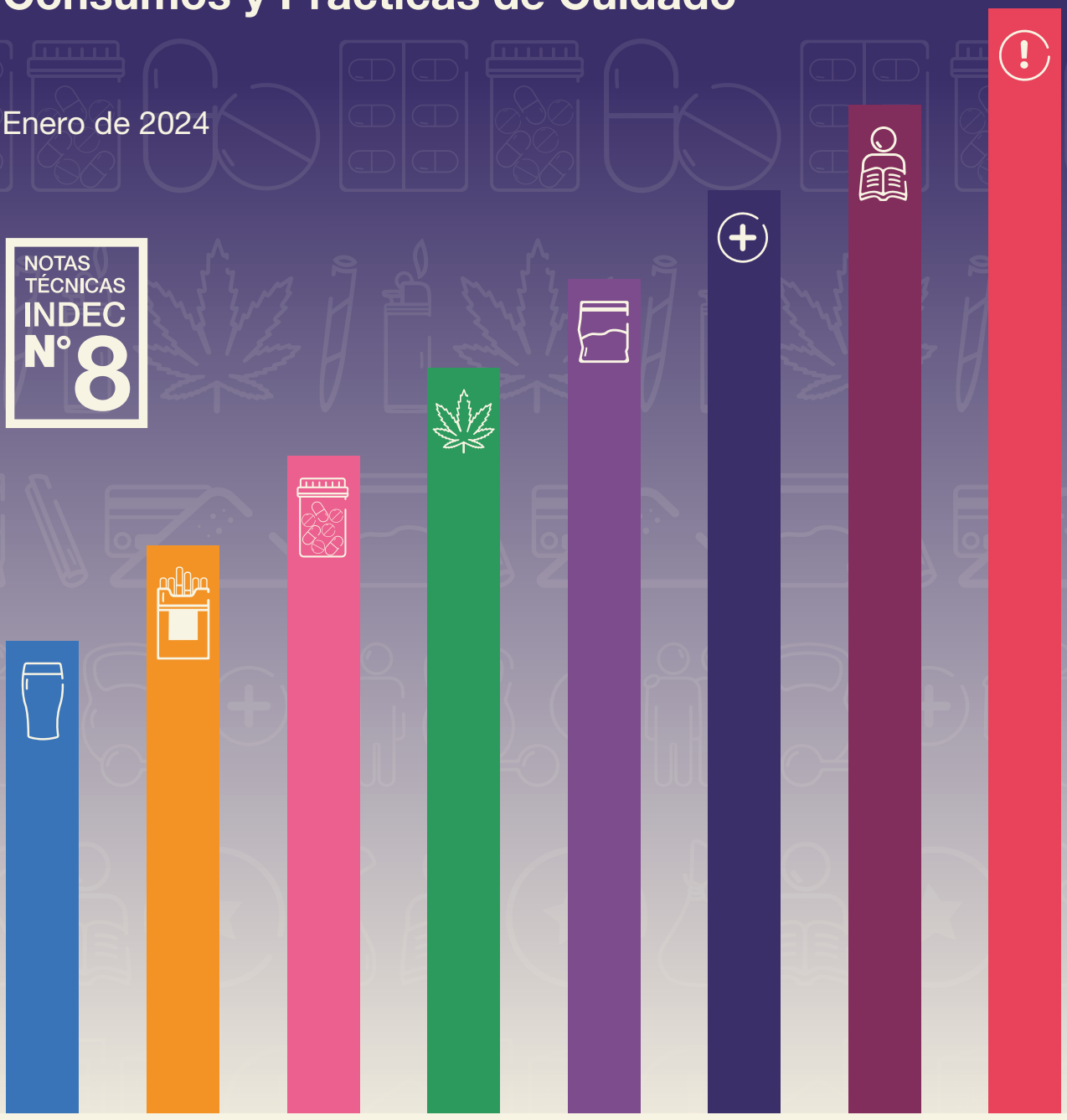
ISBN 978-950-896-664-3

NOTA TÉCNICA **ENCoPraC**

Metodología para el cálculo de los factores de expansión de la Encuesta Nacional sobre Consumos y Prácticas de Cuidado

Enero de 2024

NOTAS TÉCNICAS INDEC N° 8



Ministerio de Economía
Presidencia de la Nación

indec Instituto Nacional de Estadística y Censos
República Argentina

Metodología para el cálculo de los factores de expansión de la Encuesta Nacional sobre Consumos y Prácticas de Cuidado
Notas Técnicas INDEC. N°8

Instituto Nacional de Estadística y Censos (INDEC)

Dirección: Marco Lavagna
Dirección Técnica: Pedro Ignacio Lines
Dirección de Gestión: Santiago Tettamanti
Dirección Nacional de Metodología e Infraestructura Estadística: Gerardo Mitas
Dirección de Metodología e Innovación Estadística: Alejandra Clemente
Coordinación de Muestreo y Métodos de Estimación: Gregorio García
Dirección Nacional de Difusión y Comunicación: María Silvina Viazzi
Coordinación de Producción Gráfica y Editorial: Marcelo Costanzo

Esta publicación fue realizada por el equipo técnico de la Dirección Nacional de Metodología e Infraestructura Estadística integrado por Gonzalo Marí, Emanuel Ciardullo y Aldana Armendariz.

ISSN 2683-8478
ISBN 978-950-896-664-3

Instituto Nacional de Estadística y Censos
Metodología para el cálculo de los factores de expansión de la Encuesta Nacional sobre Consumos y Prácticas de Cuidado : notas técnicas INDEC. N°8 / 1a ed. - Ciudad Autónoma de Buenos Aires : Instituto Nacional de Estadística y Censos - INDEC, 2024.
Libro digital, PDF - (Notas Técnicas INDEC ; 8)

Archivo Digital: descarga y online
ISBN 978-950-896-664-3

1. Estadísticas. 2. Metodología de la Investigación. 3. Encuestas. I. Título.
CDD 310.2

Libro de edición argentina

Buenos Aires, enero de 2024

Índice

Índice de tablas

Índice de figuras

1. Introducción

Durante el tercer trimestre del 2022 se llevó a cabo la Encuesta Nacional sobre Consumos y Prácticas de Cuidado (ENCoPraC). Este relevamiento tuvo como objetivo caracterizar las conductas habituales de las personas de entre 16 y 75 años relacionadas con el consumo de bebidas alcohólicas, tabaco, medicamentos y otras sustancias que inciden en la salud.

La encuesta se realizó sobre la muestra previamente seleccionada para la Encuesta Permanente de Hogares (EPH) total urbano correspondiente al tercer trimestre de 2022[1], cuya cobertura alcanza a los hogares en viviendas particulares ubicadas en localidades de la República Argentina de 2.000 habitantes o más. Como es habitual, en esta instancia se relevó la información sociodemográfica, laboral y de ingresos de los hogares a través de una entrevista presencial. Al finalizar esta etapa, se utilizó una metodología de selección aleatoria que dio como resultado la elección de un miembro de cada hogar de entre 16 a 75 años para que fuera entrevistado para la Encuesta Nacional sobre Consumos y Prácticas de Cuidado, en un día y horario a convenir.

Tal como se anticipó en el informe de resultados de la ENCoPraC[2], los niveles de respuesta alcanzados en este relevamiento fueron menores a los registrados en otras encuestas realizadas por el Instituto. Este deterioro en la respuesta obligó a revisar y modificar la metodología habitual de cálculo de los factores de expansión de la encuesta, como así también los criterios para la presentación de sus resultados y su evaluación de calidad.

En esta nota técnica se presentan los resultados del análisis de los niveles de respuesta obtenidos para la ENCoPraC, sus efectos sobre la determinación de los factores de expansión para realizar la inferencia estadística, la calidad de las estimaciones y las limitaciones de cálculo de los correspondientes errores muestrales.

2. Diseño muestral de la encuesta

El diseño muestral de la ENCoPraC se basa en el diseño muestral de la EPH total urbano, el cual a su vez se apoya en el diseño de la Muestra Maestra Urbana de Viviendas de la República Argentina (MMUVRA) ajustado a los requerimientos de la encuesta.

La MMUVRA es de alcance nacional[3] y urbano, y permite seleccionar muestras para las encuestas que tienen como principales dominios de estimación las provincias y los aglomerados[4] que participan en la EPH que realiza el Instituto Nacional de Estadística y Censos (INDEC).

[1] La EPH total urbano es la ampliación de la cobertura de la EPH continua en los 31 aglomerados urbanos durante el tercer trimestre de cada año. La extensión se realiza a través de la incorporación a la muestra de viviendas particulares de localidades de 2.000 y más habitantes, no comprendidas en los dominios de estimación del operativo continuo, para todas las provincias con excepción de Tierra del Fuego, Antártida e Islas del Atlántico Sur y la Ciudad Autónoma de Buenos Aires.
[2] Ver en https://www.indec.gob.ar/indec/web/Nivel4-Tema-4-32-67.
[3] Está definida para viviendas particulares en localidades simples o compuestas de 2.000 o más habitantes.
[4] Los aglomerados de la EPH son: Gran Buenos Aires, Gran Mendoza, Gran Tucumán-Tafí Viejo, Salta, Gran Córdoba, Gran La Plata, Gran Rosario, Gran Santa Fe, Mar del Plata, Gran San Juan, Gran San Luis, Corrientes, Formosa, Gran Resistencia, Posadas, Gran Catamarca, Jujuy-Palpalá, La Rioja, Santiago del Estero-La Banda, Bahía Blanca-Cerri, Concordia, Gran Paraná, Río Cuarto, Santa Rosa-Toay, San Nicolás-Villa Constitución, Comodoro Rivadavia-Rada Tilly, Neuquén-Plottier, Río Gallegos, Ushuaia-Río Grande, Rawson-Trelew, Viedma-Carmen de Patagones.

La estructura probabilística de la EPH, heredada de la MMUVRA, consiste en 3 etapas de selección probabilística bajo un diseño complejo. En la primera etapa, se realiza una selección aleatoria de aglomerados o localidades simples, o "unidades de primera etapa de muestreo" (UPM). Los aglomerados considerados como dominios de estimación habituales de la EPH están autorrepresentados o seleccionados con probabilidad igual a 1 en la MMUVRA. El resto de las UPM de la MMUVRA son seleccionadas bajo un diseño proporcional al tamaño.

Para la segunda etapa de selección, en las UPM seleccionadas para la MMUVRA se definen las "unidades de segunda etapa de muestreo" (USM) o "Áreas MMUVRA"[5] con base en los radios censales y en la cartografía del Censo Nacional de Población, Hogares y Viviendas 2010 (CPHyV 2010). En cada UPM, todas las USM que la conforman cubren territorialmente y determinan la envolvente o el área de cobertura asociada a dicha unidad; de este modo, se conforma el marco de muestreo para la selección de la segunda etapa.

La muestra probabilística de USM para la MMUVRA emplea un diseño estratificado definido a partir del nivel educativo del jefe del hogar. La selección involucra un muestreo sistemático con probabilidad proporcional a la cantidad total de viviendas particulares ocupadas según el CNPHyV 2010 en cada estrato. Estas primeras dos etapas de selección pertenecen al diseño de la MMUVRA.

Finalmente, la tercera etapa es propia de la EPH total urbano y está constituida por una selección probabilística de viviendas, o "unidades de tercera etapa de muestreo" (UTM), a partir del listado exhaustivo de viviendas particulares en cada USM seleccionada, y que conforman en su conjunto el marco de muestreo de viviendas de la MMUVRA. 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, las tareas de campo y el recorrido del personal de la encuesta que realiza las entrevistas[6].

Tal lo adelantado en la introducción, en cada uno de los hogares que responden a la EPH se suma una cuarta etapa al seleccionar al azar un individuo perteneciente a la población objetivo de la ENCoPraC, constituida por el conjunto de personas de 16 a 75 años.

El tamaño de la muestra inicial de viviendas coincide al previsto para la Encuesta Permanente de Hogares total urbano. Para el tercer trimestre de 2022 fue de 41.688 viviendas distribuidas en 4.053 áreas de relevamiento ubicadas en los aglomerados y restos urbanos que cubre la encuesta.

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

[6] Con rigor, para la muestra definitiva de viviendas de la encuesta, se lleva a cabo una nueva etapa de selección probabilística de segmentos de viviendas. Estos están constituidos por cinco viviendas particulares contiguas o próximas entre sí dentro del listado de la MMUVRA. Su principal objetivo es concentrar los desplazamientos en terreno de quienes encuestan, 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.

Tabla 1. Distribución del tamaño de la muestra de viviendas por jurisdicción. Total país, aglomerados urbanos y resto urbano

Jurisdicción	Viviendas seleccionadas		
	Total	Aglomerados urbanos	Resto urbano
Ciudad Autónoma de Buenos Aires	1.812	1.812	-
Buenos Aires	8.127	7.457	670
Catamarca	1.186	601	585
Córdoba	2.436	1.595	841
Corrientes	1.133	603	530
Chaco	1.235	684	551
Chubut	1.782	1.252	530
Entre Ríos	2.087	1.462	625
Formosa	1.347	664	683
Jujuy	1.190	600	590
La Pampa	1.325	644	681
La Rioja	1.241	631	610
Mendoza	1.621	981	640
Misiones	1.211	601	610
Neuquén	1.243	598	645
Río Negro	1.767	481	1.286
Salta	1.506	865	641
San Juan	1.385	800	585
San Luis	1.274	613	661
Santa Cruz	1.010	460	550
Santa Fe	2.488	1.898	590
Santiago del Estero	1.259	659	600
Tucumán	1.422	862	560
Tierra del Fuego	601	601	-
Total	**41.688**	**27.424**	**14.264**

Fuente: INDEC, EPH total urbano - *Encuesta Nacional sobre Consumos y Prácticas de Cuidado 2022.*

3. Efectos de la no respuesta a nivel de las unidades de muestreo de primera y segunda etapa de la MMUVRA

Cuando el diseño muestral es complejo y el nivel de no respuesta es elevado, todas las etapas de selección se ven afectas por el fenómeno; inclusive las estratificaciones implícitas o explícitas vinculadas al diseño. Como consecuencia, el deterioro o disminución en el número de unidades en cualquiera de las etapas de diseño, o una modificación sustancial en su distribución por estrato resultante post operativo de campo, podrían afectar la integridad o validez de la inferencia estadística a partir de la muestra resultante. En rigor, la no respuesta impacta en las decisiones relacionadas a:

1) la metodología adoptada para el modelo de no respuesta (ajustes por no respuesta)

2) los criterios e información auxiliar disponible para calibrar los factores de expansión

3) la posibilidad o no de realizar y validar un cálculo de los errores muestrales para la encuesta.

Para dimensionar este fenómeno en la ENCoPraC 2022, en la tabla 2 se observa que el 50% de las áreas MMUVRA seleccionadas para el total del país presentan un nivel de no respuesta de al menos el 60% de las viviendas seleccionadas por área. Por otro lado, también se observa que este fenómeno es muy dispar a través de las provincias, con porcentajes que van del 13% al 82%. Al mismo tiempo, se observan fluctuaciones en los resultados asociados a los aglomerados urbanos EPH y los correspondientes a áreas seleccionadas en localidades del resto urbano de cada provincia.

Al cuantificar el porcentaje de áreas con al menos el 60% de viviendas seleccionadas sin respuesta como umbral de referencia, ya sea por provincia, por región o para el total del país, es posible dimensionar el porcentaje de áreas donde el nivel de no respuesta es mayor a la respuesta obtenida. Como criterio general, cuando este porcentaje supera el 50%, da cuenta del deterioro de la respuesta respecto del diseño muestral previsto y funciona como señal de alerta respecto a la confiabilidad de las estimaciones que surjan a partir de la encuesta. En este caso el problema afecta a gran parte de los aglomerados y localidades de los restos urbanos de provincia, y se manifiesta en una menor cantidad de respuestas para calcular las estimaciones que surjan de ella, y en una distorsión de la estructura probabilística subyacente.

Tabla 2. Porcentaje de las áreas de relevamiento con niveles de no respuesta del 60% o más, por jurisdicción. Resultados desagregados para aglomerados EPH y localidades del resto urbano

Jurisdicción	Aglomerados EPH	Resto urbano	Total
Ciudad Autónoma de Buenos Aires	82%	--	82%
Buenos Aires	64%	96%	69%
Catamarca	23%	10%	18%
Córdoba	27%	35%	30%
Corrientes	58%	35%	47%
Chaco	40%	35%	38%
Chubut	47%	23%	40%
Entre Ríos	60%	46%	56%
Formosa	42%	81%	63%
Jujuy	13%	12%	13%
La Pampa	55%	67%	61%
La Rioja	19%	28%	23%
Mendoza	30%	44%	35%
Misiones	30%	28%	29%
Neuquén	63%	30%	50%
Río Negro	63%	20%	32%
Salta	32%	19%	27%
San Juan	44%	15%	34%
San Luis	35%	47%	41%
Santa Cruz	63%	80%	72%
Santa Fe	38%	36%	37%
Santiago del Estero	48%	21%	38%
Tucumán	25%	0%	18%
Tierra del Fuego*	60%	-	60%
Total	50%	41%	47%

Fuente: INDEC, *Encuesta Nacional sobre Consumos y Prácticas de Cuidado 2022*.

Nota: En la provincia de Tierra del Fuego no hay áreas seleccionadas pertenecientes al resto urbano.

En un intento de sumar observaciones efectivas, cualquier estrategia válida de agrupamiento de casos para determinar los dominios de estimación (provincia, región) debe contemplar que las áreas MMUVRA fueron seleccionadas respetando una determinada distribución por estratos, que ellos no están presentes en todas las unidades de primera etapa (aglomerados/localidades), y que se distorsionan a medida que aumenta el fenómeno de no respuesta.

Si bien los criterios empleados en la estratificación de las áreas son generalmente uniformes, agruparlas cuando la no respuesta es heterogénea por estrato, u otras dimensiones del diseño, puede romper el equilibrio en las distribuciones en cada estrato, desvirtuando la estructura probabilística original de la encuesta. Las distorsiones del

diseño muestral inicial se traducen en sesgo tanto en los estimadores como en su variancia. En este último caso, se produce una pérdida de la componente de variancia entre las áreas previstas por diseño, que al reagruparlas introduce una nueva componente. En la práctica no es posible cuantificar si este cambio en la composición de la variancia permite obtener una estimación aproximadamente insesgada del error del estimador de variancia, ni el efecto final sobre este de la pérdida de grados de libertad.

El problema de no respuesta evidenciado en las áreas MMUVRA se traslada a nivel de aglomerados y en particular a la muestra de localidades que conforman el resto urbano de las provincias. Estos, a diferencia de los aglomerados EPH, tienen una etapa de selección previa a las áreas, en la que se eligen localidades bajo un diseño probabilístico, por lo que el deterioro de una de ellas implica no solo la pérdida de esa localidad sino también de aquellas a las que representa. Por otro lado, la falta generalizada de estas unidades de muestreo no permite capturar la variabilidad entre ellas, o entre localidades dentro de una provincia; todo esto impacta en la distribución de los estimadores sumando sesgo. A su vez, complejiza cualquier metodología de cálculo del error de muestreo a desarrollar en detrimento de subestimar el error. Este problema cobra mayor dimensión cuando se asume que los fenómenos que se estudian en la encuesta pudieran estar presentes de manera diferencial en estas unidades (aglomerados/localidades).

4. Determinación de los factores de expansión

La metodología estándar que adopta el INDEC para brindar estimaciones oficiales para las encuestas a hogares emplea estimadores que involucran a un ponderador o factor de expansión único para cada unidad que responde a una encuesta. Esta estrategia facilita el cálculo, de manera unívoca, de cualquier estimación que los usuarios e investigadores deseen realizar a partir de la base de microdato de una encuesta que el Instituto comparte a través de su página web.

Los factores de expansión se construyen respetando un estándar y siguiendo un conjunto de pautas que buscan asegurar, en lo posible, la validez de los resultados con una precisión esperada y minimizar sesgos para la mayoría de las estimaciones. El proceso de cálculo se inicia tomando los factores de expansión de diseño, que se obtienen de la estructura probabilística de la muestra de cada unidad seleccionada, definidos como el producto de las recíprocas de las probabilidades asignadas por el diseño a cada una de las etapas de selección.

Estos factores se caracterizan por incorporar el proceso de aleatorización que existe al seleccionar la muestra probabilística y que, en ausencia de no respuesta o deficiencias de cobertura en el operativo de campo de la encuesta, elimina cualquier sesgo de selección de las unidades.

Sin embargo, todos los operativos de diseño complejo y a gran escala, como son las encuestas oficiales, se ven afectados por la no respuesta de las unidades seleccionadas[7] y de otros errores de distinta índole denominados *no muestrales*. La mayoría son difíciles o imposibles de cuantificar y afectan a la calidad del dato en dos direcciones. Si son introducidos de manera aleatoria, la probabilidad de incrementar la variabilidad

[7] Definida como la falla de obtener información completa o respuesta de una unidad seleccionada para la muestra.

de la estimación es alta, por lo que pierde precisión. En cambio, si no son aleatorios, el principal impacto es la introducción de sesgo en los resultados.

A tal efecto, la metodología adoptada para las encuestas a hogares establece eliminar las unidades que no responden y modificar los factores iniciales de cada unidad que sí lo hace, para compensar las que se eliminaron. Este procedimiento intenta disminuir el efecto de distintos sesgos introduciendo factores de ajuste que surgen de modelizar alguno de los principales errores no muestrales bajo un conjunto de supuestos teóricos. Como resultado de este procedimiento, el factor de expansión inicial asociado a cada vivienda de la muestra recibe de manera secuencial tres tipos de ajustes:

- ajustes en base al *status* de elegibilidad de la vivienda
- ajustes por no respuesta de la vivienda
- ajustes por calibración

El primero se impone como consecuencia de las variaciones entre el estado de la vivienda del listado de la MMUVRA al momento de su selección y el detectado en campo por el operativo de la encuesta. Dependiendo de la actualización de los listados esta variación, por lo general, es marginal[8].

El tratamiento o ajuste por no respuesta es delicado, y requiere adoptar un modelo estadístico de respuesta en un intento por disminuir la transferencia de sesgo que provoca el fenómeno sobre los estimadores. El sesgo puede ser apreciable cuando las unidades que responden y las que no difieren en las características de interés de la encuesta. La bondad del modelo dependerá de la información disponible para todas las unidades de la muestra –respondan o no–, las características y magnitud de la no respuesta, y de la forma en que la no respuesta interactúa con las variables que son objeto de estudio de la encuesta.

El ajuste por calibración emplea información agregada y busca ajustar la estructura sub/sobre representada de la muestra de algunos grupos poblacionales de interés. Este proceso emplea totales poblacionales o proyectados conocidos[9] y genera una componente de ajuste que impacta en los factores de expansión surgidos de los tratamientos por elegibilidad y no respuesta. Si bien ordena la presentación de resultados en el sentido de sostener ciertas estructuras básicas de la población, su principal efecto radica en mejorar –en lo posible– la precisión de las estimaciones en términos del error por emplear una muestra de unidades de la población; y en ocasiones puede contribuir en la disminución del sesgo por no respuesta en el estimador final. El beneficio de la calibración dependerá del poder explicativo y predictivo entre las variables disponibles para la población y las que son objeto de estudio, siendo mayor cuanto mayor es la correlación entre ellas.

[8] En el caso de MMUVRA, la última actualización ocurrió entre 2018 y 2019.
[9] Las variables involucradas habitualmente en este proceso son el género o sexo, y grupos etarios de edades a niveles geográficos de región y provincia.

4.1 Factores de expansión iniciales de la ENCoPraC

Como se adelanta en la introducción, la encuesta se realizó sobre la muestra de viviendas previamente seleccionada para la Encuesta Permanente de Hogares (EPH) total urbano correspondiente al tercer trimestre de 2022. En la práctica, en cada vivienda/hogar que respondió a la EPH se seleccionó un miembro del hogar de edad entre 16 y 75 años para que fuera entrevistado. Inicialmente cada hogar con respuesta a la EPH, previo a la selección de la ENCoPraC, ya posee un factor de expansión propio de la EPH a nivel de hogar y para cada una de las personas que lo habitan. Estos factores ya cuentan con ajustes por no respuesta y un ajuste por calibración propios de la metodología que lleva adelante la EPH para brindar estimaciones. Los detalles de estos ajustes están documentados en *Encuesta permanente de hogares (EPH) total Urbano. Principales tasas de los terceros trimestres*[10].

El factor de expansión final de la EPH para cada hogar –y todas las personas que lo componen– es el que se pone a disposición de los usuarios en las bases con los microdatos y se denomina *Pondera,* tal como consta en la documentación que acompaña a las bases usuarias[11].

Como la ENCoPraC introduce una etapa adicional de selección aleatoria en cada hogar que responde a la EPH, para los procesos de ajustes se define como w_k^0 al factor de expansión inicial de la ENCoPraC para cada individuo, definido como:

$$w_k^0 = Pondera_k . \left(\frac{1}{N_k}\right)^{-1}$$

donde $\frac{1}{N_k}$ es la probabilidad de seleccionar un individuo entre los N_k pertenecientes a la población objetivo del hogar k que responde a la EPH. La inversa de esta probabilidad permite incorporar la aleatorización del proceso de selección de la ENCoPraC en el factor *Pondera*.

Como esta última etapa de selección también se ve afectada por el fenómeno de no respuesta a nivel de las personas seleccionadas, en los siguientes apartados se describen los ajustes que se llevan adelante sobre w_k^0, lo que permitirá obtener los factores de expansión finales de la ENCoPraC. En concreto, se introducen dos correcciones adicionales, a_k y λ_k, que representan los ajustes por no respuesta y por calibración específicos de la ENCoPraC, respectivamente. Como resultado de ambos ajustes se obtendrá el factor de expansión final para las personas que respondieron a la ENCoPraC 2022:

$$w_k^P = w_k^0 . a_k . \lambda_k$$

cuyo cálculo se define en las próximas secciones. Cabe mencionar que en esta notación se realiza una simplificación en la elección de los subíndices que identifican cada término para facilitar la presentación. En rigor, cada individuo que responde a la en-

[10] https://www.indec.gob.ar/indec/web/Nivel4-Tema-4-31-58
[11] Ver por ejemplo, https://www.indec.gob.ar/ftp/cuadros/menusuperior/eph/EPH_registro_4T2022.pdf

cuesta pertenece a un hogar asociado a una vivienda, seleccionada dentro de un área MMUVRA que a su vez pertenece a un aglomerado urbano.

4.2 Ajuste por no respuesta específico de la ENCoPraC

La magnitud y distribución de la no respuesta que se evidencia a partir del análisis de los resultados del apartado 3 aumenta el potencial sesgo que esta introduce en los estimadores. El impacto de los errores no muestrales sobre la estructura probabilística de la encuesta que obligan a introducir ajustes significativos sobre los pesos w_k^0, sumado a la imposibilidad de sostener la estratificación prevista por diseño, hacen impracticable replicar la metodología y el modelo de respuesta que utiliza la EPH, o el de otras encuestas a hogares. Estos modelos se caracterizan por definir una propensión o probabilidad de responder para cada unidad (vivienda o individuo) a través de un modelo que emplea información vinculada a la estratificación y a la información geográfica a la cual pertenece la vivienda seleccionada, según lo previsto por el diseño muestral de la encuesta.

El modelo estadístico subyacente define celdas o categorías de ajuste que se originan al combinar los estratos de diseño con el nivel geográfico. En cada una de estas celdas se clasifican las unidades que responden o no a la encuesta y el ajuste se realiza a través de la estimación de la propensión a la respuesta de la unidad dentro de cada una de ellas. Esta estimación se obtiene a partir del cociente entre el total de unidades que responden sobre el total de unidades seleccionadas, dando por resultado un valor único por celda que se aplica al factor de expansión de cada unidad que responde en la celda en cuestión.

La validez de este modelo requiere que la pérdida de áreas y viviendas que componen la muestra debido a la no respuesta sea marginal en los estratos que definen las celdas de ajustes, de tal forma que las hipótesis que lo sostienen no se debiliten. Este modelo de respuesta impone los siguientes supuestos: i) que en cada celda de ajuste la propensión a responder de las unidades sea independiente unas de otras; ii) que la propensión sea uniforme por celda; iii) que para cada celda la información recogida a partir de los que responden a una variable en estudio permita predecir lo que informarían para esa misma variable los que no responden. Es importante señalar que, en la metodología que sostiene los modelos de respuesta, este supuesto de *ignorabilidad* sujeta la información auxiliar disponible tanto para el conjunto de respondentes como para el de los que no responden. Esto equivale, con cierto grado de generalidad, a suponer que en una encuesta las variables en estudio no determinan la probabilidad de responder o no a ella.

En el caso de la ENCoPraC, la mayoría de las variables bajo estudio (consumos de alcohol, tabaco, psicofármacos, medicamentos opioides sin prescripción médica, entre otras) muy probablemente afectan la predisposición a responder del individuo seleccionado. Por lo tanto, es difícil sostener el supuesto de ignorabilidad antes definido.

Si las tasas de no respuesta fuesen mucho más pequeñas, el efecto de sostener un modelo bajo esta hipótesis y que ella no se cumpla generaría un sesgo marginal en los estimadores que por lo general sería despreciable; sin embargo, no es el caso de esta encuesta.

Por otra parte, para imponer modelos de respuesta bajo el supuesto de que el mecanismo es no ignorable, se necesitaría contar con información auxiliar externa a la que provee la encuesta que pueda ser empleada para explicar el fenómeno en estudio, tanto para los que responden como para los que no lo hacen, pero dicha información no se encuentra disponible.

4.3 Estimación de la propensión a la respuesta y factor de ajuste para la ENCoPraC

A partir de la magnitud y efectos de la no respuesta manifestados anteriormente, y ante la necesidad de dar una respuesta metodológica al problema de estimación en este contexto, se llevó a cabo un procedimiento de ajuste a través de un enfoque experimental. Se asume como válido el supuesto de que el mecanismo de no respuesta asociado a la etapa de selección adicional impuesta por la ENCoPraC es aleatorio (en inglés, *missing at random*, MAR). Esto significa que la propensión a responder de un individuo seleccionado en la cuarta etapa de selección, habiendo sido seleccionado previamente para participar de la EPH, depende de variables que son conocidas (y por lo tanto son observables), tanto para quienes respondieron como para quienes no lo hicieron. Estas variables corresponden a atributos de los individuos seleccionados, a características del hogar que habitan y a indicadores vinculados a los intentos realizados por el equipo de relevamiento para contactarlos.

Para formalizar la presentación del modelo utilizado se definen los siguientes elementos. Sean

R_k variable indicadora de respuesta de la unidad que toma valores (0,1), $k = 1, \dots, N(n)$

s_R conjunto de unidades de la muestra que responden $R_k = 1$, $s_R = \{k \in S: R_k = 1\}$

s_{NR} conjunto de unidades de la muestra que no responden $R_k = 0$, $s_{NR} = \{k \in S: R_k = 0\}$

v_k vector de variables auxiliares o información conocida para la unidad k seleccionada, $k \in s$

Si el mecanismo de respuesta es generado por un muestreo Poisson, entonces la probabilidad de obtener una respuesta se define como:

$$p_k = Pr(R_k = 1/Y = y_k, A = a_k, V = v_k\ )$$

donde Y representa a las variables en estudio, A es un vector de variables que reflejan las condiciones del operativo y V es un vector de variables auxiliares.

A priori, se podría considerar que $p_k > 0$ para toda unidad k, lo que equivale a asumir que todas las unidades seleccionadas tienen una propensión a la respuesta distinta de 0.

Otro supuesto que podría establecerse es que el mecanismo de respuesta no se ve afectado por el diseño muestral, en presencia de la información disponible para las unidades. Es decir, la variable indicadora de respuesta R_k y la variable indicadora que

origina el diseño muestral son independientes en términos estadísticos, condicional a un conjunto de variables auxiliares afectadas al modelo de respuesta.

La expresión definida para p_k involucra a la(s) variable(s) bajo estudio representadas por Y y habrá que asumir un mecanismo de respuesta ignorable para poder eliminar su influencia en la expresión del modelo y llegar a una forma que sea estimable con la información disponible, pudiendo ser este un supuesto fuerte. Por otro lado, cualquier expresión teórica del sesgo atribuido a la no respuesta involucrará explícitamente a la variable que está siendo estimada, sugiriendo que el ajuste debería estar apoyado, al menos en parte, en un modelo de la distribución de la característica.

Bajo el supuesto de ignorabilidad de la no respuesta, la modelización de p_k queda sujeta al supuesto de que dicha probabilidad solo depende de la información auxiliar y que alcanza con esto para definir el mecanismo de respuesta subyacente (esquema *missing at random*)

Siguiendo el trabajo de Rubin (1976), cuando se asume un esquema de estas características:

$$Pr(R_k = 1/Y = y_k, A = a_k, V = v_k\ ) = Pr(R_k = 1/\ A = a_k, V = v_k\ )$$

para todas las unidades seleccionadas ya sea que respondan o no $(s_r \cup s_{nr})$ o sea, p_k depende de los datos observados y no de aquellos no observados. Cuando esto no ocurre, el mecanismo que genera la no respuesta es MNAR (*missing not at random*). Aunque esta dicotomía es útil, en la práctica no es posible aseverar cuándo el mecanismo es MAR o MNAR si no se cuenta con datos adicionales de las unidades que no responden.

Para la estimación de la propensión de un individuo a responder se adopta un modelo paramétrico de regresión logística sin interacciones,

$$p_k = m(v_k; \beta) = \frac{exp(v_k^t\, \beta)}{1 + exp(v_k^t\, \beta)}$$

donde

β es un vector de parámetros del modelo,

p_k es la probabilidad o propensión de respuesta del individuo k seleccionado para la encuesta.

v_k es un vector de variables auxiliares empleadas para el individuo k seleccionado (ya sea que responde o no a la encuesta). Se asume que las variables que conforman este vector v_k estén vinculadas a la propensión a la respuesta y a las variables en estudio.

Sea $\hat{\beta}$ el vector estimado de β desconocido, existen distintas modalidades para estimar este vector de parámetros del modelo; se adopta el criterio de estimarlo por el método de máxima verosimilitud.

Finalmente, la probabilidad de respuesta estimada a partir del modelo para la unidad *k* se define como:

$$\hat{p}_k = m(v_k; \hat{\beta}) = \frac{exp(v_k^t \hat{\beta})}{1 + exp(v_k^t \hat{\beta})}$$

Dado que la selección de individuos de esta encuesta tiene como origen la submuestra de hogares que responden a la EPH total urbano, se dispone de un conjunto amplio de características, tanto a nivel del hogar como del informante seleccionado que responde o no a la encuesta, para modelar y predecir la probabilidad de respuesta a la encuesta bajo el supuesto MAR.

Las características elegidas para modelar la propensión a la respuesta y que componen el vector v_k son las siguientes:

Nombre	Definición de la variable
REGIÓN	Regiones estadísticas (se incluyeron 5 variables indicadoras para representar las 6 regiones habituales: Pampeana, Noreste, Noroeste, Cuyo y Patagónica. La región GBA se identifica por la categoría de control)
SEXO	Indicadora de sexo masculino de la persona respondente (1=masculino, 0=otro caso VER)
EDAD	Edad de la persona respondente (años cumplidos)
ESCOLARIDAD	Años de escolaridad de la persona respondente (cantidad de años completos)
LOG(IPCF)	Logaritmo natural del ingreso familiar per cápita (estandarizado por el desvío estándar de la variable)
PRIMERA PARTICIPACIÓN	Indicadora de primera participación del hogar al que pertenece la persona seleccionada en la encuesta (1=si, 0=otro caso)
OCUPADO	Indicadora de que la persona seleccionada es ocupada (1=si, 0=otro caso)
EXTRANJERO	Indicadora de que persona seleccionada es extranjera (1=si, 0=otro caso)
RESIDENTE_500+	Indicadora de que el hogar seleccionado pertenece a un aglomerado de 500.000 habitantes o más (1=si, 0=otro caso)
ADULTO _MAYOR	Indicadora de que en el hogar reside al menos un adulto mayor de 65 años o más (1=si, 0=otro caso)
MENOR_11A21	Indicadora de que en el hogar reside al menos un individuo de 11 a 21 años (1=si, 0=otro caso)
MENORES_10	Indicadora de que en el hogar reside al menos un niño de hasta 10 años (1=si, 0=otro caso)
INGRESOS_ LABORALES	Indicadora de que el hogar cuenta con ingresos provenientes del trabajo (1=si, 0=otro caso)
TOTAL_VISITAS	Variables indicadoras de la cantidad total de visitas para realizar la encuesta SEDRONAR (1=una visita, 2=dos visitas, 3=tres visitas, 4=cuatro o más visitas)
SENSIBILIZACION_90+	Indicadora de la cantidad de días transcurridos desde la sensibilización EPH hasta la visita SEDRONAR (0=90 días o menos,1=más de 90 días)

La probabilidad de respuesta se estima a partir de la muestra de todas las personas seleccionadas para la ENCoPraC (n=15963), respondan o no, asumiendo una especificación general del modelo a nivel de los 31 aglomerados. Los parámetros fueron estimados por el método de máxima verosimilitud sin emplear los factores de expansión.

Los resultados de la tabla 3 muestran, de forma general, que la mayoría de las variables explicativas del modelo de propensión de respuesta son significativas. En general, se observa que hay un efecto diferencial en la propensión a responder según la región estadística en la que residen los individuos, el tamaño de la localidad de residencia, la presencia de niños y jóvenes en el hogar, entre otros rasgos. También se evidencia una menor propensión a responder por parte de aquellos individuos seleccionados que participaban por primera vez del relevamiento de la EPH. Esto sugiere, por complemento, que aquellos individuos que residen en hogares que participaron en la EPH en trimestres previos y fueron seleccionados para la ENCoPraC, mostraron una mayor predisposición a responderla.

Por otro lado, también se observa que aquellos individuos dispuestos a responder lograron completar su participación en una cantidad acotada de visitas al hogar. En general, un incremento en la cantidad de intentos de contacto por parte del equipo de relevamiento se encuentra asociado a una menor propensión a la respuesta. Algo similar ocurrió al incrementarse la cantidad de días transcurridos desde el contacto inicial del hogar por parte del equipo EPH respecto al recontacto del equipo que implementó la ENCoPraC, disminuyendo la probabilidad de obtener una participación efectiva de la persona seleccionada en ese hogar.

Tabla 3. Resultados de la estimación del modelo de propensión de respuesta

Variables	Estimación	Error estándar	Valor z	Pr(>|z|)	Razones de odds	Incremento para odds
Constante	0,691	0,154	4,50	0,000		
Región_Cuyo	0,280	0,080	3,50	0,000	1,32	1
Región_NEA	0,415	0,095	4,36	0,000	1,51	1
Región_NOA	0,417	0,076	5,50	0,000	1,52	1
Región_PAMPEANA	0,443	0,066	6,74	0,000	1,56	1
Región_PATAGONIA	0,093	0,089	1,04	0,299	1,10	1
Sexo (Varón=1)	-0,146	0,041	-3,59	0,000	0,86	1
Edad	0,004	0,002	2,59	0,010	1,04	10
Escolaridad	0,027	0,006	4,55	0,000	1,14	5
Log(IPCF)	0,047	0,017	2,72	0,007	1,05	1
Primera_participación	-0,191	0,044	-4,38	0,000	0,83	1
Ocupado	0,196	0,051	3,86	0,000	1,22	1
Extranjero	-0,051	0,102	-0,50	0,621	0,95	1
Residente_500+	0,159	0,053	2,98	0,003	1,17	1
Adulto_mayor	-0,076	0,055	-1,40	0,162	0,93	1
Menor_11a21	0,184	0,046	4,03	0,000	1,20	1
Menores_10	0,156	0,048	3,28	0,001	1,17	1
Ingresos_laborales	-0,093	0,068	-1,36	0,173	0,91	1
Visitas2	0,094	0,057	1,65	0,098	1,10	1
Visitas3	-0,623	0,062	-10,06	0,000	0,54	1
Visitas4	-1,735	0,052	-33,41	0,000	0,18	1
Sensibilización_90+	-1,091	0,080	-13,72	0,000	0,34	1

Modelo	Resid. d.f.	Resid. Deviance	D.f.	Deviance	Pr(>Chi)
Modelo solo con intercepto	15.962	18.015			
Modelo propuesto	15.951	1.5951	21	2.064	< 2.2e-16

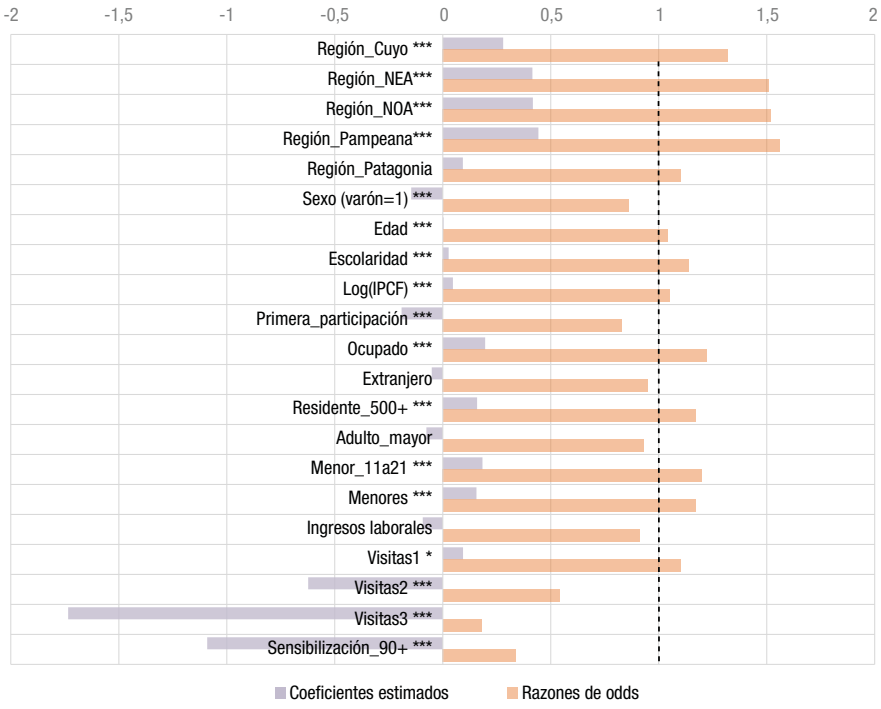
Fuente: INDEC, estimaciones con base en la *Encuesta Nacional sobre Consumos y Prácticas de Cuidado 2022*.

En la figura 2 se presentan de manera gráfica los resultados de los coeficientes estimados para el modelo de respuesta para cada una de las variables explicativas seleccionadas, y una medida de la importancia de esas variables en la propensión a responder expresada a través de razones de odds[12]. La mayoría de las variables auxiliares resultan significativas para explicar la variabilidad en la propensión a responder. A su vez, se pueden identificar aquellas variables cuya razón de odds estimada se ubica por encima

[12] De forma coloquial, la razón de odds representa una medida de asociación entre una característica o atributo de los individuos y una variable de resultados. En el contexto de la ENCoPraC, la variable de resultados corresponde a la obtención de una respuesta. Una razón de odds representa cuánto mayor (o menor) es la chance de que ocurra un resultado (obtener respuesta en la encuesta) dado que un individuo presenta una determinada característica, en comparación a la chance de que ocurra el mismo resultado en ausencia de ese mismo atributo.

o por debajo del valor 1, indicando respectivamente una asociación positiva o negativa de las variables sobre la propensión a la respuesta de los individuos seleccionados para participar de la ENCoPraC.

Figura 1. Relevancia de las variables auxiliares consideradas para modelar la propensión a responder por parte de los individuos seleccionados para participar de la ENCoPraC



Nota: *** indica que una variable es significativa al 1%, ** indica un nivel de significación del 5% y * indica un nivel de significación del 10%.

Fuente: INDEC, estimaciones con base en la *Encuesta Nacional sobre Consumos y Prácticas de Cuidado 2022*.

Para finalizar, más allá de los análisis e intentos realizados para modelizar de forma adecuada la respuesta, se advierte a los usuarios que los estimadores que resultan de este enfoque aún pueden tener sesgo en las estimaciones de totales, mientras que este problema es menor en el caso de la estimación de tasas o proporciones. Por este motivo, se sugiere a los usuarios evitar la estimación de totales a partir de los datos de esta encuesta.

4.4 Determinación del factor de ajuste

A partir de los resultados de la estimación del modelo de respuesta planteado, se definieron clases de ajuste homogéneas definidas por las probabilidades de respuesta estimadas. Si todas las unidades dentro de una celda poseen la misma propensión a la respuesta, se tiene indicios de que el supuesto MAR se satisface y el sesgo del estimador disminuye.

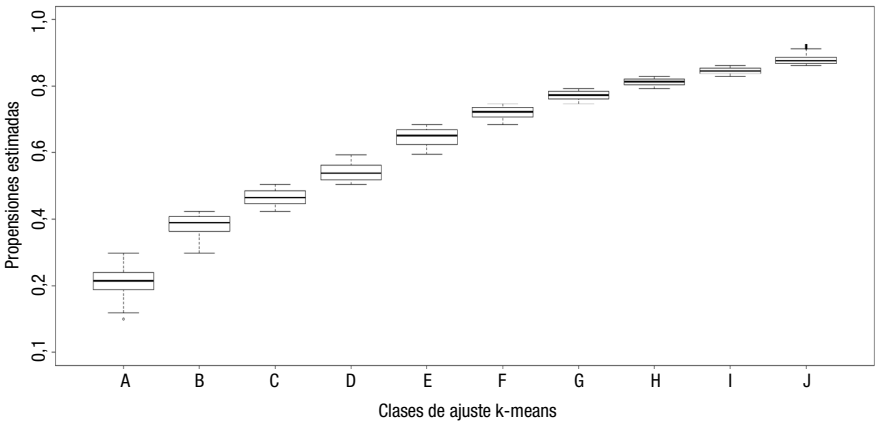
Se aplicó un método de agrupamiento denominado *k-means,* que solo requiere especificar la cantidad de clases de ajuste a crear. En el caso particular de este estudio se definieron 10 clases de ajuste y para cada una de ellas se determinó un factor de ajuste por no respuesta que toma el mismo valor para todas las unidades que pertenecen a la misma clase. Este factor fue calculado como la inversa de la tasa de respuesta de cada celda o clase de ajuste a la que pertenece cada unidad. En este trabajo, la tasa de respuesta de referencia para cada celda quedó determinada por la mediana de las propensiones de respuesta estimadas dentro de la clase a la que pertenece cada unidad que responde.

Tabla 4. Resultados del ajuste de los pesos de la ENCoPraC en base al modelo de propensión a responder

Clases de ajuste	Propensión mediana de la clase	Factor de ajuste	Cantidad de observaciones
A	0,21	4,68	477
B	0,39	2,57	541
C	0,47	2,15	981
D	0,54	1,86	784
E	0,65	1,53	541
F	0,72	1,38	953
G	0,77	1,29	1.969
H	0,81	1,23	3.054
I	0,84	1,18	3.838
J	0,88	1,14	2.825

Fuente: INDEC, estimaciones con base en la *Encuesta Nacional sobre Consumos y Prácticas de Cuidado 2022*.

Figura 2. Distribución de la propensión a la respuesta para la ENCoPraC según celdas de ajuste



Fuente: INDEC, estimaciones con base en la *Encuesta Nacional sobre Consumos y Prácticas de Cuidado 2022*.

En resumen, a partir del procedimiento descripto se obtiene el factor de ajuste por no respuesta para el individuo k que responde a la ENCoPraC, que viene dado por:

$$a_k = \frac{1}{\bar{\hat{p}}_k}$$

Donde $\bar{\hat{p}}_k = med\{\hat{p}_j, j \in C_k\}$ siendo C_k la clase homogénea definida por el agrupamiento al cual pertenece la unidad k dentro del conjunto S_r de todas las unidades (personas) que responden a la encuesta.

4.5 Factor de ajuste por calibración

Los factores de expansión de cada persona seleccionada que responde a la encuesta hasta esta instancia, $\tilde{w}_k^P = w_k^0 . a_k$, reciben una última modificación o ajuste, denominado *calibración*. Este procedimiento emplea información auxiliar de una fuente externa disponible y tiene por objetivo: i) contribuir a una mejora en los ajustes ya realizados y ii) corregir la posible sub o sobre representación en algunos grupos de la población, originadas cuando no son 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 busca definir estimadores más eficientes que el habitual estimador de expansión simple en términos del error muestral, dado que aprovechan la correlación que pueda existir entre las características indagadas por la encuesta y la información provista por la fuente externa.

El proceso de calibración que opera sobre el conjunto de personas que responden a la encuesta y genera el sistema de ponderadores definitivos, w_k^P, se puede traducir en el siguiente problema numérico de optimización:

$$\textit{minimizar, } \sum_R G(\widetilde{w}_k^P, w_k^P),$$

$$\textit{sujeto a: } \sum_R w_k^P \boldsymbol{x_k^P} = \sum_U \boldsymbol{x_q^P}$$

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, $\mathbf{x}_{\mathrm{k}}^{\mathbf{P}} = \left(x_{k1}^P, \ldots, x_{kq}^P\right)^T$ medidas en la encuesta, a partir de los factores de expansión deseados, w_k^P, reproduzcan sus totales poblacionales, $\sum_U \mathbf{x}_{\mathrm{q}}^{\mathrm{P}} = (t_{x1}^P, \ldots, t_{xq}^P)$, 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 λ_k para cada persona con respuesta.

$$w_k^P = w_k^0 . a_k . \lambda_k = \widetilde{w}_k^P \lambda_k$$

donde

w_k^0 es el factor de expansión inicial de la ENCoPraC para cada individuo,

a_k es el factor de ajuste por no respuesta para el individuo k que responde a la ENCoPraC,

λ_k es el factor de ajuste que surge de la calibración correspondiente a la persona seleccionada.

En la ENCoPraC 2022 se emplearon q variables que reflejan la estructura demográfica por sexo y grupo de edad, donde $\mathbf{x}_{\mathrm{k}}^{\mathbf{P}} = (x_{k\mathbf{1}}^P, \ldots, x_{k\boldsymbol{q}}^P)$ y cuyas componentes son:

$x_{\mathrm{k}\,\mathbf{1}}^P = 1$ si la persona tiene entre 16 y 29 años, y 0 en otro caso;

$x_{\mathrm{k}\,\mathbf{2}}^P = 1$ si la persona tiene entre 30 y 64 años, y 0 en otro caso;

$x_{\mathrm{k}\,\mathbf{3}}^P = 1$ si la persona tiene entre 65 y 75 años, y 0 en otro caso;

$x_{k\,\mathbf{4}}^P = 1$ si la persona es varón y tiene entre 16 y 75 años, y 0 en otro caso;

$x_{k\,\mathbf{5}}^P = 1$ si la persona es mujer y tiene entre 16 y 75 años, y 0 en otro caso;

$x_{\mathrm{k}\,\mathbf{6}}^P = 1$ si la persona tiene entre 16 y 75 años y pertenece a la región Gran Buenos Aires, y 0 en otro caso;

$x_{k\,7}^{P} = 1$ si la persona tiene entre 16 y 75 años y pertenece a la región Cuyo, y 0 en otro caso;

$x_{k\,8}^{P} = 1$ si la persona tiene entre 16 y 75 años y pertenece a la región Noreste, y 0 en otro caso;

$x_{k\,9}^{P} = 1$ si la persona tiene entre 16 y 75 años y pertenece a la región Noroeste, y 0 en otro caso;

$x_{k\,10}^{P} = 1$ si la persona tiene entre 16 y 75 años y pertenece a la región Pampeana, y 0 en otro caso;

$x_{k\,11}^{P} = 1$ si la persona tiene entre 16 y 75 años y pertenece a la región Patagonia, y 0 en otro caso;

$x_{k\,12}^{P} = 1$ si la persona tiene entre 16 y 75 años y es ocupada, y 0 en otro caso;

$x_{k\,13}^{P} = 1$ si la persona tiene entre 16 y 75 años y es desocupada o inactiva, y 0 en otro caso;

$x_{k\,14}^{P} = 1$ si la persona tiene entre 16 y 75 años con nivel educativo primario completo o menor, y 0 en otro caso;

$x_{k\,15}^{P} = 1$ si la persona tiene entre 16 y 75 años con secundario completo o incompleto, y 0 en otro caso;

$x_{k\,16}^{P} = 1$ si la persona tiene entre 16 y 75 años con nivel educativo superior universitario completo o incompleto, y 0 en otro caso;

$x_{k\,17}^{P} = 1$ si la persona tiene entre 16 y 75 años y pertenece a un hogar unipersonal, y 0 en otro caso;

$x_{k\,18}^{P} = 1$ si la persona tiene entre 16 y 75 años y pertenece a un hogar con dos o tres miembros, y 0 en otro caso;

$x_{k\,19}^{P} = 1$ si la persona tiene entre 16 y 75 años y pertenece a un hogar con cuatro miembros o mas, y 0 en otro caso.

Los totales de población, involucrados como marginales para estas variables en el proceso iterativo, provienen de proyecciones poblacionales y estimaciones a partir de la muestra de la EPH tercer trimestre 2022.

Para la calibración en la ENCoPraC 2022, se emplea la función de distancia *logit* (Deville y Särndal, 1992; Haziza y Beaumont, 2017) del paquete *survey* de R (Lumley, 2018) que permite controlar el rango de los w_k^P, y sus valores extremos. De esta forma, se busca limitar el riesgo de incrementar el error de muestreo en las estimaciones de la encuesta.

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_k^P finales que se emplean para todas las estimaciones oficiales de la encuesta.

La falla de los supuestos, debido a la no respuesta o porque los totales empleados para calibrar reflejan una población muy diferente de los respondentes, lleva a distorsionar los factores de expansión individuales, introduciendo probablemente más error y/o sesgo en las estimaciones (Deville & Särndall, 1992).

También pueden existir individuos cuyo ajuste por calibración resulte extremo, lo que puede generar un aumento injustificado en los errores estándares de los estimadores ajustados para los indicadores de interés. Es por ello que se aplicó un proceso de truncamiento de los factores de expansión a fin de reducir este efecto en los indicadores.

4.6 Truncamiento de los factores de expansión

Como paso final y teniendo en cuenta la distribución resultante de los pesos calibrados, se definió una cota superior con el objetivo de no permitir pesos extremos que puedan llegar a influir en el análisis o producir estimaciones inestables.

En primera instancia se analizaron los resultados de los pesos calibrados y se calcularon los cuartiles de su distribución empírica clasificados por región estadística: Q_{1r}= primer cuartil de los pesos de la región r; Q_{2r} = mediana o segundo cuartil de los pesos de la región r; Q_{3r} = tercer cuartil de los pesos de la región r. Luego, para cada región estadística se definió una cota superior para los pesos calibrados definida como $Q_{2r} + 5 \times (Q_{3r} - Q_{1r})$.

Todo peso que excede esta cota fue reemplazado por ese valor y el excedente de los pesos por encima de ella se distribuyó equitativamente entre los pesos que se encuentran por debajo. Esto permite que no se alteren los totales calibrados por región evitando la presencia de observaciones excesivamente influyentes debido a pesos extremos.

5. Estimación a partir de los datos de la encuesta

Se denomina estimación al 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. Los parámetros poblacionales que resultan de interés para estimar son, por lo general, descriptivos, y la mayoría se puede definir a partir de totales: los promedios, las proporciones y las razones o tasas. No obstante, puede haber interés en otros que involucran, por ejemplo, parámetros estadísticos de orden o más complejos.

Para alcanzar las estimaciones de esos parámetros en la ENCoPraC 2022, se emplean estimadores que recurren a los factores de expansión finales a nivel de personas w_k^P. Los factores de expansión finales utilizados son los que se obtienen como resultado de todos los ajustes que incorpora el factor *pondera,* a los que se suman los ajustes descriptos en el punto 4. Los estimadores pertenecen 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, la expresión de los estimadores más empleados es:

Parámetro	Estimador[13]
Total, $\boldsymbol{t_y}$	$\hat{t}_y = \sum_R w_k^P y_k,$
Promedio[14], $\boldsymbol{y}$	$\hat{\bar{y}} = \frac{\sum_R w_k^P y_k}{\sum_R w_k^P}$
Proporción, $\boldsymbol{p}$	$\hat{p} = \frac{\sum_R w_k^P y_k}{\sum_R w_k^P}$
Razón, $\boldsymbol{R_{yz} = \frac{t_y}{t_z}}$	$\hat{R}_{yz} = \frac{\hat{t}_y}{\hat{t}_z} = \frac{\sum_R w_k^P y_k}{\sum_R w_k^P z_k}$

5.1 Dominios de estimación

En su planteo original, la ENCoPraC se llevaría a cabo en el mismo ámbito geográfico que la EPH total urbano y contemplaría los mismos dominios de estimación para la publicación de resultados. Sin embargo, las dificultades registradas para alcanzar niveles de respuesta mínimos aceptables respecto al total de las personas seleccionadas para participar de la ENCoPraC motivaron a limitar el nivel de desagregación de los resultados. En un conjunto de ámbitos geográficos que originalmente conformarían dominios de estimación (regiones estadísticas, provincias) se observó que más del 50% de las áreas MMUVRA registraba un nivel de no respuesta superior al 60%. La magnitud de este fenómeno, que a su vez se concentra de forma dispar en las distintas regiones y localidades del país, incrementa el riesgo al sesgo por no respuesta en las estimaciones y por lo tanto no es posible brindar resultados confiables para dominios regionales, provinciales o a nivel de aglomerados urbanos particulares. En consecuencia, los resultados de la encuesta están referidos a la población de 16 a 75 años que reside en el conjunto de los 31 aglomerados urbanos de la EPH.

[13] En todos los casos, $\sum_R$ en las fórmulas hace referencia a sumar sobre las personas que responden a la encuesta.
[14] 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.

6. Indicadores de calidad de las estimaciones e implicancias de la no respuesta sobre la estimación de errores de muestreo

Uno de los rasgos distintivos del relevamiento de una encuesta a partir de una muestra probabilística es la posibilidad de cuantificar el error asociado a las estimaciones surgidas de dicha muestra, lo que permite evaluar el proceso de análisis en términos de precisión y confiabilidad. Contar con estos indicadores de calidad permite a los usuarios cuantificar el grado de confianza y conocer las limitaciones que pueden llegar a tener los resultados, para así restringir el uso de estos cuando las estimaciones no alcanzan ciertos estándares definidos para la encuesta.

El estimador habitual que emplea el INDEC para estimar varianzas de los estimadores de parámetros es el método propuesto por Rao y Wu (1988) y Rao, Wu y Yue (1992)[15]. En su formulación teórica, está propuesto para diseños estratificados multietápicos, con UPM seleccionadas mediante probabilidad proporcional a un tamaño (PPT) con reemplazo, y asumiendo una expresión para la varianza bajo un diseño con reposición con el supuesto de *último conglomerado*. Este último supuesto sostiene que la primera etapa de muestreo (UPM) brinda la información necesaria para alcanzar una estimación del error por muestra, ignorando las restantes etapas definidas en el diseño. Sin embargo, la adopción de estos supuestos habilita emplear este método como un estimador de varianza para un diseño PPT sin reemplazo, si la selección de las UPM sin reemplazo es más eficiente que la selección de UPM con reemplazo (West, 2012; Särndal, Swensson y Wretman, 1992), como es el caso de la encuesta ENCoPraC, lo que convierte el proceso inferencial en conservador y válido para la encuesta.

En primer lugar, vale destacar que el método propuesto se basa en la información contenida en la primera etapa de muestreo. En los aglomerados EPH, las unidades de primera etapa de selección (UPM) corresponden a las áreas de relevamiento, mientras que en el resto urbano de cada provincia las UPM son las localidades, y en la segunda etapa se seleccionan áreas. Cada una de estas unidades se encuentran estratificadas tal como se ha mencionado en apartados anteriores. Si bien los supuestos sólo consideran las UPM, es de esperar que estas contengan un aceptable nivel de respuesta, dado que son las que brindan la información para la estimación de varianza. Tal como fue descripto en apartados previos, y habiendo encontrado un porcentaje elevado de áreas con niveles de respuesta muy por debajo de los estándares de calidad, en la mayoría de los dominios geográficos se cuenta con un alto porcentaje de unidades deterioradas. Este resultado refleja que la información contenida en las áreas es insuficiente para estimar los errores de muestreo.

Una de las soluciones que se propone en la bibliografía, ante la ocurrencia de una alta no respuesta en un número bajo a moderado de unidades, es la unión de estas últimas hasta alcanzar un mínimo establecido de respuestas en estas nuevas pseudo unidades. Sin embargo, no es la situación de esta encuesta ya que el problema es generalizado.

La combinación de unidades resulta aceptable cuando los grados de libertad del estimador de variancia no disminuyen en forma abrupta, de otro modo, la pérdida de

[15] Para más detalles consultar: https://www.indec.gob.ar/indec/web/Institucional-Indec-Metodologias-2

grados de libertad generaría inestabilidad en estos estimadores. Vale recordar que una propuesta para determinar de forma aproximada los grados de libertad disponibles en cada ámbito de interés es calcular la suma de las UPM menos el número de estratos.

Por todos estos motivos no resulta posible aplicar algún método de estimación de varianzas, ya sea a partir del cálculo de réplicas, tal como se adopta de forma estándar para otras encuestas a hogares puntuales, u otro. No se cumplen los requisitos básicos que impone el método y las soluciones existentes para estos casos provocarían inestabilidad y subestimación en los estimadores de variancia.

7. Recomendaciones sobre las estimaciones

Dadas las limitaciones expresadas en el punto anterior, se buscó una alternativa heurística para suplir, de forma aproximada e imperfecta, la no disponibilidad de errores de muestreo. En su lugar, se utilizó un conjunto de criterios prácticos para orientar la interpretación de los resultados presentados y evaluar la pertinencia de su difusión en este informe, criterios que a su vez se ponen a disposición de los usuarios que deseen realizar sus propios cálculos a partir de la base usuaria de la encuesta.

El siguiente punteo resume un conjunto de criterios que deberán verificarse para considerar que un resultado publicado es aceptable desde el punto de vista de su solvencia estadística. Están basados en la magnitud del fenómeno que se desea medir y la cantidad de casos muestrales involucrados en los cálculos de interés. En el caso particular de prevalencias e incidencias obtenidas a través del cálculo de razones, se evalúa tanto el tamaño de la subpoblación a la que se refiere el resultado (denominador) como a la cantidad de individuos en esa subpoblación que presenta el atributo de interés (numerador).

Para que un resultado de la ENCoPraC pueda considerarse aceptable se debe verificar:

1. que en el caso de resultados asociados al cálculo de prevalencias, razones o proporciones, su resultado numérico sea superior a 0,05 (o de forma equivalente, al 5%)

2. que la cantidad total de casos involucrados en el cálculo sea superior a 200 individuos.

3. que en el caso de razones o proporciones, el atributo de interés esté presente en más de 150 individuos. Si el atributo de interés está presente en más de 50 individuos pero en menos de 150, se considerará que el resultado es de aceptabilidad dudosa.

Como resultado de la combinación de los criterios anteriores, se recomienda la siguiente estrategia:

Tabla 5. Resumen de criterios para la publicación de resultados de la ENCoPraC

Criterio 1: razones o proporciones mayores a 0.05	Criterio 2: cálculos basados en 200 casos por celda o más.	Criterio 3: frecuencia de casos que presenta el atributo de interés		
		Menos de 50 casos	*Entre 50 y 150 casos*	*Más de 150 casos*
Verdadero	*Verdadero*	No publicable	Dudoso	Publicable
	Falso	No publicable	No publicable	No publicable
Falso	*Verdadero*	No publicable	No publicable	No publicable
	Falso	No publicable	No publicable	No publicable

Fuente: INDEC, *Encuesta Nacional sobre Consumos y Prácticas de Cuidado 2022*.

Por último, en el caso de que algunas de las estimaciones sean consideradas no publicables, y si aún así la persona usuaria desea incorporarlas en una publicación, se recomienda enfáticamente la inclusión de una advertencia, y que se haga referencia a las limitaciones del caso citando el presente documento, en particular el cuadro anterior definido por el INDEC como estándar para esta encuesta. Por otro lado, a raíz de las dificultades ya mencionadas, se sugiere no utilizar los resultados de esta encuesta para la estimación de totales. A su vez, también se recomienda limitar la desagregación de resultados para grupos específicos de la población. Tal como ha sido expuesto, la determinación de los factores de expansión para esta encuesta implicó la aplicación sucesiva de cinco etapas de ajuste, las cuales no necesariamente operan en la misma dirección en todos los casos, especialmente cuando se pretende caracterizar subpoblaciones cada vez más pequeñas.

8. Referencias

American Association for Public Opinion Research (2016). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. (9° ed.). AAPOR. https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions-20169theditionfinal.pdf

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

Carlson, B. (2013). "Response Rates Revisited". *Proceedings American Statistical Associations. Survey. Research Methods Section*, JSM 2013, pp. 1200-1208. http://www.asasrms.org/Proceedings/y2013/files/308173_80404.pdf

Cassel, C. M., Särndal, C. E., & Wretman, J. H. (1983). *Some uses of statistical models in connection with the nonresponse problem. Incomplete data in sample surveys*, 3, 143-160.

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), pp. 23-37. Statistics Canada. http://www.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=12-002-X20040016890&lang=eng

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

Frankel, L. R. (1983). "The Report of the CASRO Task Force on Response Rates". En Wiseman, Frederick (ed.). *Improving Data Quality in a Sample Survey*. 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). 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, 206-226. DOI:10.1214/16-STS608

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

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

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

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

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

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

Rubin, D. B. (1976). “Inference and missing data”. *Biometrika*, Volume 63, Issue 3, December 1976, Pages 581–592. https://doi.org/10.1093/biomet/63.3.581

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

Valliant, R., Dever, J. A. y Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. Springer. DOI: 10.1007/978-1-4614-6449-5_14

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

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

9. ANEXO. Tasa de respuesta de la ENCoPraC 2022 por jurisdicción

Tabla A1. Cantidad de viviendas elegibles iniciales, personas respondentes a la ENCoPraC y tasa de respuesta global de la ENCoPraC. Resultados para el conjunto de los 31 aglomerados urbanos según jurisdicción

Jurisdicción	Hogares elegibles iniciales (**)	Hogares con personas respondentes ENCoPraC	Tasa de respuesta global ENCoPraC (*)
Ciudad Autónoma de Buenos Aires	1.541	481	31,2%
Buenos Aires	6.388	2.581	40,4%
Catamarca	500	358	71,6%
Córdoba	1.363	917	67,3%
Corrientes	517	263	50,9%
Chaco	625	340	54,4%
Chubut	1.043	587	56,3%
Entre Ríos	1.154	537	46,5%
Formosa	580	314	54,1%
Jujuy	521	410	78,7%
La Pampa	569	246	43,2%
La Rioja	521	390	74,9%
Mendoza	810	516	63,7%
Misiones	528	329	62,3%
Neuquén	514	227	44,2%
Río Negro	398	190	47,7%
Salta	763	467	61,2%
San Juan	677	379	56,0%
San Luis	542	331	61,1%
Santa Cruz	396	181	45,7%
Santa Fe	1.664	976	58,7%
Santiago el Estero	575	301	52,3%
Tucumán	757	513	67,8%
Tierra del Fuego	520	228	43,8%
Total	**23.466**	**12.062**	**51,4%**

Fuente: INDEC, *Encuesta Nacional sobre Consumos y Prácticas de Cuidado 2022*.
(*) La tasa de respuesta global para la EnCoPraC se calcula como cociente entre la cantidad de personas que respondieron la EnCoPraC con respecto a la cantidad de viviendas elegibles.
(**) Dentro del conjunto de viviendas inicialmente elegibles para este relevamiento hay un conjunto de casos en los que no se logró establecer contacto con ninguna persona y por lo tanto no es posible determinar si allí habitaba alguna persona de 16 a 75 años. A su vez, se asume que en cada vivienda habita un único hogar. Por esta razón, en algunos casos la tasa global de respuesta podría estar levemente subestimada.

10. Glosario

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

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

Área MMUVRA: unidad de área que coincide en general con el radio censal definido sobre la base cartográfica del Censo Nacional de Población y Viviendas 2010. Sin embargo, también puede estar determinada por un agrupamiento de radios contiguos para ajustarse a requerimientos de tamaño en términos de viviendas; por recortes operativos en algunos radios por baja densidad de viviendas, economía de recursos, o 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 tales como 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 conforman una población y medir una o más características de ellos. Puede brindar información con un

nivel de desagregación geográfico y detalle muy alto. Se lo puede considerar como una muestra al 100% de la población. Debido a esta característica, los resultados que se obtienen están libres de error muestral; no así de errores ajenos al muestreo (tales como no respuesta, cobertura, medición, procesamiento, u otras fuentes siempre presentes en una operación estadística).

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

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

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

Diseño muestral: marco metodológico y de trabajo que sirve de base para la selección de la muestra, y que afecta a otros aspectos importantes de un estudio o encuesta. 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 los métodos involucrados 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 que se deben emplear para obtener los resultados a partir de los datos de 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 confeccionarlo es costoso.

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

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

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

Elegibilidad: refiere a si una unidad de la muestra es parte de la población objetivo o no. Errores en la determinación de la elegibilidad afectan directamente 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 a 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. 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 variancia muestral del estimador y el sesgo al cuadrado.

Error de cobertura: error producido por diferencias entre la población objetivo y la población que cubre el marco muestral. Puede 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 deberían tener, o dicha probabilidad es más alta de la que les corresponde. 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 debido 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 las personas encuestadas 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 quien responde; 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 se emplea este término 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 tratado 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, o 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, cuando se extiende sobre más de una jurisdicción (también "aglomerado"). Para la MMUVRA, todas las localidades de 2.000 o más habitantes, según el Censo Nacional de Población y Viviendas 2010, conforman las UPM del marco de muestreo adoptado para el diseño muestral.

Marco de muestreo: cualquier lista o recurso que delimita, identifica y permite el 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 juega 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 no existe una lista de unidades de muestreo finales, 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 la 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 las réplicas, 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 el 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 realizan actualizaciones periódicas de sus listados de viviendas y de su cartografía asociada.

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

Muestra aleatoria: ver **Muestra probabilística**.

Muestra maestra: muestra aleatoria de gran tamaño donde permanecen fijas 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 con la experiencia o el juicio del investigador; es decir, no involucra un proceso de selección aleatorio.

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

Muestreo: proceso o conjunto de procesos que permiten seleccionar un número no nulo de elementos entre 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, junto 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 en cuanto a las principales características a estudiar en la población, con este tipo de muestreo las estimaciones ganan en precisión, en comparación con las alcanzadas a partir de 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 el conglomerado. En encuestas a hogares, esta alternativa de muestreo permite disminuir los costos de la encuesta, generalmente en perjuicio de la precisión en las estimaciones al depender de la homogeneidad interna entre las unidades con respecto a las características que se están estudiando.

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

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

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

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

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 (por ejemplo, totales, medias, proporciones, varianzas, etcétera).

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

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

Ponderador: ver **Factor de expansión**.

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

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

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

Radio censal: unidad de área 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 con 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 cien por ciento 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 con 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 ese 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.