ajuste y pertenecientes al tipo de estimadores “calibrados”.

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 del Paso 1, la expresión de los estimadores más empleados son:

Parámetro	Estimador ²⁵
Total, t_y	$\hat{t}_y = \sum_R w_{ijkl}^{P(1)} y_{ijkl}$
Promedio ²⁶ , \bar{y}	$\hat{y} = \frac{\sum_R w_{ijkl}^{P(1)} y_{ijkl}}{\sum_R w_{ijkl}^{P(1)}}$
Proporción, p	$\hat{p} = \frac{\sum_R w_{ijkl}^{P(1)} y_{ijkl}}{\sum_R w_{ijkl}^{P(1)}}$
Razón, $R_{yz} = \frac{t_y}{t_z}$	$\hat{R}_{yz} = \frac{\hat{t}_y}{\hat{t}_z} = \frac{\sum_R w_{ijkl}^{P(1)} y_{ijkl}}{\sum_R w_{ijkl}^{P(1)} z_{ijkl}}$

²⁵ En todos los casos, \sum_R en las fórmulas hace referencia a sumar sobre las personas que responden a la encuesta.

²⁶ La definición de los parámetros promedio y proporción coincide si Y es una variable binaria, que toma el valor de 1 cuando el individuo posee una característica dada y 0, en caso contrario.

Para lograr los mismos estimadores para los Pasos 2 y 3, en las expresiones se deben reemplazar por los factores de expansión $w_{ijkl}^{P(2)}$ o $w_{ijkl}^{P(3)}$, según sea el caso; y por $w_{ijkl}^{H(1)}$ cuando se los requiera para estimaciones a nivel de hogar.

10. Indicadores de calidad asociados al error de muestreo

Una de las etapas centrales de toda encuesta es la que evalúa la calidad de los datos, o sea, el proceso de analizar el producto final en términos de precisión y confiabilidad. Contar con indicadores de calidad de una encuesta permite a los usuarios cuantificar el grado de confianza y conocer las limitaciones que pueden llegar a tener los resultados, y 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 ENFR, la inferencia estadística sobre la población objetivo se basa en los datos recopilados de solo una parte de esta población. Es así que los resultados probablemente diferirán de los que se pueden obtener a partir de un censo completo.

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

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

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

A partir de la varianza muestral se pueden definir otras medidas más populares 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.

Al EE se lo 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, al CV se lo define como el cociente entre el EE y el estimador. No depende de las unidades en que se mide la estimación, en virtud que es una medida relativa a esta. Generalmente se lo expresa como un porcentaje, y en la práctica una estimación del CV es una de las más empleadas para informar el error de muestreo de las estimaciones de una encuesta.

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

11. 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 a base de replicaciones. La idea básica de esta estrategia es tratar al conjunto de datos de la muestra como si esta fuera la población y generar de una manera sistemática un conjunto de submuestras que pueden emplearse para estimar el error muestral en las estimaciones.

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

- incluir en la etapa de la conformación de las réplicas al conjunto de ajustes que sufren los factores de expansión iniciales (no elegibilidad, no respuesta y calibración), para incorporar la variabilidad propia de estas correcciones en los cálculos del error de muestreo y que resultan dificultosas con otros métodos;
- brindar una solución al problema de obtener estimaciones del error por muestra para un número diverso de estimadores, incluyendo a los de orden (mediana, deciles, percentiles, etc.) o los de desigualdad (índice de Gini, curva de Lorentz, etc.), que en otros métodos son complejos para implementar;
- habilitar a los usuarios a calcular por sus propios medios los errores de muestreo para sus estimaciones, con transparencia y de la misma manera 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 ENFR 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 conglomerados. Se define el peso *bootstrap* $w_{hml}^{*(b)}$ a partir de un peso inicial w_{hml} para la g -ésima unidad en el conglomerado m del estrato h en la réplica b según el siguiente ajuste:

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

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

Estos pesos replicados *bootstrap* permiten calcular la estimación de interés en cada una de las B submuestras, y con la variabilidad de los resultados obtenidos se calcula una medida del error

muestral para la estimación en cuestión. A tal efecto, se define la varianza *bootstrap* de $\hat{\theta}$ a partir de las réplicas como:

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

Donde:

$\hat{\theta}$ es el estimador²⁷ de θ calculado a partir de los ponderadores w_{hmg} definidos para la muestra; y θ , un parámetro poblacional de interés para una característica dada,

y

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

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

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

y el del coeficiente de variación:

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

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

Sin embargo, la adopción de estos supuestos habilita emplearlo como un estimador de varianza para un diseño PPT sin reemplazo, si la selección de las UPM sin reemplazo es más eficiente que la selección de UPM con reemplazo (West, 2012; Särndal, Swensson, y Wretman, 1992), como es el caso de la ENFR, 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 ENFR fueron determinadas en forma independiente en cada jurisdicción. Para ajustarse a los requerimientos del método, en las UPM autorrepresentadas de la encuesta, los estratos para el procedimiento *bootstrap* quedaron definidos por el estrato de la segunda etapa de muestreo y los “últimos conglomerados” por las USM; en cambio en la Encuesta Nacional de Factores de Riesgo, en las UPM no autorrepresentadas, los estratos *bootstrap* se corresponden con el estrato de las UPM y los “últimos conglomerados” con las UPM.

Para obtener estimaciones de varianza estables para varios tipos de análisis, deberían estar disponibles tantas réplicas como sea posible. Sin embargo, se debe alcanzar un compromiso entre garantizar la estabilidad, controlar el tamaño de la base con las réplicas y limitar el tiempo de cálculo, entre otras cuestiones. Por estos motivos, en la ENFR el total de réplicas es de 200 ($B=200$), cantidad que asegura la estabilidad del estimador de varianza para las principales estimaciones de la encuesta.

Todas las réplicas se obtienen de la muestra original, que incluye a todos los hogares y personas de las viviendas elegibles, cuyos factores iniciales vienen dados por $w_{ijkl}^{H(1)} \alpha_{1c}^{(1)}$ para los hogares en el

²⁷ Ver apartado 9.

Paso 1, y por $w_{0ijkl}^{H(1)} a_{1c}^{(p)} a_{2c}^{(p)} m_{ijkl}$ para las personas en el paso p , $p = 1, 2$ y 3 . Estos pasan a ser corregidos o reescalados según el estrato h y el “último conglomerado” m al cual pertenece el hogar, como lo requiere el procedimiento *bootstrap* descrito.

Con el fin de incorporar en la variabilidad que introducen los ajustes efectuados en los factores de expansión de la encuesta, se repiten los mismos ajustes sobre los pesos replicados. Es decir, para cada una de las 200 réplicas, los pesos *bootstrap* son ajustados nuevamente por no respuesta y calibrados por sexo y edad de manera análoga a como lo fueron los pesos originales como se detalla en las secciones 5.4, 6, 7 y 8, correspondientes a cada uno de los pasos de la encuesta. A diferencia de los pesos originales, los pesos *bootstrap* no son sometidos a un proceso de redondeo.

12. Modo de empleo de los pesos replicados

Como consecuencia de todo el procedimiento detallado en la sección anterior, la ENFR dispone de 4 conjuntos de réplicas, 3 asociados a los pasos y a las estimaciones a nivel de personas:

$$\{w_{ijkl}^{*P(p,b)}, b = 1, \dots, 200\}, p = 1, \dots, 3,$$

y otro para el bloque Hogar:

$$\{w_{ijkl}^{*H(b)}, b = 1, \dots, 200\},$$

que vinculados a la(s) base(s) con los microdatos permiten calcular los errores muestrales para las estimaciones oficiales de la encuesta.

La presente sección constituye una guía de cómo deben ser empleadas las réplicas en distintas herramientas de cálculo: R²⁸, SAS²⁹, Stata³⁰ y Wesvar³¹. En caso de no contar con ellas, se presenta un ejemplo que sugiere cómo efectuar el cálculo siguiendo la definición formulada en [1] del apartado 11, y que cualquier usuario puede poner en práctica con pocos recursos.³²

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

Se asume que el usuario puede vincular la(s) base(s) con microdatos de la encuesta con las correspondientes a las réplicas de manera unívoca, a través de los identificadores previstos e disponibles en cada una de ellas para cada paso. Como resultado, cada unidad (persona u hogar) o registro de la base que va a emplear para el cálculo de los errores muestrales posee su factor de expansión asociado y cada uno de los 200 valores de los pesos *bootstrap* replicados.

Por otro lado, solo se incluyen los códigos que brindan las estimaciones puntuales, y el que permite alcanzar una medida del error vía el error estándar o el coeficiente de variación. Se consideran en los ejemplos, la estimación de los parámetros definidos en la sección 9.

²⁸ www.r-project.org. Versión 3.6.

²⁹ www.sas.com. Versión 9.4 M3.

³⁰ www.stata.com. Versión 15.

³¹ www.westat.com/capability/information-systems-software/wesvar. Versión 5.1.

³² No se incluye a la herramienta de cálculo SPSS, ya que no cuenta oficialmente a la fecha con la posibilidad de emplear la metodología desarrollada sin recurrir a una programación *ad hoc*.

Para facilitar las indicaciones, la presentación emplea una notación genérica que no impide la correcta interpretación de los pasos a seguir por el usuario. Por lo tanto, a la base de la encuesta, se la define como **base_encuesta**, e incluye la siguiente información:³³

- **w**: factor de expansión final de la encuesta.³⁴
- **w_rep**b****: peso *bootstrap* replicado, donde *b* representa el número de réplica al cual corresponden los pesos, tomando los valores de 1 a 200.³⁵
- **Y, Z**: variables genéricas (continuas o categóricas), que hacen referencia a características para las cuales se requieren estimaciones de los parámetros poblacionales de interés (ver sección 9), y de sus respectivas estimaciones de los errores de muestreo.

12.1 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*³⁶ (Lumley, 2018). Siguiendo las indicaciones del manual,³⁷ y asumiendo que **base_encuesta** fue importada a R, se define el objeto **disenio**³⁸, que incluye las componentes que se requieren para los cálculos a través de la opción **svrepdesign**.

En **svrepdesign** se invoca al factor de expansión de la encuesta (**w**), al método que generó las réplicas (**bootstrap**), el conjunto de replicaciones (**w_rep[1-9]+**) que se encuentran en la base, y la opción **mse=T**. Estas indicaciones preparan a la herramienta para obtener las estimaciones y las 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)
```

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 incluye también la función que permite la estimación del CV correspondiente a la estimación en cuestión:

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

³³ Se sugiere al usuario leer el *Manual de uso de la base de datos usuario de la 4° ENFR*, así como los diccionarios vinculados a cada base. Disponible en: https://www.indec.gob.ar/ftp/cuadros/menusuperior/enfr/manual_base_usuario_enfr2018.pdf

³⁴ Etiqueta que hace referencia a cualquiera de los siguientes factores de expansión: $w_{ijkl}^{P(1)}$, $w_{ijkl}^{P(2)}$, $w_{ijkl}^{P(3)}$ y $w_{ijkl}^{H(1)}$ definidos en los apartados 5, 6, 7 y 8, respectivamente.

³⁵ Etiqueta que hace referencia a alguno de los 4 conjuntos con las réplicas que se empleará para la estimación del error muestral.

³⁶ <https://cran.r-project.org/web/packages/survey/index.html>. Versión 3.36.

³⁷ <https://cran.r-project.org/web/packages/survey/survey.pdf>.

³⁸ El usuario puede optar por cualquier otro nombre para el objeto.

	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))

12.2 Cálculo del error de muestreo a través de Stata

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

Asumiendo que el usuario incorporó a **base_encuesta** en el entorno de Stata, el comando **svyset** es el que se emplea para gestionar los cálculos para las estimaciones. En él se deben identificar el factor de expansión de la encuesta **w**, los pesos replicados **w_rep***, y el método para el cálculo de la varianza **bootstrap**. Asimismo, se debe incluir la opción **mse** para obtener el estimador de varianza **bootstrap** considerado en la sección 11. Para preparar a 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 y de los errores de muestreo asociados. A manera de ejemplo, se presentan los códigos correspondientes para la estimación de un total, una media, proporciones y una razón:

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

En respuesta a la primera línea del código, y para cada caso, la herramienta brinda el resultado de la estimación del parámetro, la estimación de su error estándar a través del método **bootstrap**, y los límites para el 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, CV o IC, al no contar en esa versión con la opción **bootstrap**. 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).

12.3 Cálculo del error de muestreo a través de SAS

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

Habiendo incorporado **base_encuesta** al entorno SAS, la opción a emplear en cualquiera de los procedimientos para alcanzar los errores muestrales es **varmethod=Bootstrap**, invocando los pesos replicados **w_rep1--w_rep200** vía **repweight** y al factor de expansión para las estimaciones **w** en **weight**. En particular, para la estimación de los parámetros señalados se presentan los siguientes códigos orientativos:

Estimador	Estimaciones por SAS
\hat{t}_y	<pre>proc surveymeans data=base_encuesta sum cvsum varmethod=Bootstrap; repweight w_rep1--w_rep200; weight w; var Y; run;</pre>
\hat{y}	<pre>proc surveymeans data=base_encuesta mean cv varmethod= Bootstrap; repweight w_rep1--w_rep200; weight w; var Y; run;</pre>
\hat{p}	<pre>proc surveyfreq data= base_encuesta varmethod=Bootstrap; repweight w_rep1--w_rep200; weight w; table Y; run;</pre>
\hat{R}_{YZ}	<pre>proc surveymeans data=base_encuesta varmethod=Bootstrap; repweight w_rep1--w_rep200; weight w; ratio Y/Z; run;</pre>

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).

12.4 Cálculo del error de muestreo a través de Wesvar

Wesvar³⁹ es una herramienta estadística con una opción de descarga libre al igual que R. Fue desarrollada por la empresa Westat y permite emplear la metodología de cálculo de errores por muestra con base en replicaciones (Brick, Morganstein y Valliant, 2000). A continuación, se brinda una descripción sencilla de cómo operar con ella y de las opciones básicas que hay que invocar, empleando la versión 5.1.19.

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

³⁹ www.westat.com/capability/information-systems-software/wesvar. Se puede acceder de forma gratuita a la documentación de Wesvar enviando un correo electrónico a: wesvar_tech_support@westat.com.

Wesvar (.var) a partir de la base de la encuesta (**base_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 en la carpeta o espacio de trabajo donde se encuentra.⁴⁰

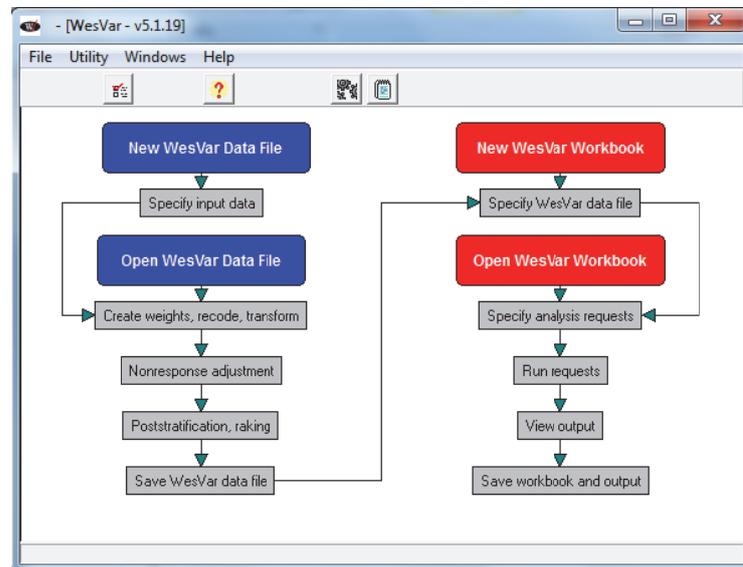


Figura 1

Al usuario le aparece una ventana como la que se ve en la figura 2, donde debe completar la información necesaria para iniciar a operar con las estimaciones. En el apartado **Variables** se deben indicar aquellas del panel **Source Variables** para las cuales se requieren estimaciones de parámetros. En **Replicates** se deben incluir las variables correspondientes a los pesos replicados de las muestras *bootstrap* de la encuesta, **w_rep1,...,w_rep200**; 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.

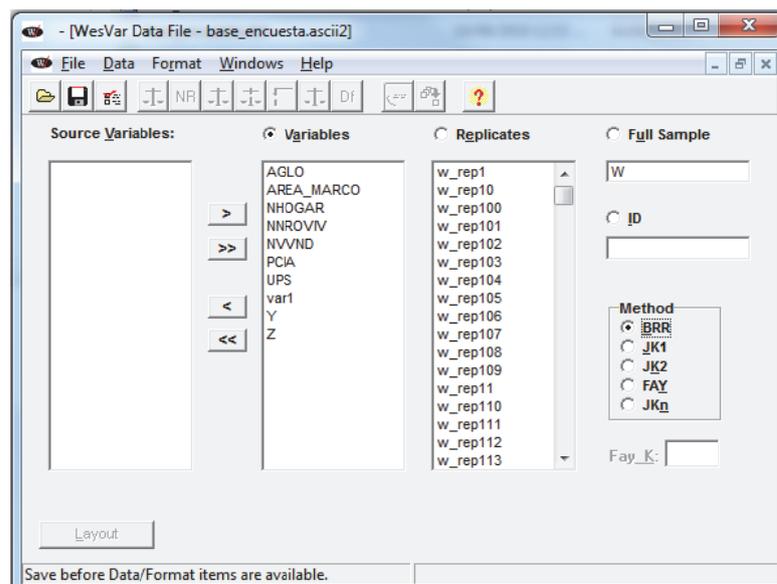


Figura 2

⁴⁰ Se advierte que la herramienta tiene la posibilidad de importar datos en formato csv/txt con delimitadores, SAS o SPSS.

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.

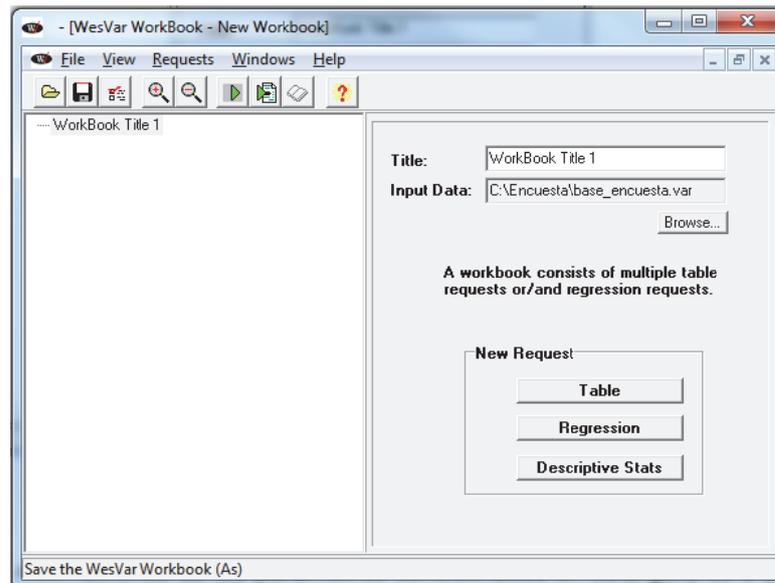


Figura 3

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

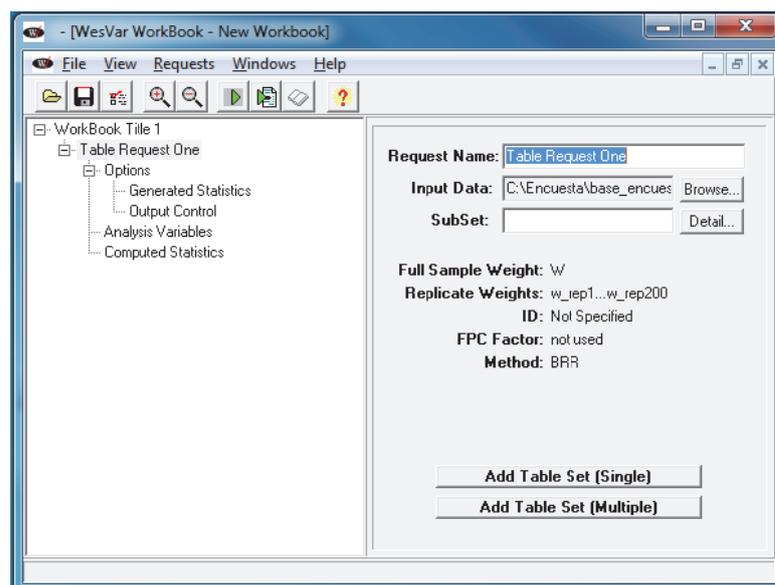


Figura 4

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

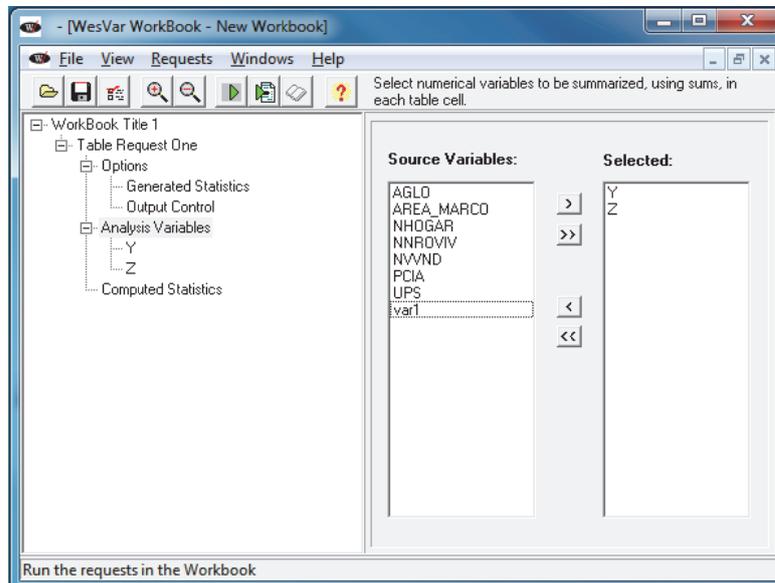


Figura 5

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

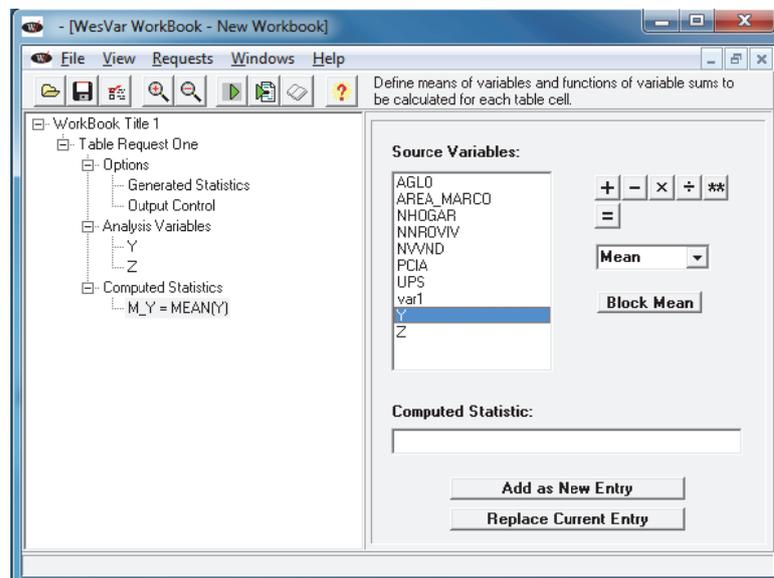


Figura 6

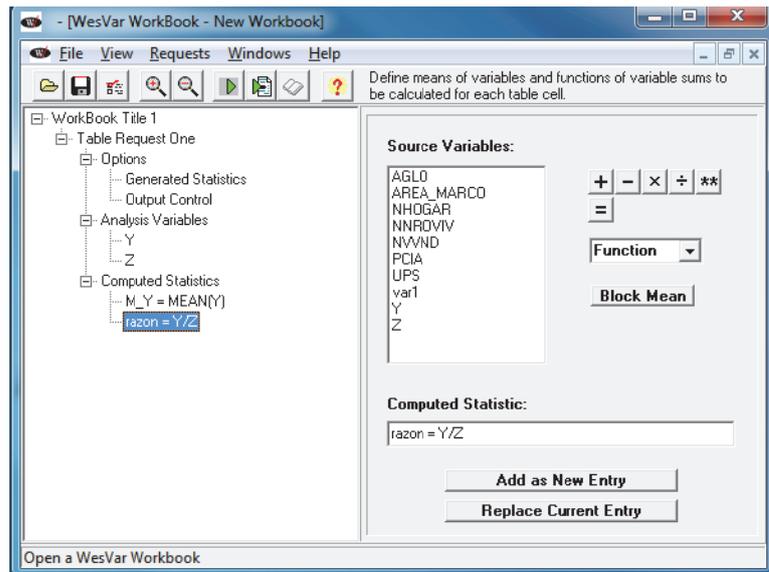


Figura 7

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

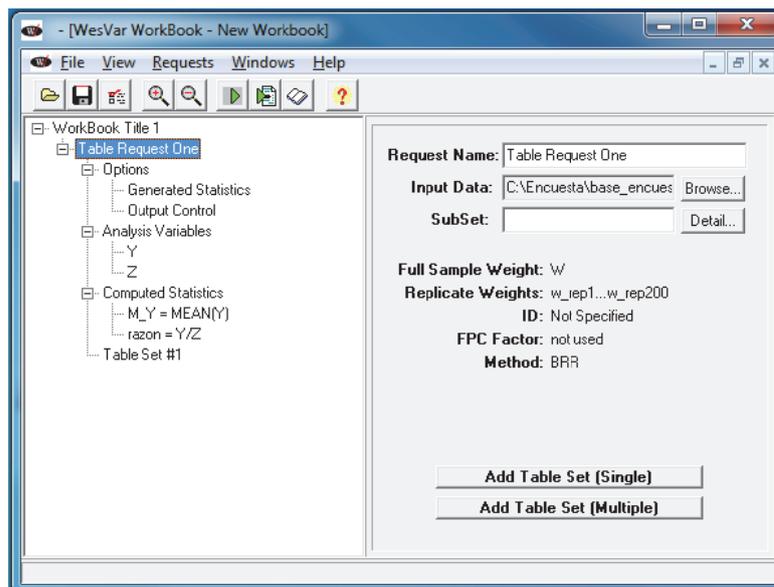


Figura 8

Haciendo clic 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.

Overall					
STATISTIC	EST_TYPE	ESTIMATE	STDERROR	CV(%)	CELL_n
SUM_WTS	VALUE	12426662.00	594707.960	4.786	26115
Y	VALUE	12426662.00	594707.960	4.786	26115
Z	VALUE	24853324.00	1189415.920	4.786	26115
M_Y	VALUE	1.00	0.000	0.000	26115
razon	VALUE	0.50	0.000	0.000	26115

Figura 9

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

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

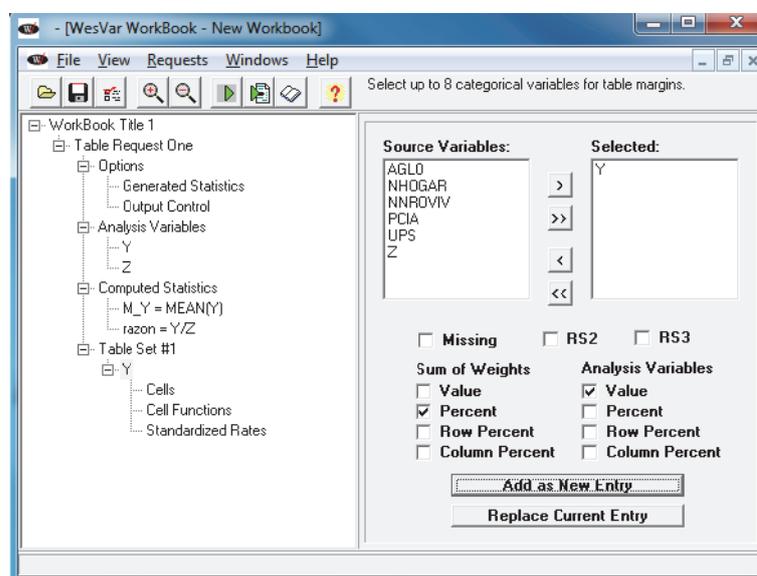


Figura 10

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

12.5 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 11 empleando las fórmulas [1] a [3].

Por ejemplo, si se asume que la variable Y está medida sobre las personas del Paso 1 de la encuesta, la expresión que se debe emplear como estimador para un total t_y , según se lo definió en la sección 9, es:

$$\hat{t}_y = \sum_R w_{ijkl}^{P(1)} y_{ijkl}$$

Si siguiendo lo señalado en la sección 11, la formulación para la varianza *bootstrap* [1] de un estimador es:

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

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

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

para el error estándar, según

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

y para del coeficiente de variación con

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

De manera análoga se procede para los casos de un promedio, una proporción, o un cociente o razón, reemplazando en [1] a $\hat{\theta}$ por \hat{y} , \hat{p}_A , o \hat{R}_{yz} , respectivamente (ver sección 9) y a las estimaciones *bootstrap* $\hat{\theta}_{(b)}^*$ que emplean a las réplicas por:

$$\hat{y}_{(b)}^* = \frac{\sum w_{ijkl}^{*P(1,b)} y_{ijkl}}{\sum w_{ijkl}^{*P(1,b)}},$$

$$\hat{p}_{A(b)}^* = \frac{\sum w_{ijkl}^{*P(1,b)} y_{ijkl}}{\sum w_{ijkl}^{*P(1,b)}},$$

o,

$$\hat{R}_{yz(b)}^* = \frac{\sum w_{ijkl}^{*P(1,b)} y_{ijkl}}{\sum w_{ijkl}^{*P(1,b)} z_{ijkl}}$$

según sea el caso, para obtener las respectivas varianzas estimadas por *bootstrap*, como también ee_B y cv_B , para cualquiera de las estimaciones del Paso 1 en cuestión. El mismo procedimiento, con el conjunto de réplicas adecuado, se puede emplear para obtener las varianzas estimadas para resultados que surgen de los otros pasos de la encuesta y para las estimaciones que se originen a partir de características de los hogares.

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

No es posible asumir en todos los resultados de la encuesta la misma confianza. Incluso en algunas situaciones no es aconsejable tomarlos como válidos para hacer inferencia estadística. Distintos motivos pueden afectar las estimaciones y, en consecuencia, la inferencia que se haga a partir de ellas. Por ejemplo, las estimaciones pueden no representar a la población objetivo de interés, cuando:

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

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

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

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

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

13.1 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 deben tener presentes a la hora de sacar sus conclusiones del fenómeno que están estudiando a partir de la encuesta.

Cuadro 2. Recomendaciones para interpretar las estimaciones

Calidad de la estimación	Condición	Recomendaciones
No confiable	Si se cumple alguna de las siguientes: a) El total de unidades involucradas en el cálculo de la estimación es menor a 100. b) La estimación de una razón es menor a 0,05. c) La estimación de una proporción es menor al 5%. d) El denominador de un cociente, razón, o proporción, tiene un CV > 10%. e) La estimación posee un CV > 33,3%.	Se recomienda no emplear a la estimación en este caso. Si existe la necesidad de publicarla, se debe advertir que las conclusiones basadas en ella no son confiables o válidas.
Poco confiable	La estimación posee un CV en el rango: $16,6\% < CV \leq 33,3\%$	La estimación debe ser considerada con precaución. Hay una alta probabilidad de que la inferencia resultante presente un nivel de error elevado. Se recomienda presentarla con alguna notación en la que se advierta de esta situación.
Confiable	La estimación posee un CV en el rango: $CV \leq 16,6\%$	La estimación puede ser considerada sin restricciones. No se requiere una notación especial.

Fuente: INDEC, 4° Encuesta Nacional de Factores de Riesgo.

Se insiste con la recomendación de que, en el caso que algunas de las estimaciones sean consideradas no confiables o poco confiables para inferir al total de la población o en subpoblaciones y el usuario desee incorporarlas en una publicación, se incluya una advertencia y se haga una referencia a las limitaciones del caso citando la presente guía metodológica, en particular el cuadro 2, definido por el Instituto como estándar para la encuesta.

13.2 Recomendaciones para estimaciones en dominios

Otro aspecto importante a tener en cuenta por los usuarios de la base de datos de la encuesta es la manera en que se calculan las estimaciones en subpoblaciones. 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 subestimarlos y en algunas circunstancias de manera grosera.

La herramienta que se emplee para la estimación del error de muestreo debe hacer uso de todas las observaciones de la muestra, para obtener una medida confiable y no estar subestimándola. Por lo general, la documentación que acompaña la herramienta contempla esta advertencia. En particular, en aquellas presentadas en los apartados 12.1 a 12.3, los usuarios que deseen calcular estimaciones en subpoblaciones o dominios, pueden recurrir a las opciones **subset**⁴¹ en *R*, **subpop** en *Stata*, y **domain** en *SAS* para obtener en forma adecuada la estimación del CV o del EE que esté calculando.⁴²

13.3 Recomendaciones sobre el cálculo de intervalos de confianza

Los intervalos de confianza (IC) brindan otro camino para evaluar la variabilidad inherente en las estimaciones provenientes de una muestra probabilística. Un intervalo de confianza es un rango de valores que tiene una probabilidad, conocida como “nivel de confianza”, de contener el valor poblacional del parámetro. En otras palabras, un intervalo de confianza al 0,95 significa que si 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 y cuenten con la estimación de su varianza o de su error estándar, un IC con un nivel de confianza del 95% se lo puede calcular en forma aproximada de la siguiente manera:

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

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

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

En la determinación de un IC juegan roles importantes la distribución probabilística del estimador y las propiedades asintóticas del estimador empleado para la varianza. A diferencia del EE y el CV, el IC obliga a adoptar algunos supuestos sobre el estimador $\hat{\theta}$ 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.

⁴¹ En el paquete Survey es posible utilizar también el comando **svyby** para obtener estimaciones en subpoblaciones.

⁴² En Wesvar no es necesario emplear una opción para advertirle que se van a realizar estimaciones en dominios o subpoblaciones; al crear una tabla donde se involucre a una variable que defina a la subpoblación (dominio) la herramienta procede correctamente al efectuar los cálculos del error por muestra.

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 e idénticamente distribuidos, requerida en este contexto para sostener el supuesto de normalidad (Heeringa, West, Berglung, 2017).

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

Referencias

- American Association for Public Opinion Research (2016). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. Recuperado de https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf.
- Brick, M., Morganstein, D. y Valliant, R. (2000). Analysis of Complex Sample Data Using Replication. Recuperado de https://www.researchgate.net/profile/David_Morganstein/publication/252297575_Analysis_of_Complex_Sample_Data_Using_Replication/links/55562a2e08ae6fd2d8235fbf/Analysis-of-Complex-Sample-Data-Using-Replication.pdf.
- Carlson, B. L. (2013). Response Rates Revisited. Proceedings American Statistical Associations. *Survey Research Methods Section - JSM*, 1200-1208. Recuperado de http://www.asarms.org/Proceedings/y2013/files/308173_80404.pdf.
- Chowhan, J. y Buckley, N. (2005). Using Mean Bootstrap Weights in Stata: A BSWREG Revision. *The Research Data Centres Information and Technical Bulletin*, 2(1), 23-37. Recuperado de <https://www150.statcan.gc.ca/n1/en/pub/12-002-x/12-002-x2005001-eng.pdf?st=LJqB8hAc>.
- Deville, J. y Särndal C. E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87. Recuperado de [DOI:10.1080/01621459.1992.10475217](https://doi.org/10.1080/01621459.1992.10475217).
- Frankel, L. R. (1983). The Report of the CASRO Task Force on Response Rates. En Frederick Wiseman (Ed.), *Improving Data Quality in a Sample Survey*. Cambridge: Marketing Science Institute.
- Gagné, C., Roberts, G. y Keown, L. (2014). Weighted Estimation and Bootstrap Variance Estimation for Analyzing Survey Data: How to Implement in Selected Software. *The Research Data Centres Information and Technical Bulletin*, 6(1). Recuperado de <https://www150.statcan.gc.ca/n1/pub/12-002-x/2014001/article/11901-eng.htm>.
- Haziza, D. y Beaumont, J. F. (2017). Construction of Weights in Surveys: A Review. *Statistical Science*, 32(2), 206-226. Recuperado de [DOI:10.1214/16-STS608](https://doi.org/10.1214/16-STS608).
- Heeringa, S., West, B. y Berglund, P. (2017). *Applied Survey Data Analysis*. Nueva York: Chapman & Hall/CRC. Recuperado de [DOI:10.1201/9781315153278](https://doi.org/10.1201/9781315153278).
- Lemaître, G. y Dufour J. (1987). An Integrated Method for Weighting Persons and Families. *Survey Methodology*, 13(2), 199-207. Recuperado de <https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X198700214607>.
- Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R*. Nueva Jersey: J. Wiley & Sons. Recuperado de [DOI:10.1002/9780470580066](https://doi.org/10.1002/9780470580066).
- Rao, J. N. K. y Wu, C. F. J. (1988). Resampling Inference with Complex Surveys Data. *Journal of the American Statistical Association*, 83, 231-241. Recuperado de [DOI: 10.1080/01621459.1988.10478591](https://doi.org/10.1080/01621459.1988.10478591).
- Rao, J. N. K., Wu, C. F. J. y Yue, K. (1992). Some Recent Work on Resampling Methods for Complex Surveys. *Survey Methodology*, 18, 209-217. Recuperado de <https://www150.statcan.gc.ca/n1/pub/12-001-x/1992002/article/14486-eng.pdf>.

- Phillips, O. (2004). Using Bootstrap Weights with WesVar and SUDAAN. *Research Data Centres, Information and Technical Bulletin*, 1(2), 6-15. Recuperado de <https://www150.statcan.gc.ca/n1/en/pub/12-002-x/12-002-x2004002-eng.pdf?st=JeakLQDY>
- Sarndall, C., Swensson, B. y Wretman, J. (1992). *Model Assisted Survey Sampling*. Nueva York: Springer-Verlag Publishing.
- SAS Institute Inc. (2017). *SAS/STAT® 14.3 User's Guide*. Cary: SAS Institute Inc.
- StataCorp (2017). *Stata Survey Data Reference: Release 15*. College Station, Texas: StataCorp LLC.
- Valliant, R., Dever, J. A. y Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. Nueva York: Springer. Recuperado de [DOI: 10.1007/978-1-4614-6449-5_14](https://doi.org/10.1007/978-1-4614-6449-5_14).
- West, B. (2012). Accounting for Multi-stage Sample Designs in Complex Sample Variance Estimation. *Survey Methodology*. Recuperado de http://www.isr.umich.edu/src/smp/asda/first_stage_ve_new.pdf.
- Wolter, K. M. (2007). *Introduction to Variance Estimation*. Nueva York: Springer-Verlag. Recuperado de [DOI: 10.1007/978-0-387-35099-8](https://doi.org/10.1007/978-0-387-35099-8).

Anexo I.A Total de UPM y USM de la MMUVRA presentes en la ENFR

Regiones	UPM	USM	
		Paso 1	Paso 2
Gran Buenos Aires	2	597	447
Noroeste	59	781	585
Noreste	59	669	502
Cuyo	25	412	309
Pampeana	98	1.626	1.221
Patagonia	45	674	505
Total del país	288	4.759	3.569

Fuente: INDEC, 4° Encuesta Nacional de Factores de Riesgo.

Dominio	UPM	USM
	Paso 3	Paso 3
Localidades 150.000 o más habitantes	26	1.621

Fuente: INDEC, 4° Encuesta Nacional de Factores de Riesgo.

Anexo I.B Localidades de 5.000 habitantes y más de la MMUVRA involucradas en la ENFR

Provincia	Localidades
	Ciudad Autónoma de Buenos Aires
Buenos Aires	Gran Bs. As. Arrecifes, Ayacucho, Bahía Blanca, Baradero, Berisso, Campana, Carlos Casares, Carmen de Patagones, Chacabuco, Chivilcoy, Coronel Pringles, Dolores, Ensenada, Escobar, General Daniel Cerri, General Rodríguez, José C. Paz, Junín, La Plata, Lincoln, Luján, Malvinas Argentinas, Mar del Plata, Marcos Paz, Máximo Paz, Mercedes, Merlo, Moreno, Morón, Necochea-Quequén, Olavarría, Pehuajó, Pergamino, Pilar, Presidente Perón, Punta Alta, Quilmes, Ramallo, Ruta Sol, Salto, San Antonio de Areco, San Fernando, San Nicolás de los Arroyos, Tandil, Tigre, Tornquist, Trenque Lauquen, Tres Arroyos, Zárate.
Catamarca	Andalgalá, Belén, Recreo, San Fernando del Valle de Catamarca, San Isidro, Santa María, Tinogasta
Córdoba	Bell Ville, Biale Massé, Córdoba, Cosquín, Cruz del Eje, Dean Funes, Estancia Vieja, Hernando, Huerta Grande, La Calera, La Carlota, La Falda, Laboulaye, Las Chacras, Las Higueras, Las Tapias, Malvinas Argentinas, Mendiolaza, Oncativo, Parque Norte-Ciudad de los Niños-Villa Pastora-Almirante Brown, Pilar, Pozo del Molle, Río Cuarto, Río Primero, Río Segundo, San Antonio de Arredondo, San Francisco, San Pedro, San Roque, Tanti, Valle Hermoso, Villa Allende, Villa Carlos Paz, Villa de las Rosas, Villa del Dique, Villa Dolores, Villa María, Villa Nueva, Villa Río Icho Cruz, Villa Rumipal, Villa Sarmiento
Corrientes	Bella Vista, Corrientes, Curuzú Cuatiá, Esquina, Gobernador Igr. Valentín Virasoro, Goya, Ituzaingó, Mercedes, Monte Caseros, Paso de los Libres, Saladas, San Luis del Palmar, San Roque, Santa Lucía, Santa Rosa, Santo Tomé
Chaco	Barranqueras, Campo Largo, Charata, Concepción del Bermejo, Coronel Du Graty, Fontana, General José de San Martín, José Castelli, La Leonesa, Las Breñas, Las Palmas, Machagai, Pampa del Indio, Presidencia de la Plaza, Presidencia Roque Sáenz Peña, Puerto Vilelas, Quitilipi, Resistencia, Tres Isletas, Villa Angela
Chubut	Comodoro Rivadavia, Esquel, Playa Unión, Puerto Madryn, Rada Tilly, Rawson, Sarmiento, Trelew, Trevelín
Entre Ríos	Basavilbaso, Chajarí, Colón, Colonia Avellaneda, Concepción del Uruguay, Concordia, Crespo, Diamante, Federación, General Ramírez, Gualeguay, Gualeguaychú, La Paz, Nogoyá, Paraná, San José, Santa Elena, Viale, Victoria, Villaguay
Formosa	Clorinda, Comandante Fontana, El Colorado, Formosa, Ibarreta, Ingeniero Guillermo N. Juárez, Laguna Blanca, Las Lomitas, Palo Santo, Pirané, Villa General Manuel Belgrano, Villa Kilómetro 213
Jujuy	Abra Pampa, Caimancito, El Carmen, La Quiaca, Libertador General San Martín, Palpalá, Perico, San Pedro, San Salvador de Jujuy, Yuto
La Pampa	25 de Mayo, Eduardo Castex, General Acha, General Pico, Intendente Alvear, Realicó, Santa Rosa, Toay, Victorica
La Rioja	Aimogasta, Chamental, Chepes, Chilecito, La Rioja, Nonogasta
Mendoza	General Alvear, Godoy Cruz, Guaymallén, La Paz, Las Heras, Luján de Cuyo, Maipú, Malargüe, Mendoza, Perdriel, Rivadavia, Rodeo del Medio, San Martín, San Rafael, Tunuyán
Misiones	25 de Mayo, Apóstoles, Aristóbulo del Valle, Concepción de la Sierra, Eldorado, Garupá, Jardín América, Leandro N. Alem, Montecarlo, Oberá, Posadas, Posadas (Expansión), Puerto Esperanza, Puerto Iguazú, Puerto Rico, San Pedro, San Vicente

Provincia	Localidades
Neuquén	Centenario, Chos Malal, Cutral Có, Junín de los Andes, Neuquén, Plaza Huinca, Plottier, Rincón de los Sauces, San Martín de los Andes, San Patricio del Chañar, Senillosa, Villa La Angostura, Zapala
Río Negro	Allen, Catriel, Cinco Saltos, Cipolletti, El Bolsón, General Conesa, General Roca, Ingeniero Luis A. Huergo, Lamarque, Luis Beltrán, Río Colorado, San Antonio Oeste, San Carlos de Bariloche, Sierra Grande, Viedma, Villa Regina
Salta	Aguaray, Apolinario Saravia, Campo Santo, Cerrillos, Colonia Santa Rosa, Embarcación, General Güemes, General Mosconi, La Merced, Las Lajitas, Misión El Cruce-El Milagro-El Jardín de San Martín, Pichanal, Profesor Salvador Mazza, Rosario de la Frontera, Rosario de Lerma, Salta, San José de Metán, San Ramón de la Nueva Orán, Tartagal, Vaqueros
San Juan	Caucete, Chimbab, Rawson, Rivadavia, San José de Jáchal, San Juan, Santa Lucía, Villa Aberastain-La Rinconada, Villa Barboza-Villa Nacusi, Villa Borjas-La Chimbera, Villa General San Martín-Campo Afuera, Villa Media Agua
San Luis	Concarán, Juana Koslay, Justo Daract, La Punta, La Toma, Merlo, Quines, San Luis, Santa Rosa del Conlara, Tilisarao, Villa Mercedes
Santa Cruz	28 de Noviembre, Caleta Olivia, Comandante Luis Piedrabuena, El Calafate, Las Heras, Pico Truncado, Puerto Deseado, Puerto San Julián, Río Gallegos, Yacimientos Río Turbio
Santa Fe	Arequito, Arroyo Seco, Avellaneda, Barrio Arroyo del Medio, Cañada de Gómez, Casilda, Coronda, El Trébol, Esperanza, Florencia, Fray Luis Beltrán, Funes, Granadero Baigorria, Pérez, Puerto General San Martín, Rafaela, Reconquista, Roldán, Romang, Rosario, San Jorge, San José del Rincón, San Lorenzo, Santa Fe, Santo Tomé, Sastre, Sauce Viejo, Teodelina, Venado Tuerto, Vera, Villa Constitución, Villa Gobernador Gálvez
Santiago del Estero	Añatuya, El Zanjón, Frías, La Banda, La Dársena, Monte Quemado, Quimilí, Santiago del Estero, Sumampa, Suncho Corral, Termas de Río Hondo, Villa Ojo de Agua, Villa San Martín (Est. Loreto)
Tucumán	Aguilares, Alderetes, Banda del Río Salí, Colombres, Concepción, Delfín Gallo, Diagonal Norte-Luz y Fuerza-Los Pocitos-Villa Nueva Italia, El Manantial, Famaillá, Ingenio San Pablo, La Florida, La Trinidad, Los Ralos, Lules, Medina, Monteros, Río Seco, San Miguel de Tucumán, Tafí Viejo, Villa Mariano Moreno-El Colmenar, Villa Quinteros, Yerba Buena-Marcos Paz
Tierra del Fuego	Río Grande, Ushuaia

Fuente: INDEC, 4° Encuesta Nacional de Factores de Riesgo.

Anexo II. Dominio de estimación del Paso 3 - Aglomerados de 150.000 habitantes y más

Aglomerados	Localidades que componen el aglomerado
Ciudad Autónoma de Buenos Aires	
Gran Buenos Aires	
Gran La Plata	Berisso, Ensenada, La Plata, Ruta Sol
Bahía Blanca-Cerri	Bahía Blanca, General Daniel Cerri
Mar del Plata	Mar del Plata
Gran Catamarca	San Fernando del Valle de Catamarca, San Isidro
Gran Córdoba	Córdoba, Mendiolaza, Parque Norte-Ciudad de los Niños-Villa Pastora-Almirante Brown, Villa Allende
Río Cuarto	Las Higueras, Río Cuarto
Corrientes	Corrientes
Gran Resistencia	Barranqueras, Fontana, Puerto Vilelas, Resistencia
Comodoro Rivadavia-Rada Tilly	Comodoro Rivadavia, Rada Tilly
Gran Paraná	Paraná
Concordia	Concordia
Formosa	Formosa
Jujuy-Palpalá	Palpalá, San Salvador de Jujuy
La Rioja	La Rioja
Gran Mendoza	Godoy Cruz, Guaymallén, Las Heras, Luján de Cuyo, Maipú, Mendoza
Posadas	Garupá, Posadas
Neuquén-Plottier	Neuquén, Plottier
Salta	Salta
Gran San Juan	Chimbas, Rawson, Rivadavia, San Juan, Santa Lucía, Villa Barboza-Villa Nacusi, Villa General San Martín-Campo Afuera
Gran San Luis	La Punta, San Luis
Gran Rosario	Fray Luis Beltrán, Funes, Pérez, Puerto General San Martín, Rosario, San Lorenzo, Villa Gobernador Gálvez
Gran Santa Fe	San José del Rincón, Santa Fe, Santo Tomé, Sauce Viejo
Santiago del Estero-La Banda	El Zanjón, La Banda, La Dársena, Santiago del Estero
Gran Tucumán-Tafí Viejo	Alderetes, Banda del Río Salí, Diagonal Norte-Luz y Fuerza-Los Pocitos-Villa Nueva Italia, El Manantial, San Miguel de Tucumán, Tafí Viejo, Yerba Buena-Marcos Paz

Fuente: INDEC, 4° Encuesta Nacional de Factores de Riesgo.

Anexo III. Distribución del total de la muestra de viviendas seleccionadas por jurisdicción y paso

Jurisdicción	Viviendas		
	Paso 1	Paso 2	Paso 3
CABA	3.160	2.360	2.360
GBA	4.390	3.290	3.290
Buenos Aires	6.660	5.000	1.640
Catamarca	1.100	820	380
Córdoba	3.290	2.470	1.100
Corrientes	1.810	1.360	430
Chaco	1.830	1.370	470
Chubut	1.610	1.210	390
Entre Ríos	2.340	1.760	940
Formosa	1.320	990	380
Jujuy	1.340	1.000	410
La Pampa	1.090	820	-
La Rioja	1.150	860	460
Mendoza	1.730	1.300	570
Misiones	1.730	1.300	420
Neuquén	1.430	1.070	410
Río Negro	1.990	1.420	-
Salta	1.760	1.320	600
San Juan	950	710	490
San Luis	1.440	1.080	420
Santa Cruz	1.150	860	-
Santa Fe	2.880	2.160	1.200
Santiago del Estero	1.060	800	460
Tucumán	1.400	1.050	570
Tierra del Fuego	560	420	-
Total	49.170	36.870	17.390

Fuente: INDEC, 4° Encuesta Nacional de Factores de Riesgo.

Anexo IV.A Total de viviendas elegibles, no elegibles y de elegibilidad dudosa por jurisdicción - Paso 1

Jurisdicción	Viviendas			
	en la muestra	elegibles	no elegibles	de elegibilidad dudosa
CABA	3.160	2.610	447	103
GBA	4.390	3.773	534	83
Buenos Aires	6.660	5.570	998	92
Catamarca	1.100	980	92	28
Córdoba	3.290	2.774	209	307
Corrientes	1.810	1.546	256	8
Chaco	1.830	1.603	221	6
Chubut	1.610	1.447	163	0
Entre Ríos	2.340	2.093	236	11
Formosa	1.320	1.116	201	3
Jujuy	1.340	1.180	151	9
La Pampa	1.090	962	128	0
La Rioja	1.150	960	178	12
Mendoza	1.730	1.495	213	22
Misiones	1.730	1.533	197	0
Neuquén	1.430	1.258	163	9
Río Negro	1.990	1.771	214	5
Salta	1.760	1.556	199	5
San Juan	950	838	112	0
San Luis	1.440	1.245	192	3
Santa Cruz	1.150	969	176	5
Santa Fe	2.880	2.541	327	12
Santiago del	1.060	888	169	3
Tucumán	1.400	1.255	145	0
Tierra del Fuego	560	492	68	0
Total	49.170	42.455	5.989	726

Fuente: INDEC, 4° Encuesta Nacional de Factores de Riesgo.

Anexo IV.B Causas de no elegibilidad o dudosa de las viviendas por jurisdicción - Paso 1

Jurisdicción	Vivienda					Local	Dirección no existente	En área insegura	Razón indeterminada
	deshabitada	demolida	de fin de semana	en construcción	usada como establecimiento				
CABA	258	2	28	18	92	6	43	100	3
GBA	239	16	38	40	53	24	124	83	0
Buenos Aires	455	33	107	53	89	65	196	58	34
Catamarca	66	3	2	1	9	3	8	0	28
Córdoba	75	4	36	16	16	5	57	7	300
Corrientes	128	5	16	19	31	17	40	8	0
Chaco	91	12	7	14	17	28	52	2	4
Chubut	76	5	16	17	9	11	29	0	0
Entre Ríos	128	4	16	16	30	18	24	11	0
Formosa	113	7	18	18	12	22	11	0	3
Jujuy	77	4	21	17	14	6	12	1	8
La Pampa	64	3	7	15	17	8	14	0	0
La Rioja	85	3	23	9	24	3	31	1	11
Mendoza	79	3	16	23	29	8	55	22	0
Misiones	109	22	11	9	23	12	11	0	0
Neuquén	63	5	5	21	37	3	29	1	8
Río Negro	97	4	19	16	37	10	31	0	5
Salta	95	10	16	11	26	18	23	1	4
San Juan	65	6	5	7	10	6	13	0	0
San Luis	99	7	17	7	12	10	40	2	1
Santa Cruz	73	2	6	15	31	13	36	1	4
Santa Fe	167	8	34	27	53	17	21	12	0
Santiago del Estero	69	6	15	8	20	10	41	0	3
Tucumán	85	8	8	7	12	15	10	0	0
Tierra del Fuego	31	7	4	7	10	1	8	0	0
Total	2.887	189	491	411	713	339	959	310	416

Fuente: INDEC, 4° Encuesta Nacional de Factores de Riesgo.

Anexo IV.C Total de hogares con y sin respuesta por jurisdicción - Paso 1

Jurisdicción	Hogares		
	elegibles	con respuesta	sin respuesta
CABA	2.638	1.534	1.104
GBA	3.816	2.554	1.262
Buenos Aires	5.590	3.582	2.008
Catamarca	991	890	101
Córdoba	2.806	2.053	753
Corrientes	1.556	1.173	383
Chaco	1.610	1.106	504
Chubut	1.448	988	460
Entre Ríos	2.098	1.688	410
Formosa	1.120	1.028	92
Jujuy	1.227	1.062	165
La Pampa	962	655	307
La Rioja	962	803	159
Mendoza	1.504	1.108	396
Misiones	1.538	1.238	300
Neuquén	1.261	825	436
Río Negro	1.771	1.469	302
Salta	1.602	1.374	228
San Juan	842	718	124
San Luis	1.248	1.001	247
Santa Cruz	969	606	363
Santa Fe	2.578	1.919	659
Santiago del Estero	890	686	204
Tucumán	1.267	1.020	247
Tierra del Fuego	492	346	146
Total	42.786	31.426	11.360

Fuente: INDEC, 4° Encuesta Nacional de Factores de Riesgo.

Anexo IV.D Total de personas seleccionadas con y sin respuesta por jurisdicción - Paso 1

Jurisdicción	Personas		
	Seleccionadas	Con respuesta	Sin respuesta
CABA	1.534	1.372	162
GBA	2.554	2.193	361
Buenos Aires	3.582	3.379	203
Catamarca	890	865	25
Córdoba	2.053	1.985	68
Corrientes	1.173	1.063	110
Chaco	1.106	1.021	85
Chubut	988	951	37
Entre Ríos	1.688	1.611	77
Formosa	1.028	1.004	24
Jujuy	1.062	991	71
La Pampa	655	621	34
La Rioja	803	738	65
Mendoza	1.108	983	125
Misiones	1.238	1.150	88
Neuquén	825	749	76
Río Negro	1.469	1.441	28
Salta	1.374	1.283	91
San Juan	718	654	64
San Luis	1.001	943	58
Santa Cruz	606	553	53
Santa Fe	1.919	1.773	146
Santiago del	686	632	54
Tucumán	1.020	948	72
Tierra del Fuego	346	321	25
Total	31.426	29.224	2.202

Fuente: INDEC, 4° Encuesta Nacional de Factores de Riesgo.

Anexo IV.E Total de personas por causa de no respuesta por jurisdicción - Paso 1

Jurisdicción	Personas sin respuesta	Causa de no respuesta			
		Ausencia	Rechazo	Otras causas	Razón indeterminada
CABA	162	81	44	28	9
GBA	361	195	90	44	32
Buenos Aires	203	101	63	26	13
Catamarca	25	11	8	4	2
Córdoba	68	27	33	2	6
Corrientes	110	55	35	18	2
Chaco	85	38	27	18	2
Chubut	37	20	12	3	2
Entre Ríos	77	39	22	12	4
Formosa	24	8	10	5	1
Jujuy	71	45	17	8	1
La Pampa	34	11	11	8	4
La Rioja	65	21	34	9	1
Mendoza	125	68	40	15	2
Misiones	88	40	29	18	1
Neuquén	76	41	26	8	1
Río Negro	28	10	12	5	1
Salta	91	50	27	13	1
San Juan	64	31	17	15	1
San Luis	58	19	19	12	8
Santa Cruz	53	22	20	9	2
Santa Fe	146	64	53	18	11
Santiago del	54	24	15	15	0
Tucumán	72	38	20	12	2
Tierra del Fuego	25	4	12	9	0
Total	2.202	1.063	696	334	109

Fuente: INDEC, 4° Encuesta Nacional de Factores de Riesgo.

Anexo V.A Total de viviendas elegibles o no y de elegibilidad dudosa por jurisdicción - Paso 2

Jurisdicción	Viviendas			
	en la muestra	elegibles	no elegibles	elegibilidad dudosa
CABA	2.360	1.932	327	101
GBA	3.290	2.809	435	46
Buenos Aires	5.000	4.153	761	86
Catamarca	820	735	69	16
Córdoba	2.470	2.081	151	238
Corrientes	1.360	1.167	185	8
Chaco	1.370	1.196	168	6
Chubut	1.210	1.083	127	0
Entre Ríos	1.760	1.567	192	1
Formosa	990	856	131	3
Jujuy	1.000	882	113	5
La Pampa	820	718	102	0
La Rioja	860	724	126	10
Mendoza	1.300	1.121	157	22
Misiones	1.300	1.147	153	0
Neuquén	1.070	932	129	9
Río Negro	1.420	1.327	159	4
Salta	1.320	1.164	153	3
San Juan	710	623	87	0
San Luis	1.080	924	154	2
Santa Cruz	860	720	135	5
Santa Fe	2.160	1.907	242	11
Santiago del	800	663	135	2
Tucumán	1.050	938	112	0
Tierra del Fuego	420	372	48	0
Total	36.870	31.741	4.551	578

Fuente: INDEC, 4° Encuesta Nacional de Factores de Riesgo.

Anexo V.B Total de hogares con y sin respuesta por jurisdicción - Paso 2

Jurisdicción	Hogares		
	elegibles	con respuesta	sin respuesta
CABA	1.956	1.124	832
GBA	2.838	1.906	932
Buenos Aires	4.170	2.704	1.466
Catamarca	744	675	69
Córdoba	2.105	1.559	546
Corrientes	1.176	896	280
Chaco	1.203	839	364
Chubut	1.084	712	372
Entre Ríos	1.571	1.270	301
Formosa	858	784	74
Jujuy	927	802	125
La Pampa	718	504	214
La Rioja	725	602	123
Mendoza	1.124	812	312
Misiones	1.152	924	228
Neuquén	934	610	324
Río Negro	1.327	1.091	236
Salta	1.201	1.025	176
San Juan	626	541	85
San Luis	925	740	185
Santa Cruz	720	468	252
Santa Fe	1.937	1.443	494
Santiago del Estero	664	495	169
Tucumán	949	769	180
Tierra del Fuego	372	261	111
Total	32.006	23.556	8.450

Fuente: INDEC, 4° Encuesta Nacional de Factores de Riesgo.

Anexo V.C Total de personas elegibles, con y sin respuesta por jurisdicción - Paso 2

Jurisdicción	Hogares con respuesta	Personas			
		no elegibles ⁽¹⁾	elegibles	que responden	que no responden
CABA	1.124	130	994	714	280
GBA	1.906	327	1579	1.239	340
Buenos Aires	2.704	196	2508	1.403	1.103
Catamarca	675	27	648	567	81
Córdoba	1.559	82	1477	1.127	350
Corrientes	896	91	805	699	106
Chaco	839	82	757	606	151
Chubut	712	35	677	562	115
Entre Ríos	1.270	78	1192	935	257
Formosa	784	35	749	649	100
Jujuy	802	65	737	640	97
La Pampa	504	37	467	355	112
La Rioja	602	62	540	424	116
Mendoza	812	119	693	522	171
Misiones	924	84	840	744	96
Neuquén	610	68	542	377	165
Río Negro	1.091	35	1056	828	229
Salta	1.025	87	938	725	213
San Juan	541	61	480	394	86
San Luis	740	53	687	478	209
Santa Cruz	468	55	413	340	73
Santa Fe	1.443	142	1301	1068	234
Santiago del	495	56	439	393	46
Tucumán	769	81	688	598	90
Tierra del Fuego	261	26	235	190	45
Total	23.556	2.114	21.442	16.577	4.865

(1) Personas que no respondieron al Paso 1 o mujeres que estuvieran embarazadas al momento de relevar el Paso 2.

Fuente: INDEC, 4° Encuesta Nacional de Factores de Riesgo.

Anexo V.D Total de personas que no responden por causa de no respuesta y jurisdicción - Paso 2

Jurisdicción	Personas que no responden	Causas de no respuesta		
		Rechazo	Ausencia personal de salud	Razón indeterminada
CABA	280	170	80	30
GBA	340	185	99	56
Buenos Aires	1.103	432	636	35
Catamarca	81	13	67	1
Córdoba	350	268	48	34
Corrientes	106	47	54	5
Chaco	151	42	107	2
Chubut	115	110	2	3
Entre Ríos	257	131	120	6
Formosa	100	5	95	0
Jujuy	97	24	73	0
La Pampa	112	50	62	0
La Rioja	116	61	20	35
Mendoza	171	121	50	0
Misiones	96	19	76	1
Neuquén	165	75	79	11
Río Negro	229	114	114	1
Salta	213	78	135	0
San Juan	86	75	10	1
San Luis	209	73	127	9
Santa Cruz	73	31	42	0
Santa Fe	234	158	69	7
Santiago del	46	15	30	1
Tucumán	90	42	48	0
Tierra del Fuego	45	30	7	8
Total	4.865	2.369	2.250	246

Fuente: INDEC, 4° Encuesta Nacional de Factores de Riesgo.

Anexo VI.A Total de personas con y sin respuesta por aglomerado - Paso 3

Aglomerado	Personas				
	en hogares con respuesta	no elegibles (1)	elegibles	con respuesta	sin respuesta
Ciudad Autónoma de Buenos Aires	1.124	210	914	486	428
Gran Buenos Aires	1.906	426	1.480	981	499
Gran La Plata	279	67	212	58	154
Bahía Blanca-Cerri	209	70	139	71	68
Mar del Plata	180	36	144	67	77
Gran Catamarca	302	29	273	239	34
Gran Córdoba	453	51	402	111	291
Río Cuarto	328	16	312	234	78
Corrientes	196	41	155	97	58
Gran Resistencia	255	97	158	51	107
Comodoro Rivadavia-Rada Tilly	190	12	178	130	48
Gran Paraná	315	85	230	117	113
Concordia	363	42	321	232	89
Formosa	303	13	290	276	14
Jujuy-Palpalá	330	28	302	203	99
La Rioja	289	49	240	148	92
Gran Mendoza	350	84	266	153	113
Posadas	265	57	208	146	62
Neuquén-Plottier	237	61	176	73	103
Salta	446	79	367	289	78
Gran San Juan	354	46	308	195	113
Gran San Luis	280	28	252	187	65
Gran Rosario	352	99	253	115	138
Gran Santa Fe	380	45	335	264	71
Santiago del Estero-La Banda	276	51	225	189	36
Gran Tucumán-Tafí Viejo	393	56	337	219	118
Total del dominio de aglomerados	10.355	1.878	8.477	5.331	3.146

(1) Personas que no respondieron al Paso 1, que estuvieran embarazadas al momento de relevar el Paso 3 o ausencia de personal de salud al momento de relevar el Paso 2.

Fuente: INDEC, 4° Encuesta Nacional de Factores de Riesgo.

Anexo VI.B Total de personas sin respuesta por causa de no respuesta y aglomerado - Paso 3

Aglomerado	Personas sin respuesta	Causas de no respuesta		
		Rechazo	Ausencia de personal de salud	Razón indeterminada
Ciudad Autónoma de Buenos	428	365	10	53
Gran Buenos Aires	499	396	10	93
Gran La Plata	154	106	9	39
Bahía Blanca-Cerri	68	19	45	4
Mar del Plata	77	72	2	3
Gran Catamarca	34	22	11	1
Gran Córdoba	291	250	9	32
Río Cuarto	78	72	0	6
Corrientes	58	52	0	6
Gran Resistencia	107	89	13	5
Comodoro Rivadavia-Rada Tilly	48	48	0	0
Gran Paraná	113	93	10	10
Concordia	89	74	9	6
Formosa	14	14	0	0
Jujuy-Palpalá	99	78	17	4
La Rioja	92	52	1	39
Gran Mendoza	113	112	1	0
Posadas	62	54	1	7
Neuquén-Plottier	103	90	2	11
Salta	78	76	2	0
Gran San Juan	113	111	2	0
Gran San Luis	65	56	7	2
Gran Rosario	138	129	9	0
Gran Santa Fe	71	60	0	11
Santiago del Estero-La Banda	36	36	0	0
Gran Tucumán-Tafí Viejo	118	109	8	1
Total del país	3.146	2.635	178	333

Fuente: INDEC, 4° Encuesta Nacional de Factores de Riesgo.

Anexo VII. Tasa de respuesta de los hogares y las personas

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

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

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

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

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

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

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

NE : cantidad de hogares o viviendas no elegibles,

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

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

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

Los siguientes cuadros presentan las tasas de respuesta con la versión $RR3$, y una cota superior o valor máximo estimado cuando se asume $e = 0$, $RR5 = \frac{R}{EL}$, para hogares; y la $RR3$ para la tasa de respuesta a nivel de personas por jurisdicción y para el total del país, en los 3 pasos de la encuesta.⁴³

Tasas de respuesta por jurisdicciones y total del país para el Paso 1

Jurisdicción	Respuesta de hogares		Respuesta de personas
	RR3	RR5	
CABA	55,0%	58,2%	89,4%
Partidos del GBA	65,7%	66,9%	85,9%
Buenos Aires	63,9%	64,1%	94,3%
Catamarca	87,5%	89,8%	97,2%
Córdoba	66,4%	73,2%	96,7%
Corrientes	75,1%	75,4%	90,6%
Chaco	68,5%	68,7%	92,3%
Chubut	68,2%	68,2%	96,3%
Entre Ríos	80,1%	80,5%	95,4%
Formosa	91,6%	91,8%	97,7%
Jujuy	86,0%	86,6%	93,3%
La Pampa	68,1%	68,1%	94,8%
La Rioja	82,6%	83,5%	91,9%
Mendoza	72,7%	73,7%	88,7%
Misiones	80,5%	80,5%	92,9%
Neuquén	65,0%	65,4%	90,8%
Río Negro	82,7%	82,9%	98,1%
Salta	85,5%	85,8%	93,4%
San Juan	85,3%	85,3%	91,1%
San Luis	80,0%	80,2%	94,2%
Santa Cruz	62,3%	62,5%	91,3%
Santa Fe	74,1%	74,4%	92,4%
Santiago del Estero	76,9%	77,1%	92,1%
Tucumán	80,5%	80,5%	92,9%
Tierra del Fuego	70,3%	70,3%	92,8%
Total del país	72,4%	73,4%	93,0%

Fuente: INDEC, 4° Encuesta Nacional de Factores de Riesgo.

⁴³ 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 cuanto representa en la población una unidad.

Tasas de respuesta por jurisdicciones y total del país para el Paso 2

Jurisdicción	Respuesta de hogares		Respuesta de personas
	RR3	RR5	
CABA	55,0%	57,5%	71,8%
Partidos del GBA	66,2%	67,2%	78,5%
Buenos Aires	63,7%	64,8%	55,9%
Catamarca	89,0%	90,7%	87,5%
Córdoba	67,0%	74,1%	76,3%
Corrientes	75,7%	76,2%	86,8%
Chaco	69,4%	69,7%	80,1%
Chubut	65,7%	65,7%	83,0%
Entre Ríos	80,8%	80,8%	78,4%
Formosa	91,1%	91,4%	86,6%
Jujuy	86,1%	86,5%	86,8%
La Pampa	70,2%	70,2%	76,0%
La Rioja	82,1%	83,0%	78,5%
Mendoza	71,0%	72,2%	75,3%
Misiones	80,2%	80,2%	88,6%
Neuquén	64,8%	65,3%	69,6%
Río Negro	82,0%	82,2%	78,4%
Salta	85,2%	85,3%	77,3%
San Juan	86,4%	86,4%	82,1%
San Luis	79,9%	80,0%	69,6%
Santa Cruz	64,6%	65,0%	82,3%
Santa Fe	74,1%	74,5%	82,1%
Santiago del Estero	74,4%	74,5%	89,5%
Tucumán	81,0%	81,0%	86,9%
Tierra del Fuego	70,2%	70,2%	80,9%
Total del país	72,5%	73,6%	77,3%

Fuente: INDEC, 4° Encuesta Nacional de Factores de Riesgo.

Tasas de respuesta por aglomerado para el Paso 3

Aglomerado	Respuesta de hogares		Respuesta de personas
	RR3	RR5	
Ciudad Autónoma de Buenos Aires	55,0%	57,5%	53,2%
Partidos del Gran Buenos Aires	66,2%	67,2%	66,3%
Gran La Plata	58,4%	58,4%	27,4%
Bahía Blanca-Cerri	46,4%	46,7%	51,1%
Mar del Plata	41,9%	46,0%	46,5%
Gran Catamarca	85,1%	86,0%	87,5%
Gran Córdoba	69,2%	74,3%	27,6%
Río Cuarto	77,9%	86,5%	75,0%
Corrientes	50,3%	51,2%	62,6%
Gran Resistencia	60,3%	60,6%	32,3%
Comodoro Rivadavia-Rada Tilly	54,0%	54,0%	73,0%
Gran Paraná	77,4%	77,4%	50,9%
Concordia	84,0%	84,2%	72,3%
Formosa	91,8%	91,8%	95,2%
Jujuy-Palpalá	83,6%	84,2%	67,2%
La Rioja	73,6%	74,1%	61,7%
Gran Mendoza	68,1%	68,1%	57,5%
Posadas	72,6%	72,6%	70,2%
Neuquén-Plottier	65,3%	65,3%	41,5%
Salta	79,1%	79,2%	78,7%
Gran San Juan	83,7%	83,7%	63,3%
Gran San Luis	76,1%	76,3%	74,2%
Gran Rosario	56,0%	56,8%	45,5%
Gran Santa Fe	82,3%	82,4%	78,8%
Santiago del Estero-La Banda	71,5%	71,7%	84,0%
Gran Tucumán-Tafí Viejo	77,4%	77,4%	65,0%
Total del dominio de aglomerados	67,2%	68,5%	62,9%

Fuente: INDEC, 4° Encuesta Nacional de Factores de Riesgo.

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

Cobertura. Grado de inclusión de los elementos de la población objetivo en el marco muestral. Si el marco no contiene 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.

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

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

Conglomerado. Conjunto de unidades o elementos de la población agrupados por naturaleza propia o sobre la base de un criterio de proximidad. El conglomerado puede ser un agrupamiento ya existente de la población (vivienda u hogar, hospital, escuela); o bien, estar definido por divisiones administrativas, operativas o geográficas del territorio 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 multietápico, en los que la selección de elementos o miembros de la población en forma directa resulta impracticable, por ausencia de listados o por motivos relacionados a los costos operativos.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Medida de tamaño. Cantidad que refleja el tamaño de una unidad de muestreo; por lo general, en encuestas a hogares es el número de viviendas o el total de población. Se la emplea para definir

probabilidades para las unidades de muestreo en métodos que seleccionan las unidades para la muestra con probabilidad proporcional al tamaño.

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

MMUVRA. Muestra maestra urbana empleada por el INDEC con alcance nacional restringido a las localidades de 2.000 o más habitantes, que se utiliza como marco secundario de selección de viviendas particulares para todas sus encuestas a hogares entre dos censos de población y viviendas. Posee un diseño muestral complejo, y se 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 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, unidad de primera etapa de muestreo, UPM; unidad de segunda etapa de muestreo, USM; etcétera.

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