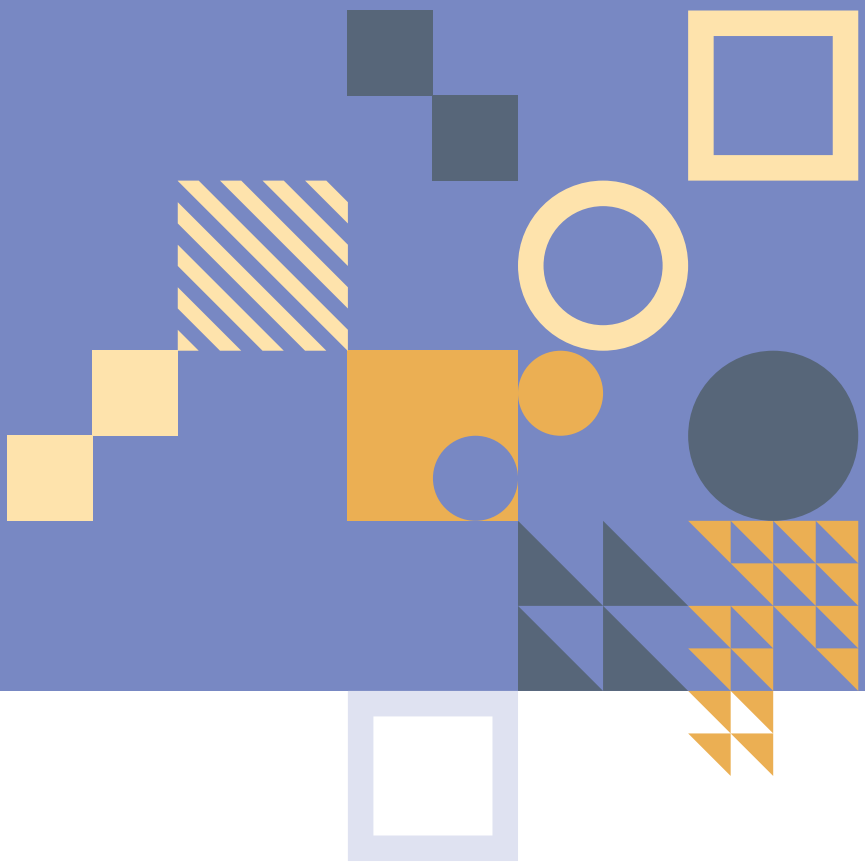
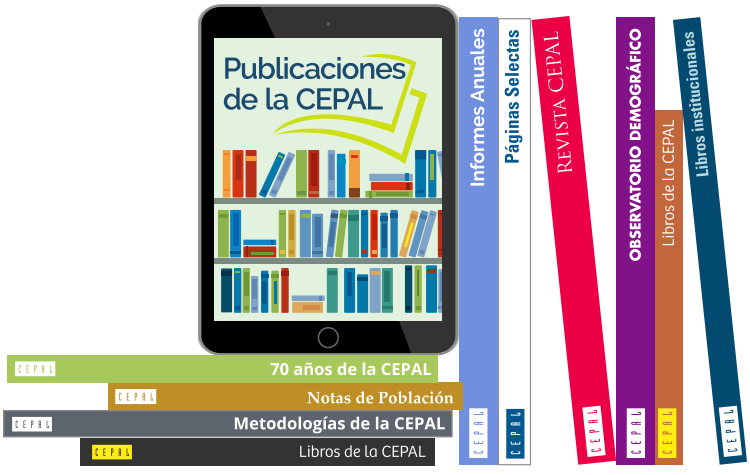
Metodologías de la CEPAL

Diseño y análisis estadístico de las encuestas de hogares de América Latina



NACIONES UNIDAS
CEPAL
75 años
Trabajando por un futuro productivo, inclusivo y sostenible

Gracias por su interés en esta publicación de la CEPAL



Si desea recibir información oportuna sobre nuestros productos editoriales y actividades, le invitamos a registrarse. Podrá definir sus áreas de interés y acceder a nuestros productos en otros formatos.

Deseo registrarme

NACIONES UNIDAS

CEPAL

www.cepal.org/es/publications

www.instagram.com/publicacionesdelacepal

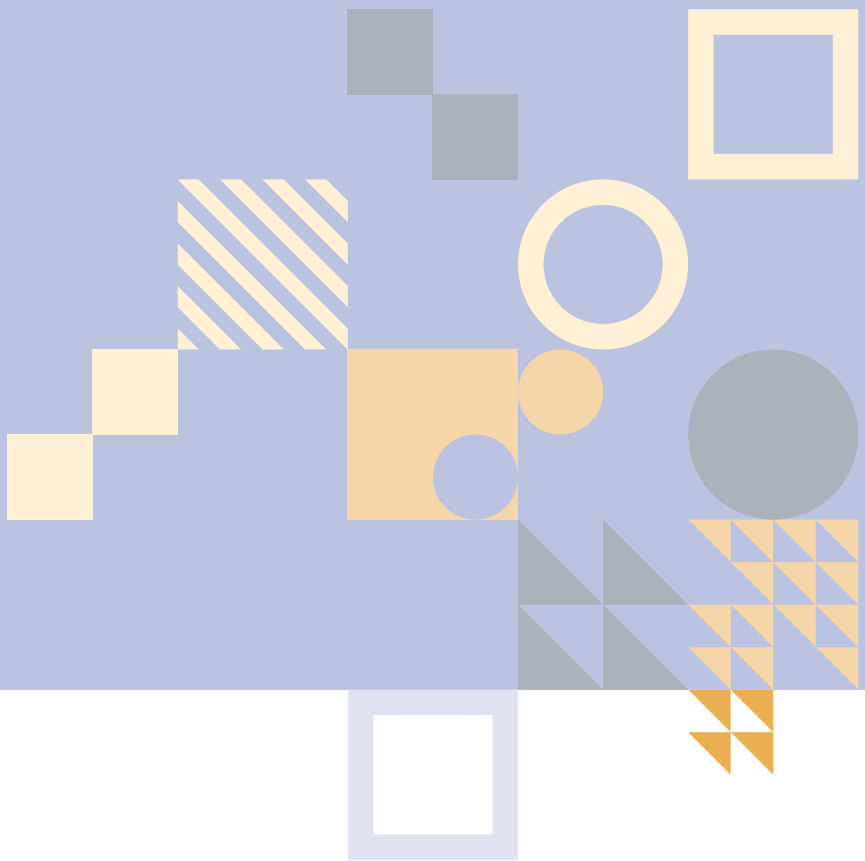
www.facebook.com/publicacionesdelacepal

www.issuu.com/publicacionescepal/stacks

www.cepal.org/es/publicaciones/apps

Metodologías de la CEPAL

Diseño y análisis estadístico de las encuestas de hogares de América Latina



NACIONES UNIDAS
CEPAL
75 años
Trabajando por un futuro productivo, inclusivo y sostenible

Este documento fue preparado por Andrés Gutiérrez, Experto Regional en Estadísticas Sociales de la División de Estadísticas de la Comisión Económica para América Latina y el Caribe (CEPAL), bajo la coordinación de Rolando Ocampo, Director, y Xavier Mancero, Jefe de la Unidad de Estadísticas Sociales, ambos de la misma División. Su elaboración contó con el apoyo del proyecto "Inequality: Innovative approaches for examining inequality through integration of different data sources in Latin America and the Caribbean" del 13er tramo de la Cuenta de las Naciones Unidas para el Desarrollo.

Se agradecen los aportes y las sugerencias del personal de las oficinas nacionales de estadística de los países receptores de las asistencias técnicas llevadas a cabo entre 2017 y 2022. Se agradece, asimismo, a Simone Cecchini, Director del Centro Latinoamericano y Caribeño de Demografía (CELADE)-División de Población de la CEPAL, y a Hanwen Zhang, Profesora Asistente de la Universidad Autónoma de Chile, por la lectura detallada de este documento y sus comentarios.

Publicación de las Naciones Unidas
ISBN: 978-92-1-122127-5 (versión impresa)
ISBN: 978-92-1-002242-2 (versión pdf)
ISBN: 978-92-1-358431-6 (versión ePub)
Número de venta: S.23.II.G.13
LC/PUB.2023/14-P
Distribución: G

Impreso en Naciones Unidas, Santiago
S.22-01034

Esta publicación debe citarse como: Comisión Económica para América Latina y el Caribe (CEPAL), *Diseño y análisis estadístico de las encuestas de hogares de América Latina,* Metodologías de la CEPAL, N° 5 (LC/PUB.2023/14-P), Santiago, 2023.

Índice

Cuadros

Gráficos

Diagramas

Introducción

Este documento contiene una revisión de algunas de las metodologías más utilizadas por las oficinas nacionales de estadística (ONE) de América Latina en cuanto al diseño y análisis estadístico de las encuestas de hogares y puede servir de guía técnica a los estadísticos de la región que participan en los procesos técnicos de este tipo de encuestas. Se consideran conjuntamente los dos principales momentos de las encuestas: el diseño y el análisis. Nótese que estos momentos están escindidos por el levantamiento de la información sobre el terreno y dividen en dos la realización de la encuesta. Los lectores que están familiarizados con la investigación social mediante las encuestas de hogares encontrarán que estas operaciones estadísticas se planean teniendo en cuenta muchos pormenores que podrían suceder sobre el terreno. Por ese motivo, el trabajo de los investigadores en las ONE consiste en hacer que el mismo diseño que se planificó se plasme en la información recopilada en la base de microdatos de las encuestas. En un segundo momento es cuando se debe asegurar que lo que se planificó quede efectivamente incorporado en el análisis de la información.

Desde esta perspectiva, el documento se divide en tres partes sustantivas que definen la planeación y el análisis de la mayoría de las encuestas de hogares en América Latina y el Caribe. La primera parte, en los capítulos I a VIII, se refiere a la planeación de una encuesta y a la definición del diseño estadístico, que comprende, entre otras cosas, la obtención de una medida de probabilidad discreta que permitirá realizar la inferencia basada en el principio de representatividad. En esta parte se abordan con más detalle los elementos básicos que se consideran por lo regular en los diseños de las encuestas de hogares. Un aspecto relevante es que, si bien aquí se considera que las encuestas de hogares tienen muchos elementos en común, se diferencian de forma cuidadosa las particularidades de cada tipo de encuesta. Por ejemplo, se trata el tema del diseño de las encuestas rotativas y se profundiza en los diferentes parámetros que se pueden considerar en este tipo de operaciones. Asimismo, se describen las características metodológicas que se deben considerar al diseñar la encuesta y se revisan los conceptos esenciales que determinarán

el tipo de aplicación que se debe considerar. También se describen los principales diseños de muestreo que se utilizan en este tipo de estudios y se exponen de forma estándar los conceptos de estratificación y aglomeración de las poblaciones. Estos conceptos se complementan con varias aplicaciones prácticas con miras a determinar el tamaño de muestra adecuado para lograr los objetivos de una investigación planeada en base a las encuestas de hogares. A pesar de que la literatura relacionada con la práctica del muestreo es relativamente abundante, existen pocos ejemplos prácticos que logren representar la problemática del tamaño de muestra, y el lector podrá encontrar herramientas ilustrativas basadas en múltiples escenarios de la problemática social.

En la segunda parte, en los capítulos del IX al XV, se abordan los principios metodológicos para el correcto procesamiento de las encuestas transversales, analizadas como representación de un momento específico en el tiempo. Se examinan con detenimiento los procesos de ponderación en la encuesta y de generación de los factores de expansión que se aplicarán a la información contenida en la base de datos para poder obtener las inferencias adecuadas a nivel nacional o regional. Si hay algo que distingue el análisis de las encuestas de cualquier otro tipo de estudio estadístico es que las propiedades importantes (como insesgamiento, consistencia y eficiencia) se basan en el diseño de muestreo y no en supuestos metodológicos ligados a algún modelo estocástico. Además de analizar las principales metodologías de estimación, se presta especial atención a la estimación del error de muestreo, que no es otra cosa que una función de la varianza de las estimaciones, y se presentan las metodologías más comunes en términos de aproximaciones teóricas y computacionales al error de muestreo. También se abordan los procesos de imputación y ausencia de respuesta, con el objetivo de recuperar tanta información como sea posible para que el investigador pueda contar con una base de datos rectangular y completa. En aquellos casos en que la imputación no resulta ser una técnica adecuada para completar la información faltante, es necesario realizar ajustes sistemáticos en los factores de expansión para que la muestra efectiva siga siendo representativa de toda la población. En el último capítulo de la segunda parte se muestran algunos enfoques útiles en la detección de datos atípicos y la mitigación del impacto del error no muestral en las respuestas obtenidas.

En la tercera y última parte, correspondiente a los capítulos XVI a XIX, se hace una aproximación a los principios del procesamiento avanzado en un sistema integrado de encuestas de hogares que permite la agregación y combinación de diferentes oleadas de las encuestas. Ello permite realizar inferencias en un lapso más amplio. De esta forma, se presentan con detalle los procesos que se surten cuando se agregan encuestas a lo largo del tiempo. En el caso de las encuestas que se definen a partir de estructuras rotativas, se presenta un enfoque metodológico que permite crear bases de datos longitudinales (tipo panel) para la estimación de flujos brutos, entre otros. Con la perspectiva de la CEPAL, también se presentan algunos de los criterios de calidad que se deberían tener en cuenta para decidir si una cifra, resultante de un proceso de estimación estadística basada en encuestas de hogares, debería ser o no dada a conocer a la sociedad.

Por último, en los anexos del documento se presenta una discusión acerca del uso presente de las encuestas de hogares y los retos que depara el futuro en materia de medición de indicadores sociales a través de las encuestas de hogares. Asimismo, se contempla una revisión del *software* que se utiliza actualmente en las ONE para llevar a cabo esta ardua tarea de diseñar y analizar las encuestas de hogares, se hace un repaso rápido de algunas de las encuestas de la región y se esbozan algunas directrices que se deberían considerar al documentar los procesos asociados a las encuestas de hogares.

Los profesionales y metodólogos de las ONE de América Latina y el Caribe podrán encontrar aquí una guía práctica sobre todas las cuestiones técnicas relacionadas con el devenir de una encuesta de hogares. Estas cuestiones abarcan la planeación, pasando por el análisis básico, y se adentran en los análisis complejos propios de una recopilación continua. Esta publicación procura recopilar las mejores herramientas técnicas y actualizar los métodos usados por las ONE en el diseño y rediseño de las encuestas de hogares. En la literatura también existen esfuerzos realizados por las Naciones Unidas para llevar la teoría básica de las encuestas de hogares al español. En ese sentido, cabe mencionar dos publicaciones: Naciones Unidas (2007 y 2008). Sin embargo, este documento complementa esos valiosos recursos al ahondar en los temas estadísticos relevantes para una mejor implementación de los procesos técnicos relacionados con las encuestas de hogares. Por tanto, aunque no es una condición necesaria, se recomienda que el lector tenga una amplia experiencia en el diseño y análisis de encuestas de hogares para su mejor aprovechamiento.

A. ¿Qué es una encuesta?

En Groves y otros (2009) se afirma que una encuesta es un método sistemático para obtener información de (una muestra de) entes, con el fin de construir descriptores cuantitativos de los atributos de una población más grande, de la que los entes son miembros.

Los datos se obtienen a partir de un conjunto de preguntas normalizadas dirigidas a una muestra representativa o al conjunto total de la población estadística objeto de estudio, formada a menudo por personas, empresas o entes institucionales, con el fin de conocer estados de opinión, características o hechos específicos. Hay una diferencia sustancial entre el sondeo (*poll*) y la encuesta (*survey*). En general, la primera expresión aparece más en el sector privado, en estudios de opinión y de consumo. Un sondeo difícilmente se utilizará para obtener estadísticas oficiales en estudios gubernamentales o en dominios científicos. Los sondeos muchas veces opacan la perspectiva científica de las cifras y pueden dar lugar a conclusiones inexactas acerca de la realidad de una problemática. Por otra parte, no todos los procesos de recopilación de información pueden llamarse "encuestas". A los efectos de este documento se aplicará la definición de Groves y otros (2009), quienes afirman que una encuesta tiene las siguientes características:

- los datos son recopilados mediante preguntas a personas;
- las respuestas son compiladas cuando: a) un encuestador pregunta y registra las respuestas del entrevistado, o b) el encuestado lee y registra sus propias respuestas, y
- los datos son recopilados de un subgrupo de personas pertenecientes a la población de interés.

Nótese que la última característica descrita implica una clara diferenciación con los censos de población y vivienda. Las encuestas de hogares son un caso particular de investigación social, en que se indaga acerca de características específicas a nivel del individuo, del hogar o de la vivienda, con el fin de obtener inferencias precisas acerca de constructos de interés. Por su naturaleza, estas investigaciones están relacionadas con variables de salud, educación, ingresos, gastos, situación laboral, acceso y uso de servicios, entre muchas otras. En algunas ocasiones, las encuestas de hogares tienen como objetivo la estimación de uno o varios indicadores que resumen un constructo económico o social. Estos pueden ser, por ejemplo, la tasa de pobreza, la tasa de desocupación o el gasto promedio en alimentación. Sin embargo, existe una tendencia creciente a extender las encuestas a constructos más diversos. Cada vez tienen más espacio las encuestas que incluyen diversos módulos en sus cuestionarios. Se conocen como encuestas de propósitos múltiples y se consideran una fuente relevante de información que permite monitorear indicadores sociales.

En esas encuestas, la unidad de análisis es el hogar, que ha sido definido por la División de Estadística de las Naciones Unidas (Naciones Unidas, 2011) como:

i) Un grupo de dos o más personas que se combinan con el fin de ocupar la totalidad o parte de una vivienda y proporcionarse alimentos y posiblemente otros artículos esenciales para la vida. El grupo puede estar compuesto solo de personas relacionadas o de personas no relacionadas o de una combinación de ambos. El grupo también puede compartir los ingresos.

ii) Una persona que vive sola en una vivienda separada o que ocupa, como huésped, una habitación (o habitaciones) separada de una vivienda, pero que no se une a ninguno de los otros ocupantes de la vivienda para formar parte de un hogar de múltiples personas.

Nótese que la anterior definición refleja una dinámica en los hogares, que constantemente se crean, desaparecen o se unen, por lo que es necesario abordar desde distintos enfoques la medición de indicadores sociales. En América Latina existe una gran variedad de encuestas que abordan diferentes problemáticas sociales. Todas ellas han sido diseñadas cuidadosamente para que respondan a las necesidades de la sociedad. En este documento se incluye una recopilación de las técnicas utilizadas tanto en su diseño como en su análisis.

No todas las encuestas se diseñan de la misma forma, por lo que debe haber una distinción entre ellas. Por ejemplo, Kalton y Citro (1993) afirman que las encuestas de hogares pueden clasificarse en varios tipos:

- **Encuestas repetidas,** definidas como una serie de encuestas transversales aplicadas en diferentes momentos con el mismo diseño metodológico, donde la selección de hogares se hace de forma independiente para cada aplicación.
- **Encuestas tipo panel,** en cuyo caso los datos son recopilados en diferentes momentos, utilizando la misma muestra de hogares en el tiempo.
- **Encuestas rotativas,** donde un porcentaje de hogares se mantiene en un período de tiempo respondiendo la encuesta y en cada aplicación algunos hogares son reemplazados por nuevos hogares de forma planificada.

El diseño de la encuesta dependerá sistemáticamente del objetivo de la medición. Por ejemplo, Kalton y Citro (1993) afirman que es prudente hacer una buena inversión en el desarrollo e implementación de un buen diseño para amortizar los costos de todo el estudio. Por lo tanto, lo que se quiere al diseñar una encuesta de hogares es que esta sea un instrumento confiable, que brinde estimaciones exactas y precisas. De lo contrario, no se podrían monitorear de manera consistente las políticas públicas y los indicadores de interés. Por ejemplo, uno de los indicadores sociales de mayor uso es la tasa de desocupación, que mide la razón entre la cantidad de personas que se encuentran desocupadas y las que forman parte del mercado de trabajo. Las encuestas de empleo tienen características particulares, diferentes a las de las encuestas que miden otro tipo de constructos. Duncan y Kalton (1987) mencionan que las encuestas de hogares pueden proveer estimaciones de los parámetros poblacionales en distintos puntos del tiempo. Entre otras cosas, es posible realizar la estimación de la tasa de desocupación mensual, proveer estimaciones del cambio neto de los parámetros poblacionales entre períodos de tiempo (por ejemplo, el cambio en la tasa de desocupación entre dos períodos consecutivos), o incluso medir varios componentes de cambio individual (por ejemplo, cambios brutos en la situación laboral de los jefes de hogar), para lo cual se requiere que la encuesta contemple un diseño de panel o de panel rotativo.

La medición de los indicadores en el mercado de trabajo es solo un pequeño componente en el vasto y amplio universo de posibilidades de medición que brindan las encuestas de hogares. Estos levantamientos se han convertido en una herramienta fundamental para medir indicadores sociales en todo el mundo. En particular, permiten a los países de América Latina hacer un seguimiento de su desarrollo económico y social. A continuación se introducen algunas temáticas de interés cuyo seguimiento depende en gran medida de la realización de encuestas de hogares.

B. Los Objetivos de Desarrollo Sostenible y otros indicadores regionales

Las encuestas de hogares pueden utilizarse como herramienta para monitorear el progreso de los países en términos de objetivos y metas comunes. En 2015, la Asamblea General de las Naciones Unidas aprobó una resolución en que se promueve un plan de acción en favor de las personas, el planeta y la prosperidad (Naciones Unidas, 2015). En esa resolución se propone el seguimiento de 17 Objetivos de Desarrollo Sostenible (ODS) y 169 metas, de carácter integrado e indivisible, que se conjugan en las dimensiones económica, social y ambiental. Para realizar el seguimiento de los ODS es posible utilizar diferentes fuentes de información, como censos, registros administrativos, registros estadísticos, proyecciones demográficas y encuestas de hogares (Naciones Unidas, 2016). En particular, cada una de las metas de los ODS contiene indicadores, muchos de los cuales no podrían estimarse de no ser por la información disponible en las encuestas de hogares.

Por ejemplo, el primer objetivo es poner fin a la pobreza en todas sus formas y en todo el mundo. La meta 1 de este objetivo es erradicar para todas las personas y en todo el mundo la pobreza extrema. Asimismo, la meta 2 es reducir al menos a la mitad la proporción de hombres, mujeres y niños de todas las edades que viven en la pobreza en todas sus dimensiones con arreglo a las definiciones nacionales. Para realizar una medición sistemática de la pobreza monetaria, las encuestas de hogares son un insumo fundamental. En una primera instancia se deben definir a nivel nacional los umbrales monetarios (líneas de pobreza) sobre los que se clasifican los hogares como pobres extremos, pobres relativos o no pobres. Estos umbrales vienen supeditados directamente a la realización de las encuestas de ingresos y gastos de los hogares, que se aplican cada cinco o diez años en los países de la región (CEPAL, 2018a). De forma sistemática, la medición de la pobreza monetaria se realiza con encuestas continuas que contienen módulos específicos de ingreso. En estos módulos se indaga sobre todas las fuentes de ingreso, tanto de las personas como del hogar. Con base en las líneas de pobreza definidas anteriormente, se clasifica a las personas en alguna de las categorías de pobreza.

De la misma manera, el objetivo 8 busca promover el crecimiento económico sostenido, inclusivo y sostenible, el empleo pleno y productivo y el trabajo decente para todos. De este objetivo se desprenden con claridad indicadores que permiten conocer la evolución de los países en la consecución de las metas. En este objetivo se encuentra la meta 8.6, que apunta a reducir considerablemente la proporción de jóvenes que no están empleados y no cursan estudios ni reciben capacitación. Esta meta se mide con el indicador 8.6.1, definido como la proporción de jóvenes (entre 15 y 24 años) que no cursan estudios, no están empleados ni reciben capacitación.

Se podrían enumerar más ejemplos en que las encuestas de hogares desempeñan un papel fundamental para la medición de los indicadores y metas de los ODS definidos en la

Agenda 2030 para el Desarrollo Sostenible. En este sentido, la División de Estadísticas de las Naciones Unidas ha establecido en un análisis preliminar que a partir de encuestas de hogares se puede obtener información sobre un total de 77 de los indicadores de los ODS, que abarcan 13 de los 17 objetivos. La mayor concentración de estos indicadores estaría en las áreas de salud, educación, igualdad de género, pobreza, hambre, trabajo y justicia.

C. Las estadísticas del mercado de trabajo

Desde otra perspectiva, en el marco de la 13a Conferencia Internacional de Estadísticos del Trabajo, celebrada en 1982, la Organización Internacional del Trabajo (OIT) adoptó algunas directrices concernientes a la medición y análisis de estadísticas oficiales de la fuerza de trabajo, del empleo y del desempleo, con miras a mejorar la comparabilidad de las cifras y mejorar su utilidad en los países (OIT, 1982). En esta resolución se hace un énfasis especial en que las encuestas de hogares constituyen un medio apropiado de recopilación de datos sobre la población económicamente activa y que la planeación de estas investigaciones en los países debería ceñirse a las normas internacionales. Por consiguiente, se afirma que las encuestas de hogares deberían:

- brindar datos de la población económicamente activa, definida como las personas en edad laboral que se han integrado al mercado de trabajo (trabajadores o personas que buscan empleo);
- proveer estadísticas básicas de sus actividades durante el año, así como las relaciones entre el empleo, el ingreso y otras características económicas y sociales, y
- proveer datos sobre otros temas particulares para responder a las necesidades a largo plazo y de índole permanente.

En 2013, la OIT decidió revisar esta resolución y propuso algunos cambios en el marco de la 19ª Conferencia Internacional de Estadísticos del Trabajo, donde se acogieron algunas modificaciones en lo que se refiere a los objetivos de medición y el alcance de los sistemas nacionales de estadísticas del trabajo, el concepto de trabajo en todas sus formas, el empleo, la medición de las personas en situación de subutilización de la fuerza de trabajo y los métodos de recopilación de datos, entre otras (OIT, 2013b). Las ONE de América Latina actualizan los instrumentos de medición de las encuestas de hogares para que puedan responder a los nuevos retos en términos de la estimación de los parámetros de interés del trabajo remunerado o no remunerado a fin de mantener la comparabilidad de las estadísticas laborales entre los distintos países. Para ello se definen nuevos y mejores indicadores que contribuyan al análisis de la dinámica del mercado laboral al brindar la información que la sociedad necesita a medida que evoluciona este constructo social.

D. Los indicadores de ingresos y gastos

Es importante resaltar que los indicadores de bienestar (en términos de ingresos y gastos) también forman parte del conjunto de parámetros que se pueden estimar desde las encuestas de hogares. La medición del ingreso a partir de las encuestas de hogares constituye un reto metodológico para los institutos nacionales de estadística (INE) en el mundo, y particularmente en América Latina. Es recomendable seguir las directrices de la Comisión Económica para Europa, que actualizan los estándares internacionales, recomendaciones y buenas prácticas en la medición del ingreso en los hogares. Por ejemplo, el llamado Grupo de Canberra ha revisado de manera exhaustiva el tema de la estimación del ingreso mediante el estudio de las prácticas de algunos países en términos del aseguramiento de la calidad y la publicación de este tipo de estadísticas oficiales, y ha provisto la siguiente definición de ingreso del hogar (Naciones Unidas, 2011):

> El ingreso del hogar se compone de las entradas monetarias, en especie o en servicios que por lo general son frecuentes y regulares, están destinadas al hogar o a los miembros del hogar por separado y se reciben a intervalos anuales o con mayor frecuencia. Durante el período de referencia en que se reciben, tales entradas están potencialmente disponibles para el consumo efectivo y, habitualmente, no reducen el patrimonio neto del hogar.

Con base en lo anterior, el uso de las encuestas de hogares para estimar el ingreso presenta retos metodológicos mayores, puesto que los entrevistados deben responder con precisión cuando se indaga sobre los ingresos personales de cada individuo en el hogar, como sueldos y salarios, ganancias, ingresos del empleo y pensiones, y también sobre los ingresos del hogar, incluidas las rentas por alquiler y los ingresos generados por el comercio. Por lo tanto, en el diseño de la encuesta se debe tener en cuenta la definición de un instrumento que sea relevante para el respondiente y le permita identificar (y, en algunas ocasiones, recordar) la información con cierto grado de exactitud.

Por ejemplo, si el respondiente es empleado regular, el instrumento de medición debería planearse de tal manera que el entrevistado pueda recordar la información de interés, como las contribuciones a la seguridad social hechas por su empleador. Por otro lado, si se requiere que el respondiente brinde información acerca de determinado período de tiempo, el planteamiento de la pregunta, la forma de indagar y el entrenamiento de los encuestadores pueden introducir sistemáticamente un sesgo en la respuesta y, por consiguiente, inducir estimaciones poco confiables. Mucho se ha investigado con respecto a cómo realizar preguntas certeras en este tipo de levantamientos. El lector interesado puede consultar los trabajos de Biemer y Lyberg (2003), Presser y otros (2004) y Groves y otros (2009).

E. Algunos desafíos de la región en cuanto a la producción de las encuestas de hogares

La forma de medición de los indicadores sociales debe estar alineada con el diseño de la encuesta. Los equipos técnicos de los distintos países deben ajustar sus metodologías a los requisitos de la encuesta para proveer estadísticas oficiales que sean no solo confiables y precisas, sino eficientes en términos de los recursos que se destinan a la recopilación de la información primaria. Ello es más pertinente aún si se tiene en cuenta que estos recursos muchas veces se deben recortar a causa de las limitaciones presupuestarias de los países.

En América Latina se observa un incremento progresivo de la tasa de ausencia de respuesta debido al aumento de las viviendas no entrevistadas, que se origina en la expansión urbana y en la desactualización de los marcos de muestreo. La continua expansión urbana en la región hace que los encuestadores afronten retos mayores cuando llegan a un área de muestreo y no encuentran la vivienda a cuyos ocupantes se supone que deberían entrevistar o cuando, en vez de una vivienda, encuentran un edificio de apartamentos. Con el uso cada vez más frecuente de dispositivos de almacenamiento electrónico, se puede realizar un análisis mejor estructurado de los instrumentos de recopilación de información. Por ejemplo, es posible programar saltos más complejos y estimar el tiempo promedio de respuesta en las preguntas del cuestionario y en los bloques de preguntas, entre otras cosas.

En el seguimiento de las metas de la Agenda 2030 y en la búsqueda del cumplimiento de los ODS, las Naciones Unidas han expresado la necesidad de contar con estadísticas oficiales, no solo a nivel nacional, sino a nivel de desagregaciones geográficas o de categorías demográficas de interés (Naciones Unidas, 2018). Por ese motivo se plantea la siguiente necesidad a los países:

> La disponibilidad de datos de alta calidad, accesibles, abiertos, oportunos y desglosados es vital para la adopción de decisiones con base empírica y la plena implementación de la Agenda 2030 para el Desarrollo Sostenible [...]. Para satisfacer estas demandas de datos es necesario fortalecer con carácter urgente la capacidad de los sistemas nacionales de estadística. La comunidad estadística mundial se esfuerza por desarrollar metodologías y tecnologías para innovar y modernizar las operaciones de producción de estadísticas, estudiar los medios para integrar todas las fuentes de datos, y analizar, visualizar y difundir los datos de manera abierta, oportuna y eficaz (Naciones Unidas, 2018).

Por supuesto, la región no es ajena a este reto y debe plantearse sin demora la necesidad de crear capacidad en los equipos técnicos de los INE que deberán utilizar como insumo las encuestas de hogares, los censos de población y los registros administrativos para poder establecer metodologías de estimación en las desagregaciones de interés. Entre estas se deberán incluir, como mínimo, las referentes a sexo, grupos etarios, ingreso, raza, etnia, estado migratorio, discapacidad y localización geográfica (Naciones Unidas, 2016).

Al respecto, uno de los primeros acercamientos al diseño de las encuestas de hogares para la publicación de estadísticas oficiales en dominios pequeños fue presentado por Singh, Gambino y Mantel (1994), que plantean algunas consideraciones sobre la estimación de indicadores sociales a nivel de dominios pequeños. Sin embargo, más allá del diseño metodológico, ahora es posible considerar otro tipo de acercamientos inferenciales orientados a la obtención de indicadores a nivel de estos dominios pequeños. Rao y Molina (2015) proveen un resumen exhaustivo de las técnicas más utilizadas en la diseminación de estadísticas oficiales en desagregaciones pequeñas. En América Latina, este tipo de metodologías, que utilizan como insumo principal las encuestas de hogares, se aplican con mayor frecuencia. Por ejemplo, Arias y Robles (2007) realizan una estimación de la pobreza monetaria en las municipalidades del Estado Plurinacional de Bolivia, utilizando los datos del censo poblacional de 2001. Por su parte, Araujo (2007) resume la experiencia ecuatoriana de la estimación de la pobreza en los municipios, cantones y provincias, y López-Calva, Rodríguez-Chamussy y Székely (2007) presentan la estimación de indicadores de desarrollo humano a partir de la estimación de áreas pequeñas en las municipalidades de México mediante la Encuesta Nacional de Ingresos y Gastos de los Hogares. Por último, Casas-Cordero Valencia, Encina y Lahiri (2016) presentan un ejercicio de estimación de la pobreza en Chile, para el que utilizan la Encuesta de Caracterización Socioeconómica Nacional (CASEN).

Capítulo I
El paradigma del error total

En este capítulo se describe muy someramente el paradigma de los errores que se cometen en una encuesta y la manera en que, al tenerlos en cuenta en la etapa de planificación, es posible medirlos de manera acertada y acotarlos sobre la base del principio de representatividad. En general, todos los procesos en una encuesta deben estar planificados de antemano, antes y después de la recopilación de los datos. Por ejemplo, el cuestionario (instrumento de medición) debe estar muy bien diseñado para que las respuestas de las personas describan acertadamente las características de los entrevistados. De la misma forma, el subconjunto de personas que participan en la encuesta debe ampliarse con precisión y confiabilidad para que represente con certeza a la población de interés.

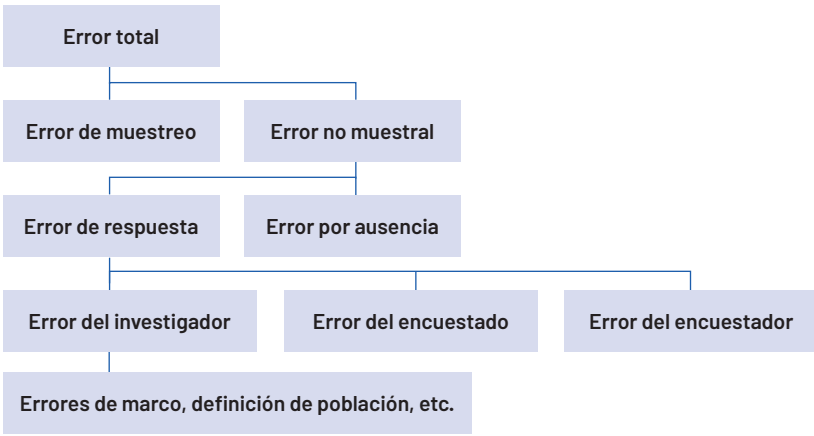
Béland y otros (2005) describen los principales elementos del diseño de una encuesta de hogares. Esta es una tarea que deben afrontar los equipos técnicos de las oficinas nacionales de estadística (ONE), en el sentido de aprender de las experiencias del pasado para mejorar los procesos operativos, metodológicos y logísticos al realizar las siguientes encuestas. Ante la nueva ronda de censos que se avecina en la próxima década, será natural actualizar los marcos de muestreo. Esto representará un reto para los equipos técnicos encargados de las encuestas de hogares en América Latina, debido a la necesidad de evaluar el impacto del cambio de los marcos de muestreo y sus efectos en la comparabilidad de las cifras oficiales.

En una encuesta, el interés no se centra en las características particulares de un individuo, sino en las características de la población a la que ese individuo pertenece. De esta forma, la inferencia siempre se realiza teniendo en mente agregados (indicadores) poblacionales. En el diagrama I.1 se presentan las dos fuentes principales de error cuando se realiza una encuesta:

i) **Error de muestreo:** ocurre porque no se incluyó a todas las personas de la población y se seleccionó una muestra.

ii) **Error no muestral:** se refiere a las posibles desviaciones de las respuestas proporcionadas por un entrevistado con respecto al verdadero atributo que se desea medir.

■ Diagrama I.1
El paradigma del error total



Fuente: Elaboración propia, sobre la base de R. Groves y otros, *Survey Methodology*, Hoboken, John Wiley & Sons, 2009.

Por ejemplo, en una encuesta de fuerza de trabajo mensual, puede generarse confusión en el respondiente si no se hace hincapié en el período de referencia. No es lo mismo indagar sobre la semana pasada que sobre el mes pasado, por lo que se debe orientar al respondiente para evitar equivocaciones. Además, pueden existir no respondientes en algún subgrupo de interés, o incluso puede suceder que el marco esté desactualizado. Uno de los objetivos de la planificación concienzuda de la encuesta es minimizar los errores no muestrales. Es necesario reducir al máximo las discrepancias encontradas entre la respuesta verdadera a una pregunta y la respuesta final.

Groves y otros (2009) observan que, durante todo el siglo pasado, surgieron una serie de teorías y principios que ofrecen un marco de referencia unificado para el diseño, la implementación y la evaluación de encuestas. Este marco se conoce comúnmente como paradigma del error total de muestreo y ha encaminado la investigación moderna hacia una mejor calidad de las encuestas.

A. Sesgos generados en las encuestas

Gutiérrez (2016) plantea que existen diferentes fuentes de sesgo en las encuestas y resume las dos fuentes de sesgo más importantes como se expone a continuación.

1. Sesgo de selección

Este tipo de sesgo ocurre cuando parte de la población objetivo no está en el marco de muestreo, o cuando el marco está incompleto y presenta deficiencias. Por ejemplo, una muestra a conveniencia es sesgada porque las unidades más fáciles de elegir o las que más probablemente respondan a la encuesta no son representativas de las unidades más difíciles de elegir. Lohr (2000) afirma que se presenta este tipo de sesgo si se cumple alguna de las condiciones siguientes:

- La selección de la muestra depende de cierta característica asociada a las propiedades de interés. Esto sucede, por ejemplo, si la encuesta se realiza ingresando a un portal web, y, precisamente, existe una diferencia significativa entre las personas que no tienen cobertura de Internet y quienes sí tienen acceso.
- La muestra se realiza a partir de una elección deliberada o un juicio subjetivo. Por ejemplo, si el parámetro de interés es la cantidad promedio de gastos en compras en un centro comercial y el encuestador elige a las personas que salen con muchos paquetes, entonces la información estaría sesgada, puesto que no se reflejaría el comportamiento promedio de las compras.
- Existen errores en la especificación de la población objetivo. Puede suceder, por ejemplo, en las encuestas electorales, cuando la población objetivo contiene a personas que no están registradas como votantes ante la organización electoral de su país.
- Existe una sustitución deliberada de unidades no disponibles en la muestra. Si, por alguna razón, no ha sido posible obtener la medición y consecuente observación de la característica de interés respecto de algún individuo en la población, la sustitución de este elemento debe hacerse siguiendo estrictos procedimientos estadísticos y no debe ser subjetiva de ningún modo.
- Existe ausencia de respuesta. Este fenómeno puede causar distorsión de los resultados cuando los que no responden a la encuesta difieren estructuralmente de los que sí responden en términos de características demográficas, sociales, educativas, laborales o de ubicación, entre otras.
- La muestra está compuesta por respondientes voluntarios. Por ejemplo, los foros radiales, las encuestas de televisión y los estudios de portales de Internet no proporcionan información confiable.

Además de lo enumerado anteriormente, en América Latina pueden existir sesgos ocasionados por la falta de cobertura en el marco de muestreo, o por la exclusión planificada de subpoblaciones de difícil acceso. Por ejemplo, en una encuesta de fuerza de trabajo, es posible que la encuesta no sea representativa de las subpoblaciones afrodescendientes o indígenas, por no cubrir exhaustivamente los territorios donde estas se ubican.

2. Sesgo de medición

Este tipo de sesgo ocurre cuando el instrumento con que se realiza la medición tiene tendencia a diferir del valor verdadero que se desea averiguar. Este sesgo debe ser considerado y minimizado en la etapa de diseño de la encuesta. Si el instrumento de medición (cuestionario) tiene defectos en su planificación, lo más probable es que el resultado de la estimación de la encuesta difiera sistemáticamente del verdadero valor respecto de cada uno de los respondientes. Ningún análisis estadístico podrá ajustar esta diferencia sistemática. Lohr (2000) cita algunas situaciones en que se presenta este sesgo de medición, las cuales se enumeran a continuación:

- Cuando el respondiente miente. Esta situación se presenta a menudo en encuestas que indagan acerca de ingresos salariales, alcoholismo y drogadicción, nivel socioeconómico e incluso edad.
- Cuando el cuestionario contiene preguntas difíciles de comprender (por ejemplo: "¿No es cierto que usted no recibe remesas desde el exterior?"). La doble negación en esta pregunta es muy confusa para el respondiente.
- Cuando hay un olvido del respondiente que impide obtener una respuesta veraz. Las personas tienden a olvidar muchas de sus experiencias, sobre todo las malas. Esta situación se debe tener en cuenta si se está trabajando en una encuesta de delincuencia, victimización, consumo de sustancias psicoactivas o módulos con preguntas sensibles.
- Cuando se brindan distintas respuestas a diferentes entrevistadores. En algunas regiones es muy probable que la raza, edad o género del encuestador influya directamente en la respuesta del entrevistado.
- Cuando se leen mal las preguntas o se polemiza con el respondiente. El encuestador puede influir notablemente en las respuestas. Por ello, es muy importante asegurarse de que el proceso de entrenamiento del entrevistador sea riguroso y completo.
- Cuando la muestra está compuesta por respondientes voluntarios. Los foros radiales, las encuestas de televisión y los estudios de portales de Internet no proporcionan, en general, información confiable. En este caso también se presenta sesgo de selección.

B. Evolución de las encuestas estandarizadas

Cuando el mundo occidental superó los grandes traumas del siglo XX (dos guerras mundiales y una recesión a gran escala), la investigación social se desarrolló de manera significativa a través de las encuestas por correo postal. Según señala el Instituto Nacional de Estadística

de la República Bolivariana de Venezuela (INE, 2013), en el caso latinoamericano, la Agencia para el Desarrollo Internacional (USAID) auspició una serie de 14 documentos denominada *Atlantida: A Case Study in Household Sample Surveys*, elaborada por la Oficina del Censo de los Estados Unidos (1965). Este estudio se realizó en el marco del programa Alianza para el Progreso y fue presentado en colaboración con la Organización de los Estados Americanos (OEA) y el Instituto Interamericano de Estadística (IASI). Fue el punto de partida para instituir un modelo que serviría de apoyo para la realización de las encuestas de hogares en América Latina.

Desde el momento en que las encuestas de hogares se instauraron como un instrumento apropiado para la investigación, se plantearon tres interrogantes que se deben responder para planificar, ejecutar y analizar una encuesta: ¿cómo se diseñarán las preguntas?, ¿cómo se seleccionará la muestra? y ¿cómo se recopilarán las respuestas?

1. Inicio de los cuestionarios estandarizados

La práctica de realizar las mismas preguntas en forma de cuestionario es relativamente reciente. Antes de adoptar un proceso estandarizado, podía suceder que cada encuestador preguntara lo mismo, pero con diferentes palabras. No era común que dos personas distintas fuesen entrevistadas con las mismas preguntas. Groves y otros (2009) mencionan que la forma en que se preguntaba y en que se recopilaba la información afectaba en gran medida los resultados de las encuestas. Por ese motivo se decidió que los encuestadores debían recibir una capacitación formal.

El formalismo del cuestionario se implementó en primera instancia en el ámbito de la psicometría. Con el fin de medir estados psicológicos, afectivos e intelectuales, se desarrollaron técnicas que permitían hacer comparables las respuestas. Likert (1932) demostró que era posible realizar este tipo de comparaciones, evadiendo los largos instrumentos de medición, al formular una sola pregunta (a todos los encuestados) con una serie de respuestas en forma de escala.

2. Inicio de los métodos de muestreo

En un principio, los investigadores trataban de recopilar datos sobre todos los elementos de la población de interés. Esta práctica resultaba logísticamente inadecuada cuando se trataba de poblaciones de gran tamaño. Calcular los indicadores sobre toda una población exigía grandes esfuerzos. Groves y otros (2009) afirman que, aunque la teoría de la probabilidad tuvo sus orígenes en el siglo XVIII, no fue hasta la segunda década del siglo XX cuando se utilizó para realizar encuestas. La primera aplicación fue la selección sistemática de un elemento en una población dada. Para realizar esta selección, los registros censales se dividían en secciones y se procedía a escoger un elemento de la sección.

Más adelante, cuando la estadística permeó la agricultura, se definieron otros tipos de muestreo (menos exigentes) y se dio origen al muestreo de áreas. Hoy en día es posible seleccionar muestras de bloques, zonas amanzanadas, secciones y sectores cartográficos, o áreas de empadronamiento censal. Se descubrió que era posible generalizar el muestreo de áreas y se creó el muestreo multietápico, que permitió la selección de grandes bloques dentro de una ciudad, y de áreas dentro de los bloques. También surgió el submuestreo sucesivo de unidades hasta llegar a la unidad de interés. Todos estos submuestreos se realizan de forma probabilística.

La Gran Depresión en los Estados Unidos y la Segunda Guerra Mundial fueron catalizadores de las encuestas a gran escala. En ese entonces, al igual que hoy, la tasa de desempleo era una cifra importante para la economía de los países. Las políticas públicas empezaron a decidirse de acuerdo con las estadísticas oficiales, puesto que las grandes encuestas comenzaron a realizarse con periodicidad mensual. Hoy en día existen cientos de encuestas mensuales que dan cuenta de la realidad de las sociedades en la región.

3. Inicio de la recopilación de datos

Debido a que en un principio no existía un cuestionario estandarizado, las respuestas abiertas eran la única opción para recopilar información. Esta práctica exigía un gran esfuerzo a la hora de resumir y sintetizar todo el corpus de palabras que los entrevistados usaban para responder.

A mediados de la década de 1970, comenzó una proliferación masiva de entrevistas por correo en los Estados Unidos. Los países con registros administrativos actualizados pueden considerar esta opción, que genera altas tasas de cobertura a precios más económicos (al prescindirse del encuestador). Las bajas tasas de respuesta (pues el encuestado debe rellenar un formulario con sus respuestas y devolverlo a la oficina postal) hicieron que paulatinamente esta forma de recopilación dejara de ser tan atractiva (Groves y otros, 2009).

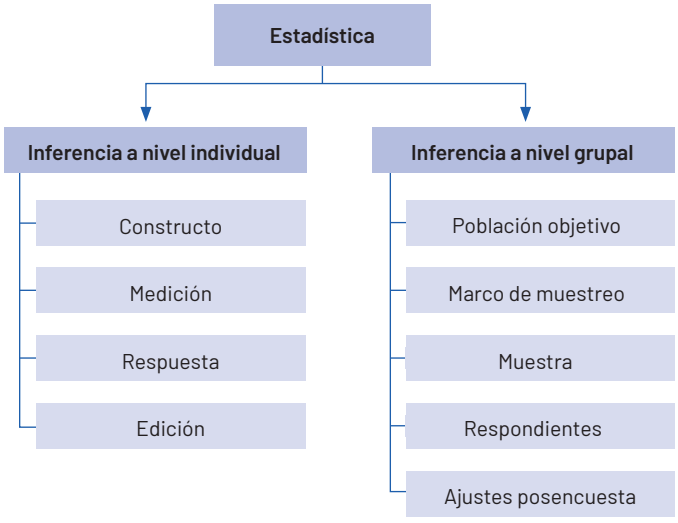
Como señala la CEPAL (1983), en el caso de América Latina, en la primera parte del decenio de 1960, varios países comenzaron a realizar encuestas de hogares periódicas con el propósito de obtener información sobre el empleo y el desempleo. En 1965, se realizó en la ciudad de México un seminario en que se presentó el estudio *Atlantida*. Posteriormente, ante la necesidad de satisfacer la demanda de información relativa a las políticas económicas y sociales, tomó gran impulso en varios países de la región la puesta en marcha de programas permanentes de encuestas de hogares, cuyo propósito era fundamentalmente obtener información sobre la fuerza de trabajo. El modelo denominado Atlántida se basó sobre todo en un tipo de encuesta empleado en los países desarrollados, cuyos mercados de trabajo poseen características propias. Sin embargo, resultó muy interesante para aquellos países que tenían poca experiencia en la realización de encuestas de hogares y constituyó la base metodológica sobre la que se han desarrollado gran parte de las encuestas de América Latina (CEPAL, 1983).

Las entrevistas telefónicas representan un camino intermedio entre la recopilación de información presencial (cara a cara) y la obtención de datos mediante formularios autoadministrados (por correo electrónico, a través de páginas de Internet o mediante correo postal). Hoy en día, la mayoría de los sondeos de investigación de medios y de mercado se realiza por teléfono. Más aún, a partir de la pandemia de enfermedad por coronavirus (COVID-19), en marzo de 2020 el mundo sufrió una paralización de las actividades sociales y económicas, debido a los esfuerzos que realizaron los gobiernos para tratar de frenar la expansión de la pandemia. Fue así como, debido a las restricciones de movilidad impuestas por los gobiernos, se suspendieron las operaciones estadísticas con recopilación presencial de datos. En América Latina, el rigor de la pandemia y de las medidas de restricción de la movilidad también afectaron la realización de las encuestas de hogares. Sin embargo, en vista de estas dificultades, la CEPAL (2020a) recomendó la continuidad de las encuestas mediante entrevistas telefónicas.

C. El ciclo de vida de una encuesta

Atendiendo al modelo de Groves y otros (2009), se puede afirmar que en todas las encuestas hay dos niveles de inferencia: el individual y el grupal (véase el diagrama I.2). El proceso de inferencia individual trata con los mismos respondientes que brindan la información primaria en el estudio. Por su parte, el proceso de inferencia grupal, basado en una aproximación inductiva, va de lo particular (la muestra) a lo general (la población).

■ Diagrama I.2
Dos niveles de inferencia en una encuesta



Fuente: Elaboración propia, sobre la base de R. Groves y otros, *Survey Methodology*, Hoboken, John Wiley & Sons, 2009.

1. Inferencia individual

a) Constructo

Gutiérrez (2016) afirma que los constructos son las ideas abstractas (ambiguas) sobre las que el investigador desea inferir y que, a su vez, dan origen a la investigación, al ser la simiente de la encuesta. Las palabras con que se describen los constructos suelen ser sencillas, pero la redacción elaborada de los constructos no siempre es precisa. Por ejemplo:

- En una encuesta de victimización, cuyo fin es medir la cantidad de incidentes relacionados con delitos en un año determinado, es necesario definir muy apropiadamente qué se entiende por delito, o cómo se define a una víctima, entre otros muchos aspectos.
- En una encuesta sobre el goce efectivo de los derechos ciudadanos en el caso de los menores de edad, se puede medir la efectividad del Estado al garantizar los derechos básicos de la primera infancia. Sin embargo, es necesario definir qué es un derecho, o cómo se define el concepto de primera infancia.

Mientras que algunos constructos son bastante abstractos (optimismo en la economía, confianza inversionista o percepción del plan nacional de desarrollo de un Gobierno), otros pueden observarse de manera más concreta (consumo de alcohol y otras drogas, nutrición en la primera infancia, productividad de una intervención en el sector agrícola o factores de riesgo asociados a una enfermedad).

b) Mediciones

La medición es una caracterización mucho más concreta que el constructo, pues representa una forma de obtener información de los constructos de interés. Lo fundamental para realizar una buena medición es realizar preguntas que induzcan respuestas que reflejen claramente los constructos que se desea medir. Groves y otros (2009) indican que estas preguntas pueden ser comunicadas de forma oral (encuestas cara a cara o telefónicas), o de forma visual (atributos de un producto desde el punto de vista del *marketing*). También pueden existir observaciones directas del encuestador (condiciones de la vivienda) u observaciones provenientes de dispositivos electrónicos o físicos (precios de productos en supermercados, muestra de agua o muestra de sangre, entre otras).

c) Respuesta y edición

El resultado de la medición es la respuesta y sus propiedades están determinadas por la naturaleza de las preguntas. Después de que los entrevistados han respondido, los datos deben someterse a un proceso de edición y validación de incoherencias.

En ese proceso, se debe examinar la distribución completa de las respuestas y buscar datos atípicos para revisarlos con detenimiento. Los datos editados constituyen el insumo para realizar todo el proceso de inferencia estadística pertinente a fin de asegurarse de que las cifras resultantes sean confiables y precisas.

2. Inferencia grupal

a) La población objetivo

De las definiciones concernientes a agregados, esta es la más abstracta. En general, la población objetivo representa el conjunto de unidades que serán estudiadas. Por ejemplo, en una encuesta es posible definir la población objetivo como los adultos nacionales. Sin embargo, esta definición no explicita el período de referencia de la medición. Tampoco aclara si se incluyen los adultos residentes en el exterior ni precisa cómo se verificará la nacionalidad de un entrevistado.

Por ese motivo, la definición de la población objetivo tiene que ser lo más precisa posible. Por ejemplo, la Gran Encuesta Integrada de Hogares de Colombia define a su población objetivo como la población civil no institucionalizada (PCNI), es decir, todas las personas que no forman parte de la fuerza pública ni residen en instituciones de aislamiento como prisiones, hospitales, sanatorios o residencias de ancianos. La PCNI contiene a la población en edad de trabajar (PET) y a la no pertenecientes a la fuerza laboral. Según la metodología de esta encuesta en particular, la edad para empezar a trabajar en las zonas rurales es de 10 años y, en las ciudades, de 12 años. A su vez, la PET contiene a inactivos y ocupados.

b) La población enmarcada

No es posible realizar una encuesta probabilística sin un marco de muestreo, definido como un dispositivo que permite ubicar e identificar (al mismo tiempo) las unidades pertenecientes a la población de interés. En América Latina, los marcos de muestreo para las encuestas de hogares en los países se crean a partir del censo, mientras que en algunos países desarrollados es posible encontrar marcos de muestreo creados a partir de líneas telefónicas o direcciones de residencia.

Es necesario señalar que todos los marcos de muestreo presentan algún nivel de desactualización con respecto a la población de interés. Por ejemplo, un marco de muestreo de áreas (basado en la cartografía del último ejercicio censal) puede estar desactualizado. Asimismo, una de las posibles desventajas del uso de este tipo de marcos es que se podría entrevistar a la misma persona en varias ocasiones (si la persona tiene múltiples residencias) o incluso no realizar nunca la entrevista a una persona que no tenga un lugar fijo de residencia. Por otro lado, un marco de muestreo de líneas telefónicas podría no contener a todos los residentes de una ciudad.

La población enmarcada está definida por el conjunto de miembros de la población objetivo que efectivamente tienen una probabilidad no nula de ser seleccionados en una muestra probabilística. En general, para definir quién pertenece a un hogar del marco existen dos alternativas:

i) Regla *de iure*: quien habitualmente reside en el hogar es miembro de ese hogar.
   - Una situación *de iure* es aquella que está reconocida por la legalidad vigente o por la autoridad competente en virtud de algún acuerdo o acto formal.
   - Evita la subcobertura de individuos que no suelen residir en su hogar, considerándolo suyo.

ii) Regla *de facto*: quien pasó la noche anterior en una residencia de un hogar es miembro de ese hogar.
   - Una situación de facto es aquella que, aunque existe en la realidad, no está reconocida formalmente.
   - Evita la sobrecobertura de individuos que tienen más de una residencia.

c) La muestra

El tamaño de la muestra define directamente la precisión y confiabilidad de las estimaciones. Este debería incrementarse a medida que lo hagan los niveles de desagregación (grupos etarios, regiones geográficas y niveles de escolaridad, entre otros). Sin embargo, dependiendo de la caracterización de la estrategia de muestreo, pueden existir escenarios en que una encuesta con un tamaño de muestra menor conlleve menores errores de muestreo que una encuesta con un mayor tamaño de muestra.

No obstante, hay ocasiones en que los esfuerzos realizados para que los individuos seleccionados en la muestra respondan no son fructíferos. De esta manera, los individuos que son efectivamente entrevistados se denominan respondientes efectivos. Por su parte, al complemento de este conjunto se les denomina no respondientes.

d) Los respondientes

Pueden existir casos de no respondientes parciales (no respondientes de determinados ítems), para los cuales debe existir un proceso de decisión en lo que se refiere a su reemplazo. Asimismo, no todas las ausencias parciales son reemplazadas. Groves y otros (2009) afirman que algunos de los factores que inciden en el aumento de la ausencia de respuesta pueden referirse a:

- Contenido: por preguntas sensibles (encuestas relacionadas con el uso de drogas, finanzas o victimización, entre otras). En este caso, se puede aumentar la tasa de respuesta si se ordenan las preguntas de manera adecuada.

- Encuestadores: se deben aplicar métodos estándar de mejoramiento de la calidad para aumentar la precisión y la tasa de respuesta de los entrevistadores que participan en el estudio.
- Método de recopilación: las encuestas telefónicas, por correo electrónico o por páginas web tienen una tasa de respuesta inferior a la de las entrevistas personales.
- Diseño de cuestionario: mala planificación en el pase de las preguntas que conforman el instrumento.
- Tiempo de la encuesta y agobio: algunas temporadas arrojan tasas de no respuesta más altas que otras. De la misma forma, algunos cuestionarios largos son propensos a inducir una mayor ausencia de respuesta parcial debido al agotamiento del respondiente. En general, las encuestas demasiado largas pueden indisponer al respondiente.

e) Los ajustes posencuesta

Toda encuesta cuenta con personas que no quisieron responder o con un marco de muestreo que no cubre a toda la población. Por ende, es necesario realizar algunos ajustes en el análisis y procesamiento para evitar, sobre todo, la subestimación de los parámetros de interés, o implementar métodos de imputación con el fin de suplir la información faltante. De esta forma, se puede utilizar una reponderación diferencial cuando sea evidente que hay un patrón de ausencia de respuesta en algunos subgrupos de la población. Esto puede ocurrir, por ejemplo, si las tasas de respuesta a nivel urbano son menores que las tasas de respuesta a nivel rural, o si los hombres responden menos que las mujeres.

También es posible imputar los valores ausentes (es, decir, asignar valores a los datos faltantes) en un subconjunto de observaciones de la muestra seleccionada. En este caso, es factible utilizar metodologías estocásticas complejas para imputar valores, o técnicas simples sistemáticas. Sin embargo, siempre será preferible obtener la respuesta directa del entrevistado.

D. El proceso de respuesta

No todas las encuestas se planean de forma tal que en todo momento exista una interacción directa entre respondiente y entrevistador. Sin embargo, los modelos de respuesta de las encuestas dan por sentado que existen, por lo menos, lo siguientes momentos en la obtención de un valor numérico que se recopila como respuesta al cuestionario:

i) La comprensión, o el momento en que el respondiente interpreta la pregunta. Groves y otros (2009) afirman que en ese momento intervienen todos aquellos procesos de atención a la pregunta y comprensión de las instrucciones. La primera tarea del respondiente es interpretar la pregunta y, al hacerlo, surgen

procesos de análisis y asignación de un significado a los elementos sustantivos de la pregunta. Además, el respondiente debe hacer una inferencia sobre el propósito de la pregunta, determinar los límites de la respuesta, así como acotar los posibles traslapes sobre las respuestas permitidas.

ii) La recuperación, momento en que el respondiente busca en su memoria la información necesaria para brindar una respuesta. En algunas ocasiones se accede a la memoria de largo plazo, que almacena todo el contenido autobiográfico y el conocimiento general. Nótese que muchas cosas pueden afectar el desempeño de la memoria de largo plazo (cuando los eventos en cuestión no se distinguen con facilidad o cuando no tienen un gran impacto personal). La memoria brinda la información relevante para que el entrevistado proporcione una respuesta adecuada. Este ciclo de recuperación de información continúa hasta que el entrevistado dé una respuesta acertada o simplemente no quiera recordar más (algunas situaciones son más difíciles de recordar) (Groves y otros, 2009). Para ayudar a la memoria de largo plazo se pueden diseñar señales o pistas autocontenidas en la pregunta. Las mejores señales son las que ofrecen un nivel de detalle más profundo.

iii) El juicio, momento en que se combina, se pondera y se resume la información recopilada. En esta etapa se originan procesos que complementan la recuperación de información que el entrevistado ha llevado a cabo anteriormente. El juicio puede llenar los vacíos de la memoria, combinar los datos recuperados o ajustarlos por omisión. Por ejemplo, en una encuesta de ingresos y gastos, las personas, por lo general, no llevan la cuenta del número de veces que compraron determinado artículo, o no tienen una respuesta predefinida al número de veces que han salido de compras. El respondiente tratará de contar el número de veces que experimentó una situación y, si ese número es muy grande, seguramente se acercará a la respuesta mediante una estimación. La estrategia de estimación del respondiente (llevar la cuenta, construir una escala mediante la recordación de eventos, realizar una estimación o adivinar al azar) depende del número de sucesos, su duración, su regularidad y el período de referencia de la encuesta (Groves y otros, 2009).

iv) La respuesta, momento en que el respondiente formula su contestación y la estandariza en el formato inducido por el cuestionario. Este es el proceso de selección y comunicación de una respuesta, que incluye el encuadre de esta dentro de las opciones que ofrece la pregunta (también implica alterar la respuesta para que se ajuste a las opciones aceptables). La forma en que se proporciona la respuesta final dependerá del ajuste que se realice en los procesos de recuperación y estimación y las restricciones que la pregunta impone. En este sentido, si para una pregunta de percepción la mayoría de las opciones de respuesta son negativas, la respuesta estará sesgada en esa dirección. Asimismo, los respondientes pueden dar mayor importancia a ciertas opciones de respuesta (Groves y otros, 2009).

El investigador debe saber que el solo hecho de haber experimentado una situación no implica que el respondiente haya recabado la información suficiente para proporcionar una respuesta. Según Groves y otros (2009), se ha visto que los testigos presenciales de una situación omiten detalles importantes acerca de esta. Además, las personas no pueden facilitar la información que no tienen. Si la gente no recaba la información necesaria, ninguna pregunta ni formulación logrará obtener la respuesta real. Por esa razón se recomienda llevar a cabo las pruebas necesarias para validar el cuestionario. Por otro lado, aunque el respondiente conozca con exactitud la respuesta a una pregunta, no será capaz de proporcionarla correctamente si no hay una buena interpretación de esa respuesta.

A su vez, es necesario advertir que la respuesta del entrevistado está supeditada a los tiempos de ocurrencia (los eventos que sucedieron hace mucho tiempo son más difíciles de recordar) y a los límites temporales y su correspondiente impacto emocional, puesto que los eventos cercanos a momentos que han tenido un impacto emocional son más fáciles de recordar (eventos catastróficos, atentados terroristas o desastres naturales). También está supeditada a las señales en las preguntas, pues la asignación de múltiples señales en la redacción de la pregunta ayuda a activar el proceso de recordación.

En cuanto a la naturaleza de las preguntas, cabe mencionar que las preguntas cerradas con escala ordenada podrían tender a producir un sesgo de respuesta positivo, pues los respondientes tienden a evadir las opciones negativas de la escala (encuestas de satisfacción). Schwarz y otros (1991) demostraron que las etiquetas numéricas afectan el proceso de respuesta, por lo que recomendaron que el encuestador no leyera los números en las opciones de respuesta y que se moderara el número de opciones en las preguntas de opinión (ni muy pocas, ni muchas).

Por otra parte, nótese que la generación de pocas opciones de respuesta hace que se pierda el poder de discriminación, mientras que la inclusión de muchas opciones puede hacer que los encuestados no distingan fácilmente entre las categorías adyacentes. Además, es posible que el respondiente no quiera esperar a que el entrevistador termine de leer todas las opciones de respuesta. En este caso se presentan dos fenómenos que es necesario evitar. En primer lugar, se ha descrito el efecto de primacía, que incrementa el riesgo de que el respondiente escoja una de las primeras opciones. En segundo lugar, existe el efecto de recencia, que implica que el respondiente siempre escogerá una de las últimas opciones.

Algunos respondientes podrán desviarse del modelo de respuesta mediante la elección de rutas alternas de evasión (en el sentido de hacer el mínimo esfuerzo para satisfacer las demandas del entrevistador). Así, se podrían encontrar respondientes que seleccionen sistemáticamente las opciones "No sabe" o "No responde", o que escojan siempre la misma opción para cada pregunta. Incluso, dependiendo de la apariencia del entrevistador, el respondiente puede presentar un sesgo a siempre estar de acuerdo (aquiescencia). De la misma manera, es posible que el respondiente quiera presentarse a sí mismo de manera favorable y omitir sus atributos no deseables (Groves y otros, 2009).

Capítulo II
Elementos estadísticos básicos en la planificación de las encuestas

El fortalecimiento continuo de las investigaciones sociales es un objetivo que los institutos nacionales de estadística procuran cumplir de forma sistemática. En el caso de aquellas operaciones que conllevan la recolección de información primaria y la selección y medición de hogares y sus miembros, contar con una documentación adecuada que describa las razones por las que se ha optado por cierta metodología de recolección en particular es un requisito fundamental para cumplir este cometido. En este apartado, se analizan diferentes métodos de recolección de la información y se discuten las particularidades de la planificación de una encuesta de hogares.

A. Universo, muestra y unidades

El concepto de encuesta se encuentra directamente relacionado con la existencia de una población finita compuesta de individuos a quienes es necesario observar y medir. Este proceso se realiza por medio de una entrevista presencial o telefónica, o de formularios electrónicos autoadministrados. El conjunto de unidades de interés recibe el nombre de "población objetivo" o "universo", y sobre ellas se obtiene la información de interés para el estudio. Por ejemplo, la Encuesta Nacional de Empleo y Desempleo del Ecuador define su población objetivo como todas las personas mayores de 10 años residentes en viviendas particulares en el Ecuador (INEC, 2018e).

Por su parte, las "unidades de análisis" corresponden a los diferentes niveles de desagregación establecidos para consolidar el diseño de la encuesta y sobre los que se presentan los resultados de interés. En México, en la Encuesta Nacional de Ingresos y Gastos de los Hogares, se definen como unidades de análisis los distintos ámbitos a que pertenece la vivienda: urbano alto, complemento urbano y rural. Por otro lado, la Gran Encuesta Integrada de Hogares de Colombia tiene cobertura nacional y sus unidades de análisis están definidas por 13 grandes ciudades, junto con sus áreas metropolitanas (DANE, 2017).

Como se explicará más adelante, es muy difícil contar con una lista actualizada de todos los hogares de cada país. Por lo tanto, para recolectar la información de la población objetivo, en el diseño de una encuesta de hogares es necesario tener en cuenta que será necesario realizar en varias etapas una selección de ciertas unidades de muestreo que servirán como medio para seleccionar finalmente a los hogares y personas que formarán parte de la muestra. Cuando lo que debe seleccionarse son personas, se hace necesario tomar un subconjunto de zonas geográficas. Con cada zona, se procede a seleccionar a su vez un subconjunto de secciones cartográficas, que antecede a la selección de hogares. Por último, el cuestionario es administrado en cada hogar a un respondiente idóneo, que proporciona la información de todos los integrantes del hogar. Dependiendo de la encuesta, en algunos casos se designan aleatoriamente respondientes individuales dentro del hogar, siendo estas las unidades de observación. Por ejemplo, se puede citar la experiencia del Brasil con la Encuesta Nacional de Hogares (PNAD), que se realiza mediante una muestra de viviendas en tres etapas: i) las unidades primarias de muestreo (UPM) son los municipios; ii) las unidades secundarias de muestreo (USM) son los sectores censales, que conforman una malla territorial definida en el último *Censo Demográfico*, y iii) las unidades que se seleccionan en último lugar son las viviendas (IBGE, 2014).

Duncan y Kalton (1987, pág. 105) afirman que la composición de la población de interés en las encuestas de hogares cambia a lo largo del tiempo, puesto que los individuos nacen, mueren, migran, e incluso pasan a formar parte de organizaciones que hacen que pierdan el estatus que les permitiría ser seleccionados como unidades de observación en una encuesta. Nótese que la población objetivo de la mayoría de las encuestas de hogares en América Latina es la población civil y se excluye a los miembros de organizaciones militares, personas recluidas en cárceles y personas que se encuentran en hospitales, entre otras. Asimismo, se debe tener en cuenta que los hogares pueden crearse o desintegrarse rápidamente. Por ende, los equipos técnicos de las oficinas nacionales de estadística (ONE) que están a cargo del diseño de las encuestas de hogares, que miden de forma transversal a la población de interés, deben tener en cuenta que, aunque los objetivos de la encuesta no cambian con el tiempo, sí lo hace la población objetivo. Por ese motivo, deben establecerse mecanismos de seguimiento y actualización que reflejen esta realidad.

B. Periodicidad

Las ONE —que son los entes encargados de administrar, diseñar, analizar y difundir los resultados de las encuestas— no realizan este tipo de operaciones estadísticas de manera aislada. De hecho, una característica fundamental de dichas operaciones es que se han convertido en un insumo fundamental para realizar un seguimiento periódico de muchos indicadores de interés. Por lo tanto, muchas encuestas de hogares se realizan de forma sistemática a lo largo del tiempo, aunque otras no tienen una periodicidad predefinida. La planificación de la encuesta debe contemplar este tipo de modelo continuo, para que la recopilación de la información primaria sobre el terreno se haga de manera más eficiente y la estimación de los indicadores de interés se pueda llevar a cabo ajustándose a los recursos de la operación. Como se mencionó anteriormente, dado que la población puede variar a lo largo del tiempo, la planificación y el análisis de este tipo de encuestas presenta desafíos. Si la composición de la población y las características de los elementos se consideraran fijas, una encuesta transversal (realizada una sola vez en un período de tiempo largo) sería suficiente para obtener estimaciones precisas y cumplir los objetivos del estudio.

En algunas ocasiones, basta con realizar una medición simple en un momento específico para completar los objetivos de la investigación. Tal es el caso de las encuestas de ingresos y gastos de los hogares. Su periodicidad es, en general, no inferior a cinco años. Entre muchos otros propósitos, se utilizan para actualizar la canasta básica familiar, de la que se derivan los insumos básicos para la medición de la pobreza (CEPAL, 2018a). Para otro tipo de cuestiones, como el seguimiento de las estadísticas derivadas del mercado de trabajo, es necesario recurrir a la medición periódica con encuestas de hogares. En este caso, los cambios naturales en las características de la población hacen que una medición simple en un punto temporal determinado sea insuficiente para el seguimiento y monitoreo de los indicadores de interés.

Por consiguiente, en el momento de planificar una encuesta continua o periódica, se debe tener en cuenta que, aunque la dificultad de su diseño sea mayor, permite obtener información más oportuna para la toma de decisiones y la formulación de políticas públicas. De esta manera, y teniendo en cuenta que el tiempo hace que la estructura de las poblaciones cambie, sin importar si las unidades de información son individuos, hogares, familias, negocios u otras entidades, estas deben considerarse parte de la población de interés cuando nacen, inmigran o alcanzan un umbral predefinido de edad. Asimismo, las unidades dejarán de formar parte de la población de interés cuando mueran, emigren o ingresen a determinadas instituciones (como el servicio militar). Si las unidades de interés son los hogares, es evidente que la población no será la misma en diferentes momentos (por ejemplo, en dos años distintos), puesto que se crean nuevas unidades cuando los jóvenes dejan a sus padres y forman nuevos hogares independientes, o cuando ocurre una separación o un divorcio en que un hogar se divide en dos. Además, cuando todos los miembros de un hogar han fallecido, este deja de ser parte de la población objetivo. De la misma forma, dos hogares dejan de ser parte de la población objetivo cuando forman un nuevo hogar al unirse a través de un matrimonio o algún otro tipo de unión civil.

Teniendo en cuenta el papel dinámico de las poblaciones y los objetivos de investigación, es posible plantear diferentes maneras de recopilar la información. A continuación se enumeran algunas categorías de encuestas que las ONE realizan en la región.

1. Encuestas transversales

Este tipo de encuesta se diseña para recolectar información únicamente sobre un punto específico del tiempo, o sobre un período de referencia, y proporcionan toda la información pertinente acerca de la población particular restringida a un período de recolección determinado. Puesto que el propósito fundamental de estas encuestas no se centra en las comparaciones intertemporales, no es posible estimar ningún cambio, a no ser que se realicen indagaciones retrospectivas. En el cuadro II.1 se muestra un modelo de este tipo de operaciones estadísticas, donde se observa una muestra de una población concreta en un período de tiempo específico (tiempo 2). Dado que se trata de una muestra transversal, no hay un patrón de repetición en los restantes períodos.

■ Cuadro II.1
Modelo de encuesta transversal

Hogar	Tiempo 1	Tiempo 2	Tiempo 3	Tiempo 4	...	Tiempo T
1		X				
2		X				
3		X				
4		X				
...		X				
n		X				

Fuente: Elaboración propia.

2. Encuestas repetidas

Cuando existe interés en realizar un seguimiento del fenómeno en observación a lo largo del tiempo, se utilizan encuestas repetidas que recolectan información de manera periódica. Este tipo de encuesta brinda información acerca de la dinámica de la composición de la población con el paso del tiempo. De esta forma, en cada recolección se observa una muestra de la población en un momento determinado. Por ejemplo, en el cuadro II.2 se muestra un acercamiento gráfico a estas encuestas, donde se pone de manifiesto su carácter sistemático. Además, se aprecia que no es posible medir cambios individuales porque las muestras son independientes en el tiempo.

■ Cuadro II.2
Modelo de encuesta repetida

Hogar	Tiempo 1	Tiempo 2	Tiempo 3	Tiempo 4	...	Tiempo T
1	X					
2		X				
3			X			
4				X		
...					X	
n						X

Fuente: Elaboración propia.

3. Encuestas de panel

Las encuestas de panel están diseñadas para recolectar información periódica sobre la misma muestra en diferentes momentos. Por definición, las unidades de muestreo son las mismas en los diferentes períodos de tiempo y, de manera general, se miden las mismas variables en cada recolección de información. Por la caracterización propia de este tipo de encuestas, es posible estimar los cambios individuales, así como los cambios netos sobre la población. Sin embargo, como la muestra no varía en ningún momento, las inferencias que se realicen estarán supeditadas a la población de la cual se seleccionó la muestra en un principio (tiempo 1). Si se modifica la estructura de la población, no será posible captar este cambio, porque las inferencias resultantes de este tipo de encuestas no son representativas de la población actual. En el cuadro II.3 se muestra un modelo propio de las encuestas de panel, donde los individuos que fueron seleccionados la primera vez son observados a lo largo del tiempo.

■ Cuadro II.3
Modelo de encuesta de panel

Hogar	Tiempo 1	Tiempo 2	Tiempo 3	Tiempo 4	...	Tiempo T
1	X	X	X	X	X	X
2	X	X	X	X	X	X
3	X	X	X	X	X	X
4						
...						
n						

Fuente: Elaboración propia.

4. Encuestas de panel dividido

Para hacer frente a las dificultades propias de las encuestas de panel, y poder observar tanto los cambios individuales como los cambios en la estructura de la población, se utilizan las encuestas de panel dividido. Estas operaciones estadísticas son una combinación del diseño de panel puro y del diseño repetido y su objetivo es realizar, por una parte, inferencias precisas acerca de los cambios de una cohorte a lo largo del tiempo y, por otra, inferencias acerca del cambio en la estructura de la población actual. De esta forma, se realiza el seguimiento continuo, periódico y sistemático de una muestra a lo largo del tiempo, pero en cada recolección de información se incluyen nuevos elementos seleccionados de la población actual. Como se señalará más adelante, este tipo de encuestas cubre con eficiencia la mayoría de los indicadores de interés en un estudio de investigación social. En el cuadro II.4 se muestra una caracterización de este tipo de encuestas que permiten fijar una muestra de panel a lo largo del tiempo, a la vez que se añaden nuevas observaciones.

■ Cuadro II.4
Modelo de encuesta de panel dividido

Hogar	Tiempo 1	Tiempo 2	Tiempo 3	Tiempo 4	...	Tiempo T
1	X	X	X	X	X	X
2	X					
3		X				
4			X			
5				X		
...					X	
n						X

Fuente: Elaboración propia.

5. Encuestas de panel rotativo

Mantener una muestra de panel es un proceso costoso desde los puntos de vista económico y logístico. También se debe tener en cuenta el desgaste de la fuente, que tenderá a brindar menos información a medida que avanza el estudio. Además, es evidente que, con el paso del tiempo, la propensión a responder será más baja porque el entrevistado se sentirá agotado al ser visitado una y otra vez. Ante esta situación, las encuestas de panel rotativo permiten realizar inferencias parciales —restringidas a períodos de tiempo específicos— del cambio individual y, a la vez, captar el cambio estructural de la población. Estas encuestas incorporan nuevos elementos de la población y mantienen elementos comunes con mediciones anteriores. Al margen de las dificultades que acarrea la ausencia de respuesta, en las encuestas de panel se produce un traslape completo entre las muestras de dos puntos cualesquiera en el tiempo. Sin embargo, en las encuestas rotativas existe un traslape parcial, por lo que se reduce el efecto de desgaste del panel (sobre la población inicial) y el efecto de la pérdida de

muestra. Además, la inclusión de nuevos elementos en la muestra proporciona información pertinente del cambio en la composición estructural de la población. En el cuadro II.5 se ejemplifica el diseño de las encuestas rotativas.

■ Cuadro II.5
Modelo de encuesta de panel rotativo

Hogar	Tiempo 1	Tiempo 2	Tiempo 3	Tiempo 4	...	Tiempo T
1	X					
2	X	X				
3	X	X	X			
4		X	X	X		
5			X	X	X	
6				X	X	X
...					X	X
n						X

Fuente: Elaboración propia.

C. Rotación de paneles

Como se describió anteriormente, algunas encuestas de hogares en América Latina prevén que un hogar sea visitado en más de una ocasión con el fin de obtener estimaciones precisas acerca de los cambios de estado que el hogar o las personas que lo habitan puedan experimentar. Por ejemplo, un hogar que en un período estuvo en situación de pobreza extrema, puede encontrarse en otro período en situación de pobreza relativa (no extrema) o incluso puede pasar a estar fuera de la pobreza. A su vez, en las encuestas de fuerza laboral, una persona puede pasar de estar empleada en un período a desempleada en otro período. Estos cambios y la dinámica propia que conllevan son de interés para los investigadores y deben contemplarse desde una perspectiva más amplia en lo que respecta al diseño. Nótese que este tipo de variaciones sobre los individuos tiene que ser captada mediante un componente de panel, por lo que las encuestas transversales o repetidas no serían viables para realizar estas estimaciones.

En América Latina hay una gran variedad de encuestas de hogares que utilizan diseños rotativos (véase el anexo A1). Por ejemplo, en la Encuesta Permanente de Hogares de la Argentina, se renueva periódicamente el conjunto de hogares que serán entrevistados mediante un diseño de rotación 2(2)2 en que las viviendas seleccionadas son entrevistadas en dos períodos consecutivos[1]. En los siguientes dos períodos, esas viviendas salen de la selección, y finalmente vuelven a ser encuestadas en los dos últimos períodos

[1] Un diseño de rotación $x(y)z$ se define como aquel en que la vivienda se incluye en el panel durante x períodos, se excluye durante los siguientes y períodos y este patrón se repite z veces a lo largo del tiempo. Los períodos pueden definirse como meses o trimestres. Además, un hogar es visitado un total de $x \times z$ veces.

(INDEC, 2018b). De esta forma, dado que la rotación es trimestral, un hogar es seguido a lo largo de 18 meses, lo que permite cumplir con los objetivos de la encuesta. Este diseño presenta algunas propiedades interesantes, que se ejemplifican mediante el cuadro II.6, en el que se representan los cuatro trimestres de los años 2016, 2017 y 2018 en cuatro grupos de muestra (A, B, C y D) compuestos por los siguientes grupos disjuntos de UPM: *a1*, *a2*, *a3*, *a4*, *b1*, *b2*, *b3*, *b4*, *c1*, *c2*, *c3*, *c4*, *c5*, *d1*, *d2*, *d3* y *d4*.

■ Cuadro II.6
Rotación de paneles en un diseño 2(2)2

Año	Trimestre	A	B	C	D
2016	T1	*a1*	*b1*	*c1*	*d1*
	T2	*a1*	*b2*	*c2*	*d1*
	T3	*a2*	*b2*	*c2*	*d2*
	T4	*a2*	*b1*	*c3*	*d2*
2017	T1	*a1*	*b1*	*c3*	*d3*
	T2	*a1*	*b2*	*c4*	*d3*
	T3	*a2*	*b2*	*c4*	*d4*
	T4	*a2*	*b3*	*c3*	*d4*
2018	T1	*a3*	*b3*	*c3*	*d3*
	T2	*a3*	*b4*	*c4*	*d3*
	T3	*a4*	*b4*	*c4*	*d4*
	T4	*a4*	*b3*	*c5*	*d4*

Fuente: Elaboración propia.

- Entre el primer y el segundo período de medición, hay un traslape del 50% de los hogares. En particular, nótese que, entre los trimestres primero y segundo de 2016, la muestra se conserva en un 50%, puesto que *a1* y *d1* se repiten. Lo mismo sucede en cada trimestre del diseño rotativo.
- En el tercer período no habrá traslape con el primer período. Nótese que entre los trimestres primero y tercero de 2016 no existe ningún elemento en común. Tampoco existe ningún elemento en común entre los trimestres segundo y cuarto de 2016. Este mismo patrón se observa a lo largo del diseño rotativo.
- En el cuarto período se tendrá un 25% de traslape con el primer período. Nótese, por ejemplo, que entre los trimestres primero y cuarto de 2017, *c3* se repite. De la misma manera, entre el cuarto trimestre de 2017 y el tercer trimestre de 2018, *d4* se repite.
- Por último, en el quinto período se volverá a tener un 50% de traslape con respecto al primer período. Por ejemplo, los primeros trimestres de 2016 y 2017 comparten el 50% de la muestra *a1* y *b1*; asimismo, los primeros trimestres de 2017 y 2018 comparten el 50% de la muestra *c3* y *d3*.

Otro ejemplo de una encuesta en la que se utiliza la rotación de paneles es la Encuesta Continua de Empleo del Estado Plurinacional de Bolivia. Esta encuesta, realizada por el Instituto Nacional de Estadística, hace uso de una metodología mixta que permite el seguimiento continuo y transversal de las tasas de desempleo y subocupación, así como el seguimiento de los cambios que se presentan entre los períodos de interés (trimestres y semestres), mediante el análisis longitudinal de los datos en el sector urbano (el diseño no es rotativo en el sector rural debido a la baja incidencia del desempleo en esta zona). En este diseño rotativo 4(0)1, una vivienda es entrevistada durante cuatro trimestres consecutivos y luego sale del panel definitivamente. Un ejemplo de este tipo de diseño se presenta en el cuadro II.7.

■ Cuadro II.7
Rotación de paneles en un diseño 4(0)1

Año	Trimestre	A	B	C	D
2016	T1	*a1*	*b1*	*c1*	*d1*
	T2	*a1*	*b2*	*c1*	*d1*
	T3	*a1*	*b2*	*c2*	*d1*
	T4	*a1*	*b2*	*c2*	*d2*
2017	T1	*a2*	*b2*	*c2*	*d2*
	T2	*a2*	*b3*	*c2*	*d2*
	T3	*a2*	*b3*	*c3*	*d2*
	T4	*a2*	*b3*	*c3*	*d3*
2018	T1	*a3*	*b3*	*c3*	*d3*
	T2	*a3*	*b4*	*c3*	*d3*
	T3	*a3*	*b4*	*c4*	*d3*
	T4	*a3*	*b4*	*c4*	*d4*

Fuente: Elaboración propia.

- Nótese que entre el primer y el segundo período de medición hay un traslape del 75% de los hogares. En particular, entre los trimestres primero y segundo de 2016, la muestra se conserva en tres cuartas partes, puesto que $a1$, $c1$ y $d1$ se repiten. Esto mismo sucede en cada trimestre del diseño rotativo.
- Por otro lado, entre el primer y el tercer período habrá un traslape del 50%. Nótese que entre los trimestres primero y tercero de 2016, la mitad de la muestra se conserva, puesto que $a1$ y $d1$ se repiten. Este mismo patrón se observa a lo largo del diseño rotativo.
- Entre el primer y el cuarto período se tendrá un 25% de traslape. Nótese, por ejemplo, que entre los trimestres primero y cuarto de 2017, $a2$ se repite. De la misma manera, entre el cuarto trimestre de 2017 y el tercer trimestre de 2018, $d3$ se repite.
- Por último, entre el primer y el quinto períodos no hay ningún tipo de traslape.

En los diseños de las encuestas de hogares, se debe tener en cuenta la rotación de los paneles y el número de veces que se visita un hogar. Esta caracterización depende directamente de los indicadores a los que la encuesta debe responder. El diseño de rotación debe ser diferente si el interés se centra en indicadores de cambio trimestral, o si se requieren indicadores de cambio anual. Por ejemplo, el diseño 4(0)1 es conveniente si el objetivo es comparar las estimaciones de la tasa de desocupación en el mismo mes de diferentes años, pero no lo será si se quiere conocer el cambio en la situación laboral de las mismas personas en dos meses iguales de diferentes años. Nótese que un aspecto importante para la definición de los diseños longitudinales es el tiempo durante el cual un hogar permanecerá en el panel. Por supuesto, hay que tener en cuenta que la tasa de ausencia de respuesta y pérdida de muestra por desgaste del respondiente aumentará en la medida en que se solicite a un hogar una participación más duradera.

La definición de los indicadores de interés debe primar sobre el diseño de las encuestas de hogares. Por ejemplo, si el objetivo de la encuesta es estimar el cambio del indicador en dos períodos de tiempo, significa que, en el cálculo de la precisión de las estimaciones, se debe tener en cuenta que las muestras pueden no ser independientes. Por lo tanto, se debe calcular la varianza de la primera y la segunda rondas y la correlación entre las dos rondas de interés. Estos tres componentes deben intervenir en el cálculo de los coeficientes de variación, así como en la determinación del tamaño de la muestra en cada ronda. Como afirman McLaren y Steel (2001, pág. 236), para la estimación de tendencias, definidas a partir de series de tiempo macroeconómicas de los parámetros de interés en los estudios de fuerza laboral, el mejor patrón encontrado es el 1(2)*m*, en el que la vivienda entra al panel en el primer mes, se excluye en los siguientes dos meses y este patrón se repite *m* veces consecutivas. A partir de entonces, la vivienda no vuelve a incluirse en el estudio. En resumen, por la naturaleza de las encuestas de hogares en la región, al pensar en incluir o cambiar la estructura rotativa en el sistema de encuestas de hogares, se debería considerar en primer lugar el diseño de repartición mensual de paneles. En otros capítulos de este documento, se puede encontrar un análisis más profundo de este tipo de estudios longitudinales.

D. Parámetros e indicadores de interés

Las encuestas se utilizan para producir estimaciones de parámetros que describen la situación de una población, respondiendo a los objetivos de la investigación. En general, es posible clasificar en dos grandes grupos los indicadores o parámetros de interés en una encuesta.

i) Indicadores descriptivos

- Medias: promedio de gasto mensual, promedio de ingreso per cápita o promedio de años de educación, entre otros.

- Proporciones: porcentaje de personas que se encuentran por debajo de la línea de indigencia, porcentaje de niños con desnutrición o porcentaje de hogares con pisos de tierra, entre otros.
- Totales: total de ingresos recibidos por concepto de remesas o total del gasto en alimentación, entre otros.
- Tamaños: cardinalidad (número de unidades) de un subgrupo poblacional, tamaño de la fuerza de trabajo, cantidad de personas inactivas o cantidad de mujeres víctimas de acoso laboral, entre otros.

ii) Indicadores analíticos

- Correlación: relación entre la cantidad de libros leídos y los años de escolaridad.
- Regresión: razón de incremento entre ingreso y años de experiencia.

Por lo general, el conocimiento de la población a cualquier nivel se refleja en forma de totales, o de funciones de totales. De ahí que en este documento se profundice en las características inferenciales de los totales, puesto que la generalización a otros parámetros es inmediata. De esta manera, un total poblacional se define como la suma de las observaciones de una variable de interés en la población, denotada como y, y se calcula mediante la siguiente ecuación:

$$t_y = \sum_{k \in U} y_k$$

Donde U hace referencia al universo de estudio, mientras que y_k hace referencia a la variable de interés en el k-ésimo individuo. Por ejemplo, en una investigación social se puede realizar una encuesta para estimar el total del gasto de los hogares de un país en productos específicos de comida y bebidas no alcohólicas. En este ejemplo, la población U corresponde a los hogares, mientras que la variable y corresponde al gasto en comida y bebidas no alcohólicas, que es observada en el k-ésimo hogar y denotada como y_k.

Un caso particular de este parámetro es el tamaño poblacional, que mide la cantidad de unidades que conforman una población y se denota como N. Por lo general, este parámetro es conocido, o al menos se tiene una aproximación de esta cantidad, gracias a la realización de los censos de población y vivienda. En una encuesta de hogares, este parámetro podría denotar el número de hogares del país —que no se conoce exactamente, aunque sí se conocen aproximaciones a esta cantidad (o proyecciones) con base en los resultados de los censos de población y vivienda— o el número de habitantes del país —que tampoco se conoce con exactitud, aunque sí se cuente con proyecciones poblacionales—. Este parámetro también toma la forma de un total poblacional:

$$N = \sum_{k \in U} 1$$

El parámetro más relevante en la investigación social tal vez sea el promedio poblacional, que describe la cantidad que debería asignarse a cada individuo de la población

si hubiese una asignación equitativa de la variable de interés. De esta forma, el promedio se define como la suma de las observaciones de la variable en la población, dividida por el tamaño poblacional N, y se calcula mediante la siguiente expresión:

$$\bar{y}_U = \frac{t_y}{N}$$

Por ejemplo, en una encuesta de hogares es posible estimar el ingreso medio por hogar de la población, definido como el total de los ingresos de todos los hogares del país dividido entre el número de hogares del país. En este caso, la variable de interés y es el ingreso del hogar. De la misma forma, también se podría estimar el gasto promedio de los hogares en educación. En ese caso, la variable de interés, denotada como y, es el gasto de todos los miembros del hogar en este concepto (sin importar la edad ni el nivel de escolaridad) y N sería el número de hogares del país.

Un parámetro que es de particular interés es el tamaño absoluto de un dominio poblacional, que mide la cantidad de unidades que conforman una subpoblación de interés U_d y se denota como N_d. Por ejemplo, en las encuestas de fuerza laboral, es muy importante estimar con gran precisión el número de personas que están desocupadas en un período determinado y comparar su evolución a lo largo del tiempo. En este caso, la subpoblación de interés, o dominio poblacional, estará definida por los desocupados. Nótese que este parámetro se define como un total sobre una variable dicotómica z_{dk}, que toma el valor de 1 si el k-ésimo individuo tiene el atributo de interés y de 0 en caso contrario. Este parámetro se calcula de la siguiente manera:

$$N_d = \sum_{k \in U} z_{dk} = \sum_{k \in U_d} 1$$

De la misma forma, la incidencia relativa de los fenómenos sociales sobre los hogares o personas puede medirse a partir de la proporción de un dominio poblacional. Por ejemplo, la proporción de personas en situación de pobreza o de pobreza extrema toma como base a toda la población, donde la variable de interés z_{dk} indica si el ingreso per cápita de un individuo es inferior a la línea de pobreza. En CEPAL (2018a), se presentan los pormenores metodológicos del cálculo de la pobreza en los países de América Latina y el Caribe. Este parámetro se calcula mediante la siguiente ecuación:

$$P_d = \frac{N_d}{N}$$

En algunos casos, resulta interesante conocer el total de una variable en una subpoblación. Por ejemplo, el total del ingreso de las mujeres, o el total de gasto en el área rural. En estas situaciones, el parámetro se conoce como total del dominio y se puede calcular mediante la siguiente expresión:

$$t_{yd} = \sum_{k \in U} y_k z_{dk} = \sum_{k \in U_d} y_k$$

Asimismo, puede ser de interés calcular medidas relativas, como la media del dominio. De esta forma, es posible calcular la media de los ingresos de hombres y mujeres, o la media de los ingresos de los ocupados, o la media del gasto en comida de la población indígena. Este parámetro puede calcularse mediante la siguiente expresión:

$$\bar{y}_{U_d} = \frac{t_{y_d}}{N_d}$$

Por último, la razón poblacional se calcula como el cociente entre dos totales. El primer total t_y se asocia a una variable de interés y, mientras que el segundo total t_x se asocia a una variable de interés x. Por ejemplo, en la medición del mercado de trabajo, la tasa de desocupación es una razón entre el total de personas desocupadas y el total de personas activas. Cabe mencionar que para clasificar a una persona como desocupada, ocupada, activa o inactiva, es necesario indagar sobre ello en la encuesta a cada uno de los miembros del hogar. Por lo tanto, ambas cantidades (el numerador y el denominador), son desconocidas de antemano. Es más, la condición de ocupación de las personas puede variar entre los períodos de observación. Este parámetro se calcula mediante la siguiente expresión:

$$R_U = \frac{t_y}{t_x}$$

Los indicadores de pobreza pueden expresarse como razones poblacionales. Tal es el caso de la incidencia, la brecha y la gravedad de la pobreza, parámetros que se expresan en términos de un umbral sobre el poder adquisitivo (Foster, Greer y Thorbecke, 1984). Este tipo de indicadores de pobreza se pueden expresar mediante la siguiente relación:

$$F_\alpha = \frac{1}{N} \sum_U \left(\frac{u-y_k}{u}\right)^\alpha I_{(y_k < u)}$$

Donde y_k determina el ingreso del individuo k, u se refiere al umbral que establece la línea de pobreza y $\alpha \geq 0$. Por ejemplo, en el caso en que $\alpha=0$, este indicador permite calcular la tasa de pobreza, que es la incidencia de este fenómeno en la población. A su vez, si $\alpha=1$, este indicador permite calcular la brecha de pobreza, que es la cantidad de dinero relativa que se necesitaría en promedio para que un país no tuviera personas en situación de pobreza. Por último, si $\alpha=2$, este indicador medirá la gravedad de la pobreza, como una combinación entre la incidencia de la pobreza de los hogares, la brecha absoluta de ingreso de los hogares en situación de pobreza y la desigualdad de ingresos entre los hogares en situación de pobreza.

En este punto, cabe resaltar que, para la definición de los parámetros básicos que se quieren estimar en una encuesta, el papel de los totales poblacionales es muy relevante. Asimismo, existen otros parámetros no lineales que pueden considerarse complejos,

pero que, al igual que los antes mencionados, resultan ser también una función de totales poblacionales. Téngase en cuenta, por ejemplo, el cambio neto de los totales de la variable de interés y en dos períodos de tiempo (t_1 y t_2) dado por la siguiente expresión:

$$\Delta_y = t_{y(2)} - t_{y(1)}$$

Donde $t_{y(2)}$ es el total de interés en el tiempo t=2, y $t_{y(1)}$ lo es en el tiempo t=1. Este tipo de parámetros son muy comunes en las encuestas que se realizan para conocer la estructura y los cambios del mercado de trabajo. Por ejemplo, en el cuadro II.8 se muestra la composición del mercado de trabajo en una población observada en dos períodos de interés. De esta forma, los totales marginales del cuadro corresponden a los cambios netos que permiten una comparación simple con el período anterior. En concreto, es posible observar que hay 313.000 empleados menos, 80.000 desempleados menos y 393.000 inactivos más en el segundo período, en comparación con el primero.

■ Cuadro II.8
Composición del mercado de trabajo en dos períodos de tiempo
(En miles de personas)

Condición	Ocupado	Desocupado	Inactivo	Total
Ocupado	9 222	128	662	**10 012**
Desocupado	221	322	151	**694**
Inactivo	256	164	5 941	**6 361**
Total	**9 699**	**614**	**6 754**	**17 067**

Fuente: Elaboración propia.
Nota: Las columnas corresponden al segundo período y las filas, al primero.

Una comparación más profunda está dada en términos de los cambios brutos, que corresponden a las entradas del cuadro II.8. De esta manera, los cambios en la fuerza de trabajo de un período a otro se explican porque el $92{,}1\%=(9.222/10.012)\times 100\%$ de los empleados conservó su empleo, el $31{,}8\%=(221/694)\times 100\%$ de los desempleados y el $4{,}0\%=(256/6.361)\times 100\%$ de los inactivos consiguió un nuevo empleo, el $6{,}6\%=(662/10.012)\times 100\%$ de los empleados está ahora inactivo en la fuerza laboral y el $1{,}3\%=(128/10.012)\times 100\%$ de los empleados perdió su empleo. Asimismo, el $46{,}4\%=(322/694)\times 100\%$ de los desempleados conservó su clasificación, el $2{,}6\%=(256/6.361)\times 100\%$ de los inactivos entró a la fuerza laboral como desempleado y el $21{,}8\%=(151/694)\times 100\%$ de los desempleados está ahora inactivo.

1. Algunos ejemplos de indicadores de interés y su relación con los diferentes tipos de encuestas

En esta sección se vinculan algunos de los parámetros antes mencionados con los tipos más comunes de encuestas. Estos ejemplos presentan algunas indicaciones sobre el tipo de encuestas que se llevan a cabo en América Latina y permiten examinar el razonamiento en

que se basan. Tomando en consideración las características generales de las encuestas de hogares, Duncan y Kalton (1987) mencionan las situaciones que se describen a continuación.

- **Estimación de parámetros poblacionales en un punto del tiempo.** Por ejemplo, supóngase que se quiere estimar el ingreso per cápita promedio por área (rural o urbana) en las regiones de un país. En este tipo de estudios, las encuestas aptas serían las transversales, las repetidas, las de panel rotativo y las de panel dividido. Nótese que las encuestas de panel puro no se consideran aptas para estimar este parámetro, puesto que la muestra no es representativa de la población en el momento actual, sino que, por el contrario, es representativa de la población en el momento en que se extrajo la muestra.
- **Estimación de cambios netos.** Si se quisiera estimar la diferencia en el número de ocupados de la fuerza de trabajo entre el primer y el segundo trimestre de 2021 en un país, las encuestas aptas serían las repetidas, las de panel rotativo y las de panel dividido. Una encuesta transversal no se consideraría apta para esta estimación, puesto que su frecuencia de realización no es trimestral. Al igual que en el caso del parámetro anterior, las encuestas de panel puro no son las adecuadas para captar este parámetro, puesto que la muestra no es representativa de la población en el momento actual.
- **Estimación de cambios brutos y componentes individuales.** Para estimar el porcentaje de personas ocupadas en el segundo trimestre de 2021 que estuvieron desocupadas en el primer trimestre de ese año en un país, es necesario que la encuesta incluya algún patrón de selección de los mismos individuos en los dos períodos. Por ello, las únicas encuestas aptas para estimar este tipo de cambios brutos son las de panel, panel rotativo y panel dividido. Las encuestas transversales o repetidas no servirían para este tipo de estimaciones, puesto que en su diseño no se considera a los mismos individuos en la muestra en dos períodos de tiempo.
- **Estimación de la incidencia de eventos en un período de tiempo.** Supóngase que se quiere estimar la proporción de mujeres que fueron víctimas de un evento de violencia en los últimos seis meses en un país. En este caso, todas las encuestas resultarían aptas si se realizaran ligeras modificaciones en el diseño. Por ejemplo, la encuesta transversal debería contener preguntas retrospectivas, las encuestas repetidas podrían ser agregadas en los últimos seis meses y en las encuestas de tipo panel rotativo y divididas se debería preguntar sobre este evento en cada medición de los últimos seis meses.
- **Estimación de la incidencia de eventos raros a lo largo del tiempo.** Por ejemplo, si se quisiera estimar la proporción de personas con una enfermedad rara, es posible que las encuestas transversales y de tipo panel no sean las más apropiadas. En el primer caso, dado que el evento es raro por definición, los requisitos de tamaño de la muestra en una encuesta transversal sobrepasarían el presupuesto y los costos de una encuesta normal. En el segundo caso, además de las consideraciones

planteadas sobre el tamaño de la muestra, por la misma definición de evento raro, tampoco sería plausible que los individuos del panel experimentaran dicho evento a lo largo del tiempo. Por otro lado, al agregar las encuestas repetidas, las de panel rotativas y la parte nueva del panel dividido, se podría llegar al tamaño de muestra adecuado para captar esta incidencia de forma precisa y eficiente.

Estos últimos ejemplos ponen de relieve la importancia de contar con procedimientos adecuados de acumulación de datos y encuestas a lo largo de un período de interés, por ejemplo, de forma anual o semestral. La acumulación de datos permite obtener una buena base inferencial para estimar todo tipo de parámetros en una ventana más amplia de tiempo. Es posible acumular datos de manera eficiente por medio de la agregación de encuestas repetidas. De esta forma, se definiría una agregación de datos vertical que añade filas, puesto que en cada recogida de datos aparecen nuevos individuos. Esto se debe a que, en el diseño de las encuestas repetidas, se selecciona a diferentes individuos en cada punto del tiempo. Este es el caso de la Gran Encuesta Integrada de Hogares de Colombia, que está diseñada para lograr la representatividad a niveles de desagregación mayores, juntando a los individuos observados en 12 mediciones continuas en un año.

Por otro lado, las encuestas de panel permiten un tipo diferente de agregación, no basado en individuos, sino en variables a lo largo del tiempo. A diferencia de las encuestas repetidas, las encuestas de panel, panel rotativo o panel dividido permiten observar a los individuos en diferentes períodos de tiempo. La agregación puede hacerse de forma horizontal, manteniendo a los individuos en las filas y añadiendo columnas cada vez que se realice una nueva medición en un período de tiempo diferente.

Capítulo III
Definición del marco de muestreo

Todo procedimiento de muestreo probabilístico requiere un dispositivo que permita identificar y ubicar a todas y cada una de las unidades pertenecientes a la población objetivo, que posteriormente participarán en el proceso de selección aleatoria definida por la muestra. Este dispositivo se conoce como "marco de muestreo".

La mayoría de las encuestas de hogares que son probabilísticas se caracterizan por utilizar marcos de muestreo de áreas (agregados cartográficos como segmentos censales, sectores censales o áreas de enumeración). Estos se derivan directamente de los censos de población y vivienda, aunque también es posible elaborar marcos de líneas telefónicas fijas y móviles. En general, sin esta herramienta no es posible realizar ningún procedimiento de muestreo probabilístico, por lo que la etapa de definir y alistar un buen marco de muestreo se aborda con bastante rigurosidad en las oficinas nacionales de estadística (ONE) durante la recopilación de datos para el censo y posteriormente. Este proceso ocurre en el marco de los trabajos censales, cuando se actualiza toda la cartografía nacional.

A. Conceptos fundamentales

Como se verá en los capítulos posteriores, dependiendo de la naturaleza del marco de muestreo se pueden proponer diferentes tipos de diseños muestrales. Por ejemplo, cuando se dispone de un marco de elementos, se puede aplicar un diseño de muestreo de elementos. En algunas ocasiones, se utilizan diseños de muestreo de conglomerados, aunque se disponga de un marco de elementos. Si no se dispone de un marco de elementos (o es muy costoso

elaborarlo), se debe recurrir a diseños de muestreo en conglomerados; es decir, se utilizan marcos de conglomerados. Por ejemplo, al realizar una encuesta cuya unidad de observación son las personas que viven en una ciudad, es muy difícil acceder a un marco de muestreo de dichas personas. Sin embargo, en una primera instancia, se puede tener acceso a la división cartográfica de la ciudad y así seleccionar algunas de sus comunas, localidades o barrios, y luego seleccionar a las personas en una segunda o tercera instancia. En el ejemplo anterior, las comunas, localidades o barrios son un ejemplo claro de conglomerados, es decir, agrupaciones de elementos que se caracterizan por aparecer naturalmente.

Cuando se dispone de listados de unidades —por ejemplo, el listado de empleados de una entidad—, es posible aplicar un diseño de muestreo de elementos, realizar la correspondiente selección aleatoria y, de acuerdo con ese mismo diseño, realizar las estimaciones necesarias. Sin embargo, al realizar la planificación de una encuesta de hogares, es muy poco probable que se utilicen marcos de elementos, a no ser que el muestreo definido se realice en dos fases: una primera fase de selección de hogares y enlistamiento de personas o unidades, y una segunda fase de selección de personas o unidades. Por ejemplo, el Instituto Nacional de Estadística y Censos (INEC) de Costa Rica realiza la Encuesta Nacional de Microempresas de los Hogares con base en la muestra de la Encuesta Nacional de Hogares (primera fase), donde se identifican las actividades económicas de los respondientes y se enlistan los trabajadores autónomos. En una segunda fase, se selecciona una submuestra con base en este marco de elementos. En general, se pueden utilizar dos tipos de marcos de muestreo, a saber:

i) **De lista:** listados físicos o magnéticos, ficheros o archivos de expedientes que permiten identificar y ubicar los objetos que participarán en el sorteo aleatorio.

ii) **De área:** mapas de ciudades y regiones en formato físico o magnético, fotografías aéreas, imágenes de satélite o similares que permiten delimitar regiones o unidades geográficas para posibilitar su identificación y su ubicación sobre el terreno.

Es una virtud del marco si contiene información auxiliar que permita aplicar diseños muestrales o estimadores que conduzcan a estrategias de muestreo más eficientes con respecto a la precisión de los resultados. También es una ventaja que la información auxiliar esté clasificada de forma sistemática y conveniente[1]. La información auxiliar discreta en el marco de muestreo permite la desagregación de la población objetivo en categorías o grupos poblacionales más pequeños. Estas desagregaciones pueden referirse, por ejemplo, a nivel socioeconómico, región o departamento. A su vez, la información auxiliar continua —en forma de una o varias características de interés de tipo continuo y positivas—, que esté muy relacionada con la característica de interés, permitirá mejorar la eficiencia de la estrategia de muestreo. Por otra parte, el marco de muestreo se considera defectuoso si presenta alguno o varios de los siguientes rasgos:

[1] Toda información disponible a nivel poblacional, o respecto de todos y cada uno de los elementos del universo, afecta directamente la estrategia empleada para alcanzar los objetivos de la investigación. Con respecto a la información auxiliar que pueda existir sobre cada elemento de la población, es deseable que esté bien correlacionada con la variable de interés.

- **Sobrecobertura:** se presenta si en el dispositivo aparecen objetos que no pertenecen a la población objetivo; es decir, si no son todos los que están.
- **Subcobertura:** se da cuando algunos elementos de la población objetivo no aparecen en el marco de muestreo, o cuando no se ha actualizado la entrada de nuevos integrantes; es decir, no están todos los que son.
- **Duplicación:** se presenta si en el dispositivo aparecen varios registros de un mismo objeto. La razón más frecuente para la presencia de este defecto es la construcción no cuidadosa del marco a partir de la unión de registros administrativos de dos o más fuentes de información.

Estos defectos ocasionan errores en el cálculo de las expresiones que se utilizarán para obtener las correspondientes estimaciones, lo que produce sesgos y pérdidas de precisión y, en algunos casos, hace que los resultados del estudio se pongan en entredicho. No obstante, una vez que se ha definido el marco de muestreo, este empieza un período de pérdida de calidad y envejecimiento, lo que plantea dificultades para la realización de las encuestas de hogares que lo utilizan. Es por esta razón por la que, a partir de la realización de los censos de población y vivienda, las ONE actualizan sus marcos de muestreo.

En resumen, el marco de muestreo es cualquier dispositivo o mecanismo que se utilice para obtener acceso observacional a la población de interés, para identificar y seleccionar una muestra, de manera que se respete el diseño de muestreo probabilístico, y para establecer contacto con los elementos seleccionados, de manera presencial, por correo postal, por correo electrónico, o mediante procedimientos automatizados como los sistemas de captura conocidos como entrevista personal asistida por computadora (*computer-assisted personal interviewing* (CAPI)) o entrevista telefónica asistida por computadora (*computer-assisted telephone interviewing* (CATI)).

Por otro lado, recordando que la población objetivo constituye el conjunto de elementos acerca del cual se desea información y sobre cuyos parámetros se requieren estimaciones exactas y precisas, la población del marco es el conjunto de los elementos que son enlistados directamente como unidades en el marco o identificados mediante un marco más complejo, como un marco para selección en varias etapas. Además, los elementos son las entidades que componen la población y las unidades de muestreo son las entidades del marco muestral. Cuando no hay uno disponible, es posible elaborarlo. Las siguientes características son deseables para un marco de muestreo:

- Que las unidades en el marco se identifiquen con una secuencia única.
- Que cualquier unidad pueda ser ubicada (dirección postal o electrónica, teléfono fijo o celular, entre otros).
- Que la muestra se pueda ordenar sistemáticamente (geografía, tamaño).
- Que el marco contenga información adicional respecto de cada unidad.
- Que se especifique el dominio geográfico o socioeconómico al que pertenece cada unidad.

- Que cada elemento de la población está presente solo una vez.
- Que no se incluyan elementos que no estén en la población.
- Que todos los elementos de la población de interés estén en el marco muestral.

La calidad del marco puede medirse a partir de la relación que existe entre la población objetivo y la población del marco. Esto quiere decir que la población enmarcada y la población de interés no siempre coinciden plenamente.

En las encuestas de hogares que precisan de un marco de áreas para su realización, el proceso de selección sistemática de los hogares necesita contar con un marco de muestreo que sirva de vínculo entre los hogares y las unidades de muestreo de las primeras etapas y que permita tener acceso a la población de interés. Como afirma Gutiérrez (2016), el marco más utilizado en este tipo de encuestas es el de áreas geográficas que vinculan directamente a los hogares o personas con un listado de divisiones cartográficas exhaustivas. Por esta razón, los diseños de muestreo de estas encuestas se apoyan en la aglomeración natural de los hogares en segmentos cartográficos, que a su vez están contenidos en agrupaciones mayores. ¿Cómo se aglomeran las personas y cómo se puede realizar un diseño de muestreo con base en esta forma de aglomeración? Las personas se aglomeran en hogares, los cuales a su vez se aglomeran en comunidades más grandes: barrios, comunas o segmentos. Estas comunidades forman ciudades, secciones administrativas o centros de poblados, entre otras, y la reunión de estas divisiones da como resultado el conjunto completo de unidades de interés en el país.

Por lo tanto, a pesar de que ningún país tiene a su disposición una lista actualizada de todos los hogares, junto con su ubicación e identificación, sí existen en todos los países listas de los segmentos cartográficos presentes en las zonas urbanas y rurales, que se actualizan en cada censo. De esta forma, si se selecciona de forma probabilística una muestra de sectores y, dentro de cada sector, se selecciona de forma probabilística una muestra de hogares, de manera indirecta se estaría seleccionando una muestra de hogares que puede representar la realidad de todo un país.

B. Los censos y su incidencia en los marcos de muestreo

Como se mencionó anteriormente, una característica esencial de los diseños de las encuestas de hogares es que la selección de las unidades finales de muestreo se compone de varias etapas, de acuerdo con las agrupaciones definidas en los marcos de muestreo, que usualmente son marcos de área obtenidos de la división geográfica del país, la región o el municipio en áreas menores mutuamente excluyentes. Los institutos de estadística de América Latina hacen grandes esfuerzos por mantener actualizados sus marcos de muestreo. Por ejemplo, en la Encuesta Nacional de Hogares de Costa Rica, se utiliza un

marco muestral construido a partir de los censos nacionales de población y vivienda de 2011, cuyas unidades son superficies geográficas asociadas a las viviendas. Este marco permite la definición de unidades primarias de muestreo (UPM) con 150 viviendas en las zonas urbanas y 100 viviendas en las zonas rurales. En esta ocasión en particular, el marco estuvo conformado por 10.461 UPM (el 64,5% de ellas urbanas y el 35,5% rurales).

Gambino y Silva (2009) mencionan que, en la práctica, la consecución de los marcos de lista de los hogares en la última etapa del muestreo puede tornarse difícil, puesto que dentro del conglomerado no es fácil observar de manera exhaustiva los hogares, sobre todo cuando la frontera del conglomerado es una línea imaginaria. Por ejemplo, en la mayoría de los casos, en el sector urbano, la distinción entre dos conglomerados está demarcada con claridad por las calles que conforman la ciudad o el centro poblado. Sin embargo, en el ámbito rural, no solamente sirven para delimitar los conglomerados los caminos existentes, sino que también se utilizan para este fin los accidentes topográficos y las señales naturales. De la misma manera, esta delimitación se vuelve compleja cuando ocurren cambios en la infraestructura del área y aparecen nuevas construcciones.

Cabe mencionar que, en general, en el estudio de un fenómeno social, las desagregaciones geográficas más amplias constituyen un interés natural para los usuarios de las encuestas. Por ello, los investigadores que planean las encuestas quieren poder desagregar la información por las regiones geográficas más grandes, que a su vez tienen cierta independencia política y administrativa. Las estadísticas nacionales que se publican a partir de las encuestas de hogares cobran mayor relevancia a nivel de las regiones, los estados o los departamentos. Este tipo de desagregaciones se conocen con el nombre de "dominios de representación", que a su vez son agregaciones de los "estratos de muestreo". Los diseños de las encuestas de hogares han ido evolucionando para permitir que este tipo de subpoblaciones tenga representatividad en la encuesta. Además, si la característica de interés con la que se planea la encuesta hace que la distribución de la población sea muy sesgada, como en el caso de los ingresos o gastos, es recomendable crear estratos de inclusión forzosa con las unidades más importantes de la población. Esta práctica garantiza un menor error de muestreo.

Siguiendo las recomendaciones internacionales, los países de la región llevan a cabo sus censos cada diez años, aunque, desafortunadamente, en algunos casos este período se extiende. En esta recopilación masiva de información se enlistan todos los hogares del país, se enumeran todos sus habitantes y se observan algunas variables de interés que servirán, a su vez, para sentar las bases de comparación de las cifras durante los siguientes diez años. El período que transcurre entre la realización de dos censos se denomina período intercensal y en él se realizan encuestas de hogares de diferentes constructos económicos y sociales. Los institutos nacionales de estadística (INE) utilizan las particiones geográficas y cartográficas generadas al realizar el censo con el fin de seleccionar muestras de hogares, mediante diseños en varias etapas. Comúnmente, estas particiones reciben el nombre de secciones cartográficas y están formadas por un número determinado de hogares contiguos. Como se mencionó anteriormente, estas particiones

se denominan unidades primarias de muestreo (UPM), que, en el área urbana, pueden ser manzanas o agregaciones de manzanas y, en el área rural, pueden ser secciones administrativas o sectores censales definidos de antemano.

Algunos países hacen uso de la información censal para definir una estratificación socioeconómica sobre los segmentos cartográficos del marco de muestreo, utilizando la información recolectada en el censo de población más reciente. Esta práctica representa una ventaja metodológica, ya que, en la mayoría de las encuestas, los parámetros de interés muestran un comportamiento estructural diferente en cada uno de los subgrupos poblacionales creados, de modo que suelen tener una mayor precisión para la estimación de los parámetros de interés. Por ejemplo, a partir del censo, es posible crear un índice de condiciones de vivienda o bienestar (teniendo en cuenta las definiciones de las necesidades básicas insatisfechas o la pobreza multidimensional) para definir grupos de viviendas mutuamente excluyentes que sean muy disímiles entre sí, aunque contengan viviendas parecidas. De esta forma, es posible estratificar los sectores cartográficos de todo un país y obtener estimaciones más precisas de los indicadores sociales (desocupación, pobreza o ingreso medio, entre otros).

En el caso de la Gran Encuesta Integrada de Hogares en Colombia, los criterios de estratificación forman dos grupos. El primero corresponde a las 24 capitales, junto con sus áreas metropolitanas, y el segundo, al resto de las cabeceras municipales, centros poblados y las áreas rurales dispersas. La encuesta también incluye criterios de estratificación económica a nivel municipal, como nivel de urbanización y estructura de la población, basada en la proporción de habitantes con necesidades básicas insatisfechas. De la misma manera, en el diseño de la muestra maestra del Instituto Nacional de Estadística y Geografía (INEGI) de México, se contempla este tipo de estratificación basada en los indicadores obtenidos con la información del Censo de Población y Vivienda 2010. Antes del proceso de estratificación sociodemográfica, fue necesario elaborar y seleccionar una serie de variables que lograran, en conjunto, separar el universo de UPM en agrupaciones que mejoraran las principales estimaciones de las diferentes encuestas usuarias del marco de muestreo (INEGI, 2012).

Ante la ausencia de un marco de muestreo de hogares y personas en los países de la región, el diseño de las encuestas de hogares se considera complejo, puesto que entraña varias etapas de selección y estratificación. Los marcos de muestreo están conformados por UPM que se definen como segmentos cartográficos individuales, como una agrupación de segmentos o incluso como una división de segmentos masivos. Esto se puede ilustrar al tomar en consideración el estrato urbano, donde las UPM corresponden a manzanas (o agregaciones o particiones de manzanas), mientras que, en el caso rural, las UPM corresponden a comunidades (o agregaciones o particiones de comunidades). En cualquier caso, la unidad de observación está constituida por las viviendas ocupadas particulares donde residen personas. En general, las UPM no tienen el mismo tamaño dentro de los estratos, incluso después de crearlas cuidadosamente; es decir, no están constituidas por un número exactamente igual de viviendas. El caso más evidente es el de las áreas rurales, donde podría ocurrir que una única UPM agrupe un conjunto de viviendas con demasiada

heterogeneidad y una alta dispersión geográfica. Así, es posible encontrar UPM con pocas viviendas o UPM con demasiadas viviendas. Esto constituye una desventaja técnica a la hora de establecer metodologías apropiadas para la recolección de la información primaria y, además, para la estimación de los errores de muestreo que se derivan de las encuestas de hogares. Por ese motivo, algunos países están considerando la redefinición de las UPM como unidades con un número uniforme de viviendas.

Como se indicó anteriormente, es usual que, tras la realización de un nuevo censo, se actualice el marco de muestreo con el que se seleccionarán las viviendas y hogares en todas las encuestas subsiguientes. Por la naturaleza de los censos, los INE deben recorrer la geografía de los países y producir una nueva cartografía que derivará en la actualización de los marcos de muestreo. Por ejemplo, si un país cuenta con un marco de muestreo que consta de 10.000 UPM, cada una de estas deberá ser clasificada por medio de una estratificación socioeconómica que se basará en la información recolectada en el último censo de población y vivienda. Kish (1965, pág. 183) afirma que la selección de UPM con tamaño desigual acarrea algunos problemas técnicos, como el hecho de que el tamaño de muestra final se convierte en una variable aleatoria, que depende de la probabilidad de selección de las UPM más grandes o más pequeñas. Ello aumenta la incertidumbre en cuanto al costo final del operativo, pues si en una primera instancia se seleccionan UPM con pocas viviendas, será necesario realizar otro proceso de selección de UPM para completar la cuota de viviendas.

Sobre esta base, y en concordancia con las recomendaciones de la CEPAL (2021), sería esperable que la actualización de la cartografía y de los marcos de muestreo se realizara, como mínimo, cada diez años. Es importante que estas actualizaciones conlleven una definición de los marcos de muestreo adecuada para lograr una mayor fluidez en los procesos logísticos de selección de hogares y mejorar la precisión de las estimaciones de los parámetros de interés. Por ejemplo, una forma muy conveniente de abordar este desafío consistiría en crear UPM que contengan, en la medida de lo posible, un mismo número de viviendas. De esta manera se mantendría una distribución uniforme en cada estrato. Siguiendo el consejo de Valliant, Dever y Kreuter (2013, pág. 212), si el equipo de planificación de la encuesta tiene la flexibilidad de definir las UPM, como suele ocurrir en las encuestas de hogares, dichas UPM deberían estar conformadas definitivamente por una cantidad igual de viviendas.

C. Definición de las unidades primarias de muestreo

La determinación del marco de muestreo para las encuestas de hogares responde básicamente a un objetivo, a saber, la definición de las unidades primarias de muestreo. A fin de optimizar esta solución, es necesario responder una pregunta fundamental: ¿cuál debe ser el tamaño

apropiado de las UPM? No es lo mismo definir las UPM como agregaciones de 20 hogares que de 1.000 hogares. Esta pregunta debe abordarse, en principio, desde el punto de vista técnico, donde confluyan diferentes perspectivas (de muestreo, logísticas, presupuestales y cartográficas). Por ejemplo, Valliant, Dever y Kreuter (2013, cuadro 9.1) mencionan el caso en que, con distintas definiciones del tamaño de las UPM, se evidencian pérdidas o ganancias significativas de eficiencia en los estimadores de las encuestas de hogares.

De esta manera, un primer acercamiento a la definición de las UPM es establecer la unión o escisión de los sectores o secciones cartográficas, o áreas de empadronamiento vinculados a los censos de población y vivienda, como insumo para la creación de las UPM. Como se discutió anteriormente, el objetivo del marco es tratar de proporcionar la mejor información para la selección de las unidades, de modo que se reduzca la variabilidad de la estrategia de muestreo. Por lo tanto, después de revisar minuciosamente los conjuntos de datos censales y la información cartográfica del censo en los niveles básicos (en adelante, y sin pérdida de generalidad, se denominarán secciones censales), es necesario elaborar un algoritmo que permita crear UPM desde la cartografía, sobre la base de uniones contiguas de secciones censales, que respeten los siguientes principios:

- La conformación de las UPM excluye todas las estructuras que no contienen hogares particulares ocupados.
- Las nuevas UPM generadas a partir de la unión de sectores censales deben estar contenidas en un solo municipio del país; es decir, no podrán definirse UPM que pertenezcan a dos o más municipios.
- De la misma forma, debe haber una diferenciación estricta entre las áreas urbanas y rurales. Ninguna UPM podrá estar definida en ambas áreas.

Nótese que siempre será necesario realizar una actualización de las viviendas con hogares particulares ocupados en las UPM seleccionadas en la primera etapa de muestreo. Esta actualización dará lugar al cálculo de las probabilidades de inclusión en la segunda etapa, sin la cual no se podrían calcular factores de expansión que favorezcan el insesgamiento de los estimadores utilizados en las encuestas de hogares. Dado que este proceso es sistemático y debe realizarse a lo largo del período intercensal, contar con UPM demasiado grandes (como pueden llegar a ser los sectores o segmentos censales, las áreas de empadronamiento o los lugares poblados) no es una alternativa viable en sentido presupuestario, puesto que se incrementarían los costos asociados a la actualización y no habría uniformidad en los procesos de muestreo.

El tamaño de las UPM en América Latina ronda las 75 a 225 viviendas. Para que exista una mayor eficiencia (logística y estadística) a la hora de realizar un muestreo en dos etapas, se recomienda que las UPM conformadas tengan algún grado de explicación con respecto a las características de interés que se quieren medir en la población. Por consiguiente, es

necesario revisar los tamaños de estas agregaciones y su comportamiento en relación con el coeficiente de correlación intraclase (CCI). Como se observa en Cochran (1977) y Gutiérrez (2016), al definir las UPM, el parámetro predominante que se debe considerar es el CCI. Para la realización de encuestas con selección en múltiples etapas, este coeficiente puede aproximarse mediante la siguiente expresión (Valliant, Dever y Kreuter, 2013):

$$CCI = \frac{SCE}{SCE+SCD}$$

Donde SCE es la suma de cuadrados relativa de los totales de la característica de interés entre las UPM y SCD es la suma de cuadrados relativa de los totales de la característica de interés dentro las UPM. Por su parte, CCI es una medida de homogeneidad entre las variables que se desean medir y la conformación de las UPM. Además de afectar la variabilidad del estimador en muestreos multietápicos, esta medida determina el tamaño de muestra necesario para satisfacer los requisitos de precisión de una encuesta de hogares. En algunos textos clásicos de muestreo, el CCI también se denota como ρ.

La magnitud del CCI está directamente ligada al tamaño de las UPM. Por lo tanto, en la conformación del marco de muestreo, es necesario ejecutar un algoritmo de control de tamaño de las UPM de tal forma que el CCI sea satisfactorio y coherente en los indicadores censales disponibles. Entre estos, cabe mencionar las dimensiones del indicador de necesidades básicas insatisfechas (NBI), los indicadores del mercado de trabajo y los indicadores demográficos.

En general, cuando el tamaño de las UPM es muy pequeño, las características de los elementos que se encuentran dentro de las UPM serán muy similares (sobre todo en el caso de los indicadores socioeconómicos). Por otro lado, si el tamaño de las UPM es demasiado grande, las características de los elementos serán más heterogéneas. Nótese que la disparidad en los tamaños de las UPM redunda en que los totales de las características de interés sean muy disímiles entre una UPM y otra. Teniendo en cuenta la forma funcional de la varianza del estimador clásico, se generará más varianza en el componente SCE; por ende, el CCI será más grande y se perderá precisión en el muestreo multietápico.

Valliant, Dever y Kreuter (2013) afirman que la práctica estándar consiste en combinar las secciones pequeñas o grupos de bloques cercanos geográficamente para que todas las UPM tengan al menos un número mínimo de personas. La variación en los tamaños de las UPM tiene un efecto marcado en el CCI (medida necesaria para diseñar una muestra) y, en el caso de las encuestas de hogares, se puede tener cierta flexibilidad en la formación de estos grupos. Por esa razón, las UPM deberían conformarse con un número casi igual de viviendas. En general, en el proceso de elaboración de las UPM, se deberían tener en cuenta las siguientes características:

- Límites geográficos y contenencia espacial, pues las UPM deben estar contenidas dentro de límites departamentales o municipales, y estar diferenciadas por su naturaleza urbana o rural.

- Tamaño y extensión, pues se debe procurar que las UPM estén dentro de rangos predefinidos en términos de número de viviendas y personas, respetando los límites geográficos, y que su extensión en kilómetros cuadrados no supere un umbral predefinido para el operativo de campo.

De esta manera, las cargas de trabajo (en los procesos de actualización, supervisión y recolección de la información primaria) serán uniformes. Además, las estimaciones resultantes serán óptimas en términos de eficiencia y precisión estadística, puesto que los pesos de muestreo serán uniformes y minimizarán la varianza de las estimaciones directas. A partir de la información contenida en los censos de población y vivienda, se podrían utilizar diferentes variables a fin de evaluar la idoneidad de las UPM con el CCI y el efecto del diseño. Por ejemplo, para evaluar la idoneidad de las UPM, es posible analizar las siguientes variables agrupadas en los correspondientes constructos:

- Variables demográficas: grupos quinquenales de edad o sexo.
- Necesidades básicas insatisfechas y sus dimensiones (acceso a la vivienda, acceso a servicios sanitarios, acceso a educación, situación de ocupación y capacidad económica).
- Variables de fuerza laboral: población en edad de trabajar, población económicamente activa, desocupados y ocupados.

En general, las medidas de correlación intraclase deben ser coherentes con las experiencias locales anteriores o con experiencias regionales que demuestren que el algoritmo de unión o escisión de los sectores censales proporciona como resultado nuevas UPM que conservan las propiedades explicativas de los grupos desde el censo, con la ventaja de controlar su tamaño en términos de viviendas.

Hansen, Hurwitz y Madow (1953) encontraron una relación marcada entre el tamaño de las UPM y la magnitud del CCI. Cuanto más pequeños sean los conglomerados, mayor será el CCI; cuanto más grandes sean los conglomerados, menor será el CCI. Esta relación tiene una repercusión directa en la forma en que se llevarán a cabo las encuestas en el período intercensal. Si se crean UPM demasiado pequeñas, se precisará un tamaño de muestra de UPM mucho mayor y, por ende, habrá un mayor costo logístico y económico. Si se crean UPM demasiado grandes, se precisará un menor tamaño de muestra, pero con UPM de dimensiones inmanejables, que implicarán operativos de actualización, supervisión y recolección demasiado costosos, junto con una gran pérdida de precisión estadística.

Para ejemplificar la relación entre el CCI y el tamaño de muestra, pueden considerarse los siguientes escenarios:

- Si el CCI es cercano a cero, las UPM serán demasiado heterogéneas por dentro y muy homogéneas entre sí; por lo tanto, se necesitarán muy pocas UPM para obtener una inferencia precisa. Esto quiere decir que existe mucha dispersión dentro de las UPM, pero, a la vez, hay muy poca variación entre ellas. En el caso de que el CCI sea idéntico a cero, solo se necesitaría una UPM en la muestra para obtener una

estimación precisa, con un submuestreo exhaustivo de todas las unidades dentro de la UPM (puesto que todas las unidades de la UPM serán diferentes).

- Si el CCI es cercano a uno, las UPM serán demasiado homogéneas por dentro y muy heterogéneas entre sí; por lo tanto, se necesitará una muestra grande de UPM para obtener una inferencia precisa. Esto quiere decir que existe poca dispersión dentro de las UPM, pero, a la vez, hay mucha variación entre ellas. En el caso de que el CCI sea idéntico a uno, para obtener una estimación precisa, se necesitaría una muestra censal de UPM, donde el submuestreo sea de una sola unidad (puesto que todas las unidades de la UPM serán idénticas).

En resumen, la definición de las UPM es un proceso que requiere una gran disponibilidad de capacidades técnicas en los equipos de las ONE para que todas las operaciones estadísticas del período intercensal estén equilibradas desde el punto de vista del presupuesto y el esfuerzo logístico. La función objetivo de este proceso es el CCI, que, como se verá en el capítulo VIII, determina el tamaño de muestra y la precisión de la inferencia.

Con la llegada de los ciclos censales, se actualizan los marcos de muestreo y, por consiguiente, la metodología de diseño y recolección de información primaria en las encuestas de hogares. En general, se debe procurar que las UPM tengan el mismo tamaño dentro de los estratos. Por ejemplo, en las áreas rurales se pueden presentar casos en que una única UPM agrupe un conjunto de viviendas con demasiada heterogeneidad. Así, es posible encontrar UPM con pocas viviendas o UPM con demasiadas viviendas. Esto constituye una desventaja técnica a la hora de establecer metodologías apropiadas para la recolección de la información primaria y, además, para la estimación de los errores de muestreo que se derivan de las encuestas de hogares. La distribución desigual de viviendas en las UPM conlleva varias consecuencias negativas. Por ejemplo, las estimaciones de las varianzas son mucho más grandes y, por lo tanto, las cifras oficiales serán menos precisas. Por ese motivo se necesita un tamaño de muestra más amplio para satisfacer un umbral de error de muestreo.

Capítulo IV
Métodos de estratificación

Para aumentar la calidad de la inferencia en las encuestas de hogares, es importante que el marco de muestreo permita clasificar las unidades primarias de muestreo (UPM) de acuerdo con su nivel socioeconómico para poder realizar selecciones independientes en cada categoría de la clasificación. De esta forma, se garantiza la homogeneidad dentro de los grupos y se disminuye la incertidumbre de la estimación. Este proceso se conoce con el nombre de estratificación.

En la literatura especializada, se describen varios métodos que clasifican cada una de las UPM del marco en grupos denominados estratos y disminuyen la varianza de los estimadores de muestreo. Si los estratos están formados por unidades homogéneas, que, a su vez, crean categorías heterogéneas entre sí, se dice que el proceso de estratificación es eficiente y el error de muestreo se reduce significativamente.

Tras definir las UPM en el marco de muestreo, es necesario agruparlas de acuerdo con sus características sociodemográficas agregadas, a fin de obtener grupos homogéneos que permitan una mayor precisión al ejecutar las estrategias de muestreo propuestas en el marco de planificación de las encuestas de hogares que realizan los institutos nacionales de estadística (INE). Es importante evaluar distintos escenarios de estratificación para encontrar una clasificación óptima de las UPM, pues esta división se utilizará en todas las encuestas de hogares que se basen en este marco de muestreo en el período intercensal.

En síntesis, el proceso de estratificación del marco de muestreo comprende los siguientes pasos:

- Aplicación de múltiples métodos de estratificación de las UPM a partir de información agregada del censo[1].

[1] También es posible añadir información geoespacial, catastral o de cualquier otra índole si se tiene una cobertura completa a nivel de las UPM.

- Realización, para cada método señalado anteriormente, de divisiones en tres, cuatro o cinco grupos a nivel nacional y evaluación de la pertinencia de llevar a cabo la estratificación en las zonas rural y urbana de forma independiente.
- Evaluación de su efectividad a partir de las pruebas y los escenarios estudiados mediante una única medida de calidad, definida como efecto de diseño generalizado, y elección del mejor escenario en función de esta medida y de la viabilidad logística con respecto al número de divisiones.

En este capítulo se presentan los diferentes pasos del proceso de estratificación y se establece la forma de agregación de las variables a nivel de las UPM para mantener una estructura uniforme que permita aprovechar mejor la discriminación entre sus estructuras y, por consiguiente, mejorar la clasificación en estratos. Asimismo, se resumen algunos de los métodos utilizados para la estratificación de marcos de muestreo, teniendo en cuenta dos enfoques: univariados sobre medidas de resumen y multivariados sobre toda la matriz de estratificación. Por último, se presentan los criterios de evaluación de los métodos de estratificación y se ilustran los resultados finales de la estratificación de un marco de muestreo, para luego exponer las consideraciones más importantes.

A. Dimensiones estructurales del marco de muestreo

El proceso de definición de un diseño de muestreo para las encuestas nacionales que responda a las necesidades de información de un país con miras a la elaboración de políticas públicas conlleva varios procesos basados en los censos nacionales de población y en el uso de una cartografía detallada del territorio nacional.

Como se indicó en el capítulo III, un aspecto fundamental para el diseño y el desarrollo de las encuestas de hogares consiste en la definición de las UPM, que son las unidades cartográficas que dividen el territorio nacional y permiten llevar a cabo los procesos de recolección de información y trabajo de campo de la mejor manera posible. Se construyen con el fin de facilitar la obtención de estimaciones precisas y confiables de los indicadores y parámetros de interés para los responsables de la toma de decisiones y los expertos en políticas públicas.

Según la planificación de las diferentes encuestas, las UPM pueden dar lugar a unidades secundarias de muestreo o permitir la selección directa de unidades de análisis como las viviendas, los hogares o las personas. Independientemente de las unidades de muestreo y las jerarquías que se definan para implementar el diseño de muestreo para las encuestas, es fundamental llevar a cabo un proceso de estratificación de las UPM en grupos que sean lo más homogéneos posible en cuanto a sus características socioeconómicas y de bienestar y definan una división del territorio nacional (Gutiérrez, 2016).

En la literatura estadística, estos grupos se denominan "estratos" y, en conjunto, deben cubrir todo el territorio nacional. Como estos grupos determinan una división, dos estratos cualesquiera deben ser mutuamente excluyentes. Los INE utilizan las divisiones geográficas y cartográficas generadas en la realización del censo para seleccionar muestras de hogares, mediante la ejecución de diseños de muestreo probabilísticos, estratificados y en varias etapas. En particular, para aumentar la calidad de la inferencia en las encuestas de hogares, es importante que el marco de muestreo permita clasificar las UPM de acuerdo con su estructura socioeconómica, a fin de poder realizar selecciones independientes en cada categoría de la clasificación.

Al garantizar la homogeneidad dentro de los estratos, se disminuye la incertidumbre de la estimación y se minimizan los errores de muestreo que se obtienen al realizar encuestas con procedimientos de muestreo probabilístico. En los países latinoamericanos, este proceso se lleva a cabo utilizando información censal a nivel de personas, hogares y viviendas, en diferentes constructos o dimensiones relacionadas con la calidad de vida y el bienestar (demografía, características de la vivienda, propiedad de determinados bienes y acceso a servicios públicos, entre otras). Las variables definidas sobre estos constructos se agregan a partir de variables binarias que toman el valor 1 si el fenómeno en cuestión se asocia de forma positiva con mejores condiciones socioeconómicas y 0 en caso contrario. Por ejemplo:

- El acceso del hogar a una conexión de Internet puede ser una variable de interés en la estratificación, pues ayuda a determinar cuáles son los hogares con mejores condiciones de bienestar. En este caso, la variable toma el valor 1 si el hogar dispone de servicio de Internet y 0 si no dispone de dicho servicio.
- El material de los pisos, las paredes y los techos de la vivienda también puede ser una variable importante en la estratificación de las UPM. Los mejores materiales se asocian a una mayor capacidad económica y mejores condiciones habitacionales. Esta variable toma el valor 1 si la vivienda no tiene materiales precarios y 0 en caso contrario.
- El nivel de educación de los jefes del hogar y su ocupación también son variables relevantes para la clasificación de los hogares (y las correspondientes UPM) con mejores condiciones de vida.

El proceso anterior se basa en un análisis exploratorio de los datos recopilados con respecto a las diferentes variables que podrían utilizarse en el proceso de estratificación. En primer lugar, es necesario tener en cuenta que la estratificación que se pretende realizar debe llevarse al nivel de las UPM. Esto significa que, cuando las UPM estén categorizadas en un estrato, todos sus componentes también estarán clasificados en la misma categoría. Por consiguiente, las personas, los hogares y las viviendas de la UPM pertenecerán al estrato en el que dicha UPM fue clasificada.

A partir de la información del censo, se deben seleccionar y definir las variables directamente relacionadas con los fenómenos que se estudiarán en las diferentes encuestas de hogares a lo largo del período intercensal. Una vez construidas las UPM, se

calculan los agregados de las variables seleccionadas en las dimensiones examinadas mediante los censos, que por lo general son las siguientes:

- Demografía y estructura de la población: sexo, edad, parentesco, origen extranjero, pertenencia a pueblos indígenas, número de hijos y número de dependientes, entre otros.
- Educación: analfabetismo, asistencia escolar, años de estudio y nivel de escolaridad, entre otros.
- Mercado de trabajo: población en edad de trabajar, participación en la fuerza de trabajo por sexo, condición de ocupación por sexo y rama de actividad, entre otros.
- Características de la vivienda: tipo de vivienda, materiales de construcción, situación de hacinamiento y equipamiento, entre otros.
- Acceso a servicios: agua potable, alcantarillado, Internet, salud y seguridad social, entre otros.
- Propiedad de determinados bienes en el hogar: en algunos países, la propiedad de bienes como televisor, microondas, aire acondicionado, automóvil o lavaplatos automático, entre otros, puede utilizarse como criterio de clasificación de los hogares.
- Necesidades básicas insatisfechas (NBI) o pobreza multidimensional: situación de hacinamiento crítico en la vivienda, servicios inadecuados, alto nivel de dependencia económica, presencia de niños en edad escolar que no asisten a la escuela y falta de acceso o acceso limitado a servicios de salud, entre otras.

La caracterización de estas dimensiones permite clasificar las UPM en el marco de muestreo. Por ejemplo, en la dimensión demográfica, se observa que, en América Latina, las UPM con mayor número de personas que se identifican como indígenas o afrodescendientes a menudo presentan niveles de bienestar o calidad de vida más bajos con respecto a las demás. De forma análoga, en el marco de los recientes fenómenos migratorios en la región, existen datos empíricos que apuntan a que las UPM en las que se concentran los extranjeros procedentes de otros países latinoamericanos presentan niveles de bienestar inferiores con respecto a las otras. Lo mismo ocurre con las UPM con mayores porcentajes de niños en la primera infancia y hogares uniparentales con madres jefas de hogar.

Desde el punto de vista de la educación, las UPM con mayores tasas de analfabetismo (que por lo general se encuentran en las áreas rurales) y niños que no asisten a la escuela presentan niveles de bienestar inferiores con respecto a las UPM que tienen un mayor porcentaje de población con estudios de educación superior (que por lo general se encuentran en las áreas urbanas de los países).

En la dimensión ocupacional, las UPM rurales concentran una elevada proporción de población ocupada que no necesariamente goza de mejores condiciones socioeconómicas con respecto a la de las UPM urbanas. Por otra parte, las UPM que tienen una mayor proporción de población desocupada o de personas dependientes (personas de 0 a

15 años o mayores de 65 años) pueden presentar peores condiciones socioeconómicas con respecto a las UPM restantes.

Con respecto a las características de la vivienda, está bien documentado que las UPM con un alto porcentaje de viviendas cuyas paredes, techos y pisos están construidos con materiales precarios generalmente presentan bajos niveles de bienestar y están más presentes en las zonas rurales y las áreas marginales de las zonas urbanas. De la misma manera, las UPM en las que se concentra un gran número de hogares en situación de hacinamiento (por ejemplo, si el número de personas del hogar sobre el número de habitaciones es superior a tres), o con acceso inadecuado a fuentes de agua potable o servicios sanitarios y de eliminación de aguas grises deficientes, suelen presentar bajos niveles de bienestar socioeconómico.

B. Información a nivel de UPM

Cabe resaltar que, a partir de la información recolectada en el censo, también es posible clasificar a las personas o los hogares en una primera instancia, para después agregarlos hasta llegar a una clasificación única de la UPM. Sin embargo, en la práctica, este proceso puede resultar complejo y sus ventajas no son claras. En virtud de lo expuesto, este capítulo se centra en la clasificación de las UPM a partir de una matriz de información agregada a nivel de las UPM, y no de la información (a nivel de microdatos) de las personas que las habitan, ni de los hogares o viviendas que las constituyen.

Debido a que las UPM tienen tamaños diferentes, la escala y el nivel en que se miden los indicadores pueden afectar los procesos de clasificación. Si la matriz de información con la cual se realiza la estratificación se construye sobre la base del número de personas (con determinadas características) dentro de la UPM, es muy probable que, al no tener en cuenta su tamaño, los métodos de estratificación no logren agrupar las UPM de forma homogénea. Por ejemplo, supóngase que hay dos UPM con un tamaño de 100 y 300 hogares, que comprenden 200 y 400 personas en la fuerza de trabajo, respectivamente, y que una de las variables de la matriz de información se define como el número de personas ocupadas. A su vez, la primera UPM corresponde a un sector de nivel socioeconómico alto y la segunda, a un sector de nivel socioeconómico bajo. En este caso, es posible que haya 150 personas ocupadas en ambas UPM y que, por esta razón, queden erróneamente clasificadas en el mismo grupo. En consecuencia, la definición de la matriz de información en términos relativos (porcentaje de aparición de un fenómeno sobre cada categoría de las variables) constituye una alternativa mejor para que el agrupamiento esté controlado por el tamaño de la UPM y supeditado únicamente a cambios estructurales en los constructos de medición del censo.

Por último, una vez definido el conjunto de variables que se incluirán en la matriz de información, es necesario verificar que todos los indicadores de esta matriz apunten hacia el mismo horizonte del constructo censal, es decir, que todos los indicadores

estén expresados en términos de acceso al bienestar de cada uno de los constructos. Además, es necesario refinar esta matriz para eliminar las variables que puedan estar muy correlacionadas con el resto de las variables o que puedan expresarse como una combinación lineal de otras variables. De esta manera, se evitan los problemas de multicolinealidad y se asegura una estratificación parsimoniosa. Al final se debe contar con una matriz de información X compuesta por P columnas (variables de estratificación) y N_I filas (número de UPM en el marco de muestreo), en la que cada fila representará la i-ésima observación de las UPM a nivel censal para cada una de las P variables.

De acuerdo con la teoría estadística, la mejor estratificación es aquella que minimiza los errores de muestreo de los estimadores, expresados en forma de varianzas o errores estándar. Además, una de las particularidades de los procesos de estratificación es que las varianzas de estos estimadores dependen, a su vez, de la variación de los microdatos observada en el censo. Sin embargo, una estratificación óptima para un indicador puede resultar ineficiente para otros. Más aún, dado que no todas las variables de interés que se examinarán en las encuestas durante el período intercensal se han medido y observado en el censo, la estratificación que se ha de utilizar debe elegirse cuidadosamente, por medio del estudio de numerosos escenarios.

En los países de la región hay un entendimiento tácito, respaldado en mayor o menor medida por datos empíricos, de que la mayoría de los fenómenos sociales que se observan en las encuestas de hogares están supeditados a la distribución de la población en las UPM. Por ejemplo, al medir la informalidad en el mercado de trabajo, seguramente se observará que este fenómeno es mucho más frecuente en las UPM de nivel socioeconómico bajo —en las que también se registrarán otros fenómenos como menos años de educación, menores tasas de acceso a la salud, menores ingresos y gastos y mayores tasas de embarazo adolescente, entre otros— con respecto a las UPM de nivel socioeconómico alto. Por ese motivo, además de la incidencia de las variables incluidas, es necesario analizar las relaciones entre ellas. Por ejemplo, puede ser de utilidad:

- Analizar si la proporción de viviendas con techos y paredes adecuadas está muy correlacionada con la proporción de viviendas con pisos adecuados.
- Tener en cuenta si la proporción de extranjeros es muy baja y si estos solo se registran en algunas UPM específicas, en cuyo caso se recomendaría excluir esta variable, dada su poca capacidad de discriminación en el proceso de estratificación.
- Determinar si la proporción de hogares con computadora y lavadora está muy correlacionada con la disponibilidad de Internet y refrigerador, por lo que estas variables no se considerarían en la matriz de estratificación.
- Evaluar si la posesión de estufa y radio presenta indicadores muy altos en todas las UPM y no incrementa la capacidad de discriminación en el proceso de estratificación.

El análisis exhaustivo de las características poblacionales indica que existe un alto nivel de correlación entre la UPM en la que se encuentra el hogar y la incidencia de fenómenos sociales y económicos. Por lo tanto, los ejercicios de estratificación que se deben estudiar presentarán un elevado nivel de coherencia interna. De esta manera, al

escoger la mejor estratificación, se garantiza que los INE dispondrán de una clasificación óptima en el período intercensal para todas las encuestas de hogares que se realicen.

En general, hay dos grandes tipos de métodos que deben evaluarse al proponer una estratificación: univariados (sobre una medida de resumen de la matriz de información) y multivariados (sobre todas las variables de la matriz de información). Cabe subrayar que, en ambos casos, el objetivo es encontrar la mejor división, que garantice que la varianza de los estimadores de muestreo sea mínima. A continuación, se presentan algunas técnicas que pueden considerarse y que están disponibles en el *software* estadístico R mediante las librerías *stratification* (Baillargeon y Rivest, 2011) y *SamplingStrata* (Barcaroli, 2014). En ambos casos se proporcionan instrucciones para el uso de las funciones de estratificación.

C. Métodos univariados sobre medidas de resumen

Es bien sabido que la mejor estratificación para una variable de interés es la que se basa en su propia variación. Durante muchos años, se desarrollaron técnicas de estratificación sobre una sola variable de interés que dejaban de lado el carácter multiuso de las encuestas de hogares. Por esta razón, se sugiere partir de la matriz de información y resumir la variación y las correlaciones entre variables mediante una técnica multivariada de reducción de datos, como las de componentes principales, análisis factorial o modelos no lineales. Dado que la matriz de información se expresa en una escala de porcentajes, es posible que la variabilidad recogida por la medida de resumen sea alta.

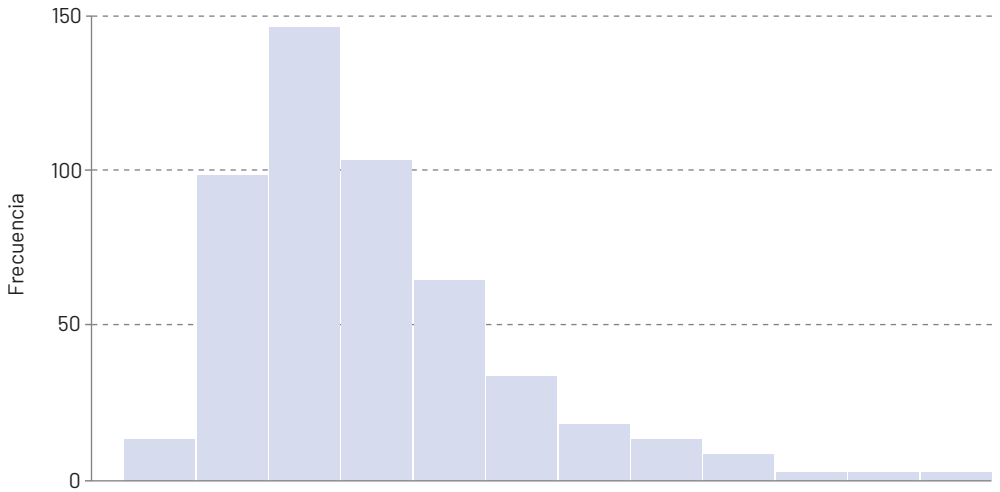
Por ejemplo, al utilizar la técnica de componentes principales, se toma como medida de resumen el primer componente, que es una función del vector propio asociado al mayor valor propio de la matriz de covarianzas ligada a la matriz de información. Por otra parte, si se utilizara un análisis factorial confirmatorio, la medida de resumen podría ser el eje principal con la carga factorial más alta. La interpretación de estas medidas de resumen constituye una parte importante de la aplicación de las técnicas de estratificación. La matriz de información se construye a partir de cinco constructos censales (demografía y estructura de la población, educación, mercado de trabajo, características de la vivienda y acceso a servicios básicos) que deberían resumirse en una medida de bienestar de la UPM, la cual, a su vez, debe tener sentido en cuanto a la relación (o contribución) de las variables al componente o factor. De ahora en adelante, se utilizará la siguiente notación para referirse a la medida de resumen como función de todas las variables incorporadas en la matriz de información:

$$y = f(x_1, \ldots, x_P)$$

Cabe señalar que es de esperar que, al estar definida como una medida de bienestar sobre las UPM, esta variable de resumen tenga un comportamiento sesgado, como se puede observar en el gráfico IV.1. Por ende, si esta característica resulta muy sesgada, puede ser

recomendable crear un estrato de inclusión forzosa con estas unidades. Esta práctica garantiza que el error de muestreo para este estrato sea nulo. A continuación se enumeran algunas técnicas de estratificación comúnmente utilizadas en la práctica estadística.

■ Gráfico IV.1
Histograma de la medida de resumen (y) sobre las unidades primarias de muestreo (UPM)



Fuente: Elaboración propia.

1. División en cuantiles

Con este método, se divide la población de las UPM en grupos creados a partir de la división de la distribución de la medida de resumen en intervalos regulares. Los cuantiles más usados son los cuartiles (división de la población en cuatro grupos), los quintiles (división en cinco grupos) y los deciles (división en diez grupos). Sin embargo, para los fines de estratificación, también es útil considerar la división en terciles (división de la población en tres grupos).

2. Método de raíz de frecuencia acumulada

Dalenius y Hodges (1959) propusieron esta técnica de estratificación basada en la raíz cuadrada de las frecuencias acumuladas de la medida de resumen sobre las UPM. Se trata de una técnica exacta, que no requiere ningún procedimiento iterativo y cuyo objetivo principal consiste en encontrar grupos que minimicen la siguiente función:

$$D = \sum_{h=1}^{H} W_h \sqrt{S_{y_h}^2}$$

Donde $W_h=N_h/N\ (h=1,\ldots,H)$ es el tamaño relativo del estrato h y $S^2_{y_h}$ es la varianza de la medida de resumen en el estrato h.

3. Estratificación óptima

Lavallée e Hidiroglou (1988) propusieron por primera vez la construcción de una estratificación óptima para poblaciones de encuestas reales basada en la minimización de la siguiente expresión, ligada a la varianza de una estrategia de muestreo estratificada:

$$\sum_{h=1}^{H-1}\left(\frac{N_h}{N}\right)^2\left(\frac{1}{(n-N_H)\,a_h}-\frac{1}{N_h}\right)S^2_{xh}$$

Donde N_h es el número de UPM en el estrato h, n es el tamaño de muestra de las UPM, N es el número de UPM en el marco de muestreo y $S^2_{x_h}$ es la varianza de la medida de resumen en el estrato h. Por último, a_h es la regla de asignación para el tamaño de muestra, dada por la siguiente relación:

$$a_h=\frac{\gamma_h}{\sum_h \gamma_h}$$

Donde, teniendo en cuenta que $\bar{X}_h$ es la media de la medida de resumen en el estrato h, según Baillargeon y Rivest (2011), γ_h es proporcional al tamaño de muestra n y está definida por:

$$\gamma_h = N_h^{2q_1}\times \bar{X}_h^{2q_2}\times S_{xh}^{2q_3}$$

Por lo tanto, dado que $n_h=n\times\gamma_h$, si se desea una estrategia de muestreo que asigne el tamaño de muestra de manera proporcional a cada uno de los estratos, la regla de asignación debería estar determinada por:

$$q=(q_1,q_2,q_3)' = (0.5,0,0)'$$

La asignación de Neyman corresponderá a $q=(0.5,0,0.5)'$; mientras que la asignación de potencia con exponente 0.7 estará dada por $q=(0.35,0.35,0)'$. Los detalles técnicos de estos tipos de asignación se encuentran en Gutiérrez (2016).

La optimización de la función objetivo puede llevarse a cabo de diferentes formas. En efecto, Lavallée e Hidiroglou (1988) utilizaron el algoritmo de optimización de Sethi (1963) para encontrar los valores óptimos. Baillargeon, Rivest y Ferland (2007) definen los pasos necesarios para implementar el procedimiento basado en el algoritmo de Sethi. Asimismo, Kozak (2004) definió un algoritmo iterativo mediante arranques aleatorios para optimizar el proceso de minimización de esta técnica de estratificación.

4. Estratificación geométrica

Utilizando las técnicas de estratificación mencionadas anteriormente, algunos autores se percataron de que, en el caso de poblaciones de UPM con medidas de resumen sesgadas, las varianzas relativas (coeficientes de variación) de la medida de resumen en cada estrato eran similares, es decir:

$$\frac{S_{x_1}}{\bar{X}_1} \cong \frac{S_{x_2}}{\bar{X}_2} \cong \ldots \cong \frac{S_{x_H}}{\bar{X}_H}$$

Teniendo en cuenta este hecho, Gunning y Horgan (2004) desarrollaron un método con el objetivo de que los coeficientes de variación de la medida de resumen tendieran a ser iguales dentro de los estratos y, de esta forma, encontraron que los límites que definían esos grupos estaban conformados en progresión geométrica. Siendo X la variable que contiene la información de la medida de resumen para todas la UPM del marco de muestreo, los límites de los estratos estarán dados por la siguiente expresión:

$$b_h = mín(X)\left(\frac{máxX}{mínX}\right)^{h/L} \quad ; \quad h = 1,2,\ldots,H\text{-}1$$

Se puede observar que los coeficientes de variación de los estratos conformados por estos límites son equivalentes y, por consiguiente, este método resulta óptimo para encontrar mejores formas de estratificar teniendo en cuenta la variación relativa dentro los estratos como función objetivo.

D. Métodos multivariados sobre la matriz de información

A partir de la matriz de información X a nivel de las UPM, que contiene N_1 filas y P columnas, es posible considerar algunos procedimientos que no requieren la reducción a una sola dimensión, sino que admiten tantas dimensiones como indicadores definidos en las columnas de X. Teniendo en cuenta que en el período intercensal se realizarán encuestas que miden variables fuertemente ligadas a las observadas en el censo, la definición de una estratificación óptima para todo el conjunto de variables de la matriz de información asegurará una división óptima para las encuestas realizadas en el período intercensal. Los siguientes métodos permiten optimizar conjuntamente la eficiencia de la estratificación.

1. *K*-medias de Jarque

Jarque (1981) propuso utilizar una versión modificada del algoritmo de *k*-medias (Macqueen, 1967), cuyo objetivo es la minimización de la siguiente función de distancia:

$$\sum_{h=1}^{H} \sum_{k \in U_h} (X_k - \bar{X}_h)' \Lambda^{-1} (X_k - \bar{X}_h)$$

Donde x_k corresponde a la medición de las P variables de la matriz de información en la k-ésima UPM, $\bar{x}_h$ es el vector de medias de la matriz de información en el estrato h y Λ es una matriz diagonal de tamaño $P \times P$ cuyas entradas se definen como la varianza de las P variables de la matriz X, es decir $\Lambda[p,p]=S^2_{x_p}$, con $p=1,2,\ldots,P$. El objetivo de esta modificación es minimizar la relación entre la varianza de un estimador de muestreo estratificado con asignación proporcional y la de un muestreo aleatorio simple. Cuando $\Lambda=I$, el algoritmo resultante es idéntico al algoritmo clásico de k-medias propuesto por Macqueen (1967).

2. Algoritmos genéticos

Ballin y Barcaroli (2013) argumentan que la mejor estratificación es la división del marco de muestreo que asegura el mínimo costo muestral que satisfaga algunas restricciones de precisión o maximice la precisión de los indicadores de interés bajo las restricciones. De esta forma, con el algoritmo se busca minimizar la siguiente función de costos:

$$c_0 + \sum_{h=1}^{H} c_h n_h$$

Donde c_0 define un costo fijo y c_h es el costo promedio de observar un hogar en el estrato h. En principio, es posible definir $c_0=0$ y $c_1=c_2=\cdots=c_H=1$, que da como resultado que el costo es el número de encuestas que deben realizarse en cada estrato. Este problema de optimización se complementa manteniendo las siguientes restricciones:

$$\sum_{h=1}^{H} \left(\frac{N_h^2}{n_h}\right) \left(1 - \frac{n_h}{N_h}\right) S^2_{x_{h,p}} \leq V_{0p} \quad p=1,2,\ldots,P$$

Donde V_{0p} es un umbral predefinido por el usuario, que indica que la varianza de la estrategia estratificada está acotada, y $S^2_{x_{h,p}}$ es la varianza poblacional de la p-ésima variable de la matriz de información en el estrato h. Utilizando algoritmos genéticos evolutivos, esta estratificación multivariada del marco de muestreo parte de la consideración de estratificaciones univariadas independientes (una para cada variable de la matriz de información) y de la definición del producto cartesiano resultante de todas estas divisiones (estratos atómicos). Este universo de posibles estratificaciones evoluciona, uniendo grupos de forma jerárquica, sujeto a las restricciones de precisión sobre cada variable de la matriz de información, hasta converger en el número de estratos definidos de antemano H.

E. Evaluación y elección de la mejor estratificación

En la evaluación de los escenarios de estratificación se consideran las técnicas univariadas y multivariadas. Al final, el resultado de la aplicación de una u otra técnica es simplemente una clasificación de las UPM. Por lo tanto, cada una de las posibles estratificaciones debe evaluarse sobre la base de la reducción de la varianza para todos los indicadores considerados en la matriz de clasificación. La medida clásica con la que se juzgan las bondades de una estrategia de muestreo es el efecto de diseño (DEFF). Por consiguiente, la evaluación de la estratificación también debe supeditarse a esta medida, que para la variable $p=1,\ldots,P$, está dada por:

$$DEFF_p = \frac{Var_{ST}(\bar{x}_p)}{Var_{SI}(\bar{x}_p)} \qquad p=1,\ldots,P$$

Donde $Var_{ST}(\bar{x}_p)$ y $Var_{SI}(\bar{x}_p)$ denotan la varianza del diseño estratificado y la varianza de un muestreo aleatorio simple para la media poblacional (porcentaje) de la p-ésima variable de la matriz de información. Por otra parte, Gutiérrez (2016, pág. 184) demuestra que, cuando la asignación es proporcional, esta relación se puede escribir de la siguiente manera:

$$DEFF_p = \frac{\sum_{h=1}^{H} W_h \, S^2_{x_{hp}}}{S^2_{x_p}} \cong 1\text{-}R^2_p \qquad p=1,\ldots,P$$

Donde, para cada estrato $h=1,\ldots,H$, se tiene que $S^2_{x_p}$ es la varianza de la variable x_p en la población y $S^2_{x_{hp}}$ es la varianza de la variable x_p supeditada al estrato h. Cabe notar que este efecto de diseño es una función del coeficiente de determinación R^2_p en un modelo lineal con intercepto que relaciona la p-ésima variable de evaluación (respuesta) con los estratos (factores). Una ventaja de expresar el efecto de diseño como en la ecuación anterior es que este no dependerá del tamaño de la muestra. Una vez definido el criterio de evaluación de la estratificación sobre una variable x_p, es necesario definir un criterio de estratificación multivariante que contemple cada una de las P variables. De acuerdo con Jarque (1981), se propone la siguiente medida de calidad, definida como el efecto de diseño generalizado ($G(S)$) sobre todas las variables de la matriz de información:

$$G(S) = \sum_{p=1}^{P} D\,EFF_p = \sum_{p=1}^{P} \frac{1}{S^2_{x_p}} \sum_{h=1}^{H} W_h S^2_{x_{hp}}$$

Ante una estratificación pertinente, se esperaría que $Var_{ST}(\bar{x}_p) < Var_{SI}(\bar{x}_p)$, por lo tanto $0<DEFF_p<1$, que determina que $0<G(S)<P$. Luego, se debería escoger el escenario para el cual $G(S)$ fuera mínimo. Cabe subrayar que, para cada uno de los escenarios estudiados, es necesario fijar el número de estratos, que en general se establece entre tres y cinco. Esta definición del número de grupos debe examinarse en el INE con los equipos que determinan la rotación de las UPM en cada período de realización de las encuestas de hogares. Si bien la definición de un número elevado de estratos reducirá la

varianza, ello puede tener repercusiones negativas en la logística de rotación del diseño de muestreo de las encuestas, haciendo que se agoten rápidamente las UPM dentro de los estratos geográficos y socioeconómicos. Por esta razón, se recomienda restringir los escenarios de evaluación a la consideración de $H=3$ y $H=4$ estratos.

En el cuadro IV.1 se ejemplifica la evaluación de estas técnicas para dos escenarios de estratificación (tres y cuatro estratos) en una matriz de información que contiene ocho variables. A partir del cuadro se puede llegar a varias conclusiones interesantes: i) en el caso del primer indicador, la mejor estratificación corresponde al método de raíz de frecuencia acumulada con cuatro estratos; ii) para el segundo indicador, la mejor estratificación es el algoritmo genético con cuatro estratos; iii) para el último indicador, la mejor estratificación es la estratificación óptima con el algoritmo de Sethi con cuatro estratos. Cabe mencionar que para cada indicador existe un método que permite una mayor eficiencia que para otros indicadores. Esto muestra claramente que la estratificación con respecto a un solo indicador puede ser un procedimiento inadecuado. Por lo tanto, sobre la base de este ejemplo, el mejor método sería el de la raíz de la frecuencia acumulada con cuatro estratos, puesto que permite una mayor eficiencia conjunta al reducir el efecto de diseño generalizado.

■ Cuadro IV.1

Efecto de diseño ($DEFF_p$) y efecto de diseño generalizado ($G(S)$) considerando tres ($H=3$) y cuatro ($H=4$) estratos para ocho variables

DEFF	División en cuantiles (H=3)	Método de raíz de frecuencia acumulada (H=3)	Estratificación óptima (H=3)	Estratificación geométrica (H=3)	*K*-medias de Jarque (H=3)	Algoritmo genético (H=3)	División en cuantiles (H=4)	Método de raíz de frecuencia acumulada (H=4)	Estratificación óptima (H=4)	Estratificación geométrica (H=4)	*K*-medias de Jarque (H=4)	Algoritmo genético (H=4)
$\bar{x}_1$	0,87	0,85	0,81	0,82	1,00	0,88	0.8	0,70	0,76	0,72	0,71	0,77
$\bar{x}_2$	0,89	0,82	0,95	0,97	0,94	0,88	0.79	0,74	0,75	0,77	0,75	0,71
$\bar{x}_3$	0,87	0,97	0,83	0,96	0,89	0,95	0.74	0,75	0,79	0,7	0,79	0,71
$\bar{x}_4$	0,92	0,89	0,81	0,94	0,96	1,00	0.77	0,73	0,73	0,7	0,71	0,74
$\bar{x}_5$	0,85	0,83	0,96	0,96	0,83	0,81	0.8	0,73	0,8	0,78	0,8	0,79
$\bar{x}_6$	0,87	0,88	0,9	0,88	0,86	0,81	0.8	0,72	0,76	0,7	0,74	0,73
$\bar{x}_7$	0,87	0,95	0,99	0,83	0,86	0,84	0.75	0,7	0,77	0,72	0,77	0,77
$\bar{x}_8$	0,93	0,82	0,91	0,99	0,93	0,88	0.77	0,74	0,72	0,78	0,76	0,75
$G(S)$	7,07	7,01	7,16	7,35	7,27	7,05	6.22	5,81	6,08	5,87	6,03	5,97

Fuente: Elaboración propia.

Para estudiar la comparabilidad del proceso de estratificación, los algoritmos de evaluación se deberían aplicar sobre las zonas urbanas y rurales de forma independiente. De la misma forma, es posible ejecutar los algoritmos de forma independiente en cada una de las divisiones administrativas mayores (o regiones) del país. Si la ganancia de eficiencia es mayor en estos escenarios, se pueden definir los estratos de forma independiente en cada zona o región. Es posible razonar que, para mantener la comparabilidad en el proceso de estratificación, solo ha de ejecutarse una estratificación nacional con todas las UPM que componen el marco de muestreo. Sin embargo, cabe recordar que la estratificación es un proceso diseñado para aumentar la precisión (disminuir la varianza) de los estimadores utilizados en las encuestas de hogares. Por consiguiente, si la estratificación regional o por zona da mejores resultados (menores efectos de diseño), se recomienda escogerla en lugar de la estratificación nacional.

Al margen de la técnica utilizada para encontrar la mejor clasificación de las UPM, se subraya la necesidad de que la viabilidad sobre el número de estratos se examine de forma exhaustiva en todas las áreas implicadas de los INE. En general, se recomienda limitar los escenarios de evaluación a $H=3$ o $H=4$ estratos. Este último componente es importante, puesto que los diseños de muestreo deberían considerar un tamaño de muestra mínimo de dos UPM por estrato para poder estimar la varianza del estimador (Gutiérrez, 2016).

El efecto de diseño no es el único aspecto que se ha de evaluar para la elección del procedimiento de estratificación. Es necesario verificar la estabilidad del método con respecto a los otros procedimientos de estratificación. Por ejemplo, en el cuadro IV.2 se muestra la matriz de coincidencias entre las diferentes clasificaciones de los estratos.

■ Cuadro IV.2
Matriz de coincidencias
(En porcentajes)

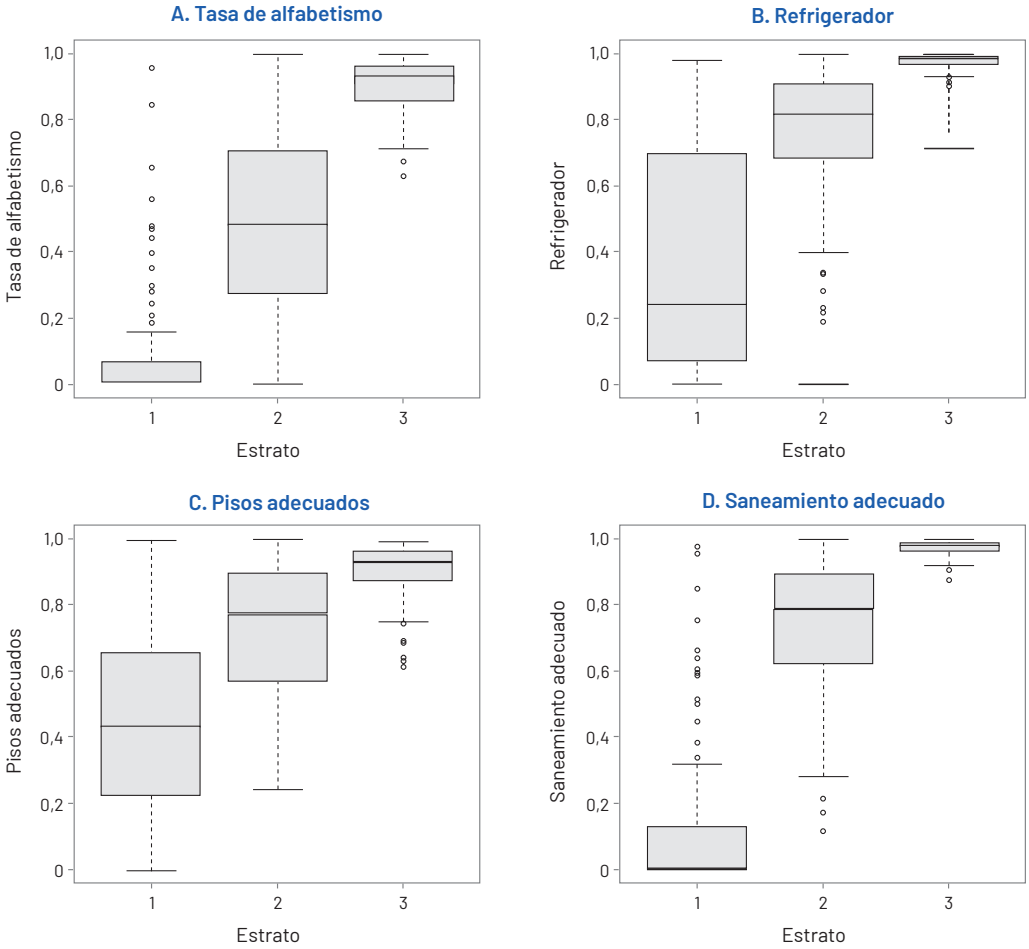
Técnica	División en cuantiles	Método de raíz de frecuencia acumulada	Estratificación óptima	Estratificación geométrica	*K*-medias de Jarque	Algoritmo genético
División en cuantiles	100	64	92	84	89	89
Método de raíz de frecuencia acumulada	64	100	68	62	71	71
Estratificación óptima	92	68	100	82	96	96
Estratificación geométrica	84	62	82	100	78	78
K-medias de Jarque	89	71	96	78	100	97
Algoritmo genético	89	71	96	78	97	100

Fuente: Elaboración propia.
Nota: Las cifras corresponden al porcentaje de unidades primarias de muestreo coincidentes en cada uno de los estratos creados mediante los métodos analizados.

Por último, también se debe evaluar la coherencia de la distribución de las diferentes variables agregadas a nivel de UPM en los estratos. Por ejemplo, la proporción de personas mayores de 15 años alfabetizadas debería ser mayor en los estratos más altos. Este patrón también se debería observar en diferentes indicadores, como la proporción de hogares con Internet, refrigerador, televisión por cable, automóvil, saneamiento adecuado o pisos adecuados, o la proporción de personas con educación superior, entre otros. En el gráfico IV.2 se muestra el comportamiento esperado de algunas variables de interés en los estratos de muestreo. Así, el estrato 1 debería presentar las peores condiciones socioeconómicas entre los tres grupos, seguido por el estrato 2, que debería tener mejores condiciones, y el estrato 3, que agruparía a las UPM con menores dificultades. Dadas sus condiciones menos favorables, en las zonas rurales debería registrarse una menor proporción de UPM en el estrato 3.

■ Gráfico IV.2

Diagrama de cajas del comportamiento esperado de algunas variables de interés en los estratos de muestreo



Fuente: Elaboración propia.

Cuando la contribución de algunas unidades al total poblacional no es significativa y dichas unidades son de difícil acceso, es común que en algunos países de la región se opte por redefinir el universo y crear un estrato de exclusión forzosa. En este estrato no se realiza ninguna encuesta y la población excluida no se tendrá en cuenta en las respectivas estimaciones. Por último, como algunos procedimientos de clasificación se basan en la generación de números aleatorios, se recomienda documentar los códigos computacionales que se utilizaron para que los resultados puedan reproducirse exactamente, por lo que se debe fijar una semilla aleatoria al comienzo del código computacional.

F. Estratificación implícita

Los estratos explícitos definidos en la sección anterior son útiles para reducir la varianza de muestreo y garantizar la representatividad de la muestra en cada uno de los subgrupos que comparten las mismas características socioeconómicas dentro de los mismos municipios. Además de los estratos socioeconómicos, algunas variables que se consideran en el proceso de estratificación explícita son:

- Los estados o las regiones de un país.
- La zona en la que se encuentra el hogar: urbana o rural. Cada país establece su definición de rural de acuerdo con sus propias definiciones nacionales.

También es posible realizar una selección ordenada que genera una estratificación implícita, sin que necesariamente se tenga control sobre el tamaño de muestra final y sin asumir independencia en la selección. Este tipo de estratificación es una forma de garantizar una asignación estrictamente proporcional de los hogares en todos los estratos implícitos. También puede incrementar la confiabilidad de las estimaciones de la encuesta, siempre que las variables de estratificación implícita que se consideren estén correlacionadas con los indicadores de interés (por ejemplo, las tasas de desocupación, subocupación o informalidad).

La estratificación implícita se recomienda ampliamente cuando la encuesta se concentra en un tema particular (como el mercado de trabajo) y requiere el uso de muestreo sistemático (con probabilidades simples o desiguales) en la selección de las UPM. Según Naciones Unidas (2008, pág. 46), en la mayoría de los países, la secuencia podría empezar con la zona urbana, desagregada por departamento —a su vez, desagregado por municipio—, seguida de la zona rural, desagregada por departamento —a su vez, desagregado por municipio o sección administrativa—. La selección sistemática de UPM deberá estar supeditada al ordenamiento de las UPM por el número de viviendas.

La estratificación implícita puede constituir un método objetivo de selección de reemplazos de las UPM a las que no se puede acceder en la operación de campo. De esta forma, si una UPM seleccionada originalmente no puede empadronarse por una u otra razón operativa, se reemplazará por aquella inmediatamente anterior (o posterior) en la

lista estratificada implícitamente. Este procedimiento permitirá el reemplazo con una UPM situada en el mismo municipio, dentro del mismo departamento, en la misma zona y con un número similar de viviendas.

Aunque la estratificación implícita permite limitar el sesgo generado por la falta de respuesta de las UPM, Vehovar (1999, págs. 348 y 349) advierte que se debe tener precaución en el recurso a esta práctica, porque puede causar sesgos importantes en las estimaciones de interés. Esto se debe a que los individuos que se encuentran en zonas a las que es posible acceder pueden diferir significativamente de aquellos que se encuentran en las zonas de difícil acceso, las cuales difícilmente serán seleccionadas por los algoritmos de muestreo que utilizan la estratificación implícita.

Por esta razón, si después de valorar los posibles sesgos, se decide realizar las sustituciones sobre las UPM de difícil acceso, se recomienda realizar un seguimiento exhaustivo en cada recolección de datos que permita clasificar el sistema de recopilación de información primaria y evaluar su impacto en la precisión de los estimadores resultantes.

Capítulo V
Diseño y mecanismo de selección de la muestra

Todas las encuestas de hogares de la región comparten el mismo principio inferencial: la selección de una muestra puede representar a la población de todo un país. Por supuesto, ante este objetivo tan ambicioso, es necesario contar con procedimientos robustos, probados y capaces de pasar los filtros más críticos y agudos. En este momento de la historia, la práctica de estos procedimientos tal vez ya no produzca ningún tipo de asombro, pero se invita al lector a contemplar todos los posibles escenarios a los que se vería enfrentada la sociedad ante la ausencia de las encuestas de hogares y sus repercusiones en materia de seguimiento del desarrollo social y económico.

Son innegables las múltiples ventajas operativas, logísticas y presupuestarias que suponen estas operaciones estadísticas sustentadas en el muestreo probabilístico. Al recolectar la información de interés sobre una fracción pequeña de la población objetivo, se logran obtener estimaciones insesgadas con niveles de precisión muy elevados, lo que garantiza la exactitud de la inferencia. Mediante el muestreo probabilístico, se realizan inferencias que proceden de lo particular a lo general, puesto que, al seleccionar una muestra, esta sirve como base para obtener conclusiones acerca de la población. Al final, la muestra será un vehículo adecuado para representar las características más importantes de la población objeto de estudio sobre la base de las variables que se incorporaron en el formulario de la encuesta. Gutiérrez (2016) afirma que el muestreo es un procedimiento que responde a la necesidad de obtener información estadística precisa sobre la población y los conjuntos de elementos que la conforman. También sostiene que el muestreo probabilístico conlleva investigaciones parciales que apuntan a realizar inferencias sobre la población completa y, en general, se basa en los siguientes principios:

- Aleatorización: las unidades incluidas en la muestra se seleccionan mediante un proceso probabilístico. De esta forma, además de eliminar los posibles sesgos de selección, la muestra resultante será válida para cualquier proceso de inferencia, puesto que se basa en el conjunto de todas las muestras que se pueden obtener con el diseño de muestreo definido.
- Inclusión: todas las unidades de la población tienen una probabilidad no nula de ser incluidas en la muestra. Ello quiere decir que el procedimiento de selección otorga a todas las unidades que componen la población la misma oportunidad de ser seleccionadas. De esta manera, la muestra final puede estar compuesta por cualquier combinación plausible de hogares o individuos.

Para que los principios anteriores se cumplan a cabalidad, es necesario contar con un instrumento que permita seleccionar los hogares del país de forma exhaustiva y completa. Es decir, el instrumento debería contener todos y cada uno de los hogares de la población. Dado que no existe una lista que permita conocer y ubicar a cada uno de los hogares de la población, se deben considerar otras posibilidades que permitan alcanzar el objetivo. Debido al principio natural de la aglomeración de las poblaciones humanas, es posible lograr este cometido de manera indirecta mediante la definición de los marcos de muestreo de áreas.

Las encuestas han tenido una gran trascendencia en la evolución de las mediciones de los indicadores sociales, que, a su vez, han conducido a que los gobiernos realicen el seguimiento y monitoreo de las cifras más importantes para la sociedad. De esta forma, es posible investigar la efectividad de las políticas públicas, a fin de concretar metas de mejora de las condiciones sociales y económicas de la ciudadanía. Dado que el objetivo del muestreo probabilístico es realizar una inferencia exacta y precisa, una muestra bien seleccionada de unos cuantos miles de individuos puede representar con gran precisión a una población de millones de personas.

En general, un preconcepto extendido sobre el significado de una muestra representativa es aquel que la define como un modelo reducido de la población. De esta interpretación se desprende un argumento de validez sobre la muestra: "una buena muestra es aquella que se parece a la población, de tal forma que las categorías aparecen con las mismas proporciones que en la población" (Gutiérrez, 2016). Sin embargo, en algunos casos es fundamental "sobrerrepresentar" algunas categorías o incluso seleccionar unidades con probabilidades desiguales (Tillé, 2006). La muestra *per se*, por consiguiente, no debe ser un modelo reducido de la población, sino una herramienta que se pueda utilizar para obtener estimaciones válidas; es decir, exactas, confiables, precisas y coherentes.

El concepto de muestra representativa no se debe utilizar para referirse a que la muestra deba parecerse a la población. La teoría de muestreo se ha ocupado de estudiar estrategias óptimas para garantizar la calidad de las estimaciones. En general, el concepto de representatividad debe estar asociado a la estrategia de muestreo y no solo a la muestra seleccionada. Consecuentemente, es la muestra expandida la que debe recibir el calificativo de representativa, pues su objetivo sí es reflejar a la población y permitir

que, mediante la correcta caracterización de una estrategia de muestreo, se puedan reproducir sus estructuras a partir de un proceso de inferencia. En resumen, la muestra no necesariamente debe ser similar a la población, pero la muestra expandida sí debe serlo.

El objetivo del equipo técnico experto en la selección de muestras debe ser lograr que la representatividad se pueda aplicar efectivamente a todo el componente de diseño y estimación. Es decir, el calificativo de representatividad es objeto de un proceso conjunto de diseño de muestreo, estimación de parámetros y acercamiento a modelos estadísticos para hacer frente a la ausencia de respuesta, entre otros factores. Uno de los objetivos de este capítulo será buscar precisión sobre las estructuras de selección de las muestras en las encuestas por muestreo. Al escoger un mecanismo apropiado para la selección de la muestra, será posible afirmar que la estrategia de muestreo es efectivamente representativa de la población de interés, puesto que cumple con altos estándares de rigurosidad y calidad en cada uno de los componentes del proceso.

A. Diseños de muestreo

Una vez que los marcos de muestreo se han refinado y se ha definido una estratificación apropiada para las unidades primarias de muestreo (UPM) que los componen, es necesario realizar el proceso de muestreo para la selección final de los hogares. Este proceso de selección debe garantizar el insesgamiento, además de ser eficiente. Ello significa que la inclusión de las unidades en la muestra estará supeditada a un modelo probabilístico libre de cualquier sesgo. Además, es necesario que este mecanismo genere la menor dispersión posible en el proceso inferencial posterior.

El procedimiento de muestreo asigna una probabilidad de selección conocida a cada posible muestra. Al diseñar un muestreo probabilístico, el investigador es el encargado de asignar estas probabilidades, mediante la definición del diseño de muestreo (Särndal, Swensson y Wretman, 2003). Aunque esta asignación de probabilidades se realiza de manera teórica, la pericia del equipo técnico determinará cuál es la mejor forma de selección y, a partir de esta, se escogerá el mejor algoritmo de muestreo. Después de establecer este conjunto de probabilidades, se seleccionará una única muestra mediante un mecanismo aleatorio que siga a cabalidad esta configuración estocástica generada por el diseño de muestreo. Las probabilidades deben ser distintas de cero. De lo contrario, no se podría garantizar una inferencia insesgada, puesto que se excluirían algunos sectores cartográficos del país. Además, estas mismas probabilidades se utilizan para crear los factores de expansión que definen todo el proceso de estimación, junto con el cálculo de los errores de muestreo, como se verá en capítulos posteriores.

Existe una clara diferencia entre un diseño de muestreo y un algoritmo de muestreo. El primero indica qué probabilidad de selección tendrán las posibles muestras en el soporte de muestreo, definido como el conjunto de todas las posibles muestras. El

segundo se define como el proceso de selección de una única muestra que respeta las probabilidades del diseño de muestreo. En la definición de una encuesta de hogares, es indispensable que se establezcan de antemano estos dos componentes. Es decir, si se ha decidido que el diseño de muestreo será en etapas, el equipo técnico deberá documentar exhaustivamente cada etapa, definiendo sus correspondientes unidades de muestreo y, por consiguiente, los diseños de muestreo en cada etapa. Asimismo, es igual de importante explicar qué algoritmos de selección se utilizarán en cada etapa. De esta forma, habrá total transparencia en la selección de las unidades, lo que redundará en la obtención de cifras oficiales confiables y precisas.

Existen muchas formas de seleccionar una muestra de hogares y cada una origina una medida de probabilidad sobre los elementos que conforman la población de interés. En general, en cada diseño particular de muestreo se establece una única función que asocia a cada hogar k con una probabilidad de inclusión en la muestra s, definida de la siguiente manera:

$$\pi_k = Pr(k \in s)$$

Si el diseño de muestreo es de tamaño fijo, estas probabilidades de inclusión de los hogares cumplirán con las siguientes propiedades:

i) $\pi_k > 0$

ii) $\sum_U \pi_k = n$

Obsérvese que la primera propiedad garantiza que ningún hogar quede excluido de la selección inicial. Si bien no todos los hogares serán seleccionados para pertenecer a la muestra s, todos tendrán una oportunidad de ser escogidos por el mecanismo de selección aleatorio. En segundo lugar, el tamaño de la muestra de hogares estará dado por la magnitud de las probabilidades de inclusión. Por esta razón, en una encuesta con un tamaño de muestra grande, se asignará una mayor probabilidad de inclusión a todos los hogares que en una encuesta de tamaño de muestra más modesto. A continuación se presenta una lista no exhaustiva de diseños de muestreo utilizados en encuestas de hogares para la publicación de estadísticas oficiales, junto con la forma particular que adoptan las probabilidades de inclusión en cada diseño.

1. Muestreo aleatorio simple

Este diseño de muestreo supone que es posible realizar una enumeración de todas las posibles muestras de tamaño fijo y escoger una de ellas mediante una selección aleatoria que asigne la misma probabilidad a cada una. Para ejecutar este diseño, es necesario contar con información suficiente y exhaustiva sobre la ubicación de todas las unidades de interés. Su uso es común en las etapas finales de selección de las encuestas, en las que los hogares o personas son seleccionadas con la misma probabilidad. Por ejemplo, una vez que se haya escogido una UPM, una parte del operativo de campo deberá estar dedicada al enlistamiento de todas las viviendas y estructuras pertenecientes a la UPM. Cuando se haya realizado este

empadronamiento, será posible asignar la misma probabilidad de inclusión a cada vivienda en la UPM. Por consiguiente, las probabilidades de inclusión en el muestreo aleatorio simple sin reemplazo son todas iguales y están dadas por la siguiente expresión:

$$\pi_k = Pr(k \in s) = \frac{n}{N}$$

Como se verá en los siguientes capítulos, cuando en este diseño de muestreo se usa el estimador de Horvitz-Thompson para estimar un total poblacional, y suponiendo que $S^2_{y_U}$ denota la varianza de la característica de interés en la población finita, las expresiones del estimador puntual y su varianza, respectivamente, toman la siguiente forma:

$$\hat{t}_{y,\pi} = \sum_s \frac{y_k}{\pi_k}$$

$$Var(\hat{t}_{y,\pi}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{y_U}$$

Una variante de este tipo de modelos de selección de muestras de hogares dentro de la UPM es el muestreo sistemático, donde se ordena el marco con algún patrón predefinido y posteriormente se selecciona un primer hogar (como arranque aleatorio). A partir de ese primer hogar seleccionado, se incluyen los restantes hogares en la muestra mediante saltos sistemáticos equiespaciados por el factor $a = \lfloor N/n \rfloor$, conocido como intervalo de salto. Por ejemplo, una muestra sistemática podría ser:

$$s = \{2, 12, 22, 32, 42\}.$$

Donde el primer hogar elegido en la UPM fue el segundo y, con saltos sistemáticos de diez hogares, se va encuestando a los restantes hogares de la lista. En este diseño, la probabilidad de inclusión también es uniforme para cada hogar en la UPM y está dada por la siguiente expresión:

$$\pi_k = Pr(k \in s) = \frac{1}{a} \approx \frac{n}{N}$$

2. Muestreo proporcional al tamaño

En este tipo de muestreo se utiliza como insumo una característica de información auxiliar cuantitativa, también conocida como medida de tamaño *(measure of size)*. En la ejecución de este diseño, necesariamente el marco de muestreo deberá contener el valor correspondiente a la medida de tamaño para cada una de sus unidades. Este muestreo se utiliza con frecuencia en las etapas iniciales de selección de las muestras, sobre todo en la selección de las UPM que formarán parte de la muestra. De esta manera, los conglomerados o UPM con más hogares o personas (medida de tamaño) tendrán una mayor probabilidad de ser seleccionados en la muestra. Por consiguiente, las probabilidades de inclusión de las UPM en la muestra serán desiguales y proporcionales a la medida de tamaño. Obsérvese que la

cantidad de individuos en las UPM es una cifra conocida, puesto que es resultado directo de los censos de población y vivienda.

Una de las ventajas de este tipo de muestreo es que hace más eficiente la estimación de los indicadores de interés. Para que esto ocurra, la medida de tamaño debe estar relacionada linealmente con la característica de interés. Esto a menudo sucede en las cuestiones sociales sobre las que se indaga en las encuestas de hogares. A mayor número de hogares, se observa una mayor incidencia de estos fenómenos. Por ejemplo, en un estrato determinado, es evidente que en las UPM con más hogares se observará un mayor número de personas pobres, o de hogares con ingresos bajos, o de personas desocupadas, entre otras cosas.

Por último, la medida de tamaño no necesariamente tiene que estar definida como el conteo simple de hogares o personas dentro de las UPM. También puede definirse como una función de estos conteos (por ejemplo, la raíz cuadrada), o incluso como una función compuesta de conteos de subpoblaciones. En el caso más simple, si N_i es la medida de tamaño de la i-ésima UPM U_i, es decir, el número de hogares que componen esa UPM, n_I es el número de UPM que serán seleccionadas en cada estrato y N la sumatoria (o total) del número de hogares en todas las UPM del estrato (es decir, el número de hogares en el estrato). En ese caso, las probabilidades de inclusión en la muestra s_I están dadas por la siguiente expresión:

$$\pi_i = Pr(U_i \in s_I) = n_I * \frac{N_i}{N}$$

3. Muestreo estratificado

Esta familia de diseños de muestreo permite realizar inferencias precisas en subgrupos poblacionales de interés, que suelen definirse como agregaciones geográficas grandes. Por ejemplo, si se quiere realizar estimaciones de la incidencia de la pobreza en las regiones geográficas de un país específico, es pertinente que esta división geográfica sea considerada para la definición de los estratos. Como se mencionó al inicio de este capítulo, estas divisiones territoriales se forman de manera natural, puesto que los estratos ya están definidos como regiones de interés en el seguimiento de los indicadores sociales. Por supuesto, es posible que la estrategia de muestreo cambie en función de los estratos. Por ejemplo, en la planificación de las encuestas de uso de tiempo, una de las características de interés sobre las que se quiere indagar es la cantidad de horas que hombres y mujeres dedican a actividades de trabajo no remuneradas. Esta realidad cambia radicalmente entre zonas rurales y urbanas. Para este tipo de encuestas de hogares, la flexibilidad de los diseños estratificados es un elemento valioso que permite definir estrategias de muestreo más precisas.

Una consecuencia directa de la estratificación es que cada subgrupo tendrá un marco de muestreo de UPM independiente y mutuamente excluyente. Esta última característica es una de las mayores ventajas del muestreo estratificado, puesto que existe independencia entre los estratos. Esto significa que, en cada estrato, se pueden ejecutar distintas

estrategias de muestreo de forma independiente. En los países de América Latina, es común que los estratos estén conformados por el cruce de las áreas geográficas grandes con la división socioeconómica (como se ilustró en capítulos anteriores). Asimismo, una desagregación común en la investigación social es la división territorial del país en zonas urbanas y rurales. Evidentemente, la realidad social del entorno urbano difiere tanto del entorno rural que vale la pena considerar esta distinción en el diseño de muestreo de las encuestas de hogares.

Las probabilidades de inclusión definidas por este diseño de muestreo variarán en función de cada estrato h $(h=1,\ldots,H)$. Por ejemplo, si se hubiese planeado un diseño aleatorio simple en cada estrato, las probabilidades de inclusión estarían dadas por la siguiente expresión:

$$\pi_k = Pr(k \in s_h) = \frac{n_h}{N_h}$$

Donde s_h define la muestra seleccionada en el estrato h, N_h sería el número de hogares en ese estrato y n_h, el tamaño de la muestra de hogares asociado a ese estrato.

En algunas ocasiones se ha sugerido que el muestreo estratificado es el mejor diseño para una encuesta de hogares, lo cual es cierto en parte. Aunque en muchas ocasiones la opción de estratificar es adecuada e incluso conveniente, en sentido estricto el muestreo estratificado no es el mejor diseño. De hecho, la varianza inducida por el diseño aleatorio estratificado puede llegar a ser más grande cuando no hay una clara homogeneidad en el comportamiento de la característica de interés dentro de los estratos.

4. Muestreo de conglomerados

Este diseño de muestreo surge como respuesta a la imposibilidad de generar una muestra de hogares directamente a partir de un marco de muestreo que enliste todos y cada uno de los hogares en un país. De hecho, de forma hipotética, si fuese posible, los costos de una muestra aleatoria simple serían tan altos que la harían inviable desde el punto de vista presupuestario. Así, ante la ausencia de un marco de muestreo de las unidades de interés, y aprovechando el principio de aglomeración de las poblaciones humanas (que forman hogares y se aglomeran en segmentos, ciudades y regiones, entre otros), la idea general de este diseño es la conformación de unidades homogéneas entre sí (conglomerados), de las que se extraerá una muestra, y la realización de un proceso exhaustivo de medición censal respecto de cada elemento del conglomerado. De esta forma, resulta natural definir las UPM como los conglomerados. Después de seleccionar una muestra de estas UPM, se realiza un censo de hogares dentro de cada una de las UPM seleccionadas. Nótese que el modelo resultante de este proceso logístico supone ventajas económicas en términos presupuestarios, ya que limita el operativo de campo a cierto número de UPM que se deben medir de manera exhaustiva.

Aunque esta estrategia resulte conveniente desde el punto de vista logístico y operativo, ciertamente no lo es desde el punto de vista de la eficiencia estadística. Los errores de muestreo que se producen al utilizar esta metodología son bastante mayores que en un diseño simple. Al realizar el proceso de aglomeración, generalmente la variación interna de los conglomerados es muy baja y la variación entre conglomerados tiende a ser muy alta, lo que crea mayor incertidumbre en la inferencia de la encuesta. Para superar estos inconvenientes, se podría pensar en un diseño de muestreo que aumente el tamaño de la muestra de conglomerados. Sin embargo, este aumento puede llegar a ser tan grande que, en algunos estratos, se deberían seleccionar todas las UPM. Por supuesto, se trata de un diseño inviable en la práctica, pero que da paso al más común en las encuestas de hogares: la selección por etapas.

5. Muestreo en varias etapas

En este diseño de muestreo, la idea general es retomar los principios del muestreo de conglomerados y realizar un submuestreo de hogares dentro de los conglomerados o las UPM que se habían seleccionado inicialmente. En general, en América Latina son muy comunes los diseños de selección en dos etapas. En la primera etapa se selecciona una muestra de UPM y en la segunda se selecciona una muestra de hogares tomada de esas UPM seleccionadas. También es posible encontrar en algunos países diseños divididos en más de dos etapas. Por ejemplo, en una primera etapa se seleccionan municipios, en una segunda etapa se seleccionan UPM dentro de los municipios seleccionados y, en la tercera, se selecciona una muestra de hogares en aquellas UPM seleccionadas en la segunda etapa pertenecientes a los municipios seleccionados en la primera etapa de muestreo. Si un municipio se incluye en la muestra, es posible realizar un proceso jerárquico y sistemático, hasta llegar a la unidad de observación. Por ejemplo, en una ciudad seleccionada, es posible hacer un submuestreo de sus secciones cartográficas, luego seleccionar sectores cartográficos (contenidos en las secciones) y, por último, seleccionar hogares o personas.

Si el diseño de muestreo incluye la selección de municipios en la primera etapa, el modelo apropiado en esta instancia deberá ser proporcional a una medida de tamaño, que puede definirse como el número de habitantes de los municipios. De esta forma, con una probabilidad muy grande, a veces igual a uno, las ciudades más importantes (con más habitantes) formarán siempre parte del estudio. Por otro lado, es posible que en algunas encuestas exista un submuestreo de personas dentro del hogar. En este caso, Clark y Steel (2007) aclaran que la selección de las personas dentro de los hogares no debería hacerse de forma aleatoria simple, puesto que ciertos grupos poblacionales podrían estar subrepresentados o sobrerrepresentados. En general, el muestreo en varias etapas tiene dos características esenciales que lo hacen robusto en términos estadísticos y eficiente a la hora de planear la logística del proceso de recopilación de información. Dichas características son:

- La independencia, que implica que no hay ninguna correlación en el diseño de muestreo de las UPM. Esto quiere decir que en cada UPM se puede ejecutar con independencia cualquier estrategia de muestreo que se considere apropiada para seleccionar la submuestra de hogares.
- La invarianza, que implica que, sin importar qué diseño de muestreo se haya ejecutado en la primera etapa para seleccionar las UPM, la segunda etapa de selección podrá ejecutarse de manera independiente de la primera. Es decir, el submuestreo de los hogares es independiente del muestreo de las UPM.

Un modelo de selección bastante utilizado en las encuestas de hogares de América Latina es el relacionado con los diseños autoponderados, en los cuales en la primera etapa de muestreo se seleccionan n_I de N_I UPM con probabilidad proporcional al número de hogares N_i que la habitan; es decir:

$$\pi_i = Pr(U_i \in s_I) = n_I \frac{N_i}{N} \quad i=1,2,\ldots,N_I.$$

En la segunda etapa de muestreo, se seleccionan hogares dentro de las UPM que se habían incluido en la etapa anterior. Esta selección se hace mediante un muestreo aleatorio simple, pero el tamaño de la submuestra es fijo para cada UPM. Es decir, no importa si una UPM es mucho más grande o pequeña que las otras. El número de hogares seleccionados será siempre el mismo. Por ejemplo, se podrían seleccionar $n_0=10$ hogares por UPM, siempre. De esta forma, en la segunda etapa, la probabilidad de que el k-ésimo hogar sea seleccionado en la submuestra s_i de la UPM U_i, que fue seleccionada en la muestra de la primera etapa s_I, está dada por la siguiente expresión:

$$\pi_{k|i} = Pr(k \in s_i | U_i \in s_I) = \frac{n_0}{N_i}$$

En los diseños autoponderados, a pesar de tenerse dos diseños de muestreo diferentes en dos etapas (proporcional al tamaño y aleatorio simple), la probabilidad de inclusión de los hogares es siempre la misma para todos los hogares, como se puede ver en la siguiente expresión:

$$\pi_k = \pi_{k|i} * \pi_i = \frac{n_0}{N_i} \frac{n_I * N_i}{N} = \frac{n_0 * n_I}{N} = \frac{n}{N}$$

Nótese que $n=n_0*n_I$ corresponderá al número total de hogares que serán seleccionados, puesto que resulta de la multiplicación del número de UPM seleccionadas en la primera etapa por el número de hogares que serán submuestreados en cada UPM en la segunda etapa. Este tipo de diseño se utiliza cuando se quiere controlar el trabajo de campo y las cuotas por ciudad o municipio. Por otro lado, una particularidad de las encuestas de hogares es que, casi siempre, las personas y los hogares comparten las mismas probabilidades de inclusión. Esto se debe a que, en la mayoría de las encuestas,

el submuestreo de las personas es exhaustivo (censo en el hogar) y, por ende, la probabilidad de inclusión en el submuestreo es forzosa.

$$\pi_k^{per} = Pr(persona \in hogar | hogar \in muestra) = 1$$

Por esta razón, se tiene que la probabilidad de inclusión de las personas en la muestra es idéntica a la del hogar, puesto que:

$$1 * \pi_{k|i} * \pi_i = 1 * \frac{n}{N} = \frac{n}{N}.$$

6. Muestreo en dos fases

En algunos casos en que el marco de muestreo contiene poca información para proponer un diseño de muestreo eficiente, el investigador puede obtener información sobre la población para construir un nuevo marco de muestreo reducido. En la primera fase, se selecciona una muestra de tamaño grande, conocida como "muestra maestra". Se debe obtener información sobre una o más variables auxiliares de cada uno de los elementos de esa muestra, con el fin de estratificar de mejor manera, recopilar información auxiliar en la muestra, o simplemente obtener muestras sucesivas y comparables a lo largo del ciclo de vida de la encuesta. En la segunda fase, con la ayuda de la información obtenida en la primera, se selecciona una submuestra mediante un diseño de muestreo conveniente, mucho más eficiente y apropiado para estimar el fenómeno objeto de estudio.

Por ejemplo, si se requieren datos estimativos precisos de distintos subgrupos poblacionales, pero no existe un marco de muestreo confiable o actualizado que permita diseñar un muestreo estratificado, es necesario realizar un diseño de muestreo en dos fases. De esta forma, se selecciona una muestra aleatoria simple de tamaño moderado. A continuación, se realiza un empadronamiento de los individuos de la muestra, a quienes se les pregunta si pertenecen a alguno de los subgrupos poblacionales de interés. En una segunda fase, con ayuda de la información recopilada en la primera fase, se realiza un diseño estratificado.

Un ejemplo de este tipo de diseños de muestreo se da en el caso de México, donde el Instituto Nacional de Estadística y Geografía (INEGI) ha planteado la elaboración de una muestra maestra que permita seleccionar submuestras para las encuestas de hogares más importantes, a la vez que se va recopilando información de los hogares pertenecientes a esta muestra maestra. En INEGI (2012), se menciona que el marco maestro de muestreo de 2012 se utilizó como punto de partida para el diseño de una muestra maestra que mantendría la información de las viviendas particulares actualizada de manera constante. En el diseño de la muestra maestra se consideraron las UPM y la estratificación utilizada en la construcción del marco de muestreo original. La muestra maestra se diseñó teniendo en cuenta la cobertura, el tamaño y la distribución de las encuestas continuas y periódicas del INEGI, utilizando los tamaños de muestra de las viviendas para estas encuestas y el

promedio óptimo de viviendas seleccionadas dentro de una UPM para determinar el número de UPM que se seleccionarían para la muestra maestra de 2012. De esta forma, la muestra maestra constituye un elemento esencial para la realización de diversas encuestas, como la Encuesta Nacional de Ocupación y Empleo, la Encuesta Nacional sobre Confianza del Consumidor, la Encuesta Nacional de Victimización y Percepción sobre Seguridad Pública o la Encuesta Nacional de Ingresos y Gastos de los Hogares.

En el caso de Costa Rica, la muestra de la Encuesta Nacional de Microempresas de los Hogares sigue un diseño en dos fases. La primera fase toma como base la Encuesta Nacional de Hogares, en la cual se identifican aquellos hogares cuyos integrantes desarrollan actividades económicas relacionadas con emprendimientos y microempresas. A partir de este listado exhaustivo, en una segunda fase, se selecciona a todas las personas que están al frente de estas microempresas y se les pide responder a un cuestionario con el fin de obtener información sobre sus características y sus actividades económicas.

Por otro lado, en Chile se realiza el Estudio Nacional de la Discapacidad, que cuenta con un marco de muestreo reducido, en una primera fase, basado en la Encuesta de Caracterización Socioeconómica Nacional (CASEN), en la que se identifican los hogares que tienen miembros con alguna discapacidad. En una segunda fase, se realiza una selección de hogares y, mediante un cuestionario estructurado, se indaga sobre las características de las personas con esta condición.

7. Muestreo balanceado

El método del cubo (Tillé, 2006) permite seleccionar muestras balanceadas, manteniendo las proporciones de la población original en la muestra en diferentes variables de equilibrio, que se espera que estén correlacionadas con las variables de interés. En general, el método del cubo permite la selección de una muestra aleatoria en que el inverso de las probabilidades de inclusión reproduce de forma exacta el total poblacional de las variables de balanceo.

Gutiérrez (2016) afirma que este es un procedimiento general y riguroso que permite la extracción de muestras probabilísticas balanceadas y la posterior estimación de las cantidades de interés, enmarcadas con métodos de inferencia basados en el diseño de muestreo. Dado que, en un diseño de muestreo balanceado, el estimador respecto de los totales de un conjunto de variables auxiliares debe ser igual al total poblacional, la varianza del estimador del total poblacional de la característica de interés se debe reducir de acuerdo con el aumento de su correlación con las variables auxiliares.

El método del cubo se compone de dos fases: la fase de vuelo y la de aterrizaje. En la primera, para que las restricciones sean satisfechas exactamente, se deben redondear a cero o uno las probabilidades de inclusión. La fase de aterrizaje consiste en el manejo adecuado del redondeo, recurriendo a la programación lineal. Esto se podría lograr, por ejemplo, aplicando el método simplex sujeto a una función de costo relacionada con la varianza del estimador.

En las encuestas de hogares es posible utilizar el algoritmo de selección del método del cubo en cada uno de los estratos conformados en el diseño de muestreo para seleccionar UPM. El método del cubo, a diferencia de los algoritmos de selección tradicionales, permite reproducir de forma exacta el número total de personas por grupos de edad y sexo a nivel de la UPM en este caso concreto.

En la Encuesta Demográfica y de Salud Familiar del Perú, se utiliza este tipo de muestreo para seleccionar las UPM. De esta manera, como variables de balanceo se podrían definir las siguientes:

- Una columna de unos para que exista balanceo en el número de UPM.
- El vector de probabilidades de inclusión iniciales.
- El total de personas por grupos de edad y sexo (a partir de la información de los censos de población y vivienda).

Si la encuesta se realiza de forma periódica, es necesario actualizar los marcos de muestreo y los tamaños poblacionales a lo largo del tiempo. Llegado el caso, el investigador puede apoyarse en las proyecciones demográficas (nacimientos esperados, muertes esperadas y población proyectada) disponibles en las fuentes oficiales como totales auxiliares.

B. El diseño de muestreo estándar en una encuesta de hogares

A continuación, se describe de manera genérica cómo es un diseño de muestreo típico de una encuesta de hogares en la región. Por supuesto, en la práctica existen variantes que se pueden alejar un poco de esta generalización, pero todas, en general, presentan la misma estructura. La mayoría de las encuestas son de naturaleza multipropósito, lo que quiere decir que existen múltiples variables de interés. Por ello, el investigador debe definir las variables más importantes de la investigación y, sobre estas, planificar el diseño de muestreo. Esto implica que, para obtener de forma simultánea la precisión requerida en todas las estimaciones, el tamaño de muestra será un poco más exigente. Asimismo, la definición de los dominios de representatividad debe estar directamente determinada por los objetivos de la encuesta y por las unidades de muestreo.

Se debe mencionar también que el diseño de muestreo de muchas de las encuestas de hogares que se realizan actualmente en la región conserva el mismo espíritu de los diseños que antes sirvieron para recolectar la información primaria. Es decir, en este campo no es usual la innovación. Más bien se podría afirmar que, cada vez que se rediseña una encuesta de hogares, el punto de partida es el diseño anterior de la encuesta. Esto resulta oportuno si se quiere mantener la comparabilidad de las cifras entre las distintas recopilaciones periódicas de información. Siempre que no haya un marco de muestreo de elementos,

es posible utilizar los principios del muestreo en varias etapas, mediante la selección de diferentes unidades que contienen los elementos de interés. Por consiguiente, el diseño de muestreo de una encuesta de hogares suele ser probabilístico, estratificado y bietápico:

- Se realiza una estratificación por zona (urbana o rural), por región, departamento o estado y por los estratos socioeconómicos definidos en los capítulos anteriores.
- De forma independiente, dentro de cada estrato se realiza un muestreo bietápico.
  - En la primera etapa, se seleccionan áreas cartográficas, denominadas UPM, siguiendo un diseño de muestreo proporcional al número de viviendas, hogares o personas del conglomerado.
  - En la segunda etapa, se escoge de manera aleatoria un número fijo de hogares dentro de cada UPM, siguiendo un diseño de muestreo aleatorio simple.

La elección de este tipo de diseño tiene una consecuencia importante en cuanto a eficiencia estadística. En la segunda etapa de muestreo, la variación que se puede presentar entre los hogares seleccionados en una misma UPM es muy baja con respecto a la que se puede presentar entre diferentes UPM. Por el principio de representatividad, las personas se aglomeran de manera natural y forman conglomerados homogéneos. Es decir, dentro de una misma UPM, los hogares tendrán características sociales bastante similares en cuanto a niveles de ingreso, gasto, desocupación, analfabetismo o educación, entre otras cosas.

Además, aunque podría suceder, no es esperable encontrar un hogar con altos niveles de ingreso y gasto, cuyos integrantes tengan un nivel de educación muy alto y que habiten una vivienda que se encuentre en un sector marginal o deprimido de la ciudad, donde el acceso al alcantarillado sea precario y existan deficiencias en los servicios de electricidad o agua potable. De la misma forma, no se esperaría que un hogar pobre, cuyo ingreso per cápita fuera bastante bajo y no alcanzara para cubrir las necesidades básicas de sus habitantes, ocupara una vivienda ubicada en un sector acaudalado.

Por lo tanto, en este tipo de investigaciones sociales, la varianza existente entre un conglomerado y otro es inmensa al compararla con la variación dentro de los conglomerados. Por esta razón, es de esperarse que existan diferencias significativas entre las UPM que componen la muestra, ya que la realidad de una UPM en un sector deprimido no es la misma que la de una UPM en un sector opulento. Este es un reflejo de las desigualdades propias de América Latina, que han ocupado la agenda política y legislativa de las últimas décadas. Se retomará esta particularidad en capítulos posteriores, al abordar el tema de la eficiencia estadística y la medición del error de muestreo.

A continuación, se definirán todos los elementos relacionados con la selección de una muestra de hogares. En general, los diseños de muestreo de las encuestas de hogares estimarán el total de cada UPM t_i mediante una submuestra seleccionada desde el marco de muestreo compuesto por los sectores cartográficos definidos en el último censo. Supóngase que la población de hogares U se divide en N_I UPM, que definen una partición de la población, llamados también conglomerados y denotados como $U_I=\{U_1,\ldots,U_{N_I}\}$

(U_I es la población de todas las UPM en un país y N_I es el número total de UPM dentro del país). Nótese que la i-ésima UPM U_i $i=1,...,N_I$ contiene N_i hogares. Luego, el proceso de selección se produce de la siguiente manera:

- Una muestra s_I de UPM es seleccionada de U_I de acuerdo con un diseño de muestreo $p_I(s_I)$. El tamaño de la muestra de UPM se denota como n_I. Nótese que s_I representa la muestra aleatoria de UPM que fue seleccionada de acuerdo con la medida de probabilidad $p_I(s_I)$.
- Para cada UPM U_i $i=1,...,n_I$ en la muestra seleccionada s_I, se realiza de forma independiente un submuestreo de hogares, de tal forma que en cada UPM existirá una muestra s_i de hogares de acuerdo con un diseño de muestreo $p_i(s_i)$. Nótese que s_i representa la muestra aleatoria de hogares que fue seleccionada en la segunda etapa, de acuerdo con la medida de probabilidad $p_i(s_i)$.

Por lo tanto, en la primera etapa se identifican todos los sectores cartográficos del país y se genera el marco de muestreo de las UPM que se separan en grupos mutuamente excluyentes, según las variables de estratificación explícita antes definidas. Dentro de cada estrato se selecciona la muestra de UPM, donde la probabilidad que tiene cada UPM de pertenecer a la muestra está determinada por el número de personas o viviendas (medida de tamaño). En esta etapa es importante tener en cuenta que se seleccionará un número mayor de UPM en los estratos más grandes. Es evidente que las regiones con más habitantes tendrán una muestra de UPM más grande, aunque esta relación no siempre es lineal. Se recomienda que el diseño de muestreo sea lo más sencillo posible[1].

A pesar de que la medida de tamaño permite que las UPM con mayor cantidad de hogares tengan una mayor probabilidad de ser escogidas, esta diferencia en las probabilidades de selección se compensa en la segunda etapa de muestreo, debido a que cada hogar tendrá la misma probabilidad de ser elegido en la muestra dentro del estrato. Cabe mencionar que, para la segunda etapa, hace falta contar con un listado exhaustivo de todos los hogares dentro de todas las UPM seleccionadas. Este proceso de selección exigirá un empadronamiento previo que permita no solo actualizar el número de hogares, sino también identificarlos y ubicarlos dentro de la UPM. De esta manera, y de forma aleatoria simple, se elige una muestra de hogares y su tamaño no varía entre una UPM y otra.

Dado que la población de interés experimenta cambios a lo largo del tiempo —debido a que los individuos nacen, mueren, migran y se unen a organizaciones, y también porque los hogares pueden formarse o disolverse—, junto con el aumento de los flujos migratorios internacionales, en los países latinoamericanos existe una migración importante desde las áreas rurales hacia las urbanas. Todo ello repercute en una desactualización constante del marco de muestreo elaborado varios años atrás. También se debe mencionar que se producen movimientos significativos entre ciudades, lo cual tiene un gran impacto sobre el marco de muestreo. Este problema de actualización del marco es común a todos los países de la región y puede abordarse a partir del ajuste constante de los pesos de muestreo de

[1] Cabe mencionar que los modelos de estimación se van volviendo más complejos a medida que en el diseño de muestra se agregan más etapas o fases.

las UPM, cada vez que se realice un operativo de campo en que se evidencie un cambio en el número de hogares de las UPM seleccionadas en la muestra de la primera etapa.

Como las UPM se seleccionan con un muestreo proporcional a su tamaño y las viviendas se seleccionan sobre el terreno mediante un muestreo simple (aleatorio simple o sistemático), previa actualización del empadronamiento y conteo de viviendas, esta actualización podría usarse para reajustar los pesos de las UPM en los nuevos procesos de recolección de información. De esta forma se reflejaría el cambio que experimenta la población de interés (que es dinámica, por definición). Sin embargo, se recomienda no modificar las probabilidades de selección de las UPM para garantizar el insesgamiento de los estimadores de muestreo.

Por ejemplo, si en un país se define un diseño de muestreo que toma en cuenta 12 viviendas dentro de cada una de las UPM seleccionadas en la primera etapa, la probabilidad de selección de la i-ésima UPM U_i estaría dada por:

$$\pi_{Ii}=Pr(U_i \in s_I)= n_I \frac{n_i}{N_i} = n_I \frac{12}{N_i}$$

Donde n_I hace referencia al número de UPM que se seleccionarán en la primera etapa, N_i representa el número de viviendas en la UPM y $n_i=12$ es el número de viviendas seleccionadas dentro de la UPM. Ahora, si el número de viviendas se actualizara en la UPM, la probabilidad de inclusión cambiaría, lo cual generaría sesgo en la estimación. Por ese motivo, las probabilidades de inclusión de las UPM deberían seguir estables de un ciclo censal a otro. El problema de subcobertura puede abordarse con el ajuste posterior de los factores de expansión en la etapa de estimación.

C. Coordinación de muestras

En la realidad, las oficinas nacionales de estadística (ONE) no realizan una sola encuesta, sino varias en un mismo año. Más aún, la información para una misma encuesta continua puede recolectarse en varios momentos en un mismo año. Por esta razón, en lo referido a la administración de los marcos de muestreo, uno de los temas más importantes es la coordinación de muestras en el tiempo y entre encuestas. Cuando se define un marco de muestreo nuevo, debido a la realización de un censo de población y vivienda, debe existir una planificación rigurosa que permita a los equipos técnicos de las ONE conocer de antemano qué UPM se seleccionarán a lo largo del siguiente período intercensal. Esta relación debe estar supeditada a todas las operaciones estadísticas basadas en encuestas de hogares.

Aunque esta planificación parezca muy exigente, puesto que un período intercensal puede durar más de diez años, es necesaria para evitar el desgaste de las UPM en las muestras maestras y el agotamiento de los respondientes. En vez de ello, la planificación rigurosa permitirá establecer de antemano los procesos logísticos, administrativos y

presupuestarios relacionados con la recopilación de la información primaria para todas las encuestas que se lleven a cabo en este período. Además, esta planificación debe atender a estrictos parámetros estadísticos en la selección de las muestras de cada encuesta. Es decir, en la selección de las UPM se debe respetar el diseño propuesto en cada operación estadística, incluidos los diseños rotativos a lo largo del período intercensal.

1. Tipos de coordinación

En esta sección se introducirán los fundamentos de los mecanismos de selección y coordinación de muestras para lograr el objetivo de planificación. En primer lugar, se establece que una muestra está coordinada positivamente con otra si ambas comparten todos los elementos. De la misma forma, dos muestras están coordinadas negativamente si no tienen ningún elemento en común. Nótese que, en el caso de las encuestas que tienen diseños rotativos complejos, existirán muestras parcialmente coordinadas y coordinadas negativamente. Por ejemplo, en un diseño rotativo 2(2)2, dos muestras cualesquiera de períodos consecutivos tendrán un traslape del 50% y estarán parcialmente coordinadas. Sin embargo, en este mismo diseño, dos muestras que estén distanciadas por dos períodos consecutivos no tendrán ningún traslape y deberán estar coordinadas negativamente.

Para lograr este cometido, es posible aplicar modelos de selección secuenciales (Gutiérrez, 2016) que utilicen números aleatorios asignados a cada UPM en el marco de muestreo. En general, existen dos tipos de números aleatorios que se pueden usar en la coordinación de muestras, incluso si se trata de diferentes diseños de muestreo. A continuación se describe cada uno de los métodos:

- Números aleatorios permanentes: cada unidad del marco recibirá un número aleatorio venido de una distribución uniforme en el intervalo unitario. Es decir, a cada unidad $i \in U_I$ se le asignará el siguiente número:

  $$\xi_i^P \sim Uniforme\ (0,1)$$

  Evidentemente, en este caso, los números aleatorios permanentes no son equidistantes.

- Números aleatorios colocados: a partir de los números aleatorios ξ_i creados en el paso anterior, es posible utilizar su rango para definir su ordenamiento y, mediante la siguiente función, crear números aleatorios equidistantes:

  $$\xi_i^C = \frac{Rango\ (\xi_i^P)\ -\ \varepsilon}{N}$$

  Donde ε es un único valor aleatorio entre cero y uno. A modo de ejemplo, considérese la población de tamaño $N=10$ del cuadro V.1, para la que se han definido números aleatorios colocados (ξ_i^C). Además, teniendo en cuenta un número aleatorio $\varepsilon=0,283$, también se definen los correspondientes números permanentes (ξ_i^P).

■ Cuadro V.1
Ejemplo reducido de la conformación de números aleatorios colocados (ξCi) **y permanentes** (ξPi)

Unidad	(ξ_i^P)	(ξ_i^C)
1	0,2875	0,1717
2	0,7883	0,6717
3	0,4089	0,2717
4	0,8831	0,7717
5	0,9404	0,9717
6	0,0455	0,0717
7	0,5281	0,4717
8	0,8924	0,8717
9	0,5514	0,5717
10	0,4566	0,3717

Fuente: Elaboración propia.

2. Coordinación de muestras aleatorias simples

Para seleccionar una muestra aleatoria simple s de tamaño n, se deberá ordenar el marco de muestreo de forma ascendente de acuerdo con los números ξ_i^P. De esta forma, la muestra s estará compuesta por lo primeros n registros del marco ordenado (o por los últimos n registros).

Es así como, para coordinar dos muestras s^1 de tamaño n_1 y s^2 de tamaño n_2, Ohlsson (1995) menciona que es posible escoger dos constantes a_1 y a_2 en el intervalo $(0,1)$. Luego, a partir del marco ordenado con los números aleatorios permanentes (o colocados), se puede definir la muestra s_1 como las primeras n_1 unidades a la derecha (o izquierda) de a_1 y la muestra s_2 como las primeras n_2 unidades a la derecha (o izquierda) de a_2. Si se quieren muestras coordinadas positivamente, entonces $a_1 = a_2$. En caso contrario, si se quieren muestras coordinadas negativamente, se deberán escoger las constantes de forma apropiada. Por ejemplo, sumar 0,5 (en el módulo 1) a la constante a_1; es decir, $a_2=(a_1+1/2) \ mod \ 1$. En general, si se quieren coordinar negativamente Q diferentes muestras, Grafstrom y Matei (2015) aconsejan añadir la cantidad de $1/Q$ (en el módulo 1) a la constante a_1.

Como continuación del ejemplo reducido, en el cuadro V.2 se presenta la selección de dos muestras coordinadas negativamente de tamaño $n_1 = n_2 = 3$, con $a_1 = 0$ y $a_2 = 0,5$.

■ Cuadro V.2
Ejemplo de la selección de dos muestras aleatorias simples coordinadas negativamente

Unidad	ξ_i^P	s^1	s^2
6	0,0455	1	0
1	0,2875	1	0
3	0,4089	1	0
10	0,4566	0	0
7	0,5281	0	1
9	0,5514	0	1
2	0,7883	0	1
4	0,8831	0	0
8	0,8924	0	0
5	0,9404	0	0

Fuente: Elaboración propia.
Nota: ξ_i^P: Números aleatorios permanentes.

3. Coordinación de muestras proporcionales

Es posible utilizar varios algoritmos de selección proporcionales a la medida de tamaño de las UPM, que corresponde generalmente al número de hogares que la conforman. El primero de ellos es el método de Poisson secuencial (Ohlsson, 1995), por el que se definen los siguientes números aleatorios permanentes respecto de cada UPM:

$$\xi_i^{pps} = \frac{\xi_i^P}{N_I \times p_i}$$

Donde N_I es el número de UPM en el marco de muestreo y $p_i = N_i / N$ es la proporción de hogares en la i-ésima UPM. De esta forma, al ordenar el marco mediante los números ξ_i^{pps} y seleccionar los primeros elementos, se obtendrá un muestreo secuencial de Poisson. En cuanto a la coordinación de muestras, es posible aplicar los mismos principios de la sección anterior. Es decir, para coordinar dos muestras s^1 de tamaño n_1 y s^2 de tamaño n_2, es posible escoger dos constantes a_1 y a_2 en el intervalo $(0,1)$. Luego, a partir del marco ordenado, se puede definir la muestra s_1 como las primeras n_1 unidades a la derecha (o izquierda) de a_1 y la muestra s_2 como las primeras n_2 unidades a la derecha (o izquierda) de a_2. En el cuadro V.3 se ejemplifica la selección de dos muestras proporcionales al tamaño de las UPM cuya coordinación es negativa.

■ Cuadro V.3
Ejemplo de la selección de dos muestreos secuenciales de Poisson coordinados negativamente

Unidad	ξ_i^P	N_I	ξ_i^{pps}	s^1	s^2
6	0,0455	198	0,0405	1	0
1	0,2875	173	0,2928	1	0
3	0,4089	184	0,3913	1	0
10	0,4566	179	0,4494	0	0
9	0,5514	195	0,4981	0	0
7	0,5281	155	0,6001	0	1
2	0,7883	162	0,8568	0	1
5	0,9404	190	0,8715	0	1
8	0,8924	166	0,9463	0	0
4	0,8831	159	0,9780	0	0

Fuente: Elaboración propia.
Nota: ξ_i^P: números aleatorios permanentes; N_I: número de unidades primarias de muestreo en el marco de muestreo; ξ_i^{pps}: números aleatorios permanentes z (método de Poisson secuencial).

En el Brasil, el Instituto Brasileño de Geografía y Estadística (IBGE) utiliza el algoritmo de Pareto (Rosén, 1997) para la selección de muestras coordinadas en la Encuesta Nacional Permanente de Hogares (PNADC) (Costa, 2007). Este algoritmo se basa en los principios de la función de distribución de Pareto con parámetros $(1,1)$ y crea los siguientes números aleatorios permanentes:

$$\xi_i^{par} = \frac{\xi_i^P/(1-\xi_i^P)}{\pi_i/(1-\pi_i)}$$

Donde $\pi_i = n_I * p_i$ es la probabilidad de inclusión de la i-ésima UPM y deberá garantizarse que sea menor que uno. Por consiguiente, al ordenar el marco mediante los números ξ_i^{par} y seleccionar los primeros elementos, se obtendrá una muestra secuencial de Poisson. Como corresponde, es posible aplicar los mismos principios de la coordinación de muestras en estos algoritmos secuenciales. En el cuadro V.4 se ejemplifica la selección de dos muestras de Pareto de tamaño $n_I = 3$, cuya coordinación es negativa.

■ Cuadro V.4
Ejemplo de la selección de dos muestras de Pareto coordinadas negativamente

Unidad	ξ_i^P	ξ_i^{par}	s^1	s^2
6	0,0455	0,0937	1	0
1	0,2875	0,9662	1	0
3	0,4089	1,5148	1	0
10	0,4566	1,9165	0	1
9	0,5514	2,4720	0	1
7	0,5281	3,1199	0	1
2	0,7883	9,7679	0	0
4	0,8831	20,3317	0	0
8	0,8924	21,0230	0	0
5	0,9404	32,9658	0	0

Fuente: Elaboración propia.
Nota: ξ_i^P: números aleatorios permanentes; ξ_i^{par}: números aleatorios permanentes (algoritmo de Pareto).

Capítulo VI
El efecto de diseño

Cuando se selecciona una muestra utilizando un diseño de muestreo complejo, que hace uso de procesos de estratificación, selección de UPM con probabilidades desiguales y múltiples etapas de selección, es muy improbable que exista independencia entre las observaciones de las unidades de interés. Además, como el muestreo de las encuestas de hogares presenta justamente este tipo de complicaciones, la distribución de la variable de interés no es la misma para todos los individuos, ni entre las UPM ni dentro de los estratos. Por ello, cuando se analizan datos que provienen de encuestas de hogares, para realizar inferencias correctas se deben tener en cuenta estas grandes desviaciones con respecto al análisis estadístico clásico, que considera muestras aleatorias simples. En la mayoría de las ocasiones, por lo tanto, se necesita aumentar el tamaño de muestra para obtener la precisión deseada.

El efecto de diseño *(DEFF)* fue definido por Kish (1965, pág. 258) como la relación entre la varianza real de una muestra y la varianza real de una muestra aleatoria simple del mismo número de elementos. Esta relación toma la siguiente expresión:

$$DEFF = \frac{Var(\hat{\theta})}{Var_{MAS}(\hat{\theta})}$$

Donde $Var(\hat{\theta})$ denota la varianza de un estimador $(\hat{\theta})$ con un diseño de muestreo complejo $p(s)$ y $Var_{MAS}(\hat{\theta})$ denota la varianza de este estimador $\hat{\theta}$ con un diseño de muestreo aleatorio simple MAS. Esta cifra da cuenta del efecto de aglomeración causado por la utilización de un diseño de muestreo complejo (p) frente a un diseño de muestreo aleatorio simple MAS, en la inferencia de un parámetro de la población finita θ (que puede ser, por ejemplo, un total, un promedio, una proporción, una razón o un percentil). En Naciones Unidas (2008, pág. 51), se concluye que hay varias formas de interpretar el efecto de diseño:

i) como el factor por el cual la varianza del diseño de muestreo complejo es mayor que la de una muestra aleatoria simple del mismo tamaño;

ii) como la medida de cuánto peor es el plan de muestreo real que la muestra aleatoria simple en términos de precisión, o

iii) como un reflejo de cuántos casos de muestra más tendrían que seleccionarse en el diseño de muestra planificado en comparación con una muestra aleatoria simple para lograr el mismo nivel de varianza de muestreo.

A. Estimación del efecto de diseño

En la expresión de efecto de diseño, se deben destacar dos hechos importantes. El efecto de diseño depende, en primer lugar, del diseño muestral $p(s)$ y, en segundo lugar, del estimador del parámetro θ. Por ello, no es correcto describirlo únicamente como una medida de eficiencia del diseño muestral, puesto que, con un mismo diseño, puede tomar diferentes valores según el parámetro que se quiera estimar.

Cabe mencionar que todos los componentes del efecto de diseño deben ser estimados, ya que se desconocen. Un estimador aproximadamente insesgado de la varianza poblacional $S^2_{y_U}$ es la varianza muestral ponderada, que está dada por la siguiente expresión:

$$\hat{S}^2_{y_U} = \left(\frac{n}{n-1}\right)\frac{\sum_s w_k\,(y_k - \hat{\theta})^2}{\sum_s w_k - 1}$$

De esta forma, en el caso en que θ corresponda a un promedio poblacional, una estimación de la varianza $Var_{MAS}\,(\hat{\theta})$ con muestreo aleatorio simple está dada por la siguiente expresión:

$$\widehat{Var}_{MAS}\,(\hat{\theta}) = \frac{1}{n}\left(1 - \frac{n}{\hat{N}}\right)\hat{S}^2_{y_U}$$

Donde $\hat{N}=\sum_s w_k$. Por lo tanto, la estimación del efecto de diseño está dada por:

$$\widehat{DEFF} = \frac{\widehat{Var}(\hat{\theta})}{\widehat{Var}_{MAS}\,(\hat{\theta})}$$

La idea central del efecto de diseño es la evaluación del mismo estimador con diferentes escenarios de muestreo. Como el estimador que se está estudiando ($\hat{\theta}$) viene ponderado por los factores de expansión de la encuesta, lo más conveniente es utilizar el mismo rasero para evaluar ambas estrategias de muestreo. Puede consultarse una discusión más profunda sobre el efecto de diseño en Gambino (2009, sec. 4.), Särndal, Swensson y Wretman (2003, pág. 188) y Gutiérrez, Zhang y Montaño (2016, pág. 101).

B. Descomposición del efecto de diseño en las encuestas de hogares

Park y otros (2003) sostienen que el efecto de diseño de cualquier encuesta puede descomponerse en tres partes que se relacionan entre sí de forma multiplicativa. En primer lugar, está el efecto debido a la ponderación desigual, $DEFF^W$. En segundo lugar, se encuentra el efecto debido a la estratificación, $DEFF^S$. Por último, se tiene el efecto debido al muestreo en varias etapas, $DEFF^C$. Por lo tanto:

$$DEFF = DEFF^W \times DEFF^S \times DEFF^C$$

La primera componente $(DEFF^W)$ del efecto de diseño general tiende a aumentar ligeramente la variación de las estrategias de muestreo. Valliant, Dever y Kreuter (2018) afirman que esta componente puede estimarse por medio de la siguiente expresión:

$$DEFF^W = 1 + cv^2\ (w_k)$$

Donde $cv(w_k)$ representa el coeficiente de variación de los pesos de muestreo w_k de las unidades en la encuesta. Si los pesos de muestreo son uniformes, no habrá un incremento significativo en la varianza de la estrategia. Por esa razón, los modelos autoponderados son deseables en los diseños de muestreo de las encuestas de hogares. Por otra parte, si los pesos de muestreo tienen una variación grande, habrá un incremento significativo en la varianza y, por ende, en el tamaño de muestra. Como se verá más adelante, los ajustes en el factor de expansión pueden provocar una alta variabilidad. Por consiguiente, se recomienda, en la medida de lo posible, crear clases o subgrupos de ajuste para mitigar y acotar la dispersión de los pesos finales de la encuesta.

Al encontrar la mejor estratificación, hay que asegurarse de que la segunda componente $(DEFF^S)$ de esta descomposición sea inferior a 1 (es decir, que la varianza se reduzca). Lamentablemente, la reducción de la varianza no suele ser tan grande y no mitiga los efectos de aglomeración debido a las múltiples etapas de las que se componen los diseños de muestreo complejos. Como indica Gutiérrez (2016), el efecto de diseño en el muestreo aleatorio estratificado sin reemplazo con asignación proporcional está dado por:

$$DEFF^S \cong \frac{Varianza\ dentro\ de\ los\ estratos}{Varianza\ total}$$

Ahora, se deduce intuitivamente que la varianza total es la suma de la varianza dentro de los estratos y la varianza entre los estratos. Por tanto, se concluye que, casi siempre, esta estrategia de muestreo arrojará mejores resultados que una estrategia aleatoria simple. Por otro lado, cabe recordar que el efecto de diseño debido a la conglomeración de la población finita en las unidades primarias de muestreo (UPM) está dado por la siguiente expresión:

$$DEFF^C = 1 + (\bar{n}_{II} - 1)\ \rho_y$$

Donde, $\bar{n}_{II}$ es el número de hogares promedio que se seleccionan en cada UPM y ρ_y es el coeficiente de correlación intraclase, calculado para la variable de interés sobre las UPM. Al haberse definido el marco de muestreo en el momento de la recopilación de la información primaria censal, ya no se tendrá control sobre el valor del coeficiente de correlación intraclase (ρ_y). Únicamente se tiene control sobre el número de viviendas que serán seleccionadas en promedio en las UPM ($\bar{n}_{II}$). Si el marco de muestreo quedó definido correctamente, el valor de ρ_y será tan pequeño como haya sido posible establecerse al proponer las UPM. De la misma manera, es recomendable que el equipo técnico de los institutos nacionales de estadística (INE) defina el menor número promedio posible de encuestas dentro de las UPM $\bar{n}_{II}$ para que el efecto de aglomeración resulte mínimo.

En general, la disminución del efecto de diseño debido a la estratificación se matiza con el aumento del efecto de diseño debido a la desigualdad de los pesos de muestreo. Es por esta razón por la que $DEFF^C$ predomina en el efecto de diseño general y, por lo tanto, se le presta mucha atención. En Naciones Unidas (2008, pág. 40), se propone que, para mitigar los efectos del muestreo multietápico, se consideren las siguientes estrategias:

i) seleccionar tantas UPM como sea posible;

ii) definir un tamaño de UPM lo más pequeño posible, en términos del número de viviendas que las componen;

iii) seleccionar un número fijo de viviendas dentro de las UPM seleccionadas, en vez de un número variable, y

iv) utilizar un muestreo sistemático en la UPM, en vez de seleccionar segmentos de viviendas contiguas.

Al encontrar la mejor estratificación, los funcionarios de los INE consiguen que la segunda componente $(DEFF^S)$ de la descomposición del efecto de diseño general sea mínima para los indicadores estudiados. También es tarea de los INE asegurarse de que los efectos de diseño dados por el efecto de conglomeración y el uso del muestreo en varias etapas $(DEFF^C)$ sea mínimo. En este caso, se deberá estudiar, para cada encuesta y operación estadística que haga uso del marco de muestreo estratificado, la relación entre UPM y hogares a la luz de los indicadores de interés. En particular, es necesario decidir cuántos hogares se seleccionarán en cada UPM y cuántas UPM se seleccionarán dentro de cada estrato.

De la misma manera, y como se verá en capítulos posteriores, el efecto debido al uso de factores de ponderación desiguales $(DEFF^W)$ puede minimizarse al decidir, a la luz de la correlación entre los indicadores particulares de cada encuesta de hogares, qué variables de control se utilizarán en la calibración de los estimadores. De esta forma, en esta estrategia tripartita, se garantiza que el efecto de diseño general de las encuestas sea pequeño.

C. Formas comunes del efecto de diseño

Suponiendo que el parámetro de interés es la media poblacional $(\bar{y})$ de una variable de interés y (por ejemplo, el ingreso per cápita mensual), es posible escribir la varianza del estimador con el diseño de muestreo complejo como:

$$Var(\hat{\bar{y}}) = \frac{DEFF}{n}\left(1 - \frac{n}{N}\right) S^2_{y_U}$$

Donde $S^2_{y_U}$ corresponde a la varianza poblacional de las características de interés, N es el tamaño de la población de interés U y n el tamaño de la muestra de individuos. Por otro lado, suponiendo que el parámetro de interés es la proporción poblacional (P) de una variable dicotómica y (por ejemplo, el porcentaje de individuos que se encuentran por debajo de la línea de pobreza en un país), es posible escribir la varianza del estimador con el diseño de muestreo complejo como:

$$Var(\hat{P}) = \frac{DEFF}{n}\left(1 - \frac{n}{N}\right) P(1-P)$$

Cuando se trata de un diseño muestral multietápico, por ejemplo, es común seleccionar UPM en la primera etapa y luego escoger hogares dentro de las áreas seleccionadas. En este contexto, el coeficiente de correlación intraclase está definido por:

$$\rho_y = 1 - \frac{N_I}{N_I - 1}\frac{SCD}{SCT}$$

Donde, apelando a la notación clásica de los análisis de varianza, $SCT = \sum_U (y_k - \bar{y}_U)^2$ hace referencia a la suma de cuadrados total, $SCT = \sum_{U_I} N_I\,(\bar{y}_{U_I} - \bar{y}_U)^2$ es la suma de cuadrados entre las UPM, y $SCD = SCT - SCE$ es la suma de cuadrados dentro de las UPM. Cuando la característica de interés y es heterogénea entre los conglomerados (UPM), pero los conglomerados son homogéneos entre sí, entonces ρ_y es cercano a 0. Mientras tanto, si los conglomerados son heterogéneos entre sí, pero homogéneos dentro de cada uno, entonces ρ_y es cercano a 1. En este tipo de escenarios, el efecto de diseño se puede expresar como $DEFF = 1 + (\bar{n}_{II} - 1)\rho_y$. En general, el efecto de diseño será mayor cuando:

i) El coeficiente de correlación crezca, lo cual no puede controlarse de antemano, puesto que se trata de la observación de la realidad. En general, ρ_y será más grande cuando la distribución de la variable de interés sea explicada por las UPM del país. Por ejemplo, si el indicador de interés es la pobreza y los hogares pobres están aglomerados, segregados y separados de los hogares más acaudalados, entonces ρ_y será más grande. Además, cuanta más segregación haya, mayor será su valor.

ii) El promedio de hogares seleccionados por UPM ascienda. Esto se controla de antemano en la etapa de diseño y será un número fijo y transversal en la encuesta.

D. Otras consideraciones

1. El efecto de diseño en subpoblaciones

La estimación del efecto de diseño es un problema común cuando se trabaja con estimaciones desagregadas en subpoblaciones de interés. Por un lado, cuando las subpoblaciones constituyen estratos (o agregaciones de estratos) planeados de antemano, cuyo tamaño poblacional se conoce previamente, se tiene el siguiente efecto de diseño:

$$DEFF_h = \frac{Var(\hat{\theta})}{Var^h_{MAS}(\hat{\theta}_h)}$$

Donde $Var^h_{MAS}(\hat{\theta}_h)$ es la varianza del estimador restringida al estrato $h(h=1,...,H)$. En el caso en que $\hat{\theta}_h$ corresponda al estimador del promedio poblacional en el estrato h, su valor es el siguiente:

$$Var^h_{MAS}(\hat{\theta}_h) = \frac{1}{n_h}\left(1 - \frac{n_h}{N_h}\right) S^2_{y_{U_h}}$$

Siendo n_h el tamaño de la muestra en el estrato h, N_h el tamaño poblacional del estrato h y $S^2_{y_{U_h}}$ la varianza poblacional de la variable de interés restringida al subgrupo h. Por lo tanto, los efectos de diseño para las medias muestrales en un diseño aleatorio estratificado serán, por definición, iguales a uno.

Por otro lado, cuando la subpoblación de interés no es un estrato o un posestrato, sino un subgrupo aleatorio (por ejemplo, las personas pobres, las personas ocupadas o cualquier otro subgrupo no planeado en el diseño de la encuesta o en la etapa de calibración), en adelante denotado con la letra g, cuyo tamaño de muestra no es fijo (o condicionalmente fijo por la calibración), sino aleatorio, la estimación correcta del efecto de diseño es la siguiente:

$$DEFF_g = \frac{Var(\hat{\theta}_g)}{Var^U_{MAS}(\hat{\theta}_g)}$$

Donde $Var^U_{MAS}(\hat{\theta}_g)$ es la varianza del estimador de interés. En el caso en que $\hat{\theta}_g$ corresponda al estimador del promedio poblacional en el subgrupo g, su varianza estaría dada por la siguiente expresión:

$$Var^U_{MAS}(\hat{\theta}_g) = \frac{1}{n}\left(1 - \frac{n}{N}\right) S^2_{y_{gU}}$$

Donde $S^2_{y_{gU}}$ es la varianza poblacional de una nueva variable calculada en toda la población, que toma el valor de y_k cuando la unidad k pertenece al subgrupo g, y toma el valor de 0 en cualquier otro caso. Por lo tanto, en ambos efectos de diseño, la estimación de la varianza del diseño de muestreo complejo $Var(_h)$ o $Var(\hat{\theta}_g)$ es la misma, pero el

denominador cambia dependiendo de si el subgrupo es un estrato o no. Por esta razón, al analizar una encuesta de hogares, hay coincidencia en las cifras relacionadas con la estimación puntual, errores estándar, intervalos de confianza y coeficientes de variación entre los diferentes programas computacionales. Sin embargo, es necesario percatarse de las opciones que estos ofrecen para calcular correctamente la cifra apropiada. En resumen, las estimaciones de $Var^{U}_{MAS}(\hat{\theta}_g)$ y $Var^{h}_{MAS}(\hat{\theta}_h)$ serán diferentes, puesto que la primera tiene que ver con toda la muestra, mientras que la segunda tiene que ver únicamente con la muestra del estrato.

Lumley (2010) afirma que el efecto de diseño compara la varianza de una media o total con la varianza de un estudio del mismo tamaño utilizando un muestreo aleatorio simple sin reemplazo, y que su cálculo será incorrecto si los pesos de muestreo se han reescalado o no son recíprocos con respecto a las probabilidades de inclusión. Por ejemplo, en el caso de las subpoblaciones, la librería *survey* de R compara la varianza de la estimación con la varianza de una estimación basada en una muestra aleatoria simple del mismo tamaño que el de la subpoblación. En el muestreo aleatorio estratificado, por ejemplo, el efecto de diseño calculado en un estrato será igual a 1.

2. El efecto de diseño general

Supóngase que el diseño muestral es estratificado, con H estratos. En ese caso, por la independencia de la selección en los estratos, la varianza del estimador de un total poblacional t_y está dada por:

$$Var(\hat{t}_{y,\pi}) = \sum_{h=1}^{H} Var_h(\hat{t}_{y,\pi})$$

Donde:

$$Var_h(\hat{t}_{y,\pi}) = DEFF_h \times Var_{MAS,h}(\hat{t}_{y,\pi})$$

Por otro lado:

$$Var(\hat{t}_{y,\pi}) = DEFF \times Var_{MAS}(\hat{t}_{y,\pi})$$

De esta forma, se tiene que:

$$DEFF = \frac{\sum_{h=1}^{H} DEFF_h \, Var_{MAS,h}(\hat{t}_{y,\pi})}{Var_{MAS}(\hat{t}_{y,\pi})} = \frac{\sum_{h=1}^{H} DEFF_h \, \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) S^2_{y,U_h}}{\frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{y,U}}$$

Es decir, el efecto de diseño puede verse como una combinación lineal de los efectos de diseño de los H estratos $(DEFF=\sum_{h=1}^{H} DEFF_h\, w_h)$, donde el peso w_h está dado por:

$$w_h = \frac{\frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) S_{y,U_h}^2}{\frac{N^2}{n}\left(1 - \frac{n}{N}\right) S_{y,U}^2}$$

Está claro que los pesos w_h son todos positivos, pero no necesariamente son menores que 1, y su suma tampoco es igual a 1. A continuación, se examina la forma de w_h en el caso especial de muestras autoponderadas en todos los estratos. En este caso, se tiene que $\frac{n_h}{N_h} = \frac{n}{N}$ para todo $h=1,\ldots,H$, y se puede ver que:

$$w_h = \frac{N_h S_{y,U_h}^2}{N S_{y,U}^2} = \frac{\sum_{U_h}(y_k - \bar{y}_{U_h})^2}{\sum_{U}(y_k - \bar{y}_{U})^2}$$

Aunque $\sum_{h=1}^{H} w_h \neq 1$, el peso del estrato h sí tiene una interpretación interesante, pues queda definido como la suma de cuadrados dentro del estrato, dividida por la suma de cuadrados totales de la variable de interés. Se puede concluir que, si los estratos están bien construidos, es decir, si la variable de interés es homogénea dentro de cada estrato y los diferentes estratos son heterogéneos entre sí, los pesos w_h serán muy pequeños y el efecto de diseño general resultará mucho más pequeño que los efectos de diseño de los estratos.

Por otro lado, si los estratos no se construyeron teniendo en cuenta la variabilidad de la característica de interés, entonces $S_{y,U_h}^2 \approx S_{y,U}^2$ y $w_h = N_h/N$. De esta forma, la suma de los pesos es igual a 1, y se puede concluir que el efecto de diseño del diseño general es un promedio ponderado de los efectos de los H estratos, y el estrato con mayor peso será aquel que tenga mayor representación del universo.

Por último, alguno de los pesos w_h puede resultar mayor que 1 cuando, para algún estrato, $\frac{n_h}{N_h} \neq \frac{n}{N}$, y también cuando los estratos no están bien construidos.

3. El efecto de diseño en las encuestas de hogares de la región

En general, para las encuestas de hogares en la región, se planean modelos de estratificación, aglomeración y selección de UPM con probabilidades desiguales. Heeringa, West y Berglund (2017) señalan que el efecto de estratificación reduce la varianza de las estrategias de muestreo, mientras que el efecto de selección desigual tiende a aumentarla. En general, estos dos efectos tienden a anularse entre sí. Por lo tanto, el efecto de diseño de una encuesta compleja

se dará únicamente en función del efecto de aglomeración, que puede llegar a ser grande en comparación con los otros dos. Como ya se había comentado antes, la expresión generalizada que da cuenta del efecto de aglomeración en los diseños de muestreo complejos de las encuestas de hogares es la siguiente:

$$DEFF \approx 1+(\bar{n}_{II}-1)\,\rho_y$$

Donde se recalca que $\bar{n}_{II}$ representa el número promedio de hogares seleccionados dentro de cada UPM y ρ_y es el coeficiente de correlación intraclase, que representa el grado de homogeneidad de la variable de interés dentro de cada hogar.

Este efecto cambiará dependiendo de si la inferencia de la encuesta de hogares se quiere realizar a nivel nacional o a nivel regional. Por ejemplo, en Naciones Unidas (2007, cap. 7), se presenta el comportamiento de esta medida en tres encuestas de hogares del Brasil: la Encuesta Nacional de Hogares (PNAD), la Encuesta Mensual de Empleo y la Encuesta de Condiciones de Vida. En general, en estas encuestas se utiliza la estratificación y selección de UPM con probabilidades desiguales. Además, el tamaño promedio de las UPM es de 250 viviendas, de las cuales se seleccionan 13 en la PNAD, 20 en la Encuesta Mensual de Empleo y 16 y 8 viviendas en la Encuesta de Condiciones de Vida en la zona rural y la urbana, respectivamente.

De acuerdo con Naciones Unidas (2007, cap. 7), los efectos de diseño no solo son diferentes para cada parámetro que se desea estimar, sino que varían de acuerdo con la subpoblación en que se realiza la estimación. Por ejemplo, al considerar el parámetro de la proporción de hogares con electricidad, se estimó que el efecto de diseño para este parámetro era de 7,92 a nivel nacional, de 1,03 en las áreas metropolitanas, de 4,43 en las ciudades grandes y de 7,27 en las áreas rurales. Por ello, y sobre la base de la expresión que define el efecto de diseño, se observó que, fijando $\bar{n}_{II'}$ el coeficiente de correlación intraclase variaba dependiendo de la zona. En efecto, se determinó que $\rho_y=0,76$ a nivel nacional, $\rho_y=0,0033$ en las zonas metropolitanas, $\rho_y=0,38$ en las ciudades grandes y $\rho_y=0,69$ en las áreas rurales. Ello supone que existe una mayor heterogeneidad de la variable de interés (acceso a la electricidad) entre las UPM a nivel nacional. Es decir, el panorama nacional ($\rho_y=0,76$) se puede entender como conformado por dos grandes grupos de UPM: aquellas con una proporción alta de hogares con acceso a la electricidad y aquellas con una proporción baja de hogares con acceso a la electricidad. Por ende, existe una gran heterogeneidad entre un grupo y otro. En la zona rural del país, se presenta una situación similar. Por otro lado, al contrario de lo que sucede en el panorama nacional o rural, dentro de las zonas metropolitanas sí se observa una gran homogeneidad entre las UPM ($\rho_y=0,0033$), lo que puede entenderse como que solo existen UPM con una proporción alta de acceso a la electricidad.

Por otra parte, en la misma encuesta PNAD, los efectos de diseño para el número promedio de cuartos usados como dormitorios son de 2,14 a nivel nacional, de 2,37 en las áreas metropolitanas, de 1,72 en las ciudades grandes y de 2,09 en las áreas rurales. Considerando que $\bar{n}_{II}=10$, el coeficiente de correlación intraclase es de $\rho_y=0,12$ a nivel nacional, $\rho_y=0,15$ en las zonas metropolitanas, $\rho_y=0,08$ en las ciudades grandes y $\rho_y=0,12$ en las áreas rurales. Ello supone que existe una alta homogeneidad de la variable de interés (dormitorios en la vivienda) entre las UPM a nivel nacional. Es decir, se puede entender que el panorama nacional ($\rho_y=0,12$) presenta un promedio de dormitorios por UPM bastante similar entre cada una de las UPM que conforman el marco de muestreo del país, de las zonas metropolitanas y de las zonas rurales. Sin embargo, en las ciudades grandes, esta similitud parece estar más marcada ($\rho_y=0,08$); es decir, las UPM son mucho más similares en lo que respecta al promedio de dormitorios en las UPM.

Como se verá en los capítulos siguientes, al conocer el valor que toma el efecto de diseño para la estimación de un parámetro de interés, es posible crear escenarios de simulación que permitirán establecer el tamaño de muestra al planificar las encuestas de hogares o al rediseñarlas después de la ronda de censos de una década en particular.

Capítulo VII
Cálculo del tamaño de la muestra

Uno de los temas más importantes en la literatura sobre diseño y análisis de encuestas de hogares es el tamaño de la muestra. Habitualmente, en los libros de estadística y muestreo se establecen las características generales de los diseños de muestreo y las propiedades estocásticas de los estimadores, sin profundizar en el hecho de que la muestra debe seleccionarse y de que esta selección depende de cuántos hogares se necesiten en el estudio. En realidad, al hablar del tamaño de la muestra en una encuesta de hogares, no solo se debe hacer referencia a los hogares, sino también a las personas.

La determinación del tamaño de la muestra debería depender del propósito de la encuesta. Por ejemplo, considérese el caso de una encuesta de propósitos múltiples que se realiza cada año con el fin de indagar acerca de fenómenos demográficos, sociales, educativos y de condiciones de vida. En este contexto, se debe tener en cuenta que el tamaño de muestra definido ha de ser útil, pertinente y apropiado para todos los indicadores que se desean medir al mismo tiempo. En este capítulo, el lector podrá encontrar una guía útil para determinar la mejor ruta a la hora de abordar el cálculo del tamaño de la muestra en las encuestas de hogares.

A. Confiabilidad y precisión

Antes de presentar las metodologías básicas para el cálculo del tamaño de muestra mínimo, es necesario definir los diferentes tipos de error de muestreo en una encuesta. En principio, se define un intervalo de confianza para el parámetro θ, generado por su estimador insesgado $\hat{\theta}$, que se supone con distribución normal de media θ y varianza $Var\,(\hat{\theta})$, como:

$$IC(1-\alpha) = \left[\hat{\theta}-z_{1-\alpha/2}\sqrt{Var(\hat{\theta})},\hat{\theta}+z_{1-\alpha/2}\sqrt{Var(\hat{\theta})}\right]$$

Donde $z_{1-\alpha/2}$ se refiere al cuantil $(1-\alpha/2)$ de una variable aleatoria con distribución normal estándar. Cuando el diseño de muestreo es complejo, es necesario reemplazar el percentil de la distribución normal estándar por el percentil de una distribución t de Student con $N_I - H$ grados de libertad, suponiendo que hay N_I unidades primarias de muestreo y H estratos. En este orden de ideas, obsérvese que:

$$1-\alpha = \sum_{Q_0 \supset s} p(s),$$

Donde Q_0 es el conjunto de todas las posibles muestras cuyo intervalo de confianza contiene el parámetro θ. Desde la expresión del intervalo de confianza, se define el margen de error como aquella cantidad que se suma y se resta al estimador insesgado. En este caso, se define como:

$$ME = z_{1-\alpha/2} \sqrt{Var(\hat{\theta})}$$

Desde esta expresión también es posible definir el error estándar, dado por:

$$EE = \sqrt{Var(\hat{\theta})}$$

Las anteriores medidas solo tienen en cuenta la precisión del estimador. Una medida que tiene en cuenta la precisión y el sesgo del estimador es el margen de error relativo, que se define como:

$$MER = z_{1-\alpha/2} \frac{\sqrt{Var(\hat{\theta})}}{E(\hat{\theta})}$$

De la misma manera, también se determina el coeficiente de variación o error estándar relativo, definido por:

$$CV = \frac{\sqrt{Var(\hat{\theta})}}{E(\hat{\theta})}$$

El tamaño de la muestra dependerá del tipo de error que se quiera minimizar. Por ejemplo, en una población particular, el tamaño de muestra requerido para minimizar el margen de error no será el mismo que el que se necesitará para minimizar el coeficiente de variación.

B. El efecto de diseño en la determinación del tamaño de la muestra

Al diseñar un estudio por muestreo con encuestas de hogares, es importante establecer el número mínimo de encuestas y entrevistas que se deben realizar. Esto es necesario para determinar el costo del estudio, y en el ámbito técnico, permite tener control desde la fase de diseño sobre la calidad estadística de los resultados esperados en el estudio. Como se mencionó anteriormente, dicha calidad puede medirse en términos de error de muestreo, con indicadores como el margen de error, el margen de error relativo o el coeficiente de variación. Todas estas medidas dependen de la varianza del estimador con el diseño muestral complejo. Por lo tanto, el hecho de contar con un valor aproximado para el efecto de diseño *(DEFF)* permite obtener una aproximación a dicha varianza, y acercarse al error de muestreo del estudio en la fase de diseño.

Uno de los primeros paradigmas con los que se debe lidiar es el de la independencia entre las observaciones. Se trata de un supuesto que gobierna gran parte de la teoría del análisis estadístico, pero que desafortunadamente no se aplica en el contexto de las encuestas de hogares. Ante los retos que se deben enfrentar y las diversas estrategias de recopilación de información, no son plausibles las fórmulas que se desprenden del supuesto de que las observaciones corresponden a una muestra de variables independientes e idénticamente distribuidas.

La estratificación, las múltiples etapas y la aglomeración de las unidades de muestreo hacen que este supuesto no se cumpla en la práctica. Por lo tanto, si se utilizan las expresiones tradicionales que se encuentran en los libros introductorios de estadística, se obtendrán tamaños de muestra insuficientes. El problema del tamaño de la muestra en las encuestas de hogares ha sido abordado por distintos autores con diferentes enfoques. Quizás uno de los más aceptados es aquel que define un factor de ajuste, llamado efecto de diseño, en función de la correlación que hay entre la variable de interés y las unidades primarias de muestreo. A partir de este efecto de diseño, se calcula el número de personas que deben ser encuestadas para minimizar un error de muestreo predefinido.

Cuando para la población de interés se selecciona una muestra a partir de un diseño de muestreo de conglomerados o en varias etapas, no es imposible afirmar que existe independencia entre las observaciones. De ahí que no sea posible utilizar las fórmulas clásicas para la determinación de un tamaño de muestra al considerar un diseño de muestreo aleatorio simple. Sin embargo, una forma sencilla de incorporar este efecto de aglomeración en las expresiones clásicas del muestreo aleatorio simple se basa en la relación de las varianzas en el efecto de diseño:

$$DEFF(\hat{\theta}) = \frac{Var_p(\hat{\theta})}{Var_{MAS}(\hat{\theta})}$$

Esta cifra da cuenta del efecto de aglomeración causado por la utilización de un diseño de muestreo cualquiera (p) frente a un diseño de muestreo aleatorio simple (MAS)

en la inferencia de un parámetro de la población finita θ (que puede ser, por ejemplo, un total, una proporción, una razón o un coeficiente de regresión). Por ello, es posible escribir la varianza del estimador en el diseño de muestreo complejo como:

$$\begin{aligned} Var_p(\hat{\theta}) &= DEFF(\hat{\theta})\, Var_{MAS}(\hat{\theta}) \\ &= DEFF(\hat{\theta})\, \frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{y_U} \end{aligned}$$

Por lo tanto, si al implementar un muestreo aleatorio simple, el tamaño de muestra n_0 es suficiente para conseguir la precisión deseada, el valor del tamaño de muestra que tendrá en cuenta el efecto de aglomeración para un diseño complejo será cercano a $n \approx n_0 \times DEFF$. Por consiguiente, un efecto de diseño $DEFF = 2{,}0$ implicaría que se deberían seleccionar casi el doble de unidades para lograr la misma confiabilidad que la producida por una muestra aleatoria simple. En Naciones Unidas (2008), se afirma que, dada esta relación, es claramente indeseable tener un plan de muestreo con valores mucho mayores que 2,5 o 3,0 para los indicadores clave de la encuesta. A partir de esta advertencia, se establece una regla precisa a la hora de escoger el escenario de muestreo más conveniente, puesto que los cuadros de muestreo deberán filtrarse en función de los casos que generen efectos de diseño inferiores a 3. Ello quiere decir que los equipos técnicos de los institutos nacionales de estadística (INE) deben plantear modelos en los que el efecto de diseño para los indicadores clave de la encuesta no sea desproporcionadamente grande.

En particular, en el caso de una proporción, la calidad del estimador se puede medir en términos de la amplitud del intervalo de confianza de al menos $(1-\alpha)\times 100\%$; es decir, la distancia entre el estimador y el parámetro no debería superar un margen de error previamente establecido (ME). Así:

$$1-\alpha \geq Pr\left(|\hat{P} - P| < ME\right)$$

Por ejemplo, el estimador de Horvitz-Thompson de la proporción $\hat{P}$ es insesgado para P y su distribución asintótica es gaussiana, con una varianza dada por:

$$Var(\hat{P}) = DEFF\, \frac{1}{n}\left(1 - \frac{n}{N}\right) P(1-P)$$

Al despejar el tamaño muestral n de la anterior expresión, se tiene que:

$$n \geq \frac{P(1-P)}{\frac{ME^2}{DEFF\, z^2_{1-\alpha/2}} + \frac{P(1-P)}{N}}$$

De la misma manera, si el interés recae en la estimación de un promedio $\bar{y}_U$, el tamaño de muestra necesario para que la amplitud relativa del intervalo de confianza no supere un margen de error relativo previamente establecido (MER) es de:

$$n \geq \frac{S^2_{y_U}\, DEFF}{\frac{MER^2\, \bar{y}^2_U}{z^2_{1-\alpha/2}} + \frac{S^2_{y_U}\, DEFF}{N}}$$

Por consiguiente, los altos valores del efecto de diseño redundarán en un mayor tamaño de muestra. Claramente, el incremento no es lineal; más aún, el tamaño de la muestra se verá más afectado en la medida en que el efecto de diseño sea más grande.

C. Algunos escenarios de interés en la asignación del tamaño de la muestra

En general, en las encuestas de hogares se parte de un marco de muestreo de áreas que agrupa a toda la población de un país. Estas áreas están definidas como agregaciones cartográficas o unidades primarias de muestreo (UPM) y contienen, a su vez, a los hogares donde se encuentran las personas que son susceptibles de ser entrevistadas. Sin embargo, debido a la agrupación natural de las personas en hogares, a veces los cálculos se hacen complejos, máxime conociendo que la población de interés es un subconjunto de los habitantes de los hogares. Por otro lado, debido a que el marco de muestreo comúnmente usado por los INE es una lista de UPM, se hace necesario, más allá de calcular el tamaño de la muestra de personas, calcular también el tamaño de la muestra de UPM y los hogares incluidos en la muestra. Por lo tanto, en este documento, el objetivo es sintetizar los mecanismos de asignación de muestra en tres escenarios que son comunes en la práctica estadística del diseño de encuestas de hogares:

i) Asignación del tamaño de la muestra en problemas de inferencia que tienen que ver con la estimación de parámetros asociados a personas. En este escenario se presenta la metodología apropiada para calcular el tamaño de la muestra de UPM, de hogares y, finalmente, de personas.

ii) Cuando la variable de diseño (y, en general, las variables más importantes de las encuestas) están presentes a nivel de hogar, no es necesario realizar un submuestreo de personas. Partiendo de la lógica del escenario anterior, se presenta la metodología adecuada para calcular el tamaño de la muestra de UPM y de hogares.

iii) Un caso menos común en los países de América Latina se presenta cuando el marco de muestreo empadrona a las personas dentro de las UPM y, además, la encuesta solo pretende observar características asociadas a los habitantes del hogar (es decir, no se intenta observar características ni del hogar ni de la vivienda). En este caso, no hay un submuestreo de hogares.

En general, al definir las expresiones de tamaño de la muestra, se debe ser cuidadoso con la notación, para lo cual se supone una población U de N elementos sobre la que se desea seleccionar una muestra s de n elementos acerca de los cuales se quiere medir una característica de interés. En algunos casos, la población U no constituye la población de interés, sino que la contiene. Es decir, si se define a U_d como la población de interés, entonces $U_d \subseteq U$. En términos de notación, se tiene lo siguiente:

- N es el tamaño de la población U.
- n es el tamaño de la muestra s.
- N_I es el número de UPM en el marco de muestreo.
- n_I es el número de UPM que se selecciona en la muestra de la primera etapa s_i.
- N_{II} es el número de hogares existentes en el país.
- n_{II} es el número de hogares seleccionados en la muestra de la segunda etapa s_{II}.
- $\bar{n}$ es el número promedio de personas que se van a seleccionar en cada UPM.
- $\bar{n}_{II}$ es el número promedio de hogares que se van a seleccionar en cada UPM.
- ρ_y es el coeficiente de correlación intraclase, calculado para la variable de interés sobre las UPM.
- b es el número promedio de personas por hogar.
- r es el porcentaje promedio de personas en el hogar susceptibles de ser observadas para medir la característica de interés.
- $z_{1-\alpha/2}$ es el percentil $(1-\alpha/2)$ asociado a una distribución normal estándar y a la confianza que se requiera en la inferencia.

Para presentar las metodologías apropiadas, junto con las expresiones adecuadas, en cada escenario se definirán las cantidades de interés, se hará una breve introducción al problema y se realizarán los cálculos de manera detallada, con ejemplos de encuestas reales. Con el fin de mantener la uniformidad de los cálculos, todos los ejemplos suponen una población de tamaño $N=50$ millones, con $N_{II}=12$ millones de hogares, y se desea obtener una muestra con una confianza del 90%. En cada escenario, se supone que el país está dividido en $N_I=30.000$ UPM, conformadas por segmentos cartográficos (agregaciones de manzanas).

Para simplificar los cálculos y mantener la atención del lector, las expresiones que se presentarán en este capítulo corresponden al número de individuos que se deberían seleccionar a nivel nacional, o para un solo subgrupo de interés. Por lo tanto, estos cálculos deben realizarse tantas veces como dominios de representatividad existan en la encuesta. Por ejemplo, si el interés está en hacer inferencia en dos estratos, el rural y el urbano, se deben calcular estas expresiones dos veces, una para cada área. Al final, el tamaño de muestra nacional será la sumatoria de los tamaños de la muestra en cada uno de los estratos del país.

D. Tamaño de la muestra para UPM, hogares y personas

Cuando la unidad de observación sean las personas, sin importar que la variable de interés esté a nivel de los hogares, será necesario siempre basar los cálculos en el tamaño de la muestra de personas. Por ejemplo, para obtener una inferencia apropiada al estimar el ingreso medio per cápita, el porcentaje de personas pobres o el porcentaje de personas con una característica particular, es necesario definir la población objetivo como todas las personas que componen un hogar y, posteriormente, medir la variable de interés que será observada en todos los casos.

Con estos elementos, es posible realizar simulaciones de algunos escenarios de muestreo, que indiquen el tamaño de muestra necesario en cada una de las etapas de la selección de la muestra. Si fuese posible sistematizar los elementos más importantes a la hora de calcular el tamaño de muestra en una encuesta de hogares, sería necesario recurrir a los siguientes pasos de manera ordenada:

- **Definir la población de interés de manera explícita.** En particular, es necesario aclarar si la unidad de análisis son las personas o los hogares. De esta forma, se deben fijar los valores para r y b. Si la unidad de análisis son todas las personas del hogar, el porcentaje de personas con la característica de interés será $r=1$; de otra forma, $r<1$. Por otro lado, el número promedio de personas por hogar b dependerá del dominio de representatividad en que se requiera el cálculo.
- **Definir el número promedio de hogares.** El número promedio de hogares que se desea encuestar en cada una de las UPM está dado por $\bar{n}_{II}$. Este proceso debería repetirse de forma iterativa en los pasos subsiguientes para poder evaluar la calidad del diseño. De las varias opciones de $\bar{n}_{II}$, será necesario escoger solo una.
- **Calcular el número promedio de personas que serán encuestadas.** Al igual que en el paso anterior, es necesario probar varios escenarios que redundarán en la elección de un número óptimo de personas por UPM. Los valores de $\bar{n}$ dependen directamente del paso anterior al escoger $\bar{n}_{II}$. Debido a que la selección de las personas está supeditada a la selección de los hogares, $\bar{n}$ se puede descomponer, manteniendo la relación con r y b, de la siguiente manera:

  $$\bar{n} = \bar{n}_{II} \times r \times b$$

- **Calcular el efecto de diseño.** Es necesario definir (o calcular con encuestas o censos anteriores) la correlación intraclase de la variable de interés con el agrupamiento por UPM ρ_y. Después se debe calcular el efecto de diseño $(DEFF)$ como función de ρ_y y de $\bar{n}$; es decir, $DEFF \approx 1+(\bar{n}-1)\,\rho_y$. Nótese que esta cifra solo se calcula sobre la población de interés.

- **Calcular el tamaño de la muestra de personas.** A partir de las expresiones de tamaño de la muestra para diseños de muestreo complejos, se debe calcular el tamaño de muestra necesario para lograr una precisión adecuada en la inferencia. En primer lugar, si lo que se quiere estimar es un promedio $\bar{y}_U$, el tamaño de muestra necesario para alcanzar un margen de error relativo máximo de $MER \times 100\%$ es de:

$$n \geq \frac{S^2_{y_U} DEFF}{\frac{MER^2 \bar{y}^2_U}{z^2_{1-\alpha/2}} + \frac{S^2_{y_U} DEFF}{N}}$$

Por otro lado, si lo que se quiere estimar es una proporción P, y se utiliza el margen de error, la expresión apropiada para calcular el tamaño de muestra estará dada por:

$$n \geq \frac{P\,(1\text{-}P)\,DEFF}{\frac{MER^2 P^2}{z^2_{1-\alpha/2}} + \frac{P(1\text{-}P)\,DEFF}{N}}$$

- **Calcular el tamaño de la muestra de hogares.** Es necesario calcular el número total de hogares que se deben seleccionar para lograr entrevistar a todas las personas que serán observadas en el punto anterior. El número de hogares que se deben seleccionar estará determinado por las cantidades n, b y r, de la siguiente forma:

$$n_{II} = \frac{n}{r \times b}$$

- **Calcular el número de UPM.** Los hogares y las personas se observan a partir de las UPM. En este paso final, es necesario calcular el número de UPM que se deben seleccionar en el muestreo a partir de la relación:

$$n_I = \frac{n}{\bar{n}} = \frac{n_{II}}{\bar{n}_{II}}$$

1. Ejemplo: proporción de personas pobres

Supóngase que el parámetro de interés es el porcentaje de personas pobres (cuyo ingreso está por debajo de un umbral preestablecido) y se quiere hacer una inferencia con un margen de error relativo máximo del 5%. De acuerdo con estudios anteriores en este país hipotético, se ha estimado que la proporción de personas pobres está alrededor de $P = 4\%$. Nótese que la población objetivo está conformada por todos los habitantes del país, puesto que $r = 100\%$. Además, en este país se ha estimado que el tamaño promedio del hogar es de alrededor de $b = 3{,}5$ personas. Por último, según estudios anteriores, la correlación intraclase de la característica de interés con las unidades primarias de muestreo es $\rho_y = 0{,}034$.

En el cuadro VII.1 se resumen los resultados del ejercicio para $\bar{n}_{II} = 10$ hogares por UPM, que implican que por cada UPM se entrevistarían en promedio a $\bar{n} = 10 * 1 * 3{,}5 = 35$ personas. Con ello, se obtendría un efecto de diseño $DEFF = 2{,}2$, para un total de personas

en la muestra de $n=55.936$ que serán observadas a partir de la selección de $n_{II}=55.936/(1*3,5)=15.982$ hogares en $n_I=55.936/35=1.598$ UPM.

■ Cuadro VII.1
Tamaño de la muestra con un submuestreo de diez hogares por unidad primaria de muestreo (UPM) en el ejemplo del apartado VII.D.1

Promedio de hogares por UPM ($\bar{n}_{II}$)	Promedio de personas por UPM ($\bar{n}$)	Efecto de diseño	Tamaño de la muestra de UPM (n_I)	Tamaño de la muestra de hogares (n_{II})	Tamaño de la muestra de personas (n)
10	35	2,2	1 598	15 982	55 936

Fuente: Elaboración propia.

Por supuesto, será posible plantear otros escenarios a medida que se evalúe el efecto que conlleva el cambio del número de hogares seleccionados en cada UPM. Por ejemplo, el investigador podría proponer que se seleccionen en promedio cinco hogares por UPM, lo que cambiaría el número de UPM que serían seleccionadas en la muestra de la primera etapa, así como el número total de personas que serían seleccionadas en todo el operativo. Debido a que el efecto de diseño es una función del número de hogares promedio que han de seleccionarse en las UPM, esta cifra también variará. A continuación se presentan algunos resultados que permiten establecer estos escenarios cuando se varía el tamaño de la muestra promedio de hogares por UPM. La elección del escenario ideal se debe dar en términos de conveniencia logística y presupuestaria en el estudio. Siguiendo las recomendaciones internacionales, se desestimarían los escenarios con efectos de diseño superiores a 3 (véase el cuadro VII.2).

■ Cuadro VII.2
Tabla de muestreo para la estimación de la proporción de personas pobres en el ejemplo del apartado VII.D.1

Promedio de hogares promedio por unidad primaria de muestreo (UPM) ($\bar{n}_{II}$)	Promedio de personas por UPM ($\bar{n}$)	Efecto de diseño	Tamaño de la muestra de UPM (n_I)	Tamaño de la muestra de hogares (n_{II})	Tamaño de la muestra de personas (n)
5	18	1,6	2 315	11 575	40 512
10	35	2,2	1 598	15 982	55 936
15	52	2,8	1 359	20 386	71 351
20	70	3,4	1 239	24 787	86 756
25	88	3,9	1 167	29 186	102 152
30	105	4,5	1 119	33 582	117 538
35	122	5,1	1 085	37 976	132 915
40	140	5,7	1 059	42 366	148 282
45	158	6,3	1 039	46 754	163 640

Fuente: Elaboración propia.

2. Ejemplo: ingreso promedio por persona

Supóngase que se desea estimar el ingreso promedio por hogar con un margen de error relativo máximo del 2%. La variable de interés (ingreso) es continua y se estima que la media se sitúa en alrededor de y $\bar{y}_U = 1.180$ dólares, con una varianza de $S^2_{\bar{y}_U} = 1.845{,}94^2$. En este caso, la población objetivo son todos los habitantes del hogar, por lo que $r = 100\%$. La composición del hogar se calcula en $b = 3{,}79$ personas por hogar. Por último, según estudios anteriores, la correlación intraclase de la característica de interés es $\rho_y = 0{,}035$. Nótese que la correlación intraclase cambia con respecto a la característica que se desee medir.

En el cuadro VII.3 se muestran los resultados del ejercicio al seleccionar $\bar{n}_{II} = 15$ hogares por UPM, que a su vez implica que se encontrarían en promedio $\bar{n} = 15 * 1 * 3{,}79 \cong 57$ personas por UPM. Con ello se obtendría un efecto de diseño $DEFF = 3$ para un total de personas en la muestra de $n = 48.861$, que serán observadas a partir de la selección de $n_{II} = 48.861 / (1 * 3{,}79) = 12.892$ hogares en $n_I = 859$ UPM.

■ Cuadro VII.3
Tamaño de la muestra con un submuestreo de 15 hogares por unidad primaria de muestreo (UPM) en el ejemplo del apartado VII.D.2

Promedio de hogares por UPM ($\bar{n}_{II}$)	Promedio de personas por UPM ($\bar{n}$)	Efecto de diseño	Tamaño de la muestra de UPM (n_I)	Tamaño de la muestra de hogares (n_{II})	Tamaño de la muestra de personas (n)
15	57	3	859	12 892	48 861

Fuente: Elaboración propia.

En el cuadro VII.4 se presentan algunos resultados que permiten establecer otros escenarios de muestreo cuando se varía el tamaño de muestra promedio de hogares por UPM. Cualquiera de estos escenarios es válido desde el punto de vista de la eficiencia estadística, aunque no todos lo serán si se tienen en cuenta otros aspectos, como los logísticos o presupuestarios. Por ejemplo, si se escogiera el penúltimo escenario, para 50 hogares por UPM se debería encuestar en promedio a 190 personas, lo que reduciría el número de UPM a 662, pero aumentaría el tamaño de muestra general a 33.098 personas. Esto supondría mayores costos de contratación de encuestadores y supervisores y, seguramente, un operativo de campo de más días de duración. Siguiendo las recomendaciones internacionales, se desestimarían los escenarios con efectos de diseño superiores a 3.

■ Cuadro VII.4
Tabla de muestreo para la estimación del ingreso promedio por persona en el ejemplo del apartado VII.D.2

Promedio de hogares promedio por unidad primaria de muestreo (UPM) ($\bar{n}_{II}$)	Promedio de personas por UPM ($\bar{n}$)	Efecto de diseño	Tamaño de la muestra de UPM (n_I)	Tamaño de la muestra de hogares (n_{II})	Tamaño de la muestra de personas (n)
5	19	1,6	1 422	7 108	26 938
10	38	2,3	1 000	10 001	37 902
15	57	3,0	859	12 892	48 861
20	76	3,6	789	15 783	59 816
25	95	4,3	747	18 672	70 766
30	114	4,9	719	21 560	81 711
50	190	7,6	662	33 098	125 443
100	379	14,2	619	61 857	234 439

Fuente: Elaboración propia.

3. Ejemplo: tasa de desocupación de las personas mayores

Supóngase que la incidencia de la desocupación es de alrededor de $P=5,5\%$ en la población objetivo, es decir, las personas mayores de 60 años que forman parte de la población económicamente activa (PEA). En este país hipotético, se ha estimado que, en promedio, hay $r=4,6\%$ de personas mayores por hogar que pertenecen a la PEA, cuyo tamaño promedio es de alrededor de $b=5$ personas. Además, se quiere realizar una inferencia con un margen de error relativo máximo del 15%. Por último, según estudios anteriores, la correlación intraclase de la característica de interés es $\rho_y=0,7$.

En el cuadro VII.5 se presentan los resultados del ejercicio, según el cual seleccionar $\bar{n}_{II}=$ 20 hogares por UPM implicaría un promedio de $\bar{n}=20*0,046*5=4,6$ personas mayores en la PEA (personas de interés) por UPM. Con ello se obtendría un efecto de diseño $DEFF=3,5$, para un total de $n=7.272$ personas mayores en la PEA, que serán observadas en la muestra a partir de la selección de $\bar{n}_{II}=7.272/(0,046*5) \cong 31.617$ hogares en $n_I=7.272/4,6 \cong 1.581$ UPM.

■ Cuadro VII.5
Tamaño de la muestra con un submuestreo de 20 hogares por unidad primaria de muestreo (UPM) en el ejemplo del apartado VII.D.3

Promedio de hogares por UPM ($\bar{n}_{II}$)	Promedio de personas por UPM ($\bar{n}$)	Efecto de diseño	Tamaño de la muestra de UPM (n_I)	Tamaño de la muestra de hogares (n_{II})	Tamaño de la muestra de personas (n)
20	4,6	3,5	1 581	31 617	7 272

Fuente: Elaboración propia.

En este caso, la muestra de 31.617 hogares conlleva un operativo muy grande que implicaría la observación de $31.617*5=158.085$ personas en los hogares, de las cuales $n=7.272$ serían los casos de interés. Como se ha visto en los ejemplos anteriores, es posible plantear otros escenarios a medida que se evalúa el efecto del cambio del número de hogares que se seleccionan en cada UPM. En el cuadro VII.6 se presentan algunos resultados que permiten establecer estos escenarios cuando se varía el tamaño de muestra promedio de hogares por UPM. Siguiendo las recomendaciones internacionales, se desestimarían los escenarios con efectos de diseño superiores a 3.

■ Cuadro VII.6
Tabla de muestreo para la estimación de la tasa de desocupación de las personas mayores en el ejemplo del apartado VII.D.3

Promedio de hogares promedio por unidad primaria de muestreo (UPM) ($\bar{n}_{II}$)	Promedio de personas por UPM ($\bar{n}$)	Efecto de diseño	Tamaño de la muestra de UPM (n_I)	Tamaño de la muestra de hogares (n_{II})	Tamaño de la muestra de personas (n)
5	1,1	1,1	1 985	9 926	2 283
10	2,3	1,9	1 716	17 157	3 946
15	3,5	2,7	1 626	24 387	5 609
20	4,6	3,5	1 581	31 617	7 272
25	5,8	4,3	1 554	38 848	8 935
30	6,9	5,1	1 536	46 074	10 597
50	11,5	8,3	1 500	74 983	17 246
100	23,0	16,4	1 472	147 222	33 861

Fuente: Elaboración propia.

E. Tamaño de la muestra para UPM y hogares

En algunas ocasiones, la variable de interés y la unidad de observación están a nivel del hogar. Esto sucede, por ejemplo, cuando todas las variables de interés se miden a nivel de la vivienda o del hogar. En este caso, es posible modificar el algoritmo de la sección anterior para seleccionar únicamente las viviendas u hogares de la muestra, sin necesidad de realizar un submuestreo de personas. Algunas cantidades desaparecen porque no son objeto de la población de hogares, como r y b. Algunas otras expresiones deben ser redefinidas en el contexto de los hogares, como el coeficiente de correlación intraclase $\rho_{y'}$ el efecto de diseño y todas las expresiones de tamaños de la muestra. En todo caso, la adaptación del algoritmo se describe a continuación.

- **Definir el número promedio de hogares.** El número promedio de hogares que se desea encuestar en cada una de las UPM está dado por $\bar{n}_{II}$. Esta cifra continúa siendo el insumo más importante del algoritmo: se propone crear escenarios de muestreo a partir de su modificación y de la evaluación del tamaño de muestra final.

- **Calcular el efecto de diseño**. Es necesario definir (o calcular con encuestas o censos anteriores) la correlación intraclase ρ_y de la variable de interés a nivel del hogar con el agrupamiento por UPM definido por la división cartográfica del último censo. De igual forma, el efecto de diseño $DEFF \approx 1+(\bar{n}_{II} -1)\ \rho_y$ continúa siendo función del tamaño de la muestra promedio de hogares por UPM $(\bar{n}_{II})$.
- **Tamaño de la muestra de hogares.** Partiendo de las expresiones de tamaño de la muestra generales para muestreos complejos y teniendo en cuenta que la población de interés son los hogares y que la variable de interés está a nivel del hogar, el tamaño de muestra necesario para alcanzar un margen de error relativo máximo de $MER\ \%$ es de:

$$n_{II} \geq \frac{S_y^2 DEFF}{\frac{MER^2 \bar{y}^2}{z^2_{1-\alpha/2}} + \frac{S^2_{y_U} DEFF}{N_{II}}}$$

La expresión apropiada para calcular el tamaño de muestra para una proporción estará dada por:

$$n_{II} \geq \frac{P\ (1-P)\ DEFF}{\frac{MER^2 P^2}{z^2_{1-\alpha/2}} + \frac{P(1-P)\ DEFF}{N_{II}}}$$

- **Cálculo del número de UPM.** Los hogares se observan a partir de las UPM. En este paso final, es necesario calcular el número de UPM que se deben seleccionar en el muestreo a partir de la relación:

$$n_I = \frac{n_{II}}{\bar{n}_{II}}$$

1. Ejemplo: gasto promedio del hogar

Supóngase que se desea estimar el promedio de gasto anual en dólares en los hogares del país con un margen de error relativo máximo admisible del 3,5%. La variable de interés (gasto) es continua y se estima que la media se sitúa en alrededor de $\bar{y}_U = 1.407$ dólares, con una varianza de $S^2_{y_U} = 2.228^2$. En este ejemplo, se supone que el país está dividido en $N_I = 10.000$ UPM y la correlación intraclase de la característica de interés, medida a nivel del hogar, con las UPM, es de $\rho_y = 0,173$.

En el cuadro VII.7 se presentan los resultados del ejercicio para $\bar{n}_{II} = 12$ hogares promedio por UPM, que serán observados a partir de la selección de $n_{II} = 16.056$ hogares y $n_I = 16.056/12 = 1.338$ UPM, lo que conlleva un efecto de diseño $DEFF = 2,9$.

■ Cuadro VII.7
Tamaño de la muestra con un submuestreo de 12 hogares por unidad primaria de muestreo (UPM) en el ejemplo del apartado VII.E.1

Promedio de hogares por UPM ($\bar{n}_{II}$)	Efecto de diseño	Tamaño de la muestra de UPM (n_I)	Tamaño de la muestra de hogares (n_{II})
12	2,9	1 338	16 056

Fuente: Elaboración propia.

En el cuadro VII.8 se presentan algunos resultados que permiten establecer otros escenarios de muestreo al variar el tamaño de la muestra promedio de hogares por UPM. Nótese que, por ejemplo, en caso de seleccionar 20 hogares por UPM, se debería seleccionar una muestra de 23.695 hogares en tan solo 1.185 UPM. Por otro lado, si solo se seleccionaran 2 hogares por UPM, se tendrían que visitar 3.246 UPM en todo el país, aunque el número de encuestas totales descendería a 6.493. Para este tipo de operativos, donde los cuestionarios de gasto de los hogares van acompañados de un operativo exhaustivo que permite al investigador conocer los hábitos de consumo del hogar de forma desagregada, y en que se visita el hogar durante un período de tiempo determinado, podría resultar más conveniente estudiar la viabilidad de seleccionar más hogares por UPM y menos UPM para que el operativo de campo no exija la contratación de demasiado personal sobre el terreno. Al estar agrupados en menos UPM, se podría realizar una mejor recopilación de la información con un equipo mediano de personas. De lo contrario, se debería contratar bastante más personal que cubra las UPM dispersas a lo largo del país. Siguiendo las recomendaciones internacionales, se desestimarían los escenarios con efectos de diseño superiores a 3.

■ Cuadro VII.8
Tabla de muestreo para la estimación del gasto promedio del hogar en el ejemplo del apartado VII.E.1

Promedio de hogares promedio por unidad primaria de muestreo (UPM) ($\bar{n}_{II}$)	Efecto de diseño	Tamaño de la muestra de UPM (n_I)	Tamaño de la muestra de hogares (n_{II})
2	1,2	3 246	6 493
4	1,5	2 102	8 407
6	1,9	1 720	10 320
8	2,2	1 529	12 233
10	2,6	1 414	14 145
12	2,9	1 338	16 056
14	3,2	1 283	17 967
16	3,6	1 242	19 877
18	3,9	1 210	21 787
20	4,3	1 185	23 695

Fuente: Elaboración propia.

2. Ejemplo: proporción de hogares sin agua potable

Supóngase que se desea obtener una muestra con un margen de error relativo máximo admisible del 10% sobre la variable de interés (necesidades básicas insatisfechas de agua) y que el parámetro de interés es el porcentaje de hogares con esta carencia. En este país, se estima que la proporción de hogares con dicha condición asciende a $P=7,5\%$. En este ejemplo, se supone que la correlación intraclase de la característica de interés con las UPM es de $\rho_y=0,045$.

En el cuadro VII.9 se muestran los resultados del ejercicio para $\bar{n}_{II} = 10$ hogares por UPM, que serán observados a partir de la selección de $n_{II} =4.360$ hogares en $n_I = 4.360/10=436$ UPM, de lo que se obtiene un efecto de diseño $DEFF=1,3$.

■ Cuadro VII.9
Tamaño de la muestra con un submuestreo de diez hogares por unidad primaria de muestreo (UPM) en el ejemplo VII.E.2

Promedio de hogares por UPM ($\bar{n}_{II}$)	Efecto de diseño	Tamaño de la muestra de UPM (n_I)	Tamaño de la muestra de hogares (n_{II})
10	1,3	436	4 360

Fuente: Elaboración propia.

En el cuadro VII.10 se presentan algunos resultados que permiten establecer otros escenarios de muestreo al variar el tamaño de la muestra promedio de hogares por UPM. Obsérvese que el efecto de diseño es directamente proporcional al número de hogares entrevistados por UPM y al tamaño final de la muestra de hogares. De la misma manera, es inversamente proporcional al número de UPM.

■ Cuadro VII.10
Tabla de muestreo para la estimación de la proporción de hogares sin agua potable en el ejemplo del apartado VII.E.2

Promedio de hogares promedio por unidad primaria de muestreo (UPM) ($\bar{n}_{II}$)	Efecto de diseño	Tamaño de la muestra de UPM (n_I)	Tamaño de la muestra de hogares (n_{II})
5	1,1	758	3 790
10	1,3	436	4 360
15	1,5	328	4 924
20	1,6	274	5 490
25	1,8	242	6 057
30	2,0	221	6 624
35	2,2	205	7 190
40	2,3	194	7 757
45	2,5	185	8 323

Fuente: Elaboración propia.

F. Tamaño de la muestra para UPM y personas

En algunos casos en los que la variable de interés esté a nivel de persona, el cuestionario de la encuesta no incluya preguntas acerca del hogar y, además, exista un inventario detallado de las personas que residen en cada UPM, será posible evadir la selección de los hogares e ir directamente a la selección de las personas. En este caso, la lógica de cálculo del tamaño de la muestra se mantiene al modificar en cierta manera el algoritmo de las secciones anteriores, como se ilustra a continuación.

- **Definir la población de interés de manera explícita.** En este caso, solo se mantiene la expresión correspondiente a r, que denota el porcentaje de personas con la característica de interés en la población.
- **Definir el número promedio de personas.** El número promedio de personas (casos de la población objetivo) que se desea encuestar por cada UPM está dado por $\bar{n}$. Al igual que en las secciones anteriores, se recomienda hacer una evaluación sobre esta cantidad para determinar posibles escenarios de muestreo.
- **Calcular el efecto de diseño.** Es necesario definir el efecto de diseño $(DEFF)$ como una función de la correlación existente entre la variable de interés y la conformación de las UPM. De esta forma, $DEFF \approx 1+(\bar{n}-1)\,\rho_y$. Nótese que esta cifra solo podrá ser calculada sobre la población de interés.
- **Tamaño de la muestra de personas.** A partir de las expresiones de tamaño de la muestra para muestreos complejos, se debe calcular el tamaño de muestra necesario para lograr una precisión adecuada en la inferencia. En este caso, las expresiones de tamaño de la muestra coinciden con las del primer escenario.
- **Tamaño de muestra final.** Es necesario calcular el número total de personas que deben ser seleccionadas para lograr observar a quienes forman parte de la población objetivo. Esta cantidad está dada por n/r.
- **Cálculo del número de UPM.** Por último, las personas se observan a partir de las UPM. En este paso final, es necesario calcular el número de UPM que se deben seleccionar en el muestreo a partir de la relación:

$$n_I = \frac{n}{\bar{n}}$$

1. Ejemplo: ingreso promedio de las personas empleadas

Supóngase que se desea estimar el ingreso promedio de las personas empleadas con un margen de error relativo máximo admisible del 2%. La variable de interés (ingreso) es continua y se estima que la media se sitúa en alrededor de $\bar{y}_U = 1.458$ dólares, con una varianza de $S^2_{y_U} = 2.191$. La población objetivo son todas las personas empleadas, cuya proporción se estima en $r = 46\%$. La correlación intraclase de la característica de interés es $\rho_y = 0{,}038$.

En el cuadro VII.11 se presentan los resultados del ejercicio al seleccionar $\bar{n}=23$ personas de la población de interés por UPM, lo que, a su vez, implica que se deberían seleccionar y enlistar en promedio $23/0,46=50$ personas por UPM. Con ello se esperarían $n=28.029$ personas empleadas en la muestra, repartidas en $n_I=28.029/23 \cong 1.219$ UPM. En ese caso, se obtendría un efecto de diseño $DEFF=1,8$. En este escenario, el operativo de campo abarcaría la selección y enlistamiento de $28.029/0,46 \cong 60.933$ personas, de las que se esperaría que 28.029 pertenecieran a la población de interés (personas empleadas).

■ Cuadro VII.11
Tamaño de la muestra con un submuestreo de 50 personas por unidad primaria de muestreo (UPM) en el ejemplo del apartado VII.F.1

Personas seleccionadas por UPM ($\bar{n}/r$)	Personas empleadas por UPM ($\bar{n}$)	Efecto de diseño	Tamaño de la muestra de UPM (n_I)	Tamaño de la muestra de personas empleadas (n)	Tamaño de la muestra de personas (n/r)
50	23	1,8	1 219	28 029	60 933

Fuente: Elaboración propia.

En el cuadro VII.12 se presentan algunos resultados que permiten establecer otros escenarios de muestreo cuando se varía el tamaño de la muestra promedio de hogares por UPM. Siguiendo las recomendaciones internacionales, se desestimarían los escenarios con efectos de diseño superiores a 3.

■ Cuadro VII.12
Tabla de muestreo para la estimación del ingreso promedio de las personas empleadas en el ejemplo del apartado VII.F.1

Personas seleccionadas por unidad primaria de muestreo (UPM) ($\bar{n}/r$)	Personas empleadas por UPM ($\bar{n}$)	Efecto de diseño	Tamaño de la muestra de UPM (n_I)	Tamaño de la muestra de personas empleadas (n)	Tamaño de la muestra de personas (n/r)
25	12	1,4	1 857	21 360	46 435
50	23	1,8	1 219	28 029	60 933
75	34	2,3	1 006	34 695	75 424
100	46	2,7	899	41 360	89 913
125	58	3,1	835	48 023	104 398

Fuente: Elaboración propia.

2. Ejemplo: proporción de personas analfabetas pobres

Supóngase que se quiere estimar la proporción de incidencia de la pobreza en la población analfabeta, con un margen de error relativo máximo admisible del 15%. Se ha estimado que alrededor del $r=14\%$ de las personas del país no saben leer ni escribir. Por otro lado,

como se vio en un ejemplo anterior, el fenómeno de la pobreza está estimado en $P=4\%$, y la correlación intraclase de la característica de interés es $\rho_y = 0{,}0454$.

En el cuadro VII.13 se presentan los resultados del ejercicio al seleccionar un promedio de $\bar{n}=14$ personas por UPM que no saben leer ni escribir. Esto implica la selección y enlistamiento de $14/0{,}14=100$ personas por UPM. Con ello se obtendría un efecto de diseño $DEFF=1{,}6$, para un total de $n=4.574$ personas analfabetas, de una muestra de 32.671 personas enlistadas y repartidas en $n_l=4.574/14 \cong 327$ UPM.

■ Cuadro VII.13
Tamaño de la muestra con un submuestreo de 100 hogares por unidad primaria de muestreo (UPM) en el ejemplo del apartado VII.F.2

Personas seleccionadas por UPM ($\bar{n}/r$)	Personas empleadas por UPM ($\bar{n}$)	Efecto de diseño	Tamaño de la muestra de UPM (n_l)	Tamaño de la muestra de personas empleadas (n)	Tamaño de la muestra de personas (n/r)
100	14	1,6	327	4 574	32 671

Fuente: Elaboración propia.

Es posible plantear otros escenarios a medida que se evalúa el efecto que conlleva el cambio del número de hogares que se seleccionan en cada UPM. En el cuadro VII.14 se presentan algunos resultados que permiten establecer estos escenarios cuando se varía el tamaño de la muestra promedio de hogares por UPM.

■ Cuadro VII.14
Tabla de muestreo para la estimación de la proporción de personas analfabetas pobres en el ejemplo del apartado VII.F.2

Personas seleccionadas por unidad primaria de muestreo (UPM) ($\bar{n}/r$)	Personas analfabetas por UPM ($\bar{n}$)	Efecto de diseño	Tamaño de la muestra de UPM (n_l)	Tamaño de la muestra de personas analfabetas (n)	Tamaño de la muestra de personas (n/r)
25	3,5	1,1	917	3 211	22 936
50	7,0	1,3	524	3 665	26 179
75	10,5	1,4	392	4 120	29 429
100	14,0	1,6	327	4 574	32 671
125	17,5	1,7	287	5 029	35 921

Fuente: Elaboración propia.

G. Tamaño de la muestra para otros parámetros de interés

En las encuestas de hogares también surgen escenarios particulares que permiten sugerir distintos caminos para la adopción de un determinado tamaño de muestra. En esta sección se analizarán los casos en que los parámetros de interés son diferencias de proporciones y dobles diferencias. También se revisará el caso del planteamiento de pruebas de hipótesis y su relación con el tamaño de la muestra.

1. Tamaño de la muestra para la estimación de la diferencia de dos proporciones

Supóngase una población U, que se encuentra dividida en dos subpoblaciones (U_1, de tamaño N_1, y U_2, de tamaño N_2)[1]. El interés del investigador está en conocer la diferencia de algunas proporciones entre estos grupos. Por ejemplo, supóngase que se quiere conocer la diferencia entre las proporciones de niños desnutridos por sexo. Se espera que la proporción de niños desnutridos no supere la proporción de niñas desnutridas para verificar que no existen brechas de género en este aspecto. Por lo tanto, el parámetro de interés se escribe como:

$$\theta = P_1 - P_2 = \frac{N_{d1}}{N_1} - \frac{N_{d2}}{N_2}$$

Donde $N_{di} = \sum_{k \in U_i)} z_{dik}$ $(i=1,2)$ y z_{dik} es una característica dicotómica que indica si el individuo k-ésimo de la subpoblación U_i está en estado de desnutrición. Por supuesto, con un muestreo aleatorio simple, un estimador insesgado para θ es:

$$\hat{\theta} = \hat{P}_1 - \hat{P}_2 = \frac{\hat{N}_{d1}}{N_1} - \frac{\hat{N}_{d2}}{N_2}$$

Donde $\hat{N}_{di} = \frac{N_i}{n_i} \sum_{k \in s_i)} z_{dik}$ y s_i es la muestra asociada a la población U_i. Luego, la varianza del anterior estimador es:

$$Var(\hat{\theta}) = Var(\hat{P}_1) + Var(\hat{P}_2) - 2Cov(\hat{P}_1, \hat{P}_2)$$

Por otro lado, siendo $|U_i|$ la cardinalidad del conjunto U_i, se definen las siguientes relaciones:

$$T_i = \frac{|U_1 \cap U_2|}{|U_i|} \quad i=1,2$$

De esta forma, T_1 y T_2 corresponden al porcentaje de traslape de las subpoblaciones. De la misma manera, definiendo $R_{1,2}$ como la correlación de Pearson entre los datos

[1] Esta metodología también se aplica en el caso en que $U \supset (U_1 \cup U_2)$.

observados de ambas subpoblaciones, la covarianza entre este par de estimadores estaría determinada por la siguiente relación (Kish, 2004):

$$Cov(\hat{P}_1,\hat{P}_2) = \sqrt{Var(\hat{P}_1)}\sqrt{Var(\hat{P}_2)}\sqrt{T_1}\sqrt{T_2}R_{1,2}$$

En esta instancia, es útil recordar que, si las poblaciones U_1 y U_2 son estratos (o agregaciones de estratos) que inducen conjuntos disjuntos y la selección de la muestra en cada uno es independiente por diseño, entonces $Cov(\hat{P}_1,\hat{P}_2)=0$. Si, por otro lado, no existe independencia en el muestreo de ambas poblaciones, entonces, necesariamente, $R_{1,2}\neq 0$. Es útil recordar que esta correlación se debe evaluar a través de las UPM. Siguiendo con el ejemplo, a pesar de que las subpoblaciones son de niños y niñas, $R_{1,2}\neq 0$. Por otro lado, para encontrar el tamaño de muestra óptimo, es útil tener en cuenta los siguientes supuestos:

i) Asumir que las subpoblaciones son grandes y por ende $N_1=N_2=N$.

ii) Dado el punto i), asumir que los tamaños de muestra pueden ser iguales, tales que $n_1=n_2=n$.

Nótese a su vez que, si las observaciones no pueden realizarse utilizando un muestreo aleatorio simple, sino que, por el contrario, la muestra aleatoria fue seleccionada mediante un diseño de muestreo complejo con un efecto de diseño *(DEFF)* no ignorable y mayor que uno, la varianza tomaría la siguiente forma[2]:

$$Var(\hat{\theta}) = \frac{DEFF}{n}\left(1-\frac{n}{N}\right)S_\theta^2$$

Donde, definiendo que $Q_i=1-P_i$, se tiene que:

$$S_\theta^2 = P_1Q_1+P_2Q_2-2\sqrt{T_1}\sqrt{T_2}R_{1,2}\sqrt{P_1Q_1}\sqrt{P_2Q_2}$$

De esta manera, un intervalo de confianza del 95% para la diferencia de proporciones está dado por:

$$IC(95\%)_\theta=\hat{\theta}\pm z_{1-\alpha/2}\sqrt{\frac{DEFF}{n}\left(1-\frac{n}{N}\right)S_\theta^2}$$

Ello quiere decir que el margen de error *(ME)* de la encuesta debe ser tal que:

$$ME< z_{1-\alpha/2}\sqrt{\frac{DEFF}{n}\left(1-\frac{n}{N}\right)S_\theta^2}$$

Por lo tanto, despejando n, se tiene que la muestra en cada subgrupo debe ser mayor que:

$$n \geq \frac{DEFF\ S_\theta^2}{\frac{ME^2}{z_{1-\alpha/2}^2}+\frac{DEFF\ S_\theta^2}{N}}$$

[2] Si el muestreo es aleatorio simple, el efecto de diseño es $DEFF=1$.

Dependiendo de los porcentajes de traslape $\sqrt{T_1}$, $\sqrt{T_2}$ y de la correlación de la característica de interés en ambas subpoblaciones $R_{1,2}$, la varianza S_θ^2 tomará diferentes formas, como se detalla a continuación:

i) Si no hay traslape, $T_1 = T_2 = 0$, y $S_\theta^2 = P_1 Q_1 + P_2 Q_2$.

ii) Si hay traslape completo, $T_1 = T_2 = 1$ y $S_\theta^2 = P_1 Q_1 + P_2 Q_2 - 2R_{1,2} \sqrt{P_1 Q_1} \sqrt{P_2 Q_2}$.

iii) Si hay traslape parcial y balanceo, $T_1 = T_2 = T$. Si además se considera que las varianzas en cada subgrupo o período son similares, $P_1 Q_1 = P_2 Q_2 = PQ$, entonces $S_\theta^2 = 2PQ\,(1 - TR_{1,2})$.

a) Covarianza en comparaciones mensuales

Supóngase que se quiere comparar la tasa de desempleo nacional entre dos meses consecutivos. En este escenario, suponiendo que existe independencia en el muestreo de los dos meses consecutivos, el porcentaje de traslape de la muestra entre los dos meses (que por diseño es nulo) es igual a cero. Por lo tanto, $T_1 = T_2 = 0$. Luego, el término de la covarianza se anula. En resumen, la varianza del estimador en este caso sería igual a:

$$Var(\hat{P}_1 - \hat{P}_2) = Var(\hat{P}_1) + Var(\hat{P}_2)$$

b) Covarianza en comparaciones trimestrales o anuales

Partiendo de un diseño rotativo 2(2)2, supóngase que se quiere comparar la tasa de desempleo nacional entre trimestres consecutivos o entre el mismo mes de dos años consecutivos. En este escenario no existe independencia en el muestreo de los dos trimestres consecutivos, puesto que la estructura del panel garantiza un traslape del 50%. En este caso, $T_1 = T_2 \approx 0.5$.

Por otro lado, existe una correlación natural entre las viviendas comunes del panel que se midieron en los períodos de interés, por lo tanto, $R_{1,2} \neq 0$. Esta correlación se calcula sobre los individuos comunes en el panel y sobre la variable dicotómica que induce la tasa de desempleo (perteneciente a la población económicamente activa). En resumen, el término de covarianza en este caso sería igual a:

$$Cov(\hat{P}_1, \hat{P}_2) = \frac{1}{2} \sqrt{Var(\hat{P}_1)} \sqrt{Var(\hat{P}_2)}\, R_{1,2}$$

c) Covarianza en comparaciones de un mismo mes

En primer lugar, supóngase que se quiere comparar la tasa de desempleo entre hombres y mujeres en un mismo mes. En este escenario no existe independencia en el muestreo de hombres y mujeres, puesto que estos grupos no son estratos de muestreo. En este caso, T_1 es la proporción de hombres y T_2 es la proporción de mujeres. Nótese que $T_1 \neq T_2$.

Como se comentó anteriormente, existe una correlación natural entre las UPM que fueron seleccionadas y que contienen tanto a hombres como a mujeres; por lo tanto, $R_{12} \neq 0$. Esta correlación se calcula sobre todos los individuos pertenecientes a la fuerza de trabajo y sobre la variable dicotómica que induce la tasa de desempleo. En resumen, el término de covarianza en este caso sería igual a:

$$Cov(\hat{P}_1, \hat{P}_2) = \sqrt{Var(\hat{P}_1)}\sqrt{Var(\hat{P}_2)}\sqrt{T_1}\sqrt{T_2}R_{1,2}$$

Por otro lado, supóngase que se quiere comparar la tasa de desempleo entre dos regiones del mismo país en un mismo mes. En este escenario existe independencia en el muestreo de las dos regiones porque la selección es independiente en cada región. Esta independencia se obtiene por definición del diseño de muestreo, puesto que ambas regiones son agrupaciones disjuntas entre estratos de muestreo. En este caso, T_1 es la proporción de personas de la primera ciudad y T_2 es la proporción de personas de la segunda ciudad. Además, tampoco existe una correlación entre las UPM que fueron seleccionadas entre estas regiones, porque la selección fue independiente, por lo tanto, $R_{12} = 0$. En resumen, el término de covarianza es nulo y, por ende, la varianza del estimador sería igual a:

$$Var(\hat{d}) = Var(\hat{P}_1) + Var(\hat{P}_2)$$

2. Tamaño de la muestra para la estimación del impacto en dos mediciones longitudinales

En el caso de las encuestas que planean un seguimiento de panel o de panel rotativo, es posible contemplar escenarios en los que se quiera estimar el efecto de una intervención, definido como la diferencia en diferencias de las proporciones de interés. De esta forma, el efecto se define como:

$$\theta = (P_{1,1} - P_{2,1}) - (P_{1,2} - P_{2,2})$$

Donde $P_{i,j}$ $(i,j=1,2)$ corresponde a las proporciones del grupo i en la medición j. Entonces, el tamaño de muestra mínimo necesario para lograr una estimación confiable de esta diferencia, con menos del $ME \times 100\%$ de margen de error, es el siguiente[3]:

$$n \geq \frac{DEFF\ S_\theta^2}{\frac{ME^2}{z_{1-\alpha/2}^2} + \frac{DEFF\ S_\theta^2}{N}}$$

[3] Nótese que el tamaño de la muestra de toda la encuesta es $4n$, en las dos rondas, puesto que se deben seleccionar n elementos en cada grupo y en cada ronda.

Donde:

$$S_{\theta}^{2}= (P_{1,1}\, Q_{1,1}+ P_{1,2}\, Q_{1,2} +P_{2,1}\, Q_{2,1} +P_{2,2}\, Q_{2,2})(1- TR)$$

Donde T corresponde a la tasa de traslape ($T=1$ corresponde a un panel completo, $T=0,5$, a un semipanel con traslape del 50% y el caso extremo $T=0$, a una encuesta en que no hay traslape). Por su parte, R se define como la correlación entre las dos mediciones o rondas ($R=0$ implica que no hay correlación entre los dos momentos, $R=-1$ implica una máxima correlación negativa entre los dos momentos y $R=1$ implica una correlación positiva máxima entre los dos momentos).

Por ejemplo, en una encuesta de fuerza laboral intermediada por alguna intervención gubernamental, puede ser de interés evaluar el efecto de esa política de asistencia laboral entre hombres y mujeres en dos períodos de tiempo.

3. Tamaño de la muestra para el contraste de hipótesis en la diferencia de proporciones

Supóngase que el investigador desea realizar el contraste de una hipótesis de interés. En particular, podría suponerse que hay dos grupos de interés en la población finita y que la hipótesis surge de la diferencia de las proporciones en las dos poblaciones. El investigador considera que la diferencia es significativa para el fenómeno en cuestión si es mayor que un valor D definido de antemano y conocido como el tamaño del efecto que el investigador desea detectar.

Nótese que la significación estadística, definida por un valor p, no siempre tiene la misma connotación de significación científica o económica que puede presentarse en fenómenos raros, para los que no necesariamente se gozaría de significación estadística. Por lo tanto, el sistema de hipótesis que se quiere contrastar es el siguiente:

$$H_o: P_1\text{-}P_2=0 \text{ vs. } H_a: P_1\text{-}P_2=D>0$$

Acudiendo a la distribución normal de los estimadores de las proporciones, y suponiendo independencia en el muestreo de los subgrupos, la regla de decisión en este caso lleva a rechazar la hipótesis nula cuando:

$$\frac{\hat{P}_1 - \hat{P}_2}{\sqrt{Var(\hat{P}_1 - \hat{P}_2)}} > z_{1-\alpha}$$

Si las características del estudio implican que el diseño de muestreo es complejo con un $DEFF>1$, esta regla de decisión lleva a rechazar la hipótesis nula si:

$$\frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\frac{DEFF}{n}\left(1 - \frac{n}{N}\right)(P_1\, Q_1) + (P_2\, Q_2)}} > z_{1-\alpha}$$

En este caso, es necesario controlar la probabilidad de cometer el error de tipo 2 (aceptar una hipótesis nula, dado que esta es falsa). Esta probabilidad se conoce como potencia y, suponiendo que el interés está en detectar un tamaño del efecto $P_1 - P_2 = D$, la potencia está dada por:

$$\beta \leq Pr\left(\frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\frac{DEFF}{n}\left(1 - \frac{n}{N}\right)(P_1 Q_1 + P_2 Q_2)}} > z_{1-\alpha} \mid P_1 - P_2 = D\right)$$

$$= Pr\left(\frac{(\hat{P}_1 - \hat{P}_2) - D}{\sqrt{\frac{DEFF}{n}\left(1 - \frac{n}{N}\right)(P_1 Q_1 + P_2 Q_2)}} > z_{1-\alpha} - \frac{D}{\sqrt{\frac{DEFF}{n}\left(1 - \frac{n}{N}\right)(P_1 Q_1 + P_2 Q_2)}} \mid P_1 - P_2 = D\right)$$

$$= 1 - \Phi\left(z_{1-\alpha} - \frac{D}{\sqrt{\frac{DEFF}{n}\left(1 - \frac{n}{N}\right)(P_1 Q_1 + P_2 Q_2)}}\right)$$

Por ello:

$$1 - \beta \geq \Phi\left(z_{1-\alpha} - \frac{D}{\sqrt{\frac{DEFF}{n}\left(1 - \frac{n}{N}\right)(P_1 Q_1 + P_2 Q_2)}}\right)$$

Entonces, dado que la función $\Phi(\ )$ es creciente, se tiene que:

$$z_{1-\beta} \geq z_{1-\alpha} - \frac{D}{\sqrt{\frac{DEFF}{n}\left(1 - \frac{n}{N}\right)(P_1 Q_1 + P_2 Q_2)}}$$

En consecuencia, al despejar n, se tiene que la muestra en cada subgrupo debe ser mayor que:

$$n \geq \frac{DEFF\,(P_1 Q_1 + P_2 Q_2)}{\frac{D^2}{(z_{1-\alpha} + z_\beta)^2} + \frac{DEFF\,(P_1 Q_1 + P_2 Q_2)}{N}}$$

H. Algunas relaciones de interés para proporciones

Cuando se trata de estadísticas de la fuerza laboral, una variable clave para el diseño de una encuesta de hogares que mida la dinámica del mercado de trabajo es el estado de los individuos en la fuerza laboral. Para los gobiernos, es de interés proporcionar un conjunto de indicadores destinados a medir y rastrear la ocupación de los ciudadanos del país (o la región). Por ejemplo, se pueden obtener estimaciones de la tasa de desempleo actual (medida de manera mensual o trimestral). También resultan relevantes la variación neta entre dos períodos y los flujos brutos entre categorías de empleo de un período a otro.

Existen tres modalidades de planificación de las encuestas de hogares que permiten abordar adecuadamente las características particulares de los estudios de la fuerza laboral. La primera se basa en las encuestas repetidas, en las que se realizan mediciones similares en distintos momentos a diferentes personas cada vez. La segunda son las encuestas de panel, en las que se realizan mediciones en distintos momentos a las mismas personas cada vez. La tercera son las encuestas rotativas, en las que se incluyen elementos que se siguen en la muestra durante un período específico y, a medida que estos salen de la muestra, se agregan nuevos elementos.

Según una regla general común para calcular el tamaño de la muestra, como la variable de diseño es dicotómica (dependiendo del estado de empleo), la varianza de ese tipo de variables encuentra su máximo cuando la probabilidad de éxito es 0,5. Sin embargo, si las políticas públicas en un país se centran en lograr que la tasa de desempleo sea baja mediante determinadas intervenciones gubernamentales que afectan (positivamente) a la fuerza laboral, y si esas estrategias son efectivas, la probabilidad de éxito de la variable de diseño cambia y puede afectar el tamaño de la muestra de las encuestas de hogares.

En esta sección, se centra la atención en el tamaño de la muestra que surge del control del margen de error. A medida que la proporción disminuye, el tamaño de la muestra aumenta sustancialmente. Sin embargo, al controlar el margen de error, debido a que la función de varianza en que se basa este enfoque es simétrica (alrededor de 0,5), se puede encontrar que el tamaño de muestra necesario para cumplir con los requisitos de calidad para cualquier proporción (P_d) es el mismo que el requerido para satisfacer los requisitos de calidad para su complemento aditivo $(1- P_d)$.

A continuación se proporcionan varios ejemplos de escenarios que se pueden encontrar en la práctica. Los cálculos se pueden reproducir empleando el *software* estadístico R (R Core Team, 2020a), mediante el uso de la biblioteca *samplesize4surveys* (Rojas, 2020), utilizando específicamente las funciones *ss4p* y *ss4dp*.

1. Estimación de proporciones

- Primer escenario: si la tasa de desempleo es baja, por ejemplo, $P=0,05$ y el margen de error se fija en $ME=0,0025$, el intervalo de confianza esperado sería $IC = 0,05 \pm 0,0025 = (0,0475,0,0525)$. En este caso, el tamaño de muestra requerido es de alrededor de 55.169.
- Segundo escenario: si la tasa de desempleo es alta, por ejemplo, $P=0,2$, y el margen de error de error se fija en $ME=0,01$, el intervalo de confianza sería $IC = 0,2 \pm 0,01 = (0,19,0,21)$, y el tamaño de muestra requerido es 12.144.

Nótese que ambos escenarios dan lugar al mismo margen de error relativo (MER), definido como $MER=\frac{ME}{P}$. En efecto, para el primero, se tiene $MER=(0,0025/0,05)\times 100\%=5\%$, y para el segundo, se tiene $MER=(0,01/0,2)\times 100\%=5\%$. Por lo tanto, incluso para el mismo margen de error relativo, el tamaño de la muestra debe ser mayor si el fenómeno de interés tiene una baja incidencia en la población finita. De hecho, es posible definir una función de información para saber si el tamaño de su muestra es suficiente para cumplir los requisitos de calidad respecto de una proporción determinada. Esto es útil porque no se sabe exactamente qué valor tomará la proporción. Además, si la encuesta de hogares intenta estimar otras proporciones (como en una encuesta multipropósito), se determinará rápidamente si su tamaño de muestra actual es adecuado para todo el estudio.

- Tercer escenario: si el tamaño de la muestra se define como $n=10.000$, y la proporción es $P=0,2$, el coeficiente de variación será del 2,8% y el margen de error será del 1,1%. Todas las proporciones estimadas tendrán un margen de error inferior al 1,4%.
- Cuarto escenario: si el tamaño de la muestra se define como $n=40.000$ y la proporción se centra en $P=0,05$, el coeficiente de variación será del 3% y el margen de error será del 0,2%. Todas las proporciones estimadas tendrán un margen de error inferior al 0,7%.

Para una proporción P dada, el tamaño de muestra requerido para lograr un margen de error particular es el mismo que para su complemento aditivo $1\text{-}P$. Por lo tanto, como es de esperar, si un tamaño de muestra alcanza los requisitos para una proporción establecida, también alcanzará los requisitos de calidad para cualquier proporción superior.

Sin embargo, para una proporción P, el tamaño muestral requerido para lograr un coeficiente de variación particular no es el mismo que para su complemento aditivo $1\text{-}P$. Luego, para una proporción baja, se puede encontrar que, con un tamaño de muestra dado, el coeficiente de variación será mayor que para su complemento aditivo. Sobre la base de los resultados encontrados en esta sección, y con un MER fijo (5% en todos los casos), se encuentra lo siguiente al intentar estimar proporciones (como la tasa de desempleo):

- Si la proporción es baja, se anticipa un gran tamaño de muestra.
- Si la proporción es alta, se espera un tamaño de muestra pequeño.

2. Estimación de cambios netos

En este apartado se dirige la atención a los cambios netos en la tasa de desempleo durante dos períodos, $\Delta=|P_1 - P_2|$. Este tipo de parámetro se puede estimar utilizando una encuesta repetida, rotativa o de panel. Sin embargo, como se explicó en las anteriores secciones, se produce una reducción en el tamaño de la muestra si se intentan estimar los cambios netos desde una encuesta rotativa o de panel. Como no se está estimando una proporción, sino un cambio neto, hay que considerar qué valores son adecuados para establecer el margen de error absoluto.

- Quinto escenario: si no se esperan cambios significativos entre ambos períodos, las tasas de desempleo son altas (por ejemplo, $\Delta\approx|0,22-0,20|=0,02)$ y el margen de error se fija en $ME=0,001$, el intervalo de confianza sería $IC=0,02\pm0,001=(0,019,0,021)$ y el tamaño de muestra requerido sería de alrededor de 96.224.
- Sexto escenario: si no se esperan cambios significativos entre períodos, las tasas de desempleo son bajas (por ejemplo, $\Delta\approx|0,05-0,03|=0,02)$, y el margen de error se fija en $ME=0,001$, el intervalo de confianza sería $IC=0,02\pm0,001=(0,019,0,021)$ y el tamaño de muestra requerido debería ser de 59.536.
- Séptimo escenario: si se esperan cambios significativos entre períodos, las tasas de desempleo difieren (por ejemplo, $\Delta\approx|0,05-0,20|=0,15)$ y el margen de error se fija en $ME=0,0075$, el intervalo de confianza sería $IC=0,15\pm0,0075=(0,1425,0,1575)$ y el tamaño de muestra requerido sería de alrededor de 22.083.

Nótese que los escenarios quinto, sexto y séptimo dan como resultado el mismo MER, definido como $MER=\frac{ME}{\Delta}$. En efecto, para el séptimo escenario se tiene $MER=(0,0075/0,15)\%=5\%$, y, para el quinto y sexto escenarios, se tiene $MER=(0,001/0,02)\%=5\%$. Por lo tanto, incluso para el mismo valor del cambio neto, el tamaño de la muestra no será el mismo y variará dependiendo de la configuración de las proporciones. Por supuesto, hay que esperar cambios más drásticos si varía la porción del traslape y la correlación entre períodos.

Además, se debe tener en cuenta que es posible encontrar diferentes configuraciones de proporciones en ambos períodos que produzcan el mismo valor en el cambio neto. En sentido contrario a lo que se esperaría, si un tamaño de muestra alcanza los requisitos de calidad para un parámetro Δ, no necesariamente cumplirá los requisitos de calidad para el mismo valor nominal del cambio neto con una configuración diferente en las proporciones correspondientes.

Para cumplir los requisitos de calidad, con el mismo MER, se necesitará un mayor tamaño de muestra si no se esperan cambios significativos en las tasas de desempleo entre los correspondientes períodos. Si el cambio neto sigue siendo el mismo para ambos períodos, para cumplir los requisitos de calidad, con el mismo MER, se necesitará un mayor tamaño de muestra si la cifra de desempleo es alta. Ahora, al intentar estimar los cambios netos de proporciones (como el cambio anual o mensual en las tasas de desempleo), se encuentra que:

- Si las tasas son significativamente diferentes, se espera un tamaño de muestra pequeño.
- Si las tasas son similares y las proporciones son bajas, se requiere un tamaño de muestra moderado.
- Si las tasas son similares y las proporciones son grandes, se espera un tamaño de muestra grande.

I. Algunas consideraciones adicionales sobre el tamaño de la muestra

Cuando la encuesta se ha planeado de modo que sea representativa de algún conjunto de estratos, es necesario reproducir estas mismas expresiones en cada uno de los subgrupos de interés. Por otro lado, las anteriores aproximaciones al cálculo de tamaño de muestra son insuficientes ante la realidad de la ausencia de respuesta y las desactualizaciones de los marcos de muestreo. En esta sección se profundizará en estos temas.

1. Asignación del tamaño de la muestra en los estratos de muestreo

Como se aclaró anteriormente, todas las encuestas de hogares en América Latina tienen un componente explícito de estratificación. Por lo tanto, una pregunta que surge inmediatamente es: después de determinar el tamaño de muestra general, ¿cómo asignarlo apropiadamente a todos los estratos de muestreo? Cabe suponer que el tamaño de muestra general es n y que hay H estratos fijos. Por ende, se quieren determinar los tamaños de muestra n_h para cada estrato *(h=1,...,H)*, de manera que se garantice la ganancia de precisión de la estrategia de muestreo.

Existen varios tipos de asignación que pueden estudiarse para determinar la más apropiada en términos de eficiencia. A continuación se presenta una lista no exhaustiva de las distintas asignaciones:

i) Asignación proporcional: donde se selecciona una proporción de elementos en cada estrato, siguiendo la estructura poblacional. Lohr (2020) afirma que este tipo de asignación se utiliza cuando es deseable que la muestra se pueda ver como una versión en miniatura de la población. Gutiérrez (2016) señala que, si se define la fracción de muestreo como $f_h=n_h/N_h$ en el estrato h, entonces, al utilizar la asignación proporcional, la fracción de muestreo será la misma para todos los estratos, tal que $f_h=f$. En este caso, la probabilidad de inclusión de cualquier elemento en la población $\pi_k = f_h = f$ es constante y fija. De esta manera, cada unidad de la muestra representará el mismo número de elementos en la

población, independientemente del estrato al que pertenezca. Con la asignación proporcional, el tamaño de muestra en cada estrato está dado por:

$$n_h = f \times N_h$$

ii) Asignación de Neyman: donde se selecciona una muestra de elementos en cada estrato de tal forma que se maximice la eficiencia estadística de la estrategia de muestreo de la estructura poblacional. Groves y otros (2009) mencionan que, con este método, se producen las menores varianzas para la media muestral, en comparación con otras técnicas de asignación de tamaño de la muestra. Con la asignación de Neyman, el tamaño de la muestra que minimiza la varianza de la estrategia de muestreo está dado por:

$$n_h = n \frac{N_h S_{yU_h}}{\sum_{h=1}^{H} N_h S_{yU_h}}$$

iii) Donde $S_{yU_h} = \sqrt{S^2_{yU_h}}$ es la raíz de la varianza de la característica de interés en cada estrato. Gutiérrez (2016) afirma que, con respecto a la asignación de Neyman, es recomendable redondear el tamaño de la muestra en cada estrato al entero más próximo.

iv) Asignación de Kish: al usar la asignación proporcional en los estratos pequeños, la muestra puede resultar muy pequeña, lo que crearía problemas de eficiencia y pérdida de precisión. Por otro lado, el hecho de utilizar una asignación uniforme (selección del mismo número de elementos en cada estrato $n_h = c$) tendrá como consecuencia una variación sustancial en las fracciones de muestreo entre los estratos y, por ende, habrá una fracción de muestreo muy grande del estrato más pequeño. Una solución intermedia entre la asignación proporcional y la asignación uniforme es la asignación propuesta por Kish, que adopta la siguiente expresión:

$$n_h = n \frac{\sqrt{\frac{1}{H^2} + I W_h^2}}{\sum_{h=1}^{H} \sqrt{\frac{1}{H^2} + I W_h^2}}$$

Donde $W_h = N_h / N$, e $I \geq 0$ es el índice de asignación de Kish, que denota la importancia relativa entre las estimaciones nacionales y las de cada estrato. A medida que este índice se hace más pequeño, se otorga menor importancia a las estimaciones nacionales. La asignación de Kish proporciona un balance entre la asignación uniforme y la proporcional. Cuando $I = 0$, se reduce a la asignación uniforme. Por su parte, si $I \to \infty$, tiende a un enfoque de asignación proporcional. Por lo general se utiliza $I = 1$ para garantizar que la precisión de las características de interés en lo nacional y en los estratos sea aproximadamente la misma.

2. Ajustes por subcobertura

Debido a las características propias de este tipo de encuestas, siempre se presentará una realidad ineludible: en las encuestas de hogares existe ausencia de respuesta. Por ello, los institutos nacionales de estadística deben tomar medidas preventivas al adjudicar los tamaños de la muestra en cada estrato. El hecho de contar con un tamaño efectivo de muestra muy inferior al planeado inicialmente puede conllevar problemas de sesgo y de precisión en las estimaciones de las cifras nacionales o regionales con las que se aborda la política económica y de desarrollo de los países de la región.

En encuestas de hogares cuyo diseño es longitudinal, el problema de la ausencia de respuesta no solo se debe abordar en el momento de realizar la encuesta. Se trata de una cuestión que debe tratarse de manera integral y más general, debido a que un hogar que pertenezca a un panel puede decidir no participar más después de un par de visitas. De este modo, el desgaste de los respondientes se convierte en un problema, puesto que contribuye al fenómeno de la ausencia de respuesta, al que se debe prestar atención para evitar problemas de sesgo y baja confiabilidad.

Kalton (2009) advierte que, al diseñar la encuesta, se debe tener en cuenta el ajuste de submuestras. Por ejemplo, para estimar el cambio de la situación de pobreza o indigencia en los hogares, es necesario realizar un ajuste del tamaño de muestra inicial. De este modo, al final de la aplicación de la encuesta, el tamaño de muestra efectivo cumplirá con los requisitos de precisión de la inferencia estadística. Los INE pueden estimar, con base en su vasta experiencia en la realización de encuestas, la probabilidad de que una persona (o jefe de hogar) responda al instrumento. Si esta probabilidad es denotada como $\phi = Pr(k \in s_r)$, donde s_r denota el subconjunto de respondientes efectivos, los tamaños de muestra de individuos y hogares serán ajustados al dividirlos por ϕ.

$$n_{final} = \frac{n_{inicial}}{\phi}$$

Por ejemplo, si esta probabilidad fue estimada en $\phi = 0,8$, todos los tamaños de muestra calculados en los pasos anteriores deberán ajustarse como $n_{final} = \frac{n_{inicial}}{0,8} = 1,25 \times n_{inicial}$. Por último, si la información auxiliar lo permite, este ajuste debería llevarse a cabo de manera diferenciada en cada uno de los estratos. Por ejemplo, si se conoce que este fenómeno de ausencia de respuesta tiene una mayor incidencia en las zonas rurales que en las urbanas, el ajuste debería realizarse de forma diferenciada.

3. Sustituciones y reemplazos

Una práctica común en los operativos de campo de las encuestas de hogares en América Latina consiste en sustituir las UPM y viviendas de las que no se ha obtenido respuesta. Se consideraría el reemplazo de las UPM cuando no se puede acceder al sitio geográfico por diferentes razones; por ejemplo, problemas de orden público o seguridad, algún cambio

importante en la infraestructura de la zona, o porque no se tiene el consentimiento informado de las autoridades de la comunidad. En este caso, si no se puede acceder a la UPM, tampoco se puede acceder a ninguno de los hogares que la integran. Para llevar a cabo sustituciones y reemplazos en las encuestas de hogares, se utiliza, por lo general, la metodología de estratificación implícita, que permite seleccionar de manera automática los reemplazos adecuados de acuerdo con la conformación de subgrupos poblacionales similares.

La estratificación implícita se utiliza cuando la encuesta está enfocada en un tema particular y, para su ejecución exitosa, requiere el uso del muestreo sistemático con probabilidades desiguales en la selección de las UPM, es decir, en la definición del diseño de muestreo de la primera etapa. Según Naciones Unidas (2008, pág. 47), en la mayoría de los países la secuencia podría empezar con el estrato urbano, desagregándolo por departamento y, a su vez, desagregando los departamentos por municipio. De forma similar, el estrato rural se desagrega por departamento y, a su vez, los departamentos se desagregan por comuna o sección administrativa. Obsérvese que la selección sistemática de UPM está condicionada a la medida de tamaño utilizada en la primera etapa, es decir, al número de viviendas que la componen. De esta forma, la estratificación implícita consiste en que, para cada estrato explícito (urbano, rural o regional, entre otros) se crea una lista ordenada de UPM. El orden de la lista se determinará por los estratos implícitos definidos en la planificación de la encuesta (departamento, municipio) y, dentro de cada subgrupo, se ordenarán las UPM en orden descendente (o ascendente). De esta forma, esta metodología constituye un método objetivo de selección de reemplazos, puesto que, si no se puede acceder a la UPM seleccionada en un inicio, su reemplazo será la inmediatamente anterior (o posterior) en la lista estratificada de manera implícita. Con este procedimiento se seleccionará como reemplazo la UPM ubicada en el mismo municipio, dentro del mismo departamento, en la misma zona y con un número similar de viviendas, respetando el principio de representatividad. De otra forma, si no se considerara un procedimiento similar a la estratificación implícita, los reemplazos de las UPM podrían seleccionarse de forma aleatoria en otro departamento y con un número de viviendas mucho más grande o mucho más pequeño, lo que añadiría sesgo a la selección inicial.

Aunque la estratificación implícita permite limitar el sesgo introducido por la ausencia de respuesta de las UPM, Vehovar (1999, págs. 348-349) advierte que se debe tener precaución al utilizar esta práctica, puesto que también puede causar sesgos importantes en las estimaciones de interés. Ello se desprende del hecho de que los individuos ubicados en zonas a las que sí es posible acceder diferirán significativamente de los que estén en las zonas de difícil acceso, pues es evidente que pertenecen a dos realidades diferentes. Por esta razón, es útil que, después de haber valorado los posibles sesgos, si se ha tomado la determinación de realizar las sustituciones sobre las UPM de difícil acceso, se realice un seguimiento exhaustivo en cada aplicación de la encuesta que permita clasificar el mecanismo de recolección de información primaria y evaluar su impacto en la precisión de los estimadores resultantes.

Capítulo VIII
Estimación de parámetros

Un estimador se define como una función de la muestra aleatoria, que toma valores en el conjunto de los números reales y solo depende de los elementos pertenecientes a la muestra. El diseño de muestreo se define por el soporte, que es el conjunto Q de todas las posibles muestras. Las propiedades estadísticas de un estimador están determinadas por la medida de probabilidad discreta generada por el diseño de muestreo. Es decir, dada la probabilidad de selección de cada muestra $s \in Q$, la esperanza, la varianza y otras propiedades de interés de los estimadores se definen a partir de $p(s)$. En particular, la esperanza de un estimador $\hat{\theta}$ se traduce en la siguiente expresión:

$$E(\hat{\theta}) = \sum_{s \subset Q} \theta\,(s) p(s)$$

Las propiedades más buscadas en un estimador $\hat{\theta}$ son: el insesgamiento, la eficiencia y la consistencia. El sesgo está definido por la siguiente expresión:

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta$$

y el error cuadrático medio está dado por:

$$ECM(\hat{\theta}) = E[\hat{\theta} - \theta]^2 = Var(\hat{\theta}) + B^2\,(\hat{\theta})$$

Si el sesgo de un estimador es nulo, se dice que el estimador es insesgado y, cuando esto ocurre, el error cuadrático medio se convierte en la varianza del estimador. Además, si la varianza del estimador es pequeña en relación con otros estimadores, se dice que el estimador es eficiente. Por último, si, cuando el tamaño de la muestra crece, el valor del estimador se acerca al parámetro desconocido, se dice que el estimador es consistente. Särndal, Swensson y Wretman (2003) afirman que el objetivo de un estudio por muestreo es estimar uno o más parámetros poblacionales. De acuerdo con Gutiérrez (2016), las decisiones más importantes a la hora de abordar un problema de estimación por muestreo son:

i) La elección de un diseño de muestreo y un algoritmo de selección que permita implementar el diseño.

ii) La elección de una fórmula matemática o estimador que calcule una estimación del parámetro de interés en la muestra seleccionada.

Estas decisiones no son independientes. Por lo general, la elección de un estimador dependerá del diseño de muestreo utilizado. De hecho, si $\hat{\theta}$ es un estimador del parámetro θ y $p_s(\cdot)$ es un diseño de muestreo definido sobre un soporte Q, la estrategia de muestreo será la dupla $(p(\cdot), \hat{\theta})$. A continuación se presentan algunos estimadores comúnmente utilizados en el procesamiento de las encuestas de hogares en América Latina.

Aunque el marco de referencia de la teoría de muestreo es la estimación de un parámetro de interés, en la práctica no solo se necesitan estimaciones que comprendan a toda la población, sino también estimaciones relativas a distintos subgrupos poblacionales, que determinan una partición de la población de interés. Es bien sabido que, cuando se habla de subgrupos poblacionales, se hace referencia a dominios de interés, estratos o posestratos. Cuando el investigador se enfrenta a una encuesta que tiene en cuenta subgrupos poblacionales (es decir, siempre), es indispensable saber en qué se diferencia cada uno de ellos de los demás, pues de esto depende que la investigación arroje resultados confiables mediante el planteamiento de la mejor estrategia de muestreo.

Asumiendo que $U_1, \ldots, U_g, \ldots, U_G$ denotan subgrupos poblacionales, de manera que $\bigcup_{g=1}^{G} U_g = U$, y siendo N_g el tamaño absoluto del subgrupo U_g, se obtiene que $\sum_{g=1}^{G} N_g = N$. A partir de estas definiciones, se pueden plantear algunas diferencias y similitudes entre dichos subgrupos, que se resumen a continuación.

- Dominios de interés: este tipo de subgrupos poblacionales requiere estimaciones separadas. Los requisitos se planean en la etapa de diseño para asegurar que el diseño de la muestra garantice una cobertura apropiada en cada uno de los dominios de interés al recopilar la información. Por lo general, esto solo se puede lograr ampliando el tamaño de la muestra (n), pues el marco de muestreo no informa de la pertenencia de las personas a los dominios de interés. Un aspecto importante de esta clase de subgrupos poblacionales es que el número de personas de la muestra que pertenecen a un dominio de interés es siempre aleatorio y, en el caso de algunos dominios particulares, puede llegar a ser muy pequeño. Por otra parte, el tamaño absoluto de cada dominio no se conoce ni antes de la etapa de diseño ni después de la etapa de estimación. Algunos ejemplos claros de estos subgrupos son aquellos basados en la condición de ocupación, la situación de pobreza o la rama de actividad, entre otros.
- Estratos: se llama así a los subgrupos que se definen cuando el marco de muestreo permite conocer la pertenencia de todas las unidades de la población a un subgrupo poblacional determinado. Cuando se sabe que la característica de interés tiene un comportamiento distinto en cada uno de los estratos y se planea un diseño de muestreo que tenga en cuenta este aspecto mediante la selección aleatoria de unidades en cada uno de los estratos, se dice que el diseño de muestreo es

estratificado. El aspecto fundamental de esta clase de subgrupos poblacionales es que el conocimiento de la pertenencia de las personas a los estratos se incorpora en la etapa de diseño de la muestra. A diferencia de los dominios, en los estratos se conoce el tamaño poblacional y se controla el tamaño de la muestra antes de la etapa de estimación. Algunos ejemplos claros de estos subgrupos corresponden a las zonas urbanas o rurales, las regiones y los municipios.

- Posestratos: la propiedad que caracteriza a este tipo de subgrupos poblacionales es que, aunque se conoce su tamaño en la etapa de diseño, se desconoce el número de personas que pertenecerán al posestrato en la muestra seleccionada. Algunos ejemplos claros de estos subgrupos son los que se definen a partir del grupo etario, el sexo o la etnia. Si bien no siempre se utilizan en la fase de diseño, sus proyecciones demográficas se utilizan en la fase de análisis para analizar y mejorar la eficiencia de los estimadores. De acuerdo con Särndal, Swensson y Wretman (2003), existen dos casos en los que se presenta esta situación, llamada comúnmente posestratificación:

  i) Cuando el marco de muestreo permite conocer la pertenencia de todos los elementos a los subgrupos poblacionales, pero el investigador decide no utilizar esta información en la etapa de diseño. Las principales razones para obviar este tipo de información suelen ser de carácter práctico o logístico. Una vez realizada la selección de la muestra, se observa la característica de interés yk en las personas que la componen. El investigador utiliza la información auxiliar de pertenencia a los posestratos en la etapa de estimación para mejorar la eficiencia de la estrategia de muestreo; en particular, del estimador propuesto.

  ii) Cuando, mediante una fuente de información confiable, se conoce el tamaño absoluto N_g de cada subgrupo poblacional, pero se desconoce la pertenencia de las personas a los subgrupos debido a un defecto del marco de muestreo. Después de la etapa de diseño, se observa la característica de interés y se analiza la pertenencia de las personas seleccionadas a los posestratos, a fin de utilizar esta información en la etapa de estimación para mejorar la eficiencia de los estimadores de los parámetros de interés.

Para llevar a cabo el análisis apropiado de una encuesta, no se pueden pasar por alto las diferencias entre estos subgrupos. Es más, el diseño y rediseño de las encuestas se basan fundamentalmente en la búsqueda de estos subgrupos de la población. En todas las encuestas de hogares que se realizan en América Latina , el objetivo es investigar fenómenos asociados a subgrupos poblacionales dispersos en el territorio de cada país. Por ejemplo, en una encuesta de fuerza de trabajo, resulta de interés estimar las tasas de desocupación de los hombres y las mujeres, de las diferentes etnias o de las zonas urbanas y las rurales, entre otras. En general, cabe especificar los siguientes aspectos:

i) El tamaño de la muestra de una encuesta casi siempre se basa en la incidencia de un fenómeno que clasifica a la población en un determinado dominio de interés.

ii) El tamaño de la muestra se reparte de antemano entre los diferentes estratos geográficos para mejorar la eficiencia de la recolección de información y del diseño de muestreo.

iii) A menudo, las proyecciones demográficas sobre los posestratos se utilizan en la fase de estimación para mejorar la precisión de los estimadores.

A. Estimador de Horvitz-Thompson de totales y tamaños poblacionales

1. Estimación de totales

En la mayoría de los casos, los indicadores sociales a nivel nacional pueden verse como funciones de los totales de una o más variables de interés. Por ejemplo, si se desea estimar un total $t_y = \sum_U y_k$, el estimador tradicionalmente utilizado en las encuestas de hogares es el de Horvitz-Thompson, el cual es insesgado (sesgo nulo) por definición, sin necesidad de acoger ningún tipo de supuestos. Este estimador toma la siguiente expresión:

$$\hat{t}_{y,\pi} = \sum_s d_k \, y_k$$

Donde la muestra s hace referencia al subconjunto de la población seleccionado siguiendo un diseño de muestreo probabilístico que determina los pesos de muestreo d_k, los cuales expanden el valor de la variable de interés y_k para el k-ésimo individuo. A su vez, d_k es el inverso multiplicativo de la probabilidad de inclusión del k-ésimo individuo en la muestra, $d_k = \pi_k^{-1}$.

Como se verá más adelante, cuando el diseño de muestreo contiene sistemas de estratificación y selección de conglomerados en varias etapas, esta probabilidad es el producto de las probabilidades condicionales que surgen en los subsiguientes procesos de selección probabilística. Por tanto, el peso final de muestreo suele ser una multiplicación de factores de expansión en cada etapa o fase del diseño de muestreo. En general, este estimador toma diferentes formas a medida que el diseño de muestreo cambia. A continuación se presenta una lista no exhaustiva de los diseños más importantes de la teoría de muestreo para encuestas de hogares.

a) Muestreo aleatorio simple

En este caso, las probabilidades de inclusión son uniformes en cada unidad incluida en la muestra:

$$\pi_k = \frac{n}{N}$$

Por lo tanto, el estimador toma la siguiente forma:

$$\hat{t}_{y,\pi} = \frac{n}{N}\sum_s y_k$$

b) Muestreo proporcional al tamaño

Este diseño de muestreo determina probabilidades de inclusión proporcionales al valor de una característica de información auxiliar estrictamente positiva y disponible en el marco de muestreo[1]. Así, las probabilidades de inclusión obedecen a la siguiente relación:

$$\pi_k = \frac{n\, x_k}{t_x} \qquad 0 < \pi_k \leq 1$$

En consecuencia, el estimador toma la siguiente forma:

$$\hat{t}_{y,\pi} = t_x \sum_s \frac{y_k}{n\, x_k}$$

Por último, no es cierto que la asignación de probabilidades desiguales en las unidades de muestreo genere sesgo en la encuesta. De hecho, cuando en estas condiciones se utiliza el estimador de expansión (estimador de Hansen-Hurwitz, en el caso de muestreos con reemplazo, y estimador de Horvitz-Thompson, en el de muestreos sin reemplazo), el sesgo es nulo. Más aún, si se utilizara un estimador sin factores de expansión, la inferencia estaría sesgada. Por ende, es natural que, si el diseño supone probabilidades desiguales, estas se utilicen dentro de un estimador que considere esta desigualdad.

c) Muestreo estratificado

Si mediante $\hat{t}_{yh,\pi}$ se estima insesgadamente el total de la característica de interés t_{yh} del estrato h, el estimador insesgado para el total poblacional t_y está dado por:

$$\hat{t}_{y,\pi} = \sum_{h=1}^{H} \hat{t}_{yh,\pi}$$

[1] Tradicionalmente, en las encuestas de hogares, la característica de información auxiliar disponible en un marco de muestreo de áreas es el tamaño de las áreas de empadronamiento censales, medido en número de viviendas. De ahí que este diseño de muestreo se denomine "proporcional al tamaño" y que la característica de interés se conozca como "medida de tamaño".

Por ejemplo, en un diseño de muestreo aleatorio estratificado, las probabilidades de inclusión de primer orden están dadas por:

$$\pi_k = \frac{n_h}{N_h} \quad \text{si} \quad k \in U_h$$

En este caso, siendo s_h la muestra seleccionada en el estrato U_h, el estimador insesgado del total t_y está dado por:

$$\hat{t}_{y,\pi} = \sum_{h=1}^{H} \frac{n_h}{N_h} \sum_{k \in s_h} y_k$$

d) Muestreo por conglomerados

En el sistema general de muestreo por conglomerados, se utiliza un diseño de muestreo específico para la selección de los conglomerados en la muestra. La probabilidad de que el k-ésimo elemento esté incluido en la muestra s es idéntica a la probabilidad de inclusión del conglomerado al que pertenece π_{Ii}; es decir:

$$\pi_k = \pi_{Ii} \quad \text{si} \quad k \in U_i$$

Si se supone que la población se divide en N_I conglomerados y se selecciona una muestra de conglomerados s_I de tamaño n_I, el estimador de Horvitz-Thompson del total poblacional para un diseño de muestreo aleatorio de conglomerados estará dado por:

$$\hat{t}_{y,\pi} = \frac{n_I}{N_I} \sum_{s_I} t_{yi}$$

Donde t_{yi} hace referencia al total de la característica de interés en el conglomerado U_i. Como se mencionó en los capítulos anteriores, la definición de conglomerados con tamaños muy desiguales redunda en un aumento significativo de la varianza del estimador. Por ello, en las encuestas de hogares se intenta crear conglomerados acotados, a nivel de manzana o sección administrativa. Esta práctica es muy pertinente, pues la varianza del estimador de expansión dependerá de la varianza de los totales de los conglomerados: de haber una gran variación en los tamaños, habrá también una gran variación en los totales y, por consiguiente, la varianza del estimador será alta. De lo contrario, si se conoce una característica de información auxiliar a nivel de los conglomerados (medida de tamaño), es posible utilizar esta información del marco para reducir la varianza en el estimador.

e) Muestreo en dos etapas

En este diseño, la probabilidad de inclusión de primer orden del k-ésimo elemento está dada por:

$$\pi_k = Pr(k \in s) = Pr(k \in s_i \mid i \in s_I)\, Pr(i \in s_I) = \pi_{k|i}\, \pi_{Ii}$$

Donde s_i corresponde a la submuestra de elementos seleccionada en el conglomerado U_i. En particular, cuando el diseño de muestreo es aleatorio simple en las dos etapas y para cada unidad primaria de muestreo (UPM) seleccionada $i \in s_I$ de tamaño N_i se selecciona una muestra s_i de elementos de tamaño n_i, el estimador de Horvitz-Thompson toma la siguiente forma:

$$\hat{t}_{y,\pi} = \frac{N_I}{n_I} \sum_{i \in s_i} \frac{N_i}{n_i} \sum_{k \in s_i} y_k$$

f) Muestreo en dos fases

En la primera fase de este tipo de muestreo, se selecciona una muestra de elementos s_a y se recoge información de interés para crear una versión reducida y acotada del marco de muestreo. En la segunda fase, se utiliza esa información para realizar una nueva selección que defina una submuestra s, en la que se observa la característica de información auxiliar. Con este sistema, la probabilidad de que un elemento esté en la submuestra de la segunda fase s depende de lo que haya sucedido en la muestra de la primera fase s_a. Por lo tanto, la probabilidad de inclusión de un elemento en la muestra final no tiene una forma cerrada y es algebraicamente intratable. Por ende, se define el estimador de Horvitz-Thompson condicionado, que toma la siguiente forma:

$$\hat{t}_{y,\pi^*} = \sum_s \frac{y_k}{\pi_k^*} = \sum_s \frac{y_k}{\pi_{ak}\, \pi_{k|s_a}}$$

En esta expresión, π_{ak} denota la probabilidad de inclusión del elemento en la muestra de la primera fase y $\pi_{k|s_a}$ denota la probabilidad de inclusión del elemento en la submuestra de la segunda fase, condicionada a su inclusión en la primera fase.

2. Aplicación del estimador de Horvitz-Thompson en una encuesta de hogares estándar

Si se supone una encuesta de hogares con un diseño estándar que tenga, por ejemplo, un sistema estratificado de H estratos con dos etapas de selección dentro de cada estrato (la primera etapa con selección de UPM dentro del estrato, la segunda con selección de hogares), el peso de muestreo final y el estimador del total estarán dados por la siguiente expresión:

$$\hat{t}_{y,\pi} = \sum_s d_k\, y_k = \sum_h \sum_{i \in s_{Ih}} \sum_{k \in s_{hi}} w_{hik}\, y_{hik}$$

Por ejemplo, si dentro de cada estrato U_h $h=1,\ldots,H$ existen N_{Ih} UPM, entre las que se selecciona una muestra s_{Ih} de n_{Ih} unidades mediante un diseño de muestreo aleatorio simple, y, además, se considera que el submuestreo dentro de cada unidad primaria seleccionada es también aleatorio simple, de manera que para cada UPM seleccionada

$U_i \in s_{Ih}$ de tamaño N_i se selecciona una submuestra s_i de elementos de tamaño n_i, la forma final del estimador de Horvitz-Thompson para el total poblacional quedaría expresada de la siguiente manera:

$$\hat{t}_{y,\pi} = \sum_{h=1}^{H} \hat{t}_{yh,\pi} = \sum_{h=1}^{H} \left[\frac{N_{Ih}}{n_{Ih}} \sum_{i \in s_{Ih}} \frac{N_i}{n_i} \sum_{k \in s_i} y_k \right]$$

3. Estimación de tamaños y totales en dominios

En general, todas las expresiones aplicables a la estimación de totales son apropiadas para la estimación de tamaños poblacionales, puesto que la característica de interés tomará el mismo valor para todas las unidades en la muestra; es decir, $y_k = 1 \ \forall \ k \in s$. De esta forma, el estimador de Horvitz-Thompson de un tamaño poblacional está dado por la suma de los factores de expansión:

$$\hat{N} = \sum_{s} d_k$$

De forma generalizada en América Latina, el diseño estándar de una encuesta de hogares incluye un sistema de muestreo estratificado con dos etapas de selección. Con estas condiciones, el estimador del tamaño poblacional estará dado por la siguiente expresión:

$$\hat{N} = \sum_{s} d_k = \sum_{h} \sum_{i \in s_{Ih}} \sum_{k \in s_{hi}} w_{hik}$$

Al asumir un diseño de muestreo estratificado de dos etapas, con selección aleatoria simple en cada etapa, el estimador de Horvitz-Thompson del tamaño poblacional tendría la siguiente forma:

$$\hat{N}_{\pi} = \sum_{h=1}^{H} \left[\frac{N_{Ih}}{n_{Ih}} \sum_{i \in s_{Ih}} \frac{N_i}{n_i} \sum_{k \in s_i} 1 \right]$$

Como afirma Gutiérrez (2016), además de las estimaciones relativas a la población en general, en muchas investigaciones es necesario realizar estimaciones relativas a determinados subgrupos poblacionales (denominados "dominios" por la Subcomisión de Muestras Estadísticas de las Naciones Unidas. La definición de los dominios se logra una vez registrada la información de los elementos. Los dominios tienen que cumplir con los siguientes requisitos:

i) Ningún elemento de la población puede pertenecer a dos dominios simultáneamente.

ii) Cada elemento de la población debe pertenecer a un dominio.

iii) La unión de todos los dominios equivale a la población del estudio.

La estimación por dominios se caracteriza por el desconocimiento previo de la pertenencia de las unidades poblacionales a los dominios. Es decir, para saber qué unidades de la población pertenecen a cada dominio, es necesario realizar el proceso de medición. En primer lugar, se construye una función indicadora z_{dk} de la pertenencia del elemento al dominio, que toma el valor 1 si el elemento k pertenece al dominio U_d $(k \in U_d)$, y el valor 0 en caso contrario. Ahora se pueden utilizar los principios del estimador de Horvitz-Thompson para hallar un estimador insesgado del tamaño del dominio U_d, que estará dado por:

$$\hat{N}_d = \sum_{s_d} d_k$$

Al multiplicar la variable de pertenencia z_{dk} por el valor de la característica de interés y_k, se crea una nueva variable y_{dk} dada por $y_{dk} = z_{dk}\, y_k$, y —una vez construida— es posible definir el estimador insesgado del total de la característica de interés en el dominio U_d, dado por:

$$\hat{t}_{y_d,\pi} = \sum_{s} d_k\, y_{dk} = \sum_{s_d} d_k\, y_k$$

B. Estimador de Hájek de medias y proporciones

Es muy probable que, en el momento de estimar medias y proporciones, no se tenga un conocimiento exacto del tamaño poblacional. Por ejemplo, para la estimación de indicadores a nivel de hogar en encuestas mensuales, es difícil conocer con certeza el número de hogares en el país mes a mes. Por esta razón, cuando se definen estimadores de indicadores relativos, es necesario hacer un doble proceso de inferencia: a nivel de la característica de interés que se quiere investigar y a nivel del tamaño de la población. El enfoque más utilizado se basa en el estimador de Hájek, que define una media como la división de dos estimadores de Horvitz-Thompson, de la siguiente manera:

$$\hat{\bar{y}}_s = \frac{\hat{t}_y}{\hat{N}} = \frac{\sum_s d_k\, y_k}{\sum_s d_k}$$

En el caso de la estimación de una proporción P_d, el estimador de Hájek toma la siguiente forma:

$$\hat{P}_d = \frac{\hat{N}_d}{\hat{N}} = \frac{\sum_s d_k\, z_{dk}}{\sum_s d_k} = \frac{\sum_{sd} d_k}{\sum_s d_k}$$

En el caso de la estimación de una media en un subgrupo poblacional —por ejemplo, la media del gasto en la zona urbana—, el estimador de Hájek puede escribirse de la siguiente manera:

$$\hat{\bar{y}}_d = \frac{\hat{t}_{yd}}{\hat{N}_d} = \frac{\sum_s d_k y_k z_{dk}}{\sum_s d_k z_{dk}} = \frac{\sum_{s_d} d_k y_k}{\sum_{s_d} d_k}$$

En general, estos estimadores se pueden considerar no lineales y sus propiedades estadísticas son complejas. Puesto que tanto el numerador como el denominador son variables aleatorias, es necesario verificar algunos supuestos relacionados con el tamaño de la población y de la muestra (Gutiérrez, 2016). En el caso de un diseño de muestreo caracterizado por un sistema estratificado y tres etapas de selección, el estimador de la media poblacional estará dado por la siguiente expresión:

$$\hat{\bar{y}}_d = \frac{\sum_h \sum_{i \in s_{1h}} \sum_{j \in s_{hi}} \sum_{k \in s_{hij}} w_{hijk}\ y_{hijk}}{\sum_h \sum_{i \in s_{1h}} \sum_{j \in s_{hi}} \sum_{k \in s_{hij}} w_{hijk}}$$

Las demás expresiones para los estimadores de proporciones o medias en una subpoblación, en el marco de un diseño de muestreo estándar, pueden derivarse fácilmente siguiendo los principios expuestos anteriormente.

C. Otros estimadores de muestreo

Cuando se dispone de información auxiliar, es posible mejorar la eficiencia de la estimación recurriendo a diferentes formas funcionales que estiman el total; por ejemplo, con el estimador de razón:

$$\hat{t}_{y,r} = t_x \frac{\sum_s d_k y_k}{\sum_s d_k x_k}$$

Donde t_x denota el total poblacional, que se supone conocido para toda la población, de una variable auxiliar x incluida en la encuesta de hogares. Por supuesto, en el análisis de este tipo de encuestas es común realizar inferencias sobre parámetros que tienen una forma no lineal. Uno de los más básicos es la razón poblacional $R_U = t_{y_1} / t_{y_2}$, cuyo cálculo se lleva a cabo estimando ambos componentes de la fracción:

$$\hat{R} = \frac{\hat{t}_{y_1}}{\hat{t}_{y_2}} = \frac{\sum_s d_k y_{1k}}{\sum_s d_k y_{2k}}$$

Como se indicó anteriormente, la estimación de un promedio poblacional $\bar{y}_U = t_y / N$ se lleva a cabo de forma eficiente estimando el tamaño de la población y se puede ver como un caso particular de la estimación de una razón. Por otra parte, las encuestas de hogares con diseños de panel o rotativos suponen un mayor interés en la estimación del cambio de indicadores en dos períodos de tiempo $\Delta = t_{y(t)} - t_{y(t-1)}$. Un estimador de este parámetro está dado por:

$$\hat{\Delta} = \hat{t}_{y(t)} - \hat{t}_{y(t-1)}$$

Si se desea estimar algunas características asociadas con la pobreza, es posible utilizar estimadores más complejos. Siendo y_k el ingreso del individuo k y l el umbral de pobreza, es posible utilizar el siguiente estimador:

$$\hat{F}_\alpha = \frac{1}{n} \sum_{k \in s} d_k \left(\frac{l - y_k}{l} \right)^\alpha I(y_k < l)$$

Donde $I(y_k<l)$ es una variable indicadora que toma el valor 1 si $y_k<l$ o 0 en caso contrario. Si $\alpha=0$, se obtiene una estimación de la incidencia de la pobreza y, si $\alpha=1$, se obtiene una estimación de la brecha de pobreza (Foster, Greer y Thorbecke, 1984).

La selección del estimador está estrechamente relacionada con el diseño de la encuesta. Por ejemplo, si se pretende estimar un indicador para un período de tiempo definido, el diseño de la encuesta no debería conllevar un sistema de rotación con traslape de hogares, pues la correlación del indicador redundaría en un aumento de su varianza y, por ende, en una pérdida de eficiencia. Sin embargo, para estimar el cambio del indicador entre dos períodos, se necesita un diseño de rotación que asegure un tamaño de muestra suficiente para estimar con precisión este cambio. Cochran (1977, sec. 12.13) afirma que, cuando se desea estimar el indicador en el período actual y el cambio entre períodos, es recomendable que la tasa de traslape sea de 2/3, 3/4 o 4/5 de una ronda a otra.

D. Estimadores de calibración

La calibración se ha consolidado como un importante instrumento metodológico en la producción de grandes cantidades de datos estadísticos (Särndal, 2007). Este método integra información auxiliar en las estimaciones de la encuesta, no solo para garantizar la coherencia con las cifras oficiales divulgadas por los institutos nacionales de estadística (INE), sino para aumentar la eficiencia del proceso de estimación. Gutiérrez (2016) describe brevemente este método de la siguiente manera:

i) Se supone que se tiene acceso a un vector de información auxiliar, $\boldsymbol{x}_k=(x_{1k},x_{2k},\ldots,x_{pk})$, de p variables auxiliares, conocido para las personas seleccionadas en la muestra.

ii) Además, a partir de registros administrativos u otras fuentes confiables, se conoce el total correspondiente al vector de información auxiliar $\boldsymbol{t}_{\boldsymbol{X}}=\sum_{k \in U} \boldsymbol{x}_k$.

iii) El propósito del estudio es estimar el total de la característica de interés incorporando la información auxiliar, dada por $\boldsymbol{x}_k$ $(k \in s)$.

iv) Los pesos resultantes deben cumplir con la siguiente restricción:

$$\sum_{k \in s} w_k \boldsymbol{x}_k = \boldsymbol{t}_{\boldsymbol{X}}$$

conocida como ecuación de calibración.

v) El resultado de la calibración es un nuevo conjunto de pesos w_k que son muy cercanos al inverso de la probabilidad de inclusión del k-ésimo elemento $d_k=1/\pi_k$.

En general, la estrategia de estimación utilizada por los INE de los países de América Latina recurre a la metodología de calibración sobre proyecciones poblacionales en los dominios de representatividad de la encuesta; por ejemplo, departamento, zona urbana, zona rural, sexo y grupos de edad. Entre las ventajas de utilizar estos procedimientos, se incluye el hecho de que las estimaciones tendrán un sesgo despreciable y los errores estándar serán más pequeños en comparación con los del estimador de Horvitz-Thompson. De esta forma, se crea un sistema de ponderación que reproduce la información auxiliar disponible y estima de manera eficiente cualquier característica de interés de la encuesta. Esta coherencia entre las cifras oficiales y las que puede producir la encuesta hace que sea preferible el uso de los estimadores de calibración.

En una encuesta de hogares, las restricciones de calibración pueden establecerse con respecto a determinadas características de los hogares y de las personas simultáneamente. De esta forma, por ejemplo, es posible calibrar sobre las proyecciones demográficas y, al mismo tiempo, controlar las estimaciones del número de hogares en el país de manera conjunta. Estevao y Särndal (2006) estudian una gran variedad de casos en los que se calibra conjuntamente en distintos niveles de desagregación sobre diferentes sistemas de muestreo. Por ejemplo, en la Encuesta Continua de Empleo del Estado Plurinacional de Bolivia, la calibración se realiza a partir de una posestratificación sobre los tamaños poblacionales de los cruces resultantes entre las variables "departamento" (hay nueve departamentos), "zona" (rural y urbana) y "población en edad de trabajar" (con dos categorías: 10 años o más y menor de 10 años).

En resumen, el uso de este tipo de estimadores garantiza una coherencia estética, puesto que es deseable que las estimaciones puntuales de las encuestas coincidan con los conteos censales, las proyecciones poblacionales y los registros estadísticos o administrativos. Además, aumentará la precisión, porque en la búsqueda de la mejor estrategia de muestreo, el estadístico quiere obtener cifras precisas y confiables que se traduzcan en intervalos de variación angostos y menores errores de muestreo. Por último, cuando se realiza una integración adecuada de la información auxiliar, disminuye el sesgo generado por la falta de respuesta (debido a los individuos) o de cobertura (debido a defectos del marco de muestreo).

1. Ganancia de eficiencia

Para mostrar la manera en que los estimadores de calibración disminuyen la varianza con respecto a los estimadores comunes, se planeó el siguiente experimento de simulación empírica:

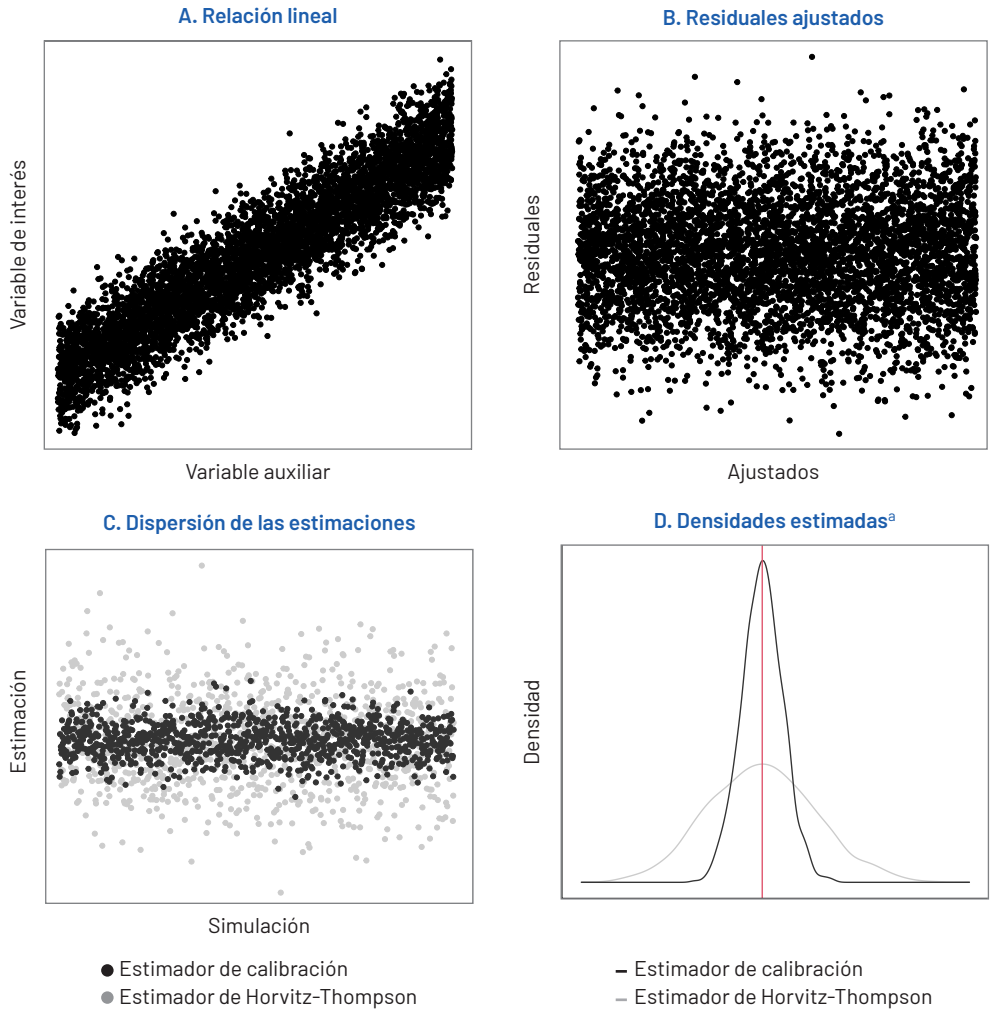
i) Se generaron cuatro conjuntos de datos que presentan una relación específica entre la variable de interés y las variables de información auxiliar.

ii) Se utilizó la metodología de calibración y se compararon empíricamente las medidas de variabilidad para 1.000 iteraciones.

En los gráficos VIII.1, VIII.2, VIII.3 y VIII.4 se muestran la relación entre las variables (gráficos VIII.1A, VIII.2A, VIII.3A y VIII.4A), los residuales ajustados en un modelo de regresión simple (gráficos VIII.1B, VIII.2B, VIII.3B y VIII.4B), la dispersión (gráficos VIII.1C, VIII.2C, VIII.3C y VIII.4C) de las estimaciones de Horvitz-Thompson (puntos grises) y de las estimaciones de calibración (puntos negros), así como la distribución empírica (gráficos VIII.1D, VIII.2D, VIII.3D y VIII.4D) del estimador de Horvitz-Thompson (línea gris) y del estimador de calibración (línea negra). La línea roja vertical indica el valor del parámetro poblacional.

■ Gráfico VIII.1

Comportamiento de los estimadores en una relación de dependencia lineal

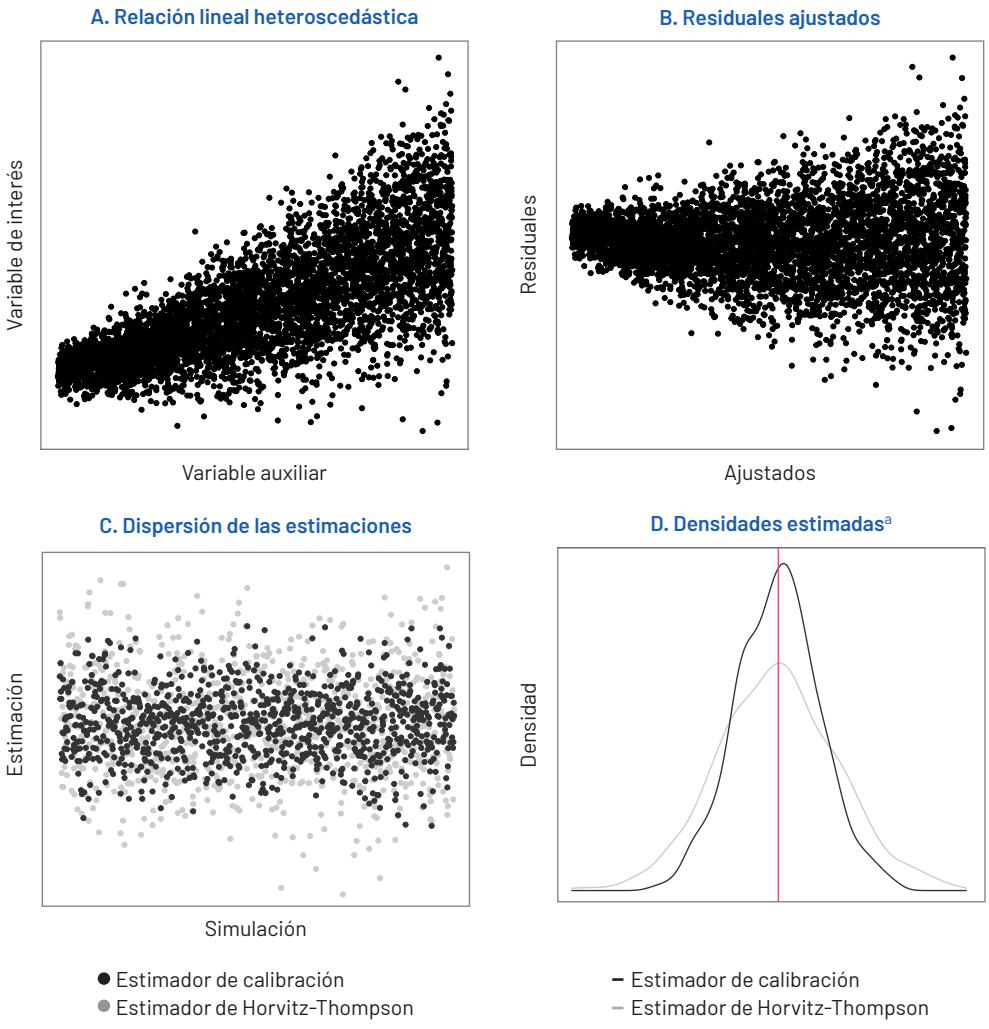


Fuente: Elaboración propia.

[a] La línea roja vertical indica el valor del parámetro poblacional.

En el caso del segundo conjunto de datos, se asume que existe una relación lineal entre la característica de interés y una variable de información auxiliar continua. Del gráfico VIII.2 surge que existe heterocedasticidad en el modelo y esto se refleja en los residuales. A pesar de que ambos estimadores resultan insesgados para el parámetro de interés, el estimador de calibración es un poco más eficiente que el de Horvitz-Thompson.

■ Gráfico VIII.2
Comportamiento de los estimadores en una relación de dependencia lineal con heterocedasticidad

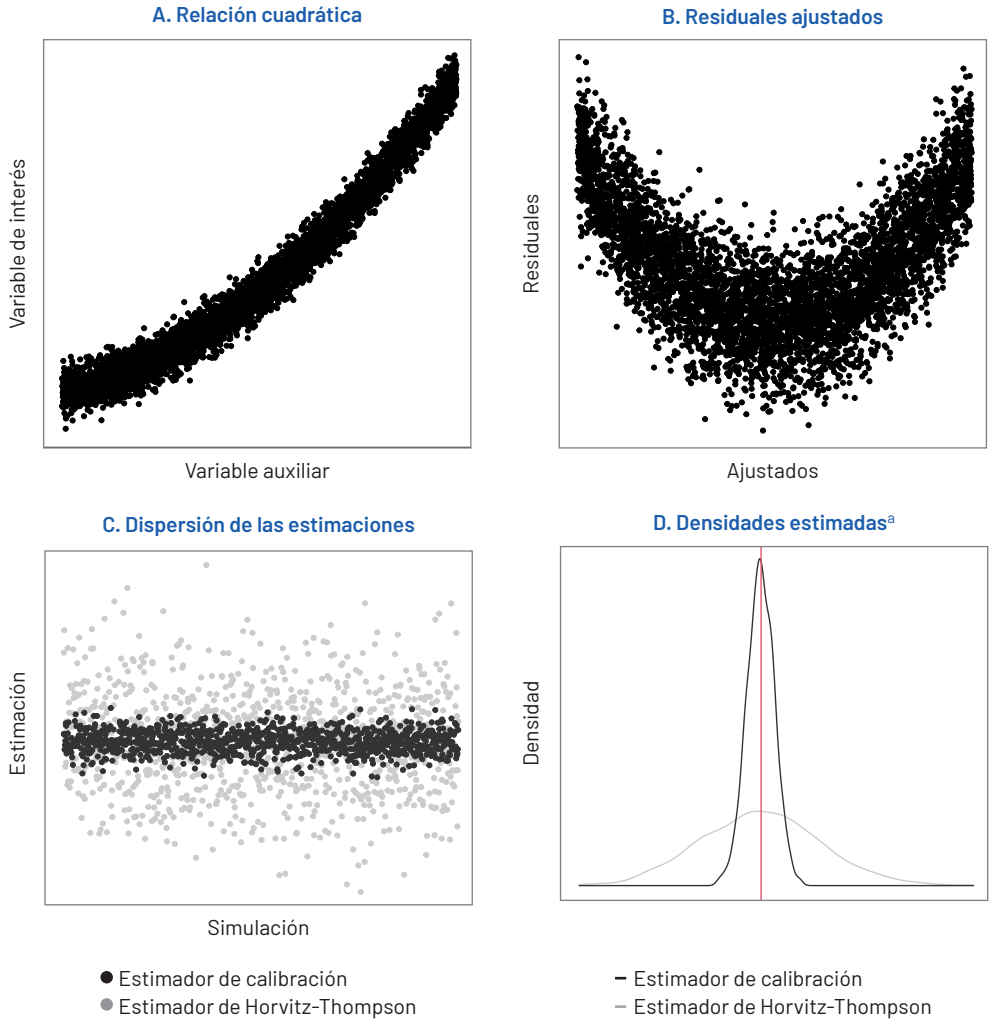


Fuente: Elaboración propia.
[a] La línea roja vertical indica el valor del parámetro poblacional.

En el tercer conjunto de datos, se asume que existe una relación cuadrática entre la característica de interés y una variable de información auxiliar continua. Al utilizar un estimador de calibración lineal, los residuales muestran un comportamiento inapropiado.

Aunque ambos estimadores resultan insesgados para el parámetro de interés, el estimador de calibración es más eficiente que el de Horvitz-Thompson (véase el gráfico VIII.3).

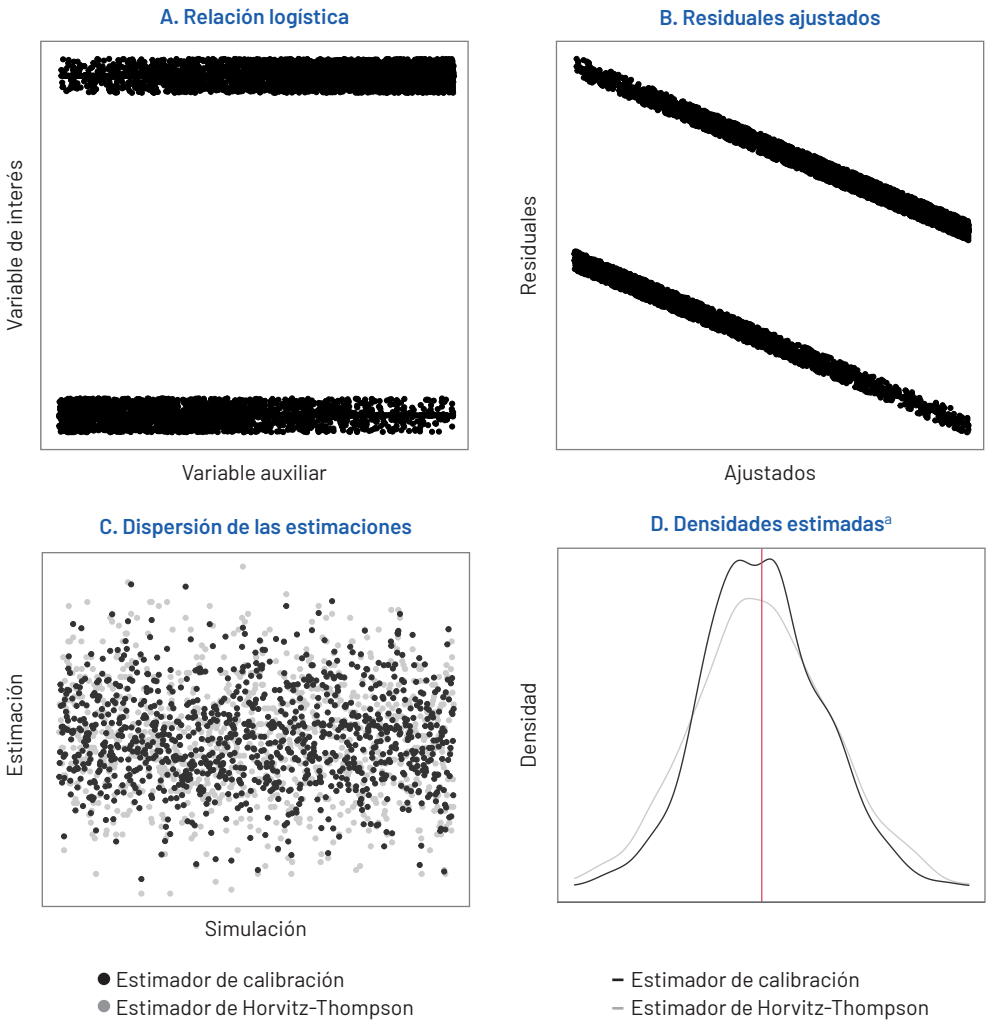
■ Gráfico VIII.3
Comportamiento de los estimadores en una relación de dependencia cuadrática



Fuente: Elaboración propia.
[a] La línea roja vertical indica el valor del parámetro poblacional.

Con respecto al último conjunto de datos, se asume que existe una relación de dependencia logística entre la característica de interés y una variable de información auxiliar dicotómica. En el gráfico VIII.4 se observa que, al utilizar un estimador de calibración lineal, los residuales muestran un comportamiento inapropiado. Ambos estimadores resultan insesgados e igualmente eficientes para el parámetro de interés.

■ Gráfico VIII.4
Comportamiento de los estimadores en una relación de dependencia logística



Fuente: Elaboración propia.
[a] La línea roja vertical indica el valor del parámetro poblacional.

Es posible demostrar que la razón entre la varianza del estimador de calibración y la varianza del estimador de Horvitz-Thompson está supeditada al coeficiente de determinación R^2_ξ en un modelo de regresión lineal simple ξ entre la característica de interés y la información auxiliar:

$$\frac{Var(\hat{t}_{y,cal})}{Var(\hat{t}_{y,HT})} = (1-R^2_\xi+o(\sqrt{n})) \approx (1-R^2_\xi)$$

Por ende, el uso de la metodología de calibración supone casi siempre una ganancia de eficiencia en la estrategia de muestreo.

2. Tipos de estimadores de calibración

La calibración es un ajuste de los pesos de muestreo para que las estimaciones de algunas variables de control reproduzcan de forma perfecta los totales poblacionales de estas variables. Sin embargo, es necesario tener en cuenta las diferencias entre los métodos de calibración, que en general corresponderán al nivel de desagregación de la información auxiliar:

i) Calibración con variables continuas: la calibración se realiza con los totales de variables continuas como ingreso o gasto, entre otras.

ii) Posestratificación con variables categóricas: la calibración se realiza con los tamaños poblacionales (basados en proyecciones demográficas o registros administrativos) de subgrupos de interés.

iii) *Raking* con variables categóricas: calibración sobre los tamaños marginales de las tablas de contingencia de subgrupos de interés. Dado que, a diferencia del caso anterior, esta calibración solo tiene en cuenta los tamaños marginales y no los tamaños de los cruces, este método supone menos restricciones.

a) Posestratificación

La posestratificación es una de las técnicas más utilizadas para el ajuste de los pesos de muestreo mediante la calibración. Dicho ajuste requiere la definición de categorías poblacionales, según criterios como el grupo de edad, la región y la etnia o raza, entre otros. Este método se implementa dentro de cada uno de los cruces definidos por las covariables de interés (edad, región, etnia o raza). Es necesario tener acceso a la información auxiliar a nivel de todos los cruces definidos por los subgrupos que, en general, corresponden a proyecciones demográficas. En este caso, la suma de los pesos ajustados reproducirá con exactitud los tamaños poblacionales en cada cruce.

Las categorías formadas para definir los pesos de muestreo se conocen como "posestratos", pues se definen después de la selección de la muestra y la recopilación de los datos. Esto constituye una ventaja, ya que esas variables no necesariamente se tienen en cuenta en la planificación del diseño de muestreo. Por ejemplo, en una encuesta de hogares, es difícil estratificar por etnia o raza, edad, sexo o nivel educativo alcanzado. Como se sabe que estas variables pueden estar correlacionadas con la pobreza, el ingreso o la ocupación, sería una buena idea contemplarlas en la calibración. Suponiendo la definición de cuatro categorías para la raza, dos para el sexo y cinco para la edad, se obtendrían 40 posestratos y sus respectivas restricciones.

De esta manera, siendo $g=1,\ldots,G$ el indicador del cruce poblacional (posestrato), el estimador de posestratificación quedaría definido de la siguiente manera:

$$\hat{t}_{y,pos}=\sum_{g=1}^{G}\frac{N_g}{\hat{N}_g}\hat{t}_{y_g}$$

Donde N_g corresponde al tamaño poblacional del posestrato, $\hat{N}_g = \sum_{sg} d_k$ y $\hat{t}_{y_g} = \sum_{s_g} d_k \, y_k$. En virtud de lo expuesto, el factor de expansión del estimador posestratificado queda definido como sigue:

$$w_k = d_k \frac{N_g}{\hat{N}_g} \qquad (k \in s_g)$$

Donde d_k corresponde al peso determinado por el diseño de muestreo, corregido por los ajustes de elegibilidad y la falta de respuesta, que se presentan en los capítulos IX y XII de este documento.

La cantidad de posestratos en la calibración depende de la cantidad de interacciones entre las variables auxiliares. En algunos casos, es posible encontrar cientos de interacciones. Aunque el tamaño de los posestratos se reproduce sin error, ello puede disminuir la eficiencia de la calibración en las variables de interés. Es decir, la existencia de muchas variables e interacciones aumenta la inestabilidad de las estimaciones, sobre todo si existen cruces con celdas vacías. Es posible que el efecto de las interacciones incida en la creación de los nuevos pesos calibrados y se obtengan datos atípicos en los pesos de calibración resultantes.

b) *Raking*

¿Qué sucede si los conteos poblacionales (información auxiliar) no están disponibles para todos los cruces de las variables de calibración? Es posible que los agregados poblacionales de las variables provengan de distintas fuentes y no se pueda llegar al nivel de cruce. En este caso, es factible calibrar los marginales de la tabla cruzada, sin necesidad de calibrar todas sus entradas. Asimismo, el número de restricciones disminuiría con respecto a la posestratificación, pues se sumaría el número de categorías, mientras que en la posestratificación se multiplican. En el escenario anterior, en el que se suponen cuatro categorías para la raza, dos para el sexo y cinco para la edad, habría únicamente 11 restricciones.

Para ajustar los marginales de la tabla cruzada, es necesario realizar un procedimiento iterativo de ajuste proporcional, que no tiene una escritura cerrada (Gutiérrez, 2016, cap. 10). Por ejemplo, si el *raking* es de dos marginales, se ajustan primero las filas, luego las columnas, y así sucesivamente, hasta alcanzar la convergencia de los pesos calibrados. El procedimiento se detiene cuando se alcanza una tolerancia prefijada. Sin pérdida de generalidad, en el marco de este enfoque, los pesos calibrados se escriben de la siguiente manera:

$$w_k = d_k \times exp\,(u_h) \times exp\,(v_g)$$

Donde u_h es una función de los totales marginales de las filas de la tabla cruzada y v_g es una función de los totales marginales de las columnas. El *raking* permite utilizar variables que pueden predecir las variables de interés o explicar la probabilidad de responder del hogar (o la persona), además de paliar los efectos negativos que las bajas tasas de cobertura del marco de muestreo tienen en la inferencia.

3. La calibración como cambio de paradigma en una teoría de estimación exhaustiva

Särndal (2007) concluye que existen algunas ideas en las cuales vale la pena profundizar un poco más. A continuación se reproducen algunos conceptos importantes para destacar el uso práctico de los estimadores de calibración:

i) La calibración no puede separarse de la práctica. Inicialmente, los métodos de ponderación de las oficinas nacionales de estadística (ONE) se basaban en la ponderación de unidades mediante el inverso de su probabilidad de inclusión y, posteriormente, pasaron a basarse en las ponderaciones derivadas del enfoque de posestratificación. Las ponderaciones de calibración constituyen una extensión de las ideas anteriores. Aunque las referencias a la calibración en la literatura son recientes, su empleo como técnica para producir ponderaciones no lo es.

ii) La calibración no puede separarse de la coherencia estética. Dado que las ecuaciones de calibración imponen esta propiedad de coherencia al vector de ponderaciones, la aplicación de dichos ponderadores a las variables auxiliares producirá un resultado idéntico al de los totales poblacionales de estas variables de información auxiliar. El deseo de promover la credibilidad de las estadísticas oficiales es una razón para que las entidades busquen este tipo de propiedades.

iii) La calibración debe ser de fácil interpretación. El enfoque de calibración ha ganado popularidad en las aplicaciones reales debido a que las estimaciones resultantes son fáciles de interpretar y motivar, pues están directamente relacionadas con los pesos determinados por el diseño de muestreo. La calibración sobre los totales conocidos brinda al usuario una forma de estimación natural y transparente. El usuario que entiende la ponderación muestral aprecia el método de calibración, ya que este modifica sutilmente los pesos originales, pero, al mismo tiempo, respeta los totales de la información auxiliar. Además, la calibración determina un único vector de ponderaciones aplicable a todas las variables incluidas en el estudio. Por esta razón, se trata de un método muy valorado en las entidades oficiales que se ocupan de encuestas muy extensas.

iv) La calibración constituye un enfoque exhaustivo y unificado, basado en los avances de teorías anteriores. En la práctica de las encuestas de hogares, es común encontrar problemas como la falta de respuesta, defectos del marco muestral y errores de medición. Aunque algunos procesos como la imputación y la reponderación se conocen y utilizan ampliamente, estos métodos no necesariamente se enmarcan en una teoría exhaustiva de inferencia en poblaciones finitas. En la mayoría de los artículos teóricos, se aborda la estimación de parámetros en un mundo ideal, en el que la falta de respuesta y otros errores no muestrales no existen. La calibración proporciona una teoría unificada para superar estos inconvenientes.

E. Estimadores compuestos

Es posible mejorar la estimación del total actual $t_y^{(l)}$ teniendo en cuenta la información generada por el traslape[2] de la encuesta en el segundo período, de manera que:

$$\hat{t}^{K}_{y(t)}=(1-K)\ \hat{t}_{y(t)}+K(\hat{t}_{y(t-1)}+\hat{\Delta}_M)$$

Donde $0<\alpha<1$ y $\hat{\Delta}_M=\hat{t}^{M}_{y(t)}-\hat{t}^{M}_{y(t-1)}$ es la diferencia entre las estimaciones actual y previa en la muestra traslapada. Por otra parte, es posible añadir un término que dé cuenta de la diferencia entre las estimaciones actuales de las muestras con y sin traslape. De esta forma, se obtiene un estimador compuesto como el siguiente:

$$\hat{t}^{AK}_{y(t)}=\hat{t}^{K}_{y(t)}+A(\hat{t}^{U}_{y(t)}-\hat{t}^{M}_{y(t)})$$

Steel y McLaren (2008) afirman que esta clase de estimadores puede generar ganancias de eficiencia porque aprovecha la correlación entre las estimaciones del mismo panel a lo largo del tiempo y puede otorgar una ventaja adicional a la estimación tradicional que utiliza únicamente la muestra actual en su conjunto, sin detenerse en su composición: parte traslapada y parte no traslapada. Estos estimadores permiten usar diferentes valores para las constantes A y K, que pueden elegirse con el objetivo de minimizar la varianza del estimador o pueden seleccionarse de forma separada para aumentar la eficiencia del estimador con respecto a estimadores de niveles, como totales o proporciones del período actual, o de cambios temporales. Gurney y Daly (1965) utilizan $A=0,4$ y $K=0,2$ en una aplicación con paneles rotativos.

Estos estimadores también pueden escribirse como estimadores de regresión o calibración. En efecto, Gambino, Kennedy y Singh (2001) proponen que, además de las covariables y las restricciones de calibración habituales, se utilice la siguiente covariable de calibración para estimar indicadores de nivel:

$$x_k^{(l)}=\begin{cases} y_k^{(t-1)}, & k\in s_U \\ \hat{y}^{(t-1)}, & k\in s_M \end{cases}$$

Por ejemplo, si se desea estimar la proporción actual de personas ocupadas, $x_k^{(l)}$ representará la covariable que se deberá añadir al sistema de calibración; $\hat{y}^{(t-1)}$ será la proporción estimada de personas ocupadas en el período anterior, e $y_k^{(t-1)}$ representará una variable dicotómica sobre la población económicamente activa, que toma el valor 1 si la persona estuvo ocupada en el período anterior o 0 en caso contrario. En este caso, el total de control correspondiente estará definido por el total nacional estimado de personas ocupadas en el período anterior. Por lo tanto, como es de esperar, la suma ponderada (con los pesos de calibración) de esta covariable deberá ser igual a su total de control.

---

2 En esta sección, se utilizan los símbolos *M (matched)* y *U (unmatched)* en subíndice o superíndice para referirse, respectivamente, a las partes traslapada s_M y no traslapada s_U de la muestra actual.

Para los estimadores de cambio, Gambino, Kennedy y Singh (2001) proponen la siguiente covariable:

$$x_k^{(c)} = \begin{cases} y_k^{(t)}, & k \in s_U \\ y_k^{(t)} - R\,(y_k^{(t)} - y_k^{(t-1)}), & k \in s_M \end{cases}$$

Donde $R=\sum_s w_k / \sum_{s_M} w_k$ es el inverso de la proporción de traslape real en la muestra. Siguiendo con el ejemplo, $y_k^{(t)}$ representa una variable dicotómica sobre la población económicamente activa que toma el valor 1 si la persona está ocupada en el período actual o 0 en caso contrario. En este caso, el total de control correspondiente sigue estando definido por el total nacional estimado de personas ocupadas en el período anterior. Por lo tanto, al sumar sobre la muestra expandida, se presenta la siguiente relación $\hat{t}_{y(t-1)} = \hat{t}_{y(t)} - \hat{\Delta}_{M}$, es decir, que la estimación del período anterior es igual a la estimación del mes actual menos una diferencia de estimaciones en la muestra traslapada.

Si bien es posible usar ambas covariables $x_k^{(c)}$ y $x_k^{(l)}$ en el sistema de calibración, es probable que los pesos resultantes sean negativos o inferiores a 1. Para evitar estos inconvenientes numéricos, y dado que para ambas variables se tiene el mismo total de control, es posible definir una combinación lineal convexa y crear una nueva covariable de calibración, de la siguiente manera:

$$x_k=(1-\alpha)\, x_k^{(l)}+\alpha x_k^{(c)}$$

Según Fuller y Rao (2001), pese a que en algunas aplicaciones particulares se ha encontrado mayor eficiencia (menor varianza) al utilizar valores de $\alpha=0,65$ o $\alpha=0,75$, se recomienda usar $\alpha=2/3$ en los sistemas de producción de las ONE.

De acuerdo con Gambino, Kennedy y Singh (2001), estos estimadores son fáciles de implementar porque concuerdan con la forma habitual del sistema tradicional de ponderación en las encuestas de hogares; es decir, se crea una nueva columna en la base de datos que define el nuevo factor de expansión, y esto se logra simplemente con la adición de nuevas covariables a la matriz de calibración.

En 2019, el Instituto Nacional de Estadística (INE) del Uruguay decidió rediseñar la Encuesta Continua de Hogares, que se había definido en 2006 sobre la base de la selección de muestras mensuales independientes. La nueva metodología prevé un sistema de panel rotativo, en el que la muestra de un determinado mes está compuesta por seis grupos de rotación y la recopilación de la información tiene un carácter mixto: la primera encuesta se realiza de forma presencial y el seguimiento, que dura cinco meses, de forma telefónica. Además, el INE implementó un sistema de estimación basado en los estimadores compuestos con mejoras en la eficiencia estadística (Macari y Ferreira, 2020).

Capítulo IX
Construcción de los factores de expansión

En todas las bases de datos de las encuestas de hogares, se encuentra una columna que contiene los pesos de muestreo o factores de expansión. Con esta columna se realizan todos los análisis requeridos en la encuesta, desde la estimación de medias, razones, tamaños y proporciones hasta el ajuste de modelos lineales y no lineales. La razón principal por la que se utilizan los factores de expansión es la necesidad de producir estimaciones que reflejen de manera precisa el comportamiento de la población objetivo. El uso correcto de los factores de expansión garantiza que la estimación sea insesgada y coherente, que el error de muestreo sea pequeño, condicionado al diseño muestral y al tamaño de la muestra, y que se corrijan las deficiencias de cobertura del marco de muestreo.

La naturaleza de los factores de expansión es intuitiva y se da en el marco del principio de representatividad que gobierna la inferencia de las encuestas de hogares y cualquier otra operación estadística basada en la selección de una muestra. De esta forma, el factor de expansión de una unidad muestral representa el número de veces que se representa a sí misma y a otras unidades similares a ella misma. En general, en condiciones de regularidad, el factor de expansión será siempre positivo y mayor que la unidad. Además, la suma de los factores de expansión de la base de datos deberá aproximarse al tamaño de la población sobre la que se desea realizar la inferencia.

Por ejemplo, un hogar en una encuesta con un factor de expansión de 500 se representa a sí mismo y a otros 499 hogares. La definición teórica del factor de expansión, que surge del inverso multiplicativo de la probabilidad de inclusión de un hogar en la muestra, hace que la inferencia sea insesgada y confiable. Sin embargo, debido a que la probabilidad de inclusión es un número real contenido en el intervalo $(0,1)$, su inverso multiplicativo también será un número real mayor o igual que uno. Asimismo, si en el país

hay alrededor de 4 millones de hogares, se espera que la suma de los factores de expansión de la muestra de hogares se sitúe en torno a esta cifra.

Los procesos de inferencia estadística establecidos en cualquier encuesta de hogares descansan sobre el principio de representatividad, que sostiene que es posible seleccionar una muestra y representar con bastante precisión y exactitud la realidad de la población de interés. A su vez, las propiedades estadísticas de la inferencia en las encuestas de hogares descansan sobre las probabilidades de inclusión determinadas por el diseño de muestreo que se implementó en la encuesta. En general, el peso de muestreo dk asociado a un individuo k en la muestra s es función de la probabilidad de inclusión del individuo, así:

$$d_k = \frac{1}{Pr\,(k \in s)}$$

Como se mencionó anteriormente, para conservar la estabilidad en los pesos de muestreo, es posible definir diseños de muestreo autoponderados, donde las unidades finales de muestreo tengan la misma probabilidad de inclusión, sin importar el tamaño de la unidad primaria de muestreo que las contiene. Este tipo de diseño es útil porque implica un mayor control sobre las estimaciones finales. Además, Valliant y Dever (2017) afirman que los pesos de muestreo se utilizan con los siguientes fines: i) incorporar las probabilidades de selección de las unidades en la muestra; ii) ajustar en casos en que no se pueda determinar si algunas unidades de la muestra son miembros de la población de interés; iii) minimizar el sesgo causado por la falta de respuesta cuando algunas unidades no responden aunque estén incluidas en la muestra; iv) incorporar información auxiliar externa para reducir los errores muestrales de las estimaciones, y v) compensar cuando la muestra no cubre correctamente a la población de interés.

Cabe destacar que la conformación de los pesos de muestreo se transforma en un reto metodológico para el investigador, pues debe ajustarse a la realidad de la región, en el sentido de que las poblaciones de los municipios se expanden cada vez más en el sector urbano y los marcos de muestreo de las áreas geográficas se desactualizan con rapidez. Se han planteado varias soluciones a este problema (Gambino y Silva, 2009) y todas requieren esfuerzos económicos, logísticos y técnicos. Por ende, los equipos de los institutos nacionales de estadística (INE) (a todo nivel) deben ser flexibles y adecuarse a esta realidad cambiante de la movilidad de las poblaciones, sobre todo en las áreas urbanas.

En condiciones ideales, el marco de muestreo debería coincidir plenamente con la población finita. Sin embargo, en general, no es posible contar con una lista de todos los elementos de la población y, en el contexto de las encuestas de hogares, no existe un listado que enumere todos los hogares de un país de manera actualizada. De ahí que la práctica estándar sea construir el marco de muestreo en varias etapas, comenzando por la selección de una muestra de áreas geográficas y un empadronamiento exhaustivo de todos los hogares en las áreas seleccionadas, seguido de la selección de hogares. Este mecanismo hace que el marco de muestreo de las encuestas de hogares presente imperfecciones.

Para hacer frente a esas imperfecciones, Valliant y Dever (2017) recomienda utilizar los códigos de disposición estandarizados por la American Association for Public Opinion Research (AAPOR), tratar la falta de respuesta de manera diferenciada y clasificar cada unidad de la muestra en alguna de las siguientes categorías:

i) ER (*eligible respondents*): unidades elegibles que fueron respondientes efectivos y que constituyen los casos elegibles sobre los que se ha recopilado una cantidad suficiente de información.

ii) ENR (*eligible nonrespondents*): unidades elegibles no respondientes que constituyen los casos elegibles sobre los que no se recopiló ningún dato o se obtuvo información parcial.

iii) IN (*ineligibles*): unidades no elegibles que constituyen los casos de miembros no elegibles que no forman parte de la población de interés.

iv) UNK (*unknown eligibility*): unidades con elegibilidad desconocida que constituyen los casos en que no se puede conocer si la unidad es elegible o no.

Para construir los factores de expansión de una encuesta se recomienda seguir los siguientes procesos en el orden indicado:

i) Creación de los pesos básicos.

ii) Ajuste por elegibilidad desconocida.

iii) Descarte de las unidades no elegibles.

iv) Ajuste por falta de respuesta.

v) Calibración por proyecciones poblacionales y variables auxiliares.

vi) Recorte y redondeo de los factores finales (opcional).

A. Creación de los pesos básicos

Este primer paso ya se ha explicado de forma detallada en el capítulo dedicado a la selección de la muestra. Obsérvese que, asociada a cada diseño particular de muestreo, existe una única función que vincula cada elemento con una probabilidad de inclusión en la muestra. De esta forma:

$$\pi_k = Pr\ (k \in s)$$

Por lo tanto, el primer paso en la reponderación de los pesos de muestreo es justamente la creación de los pesos básicos d_{1k}, que se definen como el inverso multiplicativo de la probabilidad de inclusión:

$$d_{1k} = \frac{1}{\pi_k} \qquad \forall k \in s$$

Estos pesos se crean incluso para aquellas unidades que se excluirán de la muestra porque son no elegibles o porque no proporcionaron ninguna información y después se modificarán como resulte pertinente. Se recomienda calcular las probabilidades de inclusión (si el muestreo es sin reemplazo) o selección (si el muestreo es con reemplazo) a medida que se avance en las etapas del muestreo. De esta forma, siempre se puede confirmar la coherencia de los pesos en cada etapa y de los pesos finales.

A partir de las modificaciones posteriores sobre este peso de muestreo, la distribución de los ponderadores irá sufriendo algunos cambios. Si la distribución original de los pesos básicos difiere desde el punto de vista estructural de la distribución final de los ponderadores, resultante de todos los ajustes debidos a las imperfecciones del marco, podrían desvanecerse las propiedades estadísticas de insesgamiento, consistencia y precisión. Ello significa que el nivel de desactualización del marco de muestreo tiene implicaciones directas en la calidad de la inferencia. Es decir, si el marco de muestreo es muy imperfecto, los ponderadores finales no permitirán realizar una inferencia precisa.

B. Ajuste por elegibilidad desconocida

El segundo paso consiste en redistribuir el peso de las unidades cuyo estado de elegibilidad es desconocido. Por ejemplo, si la encuesta está orientada a la población mayor de 15 años y hay personas que no proporcionan ninguna información acerca de su edad, se hace necesario distribuir estos pesos. Esta situación también se puede presentar a nivel del hogar, cuando este no puede ser contactado porque nadie atiende el llamado del encuestador ("nadie en casa"). Se acostumbra redistribuir los pesos de las unidades con elegibilidad desconocida (UNK) entre las unidades de las que sí se conoce su estado de elegibilidad (ER, ENR, IN).

Por consiguiente, si no es posible determinar la elegibilidad de algunas unidades que aparecen en el marco de muestreo, se tendrá una muestra s que contendrá el subconjunto de las unidades elegibles respondientes (s_{ER}), el subconjunto de las unidades elegibles no respondientes (s_{ENR}), el subconjunto de las unidades no elegibles (s_{IN}) y el subconjunto de las unidades con elegibilidad desconocida (s_{UKN}). En este último caso, la elegibilidad de estas unidades se mantiene desconocida, a no ser que de manera arbitraria sean clasificadas como ENR (elegibles no respondientes), o se tenga información auxiliar en el marco de muestreo que permita asignarles un estado de elegibilidad.

Se recomienda formar $B\ (b=1,\ldots,B)$ categorías basadas en la información del marco de muestreo[1]. Estas categorías pueden ser estratos o cruces de subpoblaciones. Siendo s_b la muestra de unidades en la categoría b (que incluye ER, ENR, IN y UNK), se define el factor de ajuste por elegibilidad como:

$$a_b = \frac{\sum_{s_b} d_{1k}}{\sum_{s_b \cap (s_{ER} \cup s_{ENR} \cup s_{IN})} d_{1k}}$$

[1] Valliant y Dever (2017) recomiendan formar categorías con al menos 50 casos.

En la categoría b, los pesos ajustados por elegibilidad desconocida para aquellas unidades cuya elegibilidad sí pudo ser establecida (independientemente de su estado de respuesta) estarán dados por la siguiente expresión:

$$d_{2k} = a_b * d_{1k} \qquad \forall k \in s_b \cap (s_{ER} \cup s_{ENR} \cup s_{IN})$$

C. Descarte de las unidades no elegibles

En esta etapa, tanto las viviendas con elegibilidad desconocida (UNK) como las que han cambiado su estado de ocupación y ahora no contienen ningún hogar particular (IN) se retirarán de la población objetivo. Por ende, la base de datos con la que se prosigue el proceso presentará ahora un menor número de unidades. Este tercer paso consiste en ajustar el peso de la etapa anterior de la siguiente manera:

$$d_{3k} = \begin{cases} 0 & \text{si la unidad } k \in (s_{UNK} \cup s_{IN}) \\ d_{2k} & \text{si la unidad } k \in (s_{ER} \cup s_{ENR}) \end{cases}$$

D. Ajuste por falta de respuesta

En este paso, los pesos de los respondientes efectivos (ER) se ajustan para tener en cuenta a los que no respondieron (ENR). Al final del proceso, los pesos de los ER se incrementan para compensar el hecho de que algunas unidades elegibles no proporcionaron información. Para el manejo efectivo de la falta de respuesta, se consideran las siguientes variables aleatorias:

$$I_k = \begin{cases} 1 & \text{si } k \in (s_{ER} \cup s_{ENR}) \\ 0 & \text{en caso contrario} \end{cases}$$

$$D_k = \begin{cases} 1 & \text{si } k \in s_{ER} \\ 0 & \text{si } k \in s_{ENR} \end{cases}$$

Al suponer que la distribución de las respuestas puede ser estimada, entonces la probabilidad de respuesta o puntaje de propensión (*propensity score*) viene determinada por:

$$Pr[k \in s_{ER} \mid k \in (s_{ER} \cup s_{ENR})] = Pr[D_k = 1 \mid I_k = 1] = \phi_k$$

Si el patrón de falta de respuesta es completamente aleatorio (la no respuesta no sigue ningún patrón específico) o aleatorio (el patrón de la no respuesta puede explicarse por un conjunto de covariables $\boldsymbol{z}$), entonces:

$$\phi_k = f(\boldsymbol{z}_k, \boldsymbol{\beta}) \qquad \forall k \in (s_{ER} \cup s_{ENR})$$

De esta forma, si fuese plausible tener acceso a las covariables $\mathbf{z}$ para los individuos elegibles en la muestra, se podría estimar el patrón de falta de respuesta mediante la siguiente relación funcional:

$$\hat{\phi}_k = f(\mathbf{z}_k, \hat{\boldsymbol{\beta}}) \qquad \forall k \in (s_{ER} \cup s_{ENR})$$

Por otro lado, si el patrón de falta de respuesta es no aleatorio (la misma estructura de la falta de respuesta es explicada por la variable de interés; por ejemplo, cuando en una encuesta de mercado laboral son los desempleados quienes no responden), entonces:

$$\phi_k = f(\mathbf{y}_k, \beta) \qquad \forall k \in (s_{ER} \cup s_{ENR})$$

En este caso, como no es posible tener acceso a las variables de interés para todos los individuos de la muestra de unidades elegibles (precisamente porque no todos respondieron), no es posible estimar el patrón de falta de respuesta y, por consiguiente, habrá problemas de sesgo. Por otra parte, Kim y Riddles (2012) muestran que es posible utilizar un modelo basado en la estimación de las probabilidades de respuesta o puntaje de propensión (*propensity score*). De esta forma, teniendo en cuenta que la probabilidad de que un individuo conteste es $\phi_k = Pr(k \in s_{ER})$, al suponer que existe acceso al vector de información auxiliar $\mathbf{z}_k$ conocido para todo $k \in (s_{ER} \cup s_{ENR})$, es posible estimarla, por ejemplo, mediante un modelo de regresión logística; es decir:

$$\hat{\phi}_k = \frac{exp\{\mathbf{z}_k, \hat{\boldsymbol{\beta}}\}}{1 + exp\{\mathbf{z}_k, \hat{\boldsymbol{\beta}}\}} \qquad \forall k \in (s_{ER} \cup s_{ENR})$$

Donde $\hat{\boldsymbol{\beta}}$ es el vector de coeficientes estimado de la regresión logística. Por tanto, si la falta de respuesta no depende de la variable de interés, puede definirse el siguiente estimador insesgado:

$$\hat{t}_y = \sum_{k \in s_{ER}} d_{4k} y_k$$

Donde:

$$d_{4k} = \frac{d_{3k}}{\hat{\phi}_k} \qquad \forall k \in s_{ER}$$

Es posible aumentar la eficiencia del estimador si se crean categorías homogéneas de individuos que tengan la misma probabilidad de responder. En este caso, los valores de las covariables pueden utilizarse para crear estas categorías. Por consiguiente, siempre es necesario obtener un conjunto de covariables que esté disponible para respondientes y no respondientes a la vez.

Por ejemplo, considérese un escenario simplificado en que es posible determinar que la probabilidad de responder está relacionada únicamente con las variables de edad (cinco categorías) y sexo (dos categorías). En este caso, sería posible formar $Q = 10$ $(q = 1, \ldots, Q)$ categorías de acuerdo con el cruce de estas variables para obtener una estimación de la

probabilidad de respuesta en cada clasificación y ajustar el peso de muestreo. De esta manera, siendo s_q la muestra seleccionada en la categoría q, la probabilidad de respuesta o puntaje de propensión en esta categoría se estimaría como:

$$\phi_q = \frac{\sum_{s_{ER} \cap s_q} d_{3k}}{\sum_{s_q} d_{3k}}$$

El nuevo peso ajustado por la falta de respuesta estará dado por:

$$d_{4k} = \frac{d_{3k}}{\phi_q} = d_{3k} \frac{\sum_{s_q} d_{3k}}{\sum_{s_{ER} \cap s_q} d_{3k}}$$

En un escenario más complejo, si las probabilidades de respuesta fueron estimadas con un modelo de puntaje de propensión, y teniendo en cuenta que las predicciones de estas probabilidades varían entre cero y uno, es posible crear clases de individuos (respondientes y no respondientes) con probabilidades similares. En este caso, se asumiría que las unidades dentro de una misma clase tendrán la misma configuración de covariables, o al menos, una probabilidad de respuesta estimada similar $\hat{\phi}_k$. Así, dentro de cada clase, las unidades serían tratadas como si hubiesen sido aleatorizadas para pertenecer al grupo de respondientes (tratamiento) o al grupo de no respondientes (control).

Por lo tanto, el objetivo de este proceso es asegurar que se pueda ajustar cualquier diferencia en las covariables. Si el modelo es adecuado, la estimación $\hat{\phi}_k$ resumiría los efectos de las covariables en la respuesta del individuo. Teniendo esto en cuenta, una vez que hayan sido creadas las clases, es posible realizar el ajuste mediante alguna medida de localización en cada clase y, de esta forma, todos los individuos de una misma clase se ajustarían de la misma manera. Partiendo del supuesto de que se pueden crear C clases y que s_c es la muestra de n_c unidades elegibles en la clase c $(c=1,2,\ldots,C)$, es posible utilizar alguna de las siguientes medidas (Valliant y Dever, 2017):

i) Promedio no ponderado:

$$\hat{\phi}_c = \frac{\sum_{k \in s_c} \hat{\phi}_k}{n_c}$$

ii) Promedio ponderado:

$$\hat{\phi}_c = \frac{\sum_{k \in s_c} d_{3k} \hat{\phi}_k}{n_c}$$

iii) Mediana no ponderada:

$$\hat{\phi}_c = mediana[\hat{\phi}_k] \quad \forall k \in s_c$$

iv) Tasa de respuesta no ponderada:

$$\hat{\phi}_c = \frac{\#(s_{ER} \cap s_c)}{n_c}$$

v) Tasa estimada de respuesta:

$$\hat{\phi}_c = \frac{\sum_{s_c \cap s_{ER}} d_{3k}}{\sum_{s_c} d_{3k}}$$

Cabe mencionar que, si todas las unidades dentro de una clase tienen la misma probabilidad de responder, la tasa de respuesta no ponderada es la mejor opción. Además, si, dentro de las clases, las unidades tienen una probabilidad de responder muy disímil, significa que puede usarse el promedio no ponderado (o ponderado) del puntaje de propensión. De la misma manera, la tasa estimada de respuesta puede ser ineficiente si los pesos de muestreo varían demasiado, pero la probabilidad de respuesta o puntaje de propensión es similar en cada clase. Por último, la mediana se considera si la distribución de la probabilidad de respuesta es sesgada.

E. Calibración de los factores de expansión

Después de conformar el sistema de ponderación de pesos de muestreo en la encuesta, es posible calibrar estos pesos con la información auxiliar disponible de cada país, a nivel nacional, por estratos de interés, e incluso por las variables continuas sobre las que se tenga interés. Särndal y Lundström (2005) afirman que, cuando los estudios por muestreo están afectados por la falta de respuesta, es deseable que la estructura inferencial que sustenta la encuesta favorezca estimadores con sesgo pequeño o nulo y con errores estándar pequeños. A su vez, durante décadas, los INE han preferido sistemas de pesos capaces de reproducir la información auxiliar disponible y eficientes a la hora de estimar cualquier característica de interés en un estudio multipropósito[2].

De ahora en adelante, y con el fin de simplificar la notación estadística, se denotará indistintamente la muestra de unidades elegibles respondientes como s. Como se vio en los capítulos anteriores, debido a la construcción teórica de los estimadores de calibración, los pesos calibrados responden a la siguiente restricción:

$$\sum_{k \in s} w_k \boldsymbol{x}_k = \boldsymbol{t}_{\boldsymbol{X}}$$

El ejemplo más básico se tiene cuando se desea que los pesos de muestreo reproduzcan con exactitud el tamaño de las regiones N_h de un país, o el tamaño del país N. Así pues, utilizar la metodología de calibración (Deville y Särndal, 1992) hace que se cumpla la siguiente ecuación de calibración sobre los nuevos pesos calibrados w_k para todos los estratos explícitos:

$$\sum_{sh} w_k = n_h$$

[2] Esta información se refiere, por ejemplo, al número de hogares o habitantes del país.

Gutiérrez (2016) menciona que esta coherencia entre las cifras oficiales y las que puede producir la encuesta hace que sea preferible el uso de los estimadores de calibración. Las anteriores características se satisfacen al utilizar el enfoque de calibración que conlleva una estructura inferencial robusta en presencia de información disponible, puesto que reduce tanto el error de muestreo como el error debido a la falta de respuesta. Una vez que se ha ejecutado el proceso de calibración, se crean nuevos pesos que, en general, pueden escribirse así:

$$w_k = g_k * d_{4k} \quad \forall k \in s$$

Donde los valores g_k son dependientes de la muestra seleccionada s y de la función de optimización escogida para realizar el proceso de calibración. En general, no tienen una forma cerrada. No obstante, dependiendo de la estructura en la información auxiliar, pueden tomar valores particulares. Por ejemplo, con la distancia de ji al cuadrado (χ^2), el estimador de calibración tomará la siguiente forma, dada por:

$$\hat{t}_{y,cal} = \sum_{k \in s} w_k y_k = \hat{t}_{y,\pi} + (\boldsymbol{t_x} - \hat{\boldsymbol{t}}_{\boldsymbol{x},\pi})\widehat{\boldsymbol{B}}_s$$

Donde $\hat{B}_s$ es un vector de coeficiente de regresión dependiente de la muestra s y de constantes q_k, cuya forma funcional se presenta a continuación:

$$\hat{\boldsymbol{B}}_s = \left(\sum_s w_k q_k \boldsymbol{x}_k \boldsymbol{x}_k' \right)^{-1} \sum_s w_k q_k \boldsymbol{x}_k y_k$$

En este caso particular, los ponderadores g_k se pueden escribir como sigue:

$$g_k = 1 + (\boldsymbol{t_x} - \hat{\boldsymbol{t}}_{\boldsymbol{x},\pi}) \left(\sum_s w_k q_k \boldsymbol{x}_k \boldsymbol{x}_k' \right)^{-1} \sum_s w_k q_k \boldsymbol{x}_k$$

Nótese que los estimadores de calibración son aproximadamente insesgados, pero la magnitud del sesgo está dada por la siguiente expresión:

$$Bias = (\hat{t}_{y,cal}) = E\left[\sum_{k \in s} (w_k - d_k) y_k \right]$$

Si los nuevos pesos calibrados son cercanos a los pesos originales en todas las posibles muestras, entonces el sesgo será insignificante. Ahora, si el tamaño de muestra es insuficiente, no conviene utilizar este tipo de estimadores. Además, se sugiere que el coeficiente de variación del estimador de Horvitz-Thompson para las covariables (dadas por todos los cruces y celdas considerados) sea inferior al 10% para asegurar que el sesgo de los estimadores de calibración resulte despreciable.

Por otro lado, cuando se tienen múltiples variables discretas, es posible que el cruce de categorías contenga muy pocas unidades para las cuales se deban ajustar los pesos originales. Esto puede producir un sesgo en algunos subgrupos. Si aun así se decide mantener estas múltiples restricciones de calibración, será necesario realizar una comprobación empírica del ajuste que cada modelo pueda tener con todas las variables de la encuesta.

1. Medidas de calidad en la calibración

Silva (2004) presenta algunas consideraciones respecto del sesgo que puede producirse al utilizar esta metodología en las encuestas de hogares y aborda algunos criterios para evaluar la calidad de la calibración. Estas medidas se pueden considerar como protección contra el sesgo generado por la imposición de demasiadas restricciones. Además, se resalta la importancia de que las variables aplicadas en la calibración sean estimadas de manera precisa por los estimadores clásicos de muestreo. Por ejemplo, si el número de personas en una región es una variable de calibración (utilizando como total auxiliar las proyecciones demográficas), el coeficiente de variación del estimador de Horvitz-Thompson sobre esta variable debería ser menor (por ejemplo, del 10%).

La teoría afirma que, cuantas más variables de calibración se tengan, menor será la varianza asociada a las estimaciones (no así el sesgo). Sin embargo, existen problemas computacionales cuando se deben satisfacer demasiadas restricciones. Una primera opción consiste en verificar que no haya variables que puedan tener codependencia lineal con otras. Al descartarlas, es posible conservar una varianza pequeña, puesto que se descartan combinaciones lineales de otras variables. Se recomienda hacer un análisis de cuántas variables deben utilizarse en la calibración para optimizar el error cuadrático medio de los estimadores finales en las encuestas de hogares.

En una primera instancia, no sería adecuado utilizar demasiadas restricciones de calibración para satisfacer las proyecciones demográficas de muchas variables. Es fácil equivocarse en esta definición. Por ejemplo, si la encuesta es representativa a nivel de departamento (diez niveles), sexo (dos niveles) y edad (cuatro niveles), podría ser contraproducente utilizar $10 \times 2 \times 4 = 80$ restricciones de calibración y se debería empezar por analizar una estrategia más parsimoniosa, con $10 + 2 + 4 = 16$ restricciones de calibración. Cabe tener en cuenta que, a medida que las desagregaciones sean más profundas, el nivel de error en las proyecciones poblacionales será más grande. Por otra parte, cuantas más restricciones haya, más sesgo y varianza se introducen en la estimación. La idea general del proceso es encontrar un número de restricciones parsimonioso que permita obtener estimaciones aproximadamente insesgadas, con una varianza inferior a la obtenida con los factores de expansión originales.

Por otro lado, si los pesos de calibración resultan ser menores que uno, su interpretación puede tornarse difícil (aunque ello no constituye un problema teórico). El usuario común entiende el factor de expansión como un factor de representatividad: es la cantidad de veces que una persona se representa a sí misma y a algunas otras en la población. Por ende, los pesos negativos o menores que uno no resisten esta interpretación intuitiva y natural. Además, los pesos negativos pueden conllevar estimaciones negativas en algunos dominios en que el tamaño de muestra es pequeño, lo cual resulta problemático en un contexto en que todas las variables de estudio son no negativas.

Con el fin de garantizar que los pesos se ubiquen en un intervalo determinado, se debe minimizar una distancia que, a su vez, debe conllevar pesos restringidos a este

intervalo y respetar las ecuaciones de calibración. Es posible que no se tenga una solución exacta para todas las restricciones de calibración e incluso que el algoritmo de calibración no converja.

Con base en lo anterior, es necesario analizar los pesos g_k en perspectiva en cada dominio, estrato y posestrato de interés. Una buena idea puede ser la de identificar aquellos g_k que resulten potencialmente grandes o influyentes. Se recomienda posestratificar la muestra, aplicar la calibración a aquellas unidades en que los g_k sean estables y utilizar los pesos originales en el conjunto restante.

Es posible lograr que los pesos de calibración estén restringidos a un espacio predefinido por el usuario, mediante límites (L,U) sobre los g_k. De esta forma, si $w_k \geq 1$, ello implica que $g_k \geq 1$ y, por tanto, $L=1$. Se acostumbra tomar $U > Q_3+1{,}5*(Q_3 - Q_1)$, donde Q_3 y Q_1 están dados en términos de la distribución de g_k y corresponden al tercer y primer cuartil, respectivamente.

Si el mecanismo que genera la falta de respuesta no es aleatorio (*missing at random* (MAR)) o completamente aleatorio (*missing completely at random* (MCAR)), es posible que los ponderadores de calibración produzcan un sesgo en las estimaciones finales. En general, cuando hay falta de respuesta es más probable que aparezcan pesos de calibración negativos y que los pesos de calibración no converjan hacia los pesos originales. Además, la varianza de los estimadores de calibración no convergerá hacia los resultados usuales de los estimadores de regresión.

Partiendo de que existen P variables de información auxiliar en las ecuaciones de calibración, Silva (2004) presenta algunas medidas que permiten decidir qué escenarios de calibración son los mejores. Si $\hat{t}_{x_{p,cal}}$ es el estimador de calibración para la p-ésima variable de información auxiliar cuyo total poblacional es t_{x_p}, se define el error relativo promedio sobre las variables auxiliares:

$$M1 = \frac{1}{P}\sum_{p=1}^{P}\frac{|\hat{t}_{x_{p,cal}} - t_{x_p}|}{t_{x_p}}$$

Se esperaría que $M1$ fuese nulo, si es que la distancia utilizada en la optimización de la calibración es la de χ^2. Sin embargo, como esta distancia puede arrojar valores negativos para los pesos de muestra, es preferible utilizar otro tipo de distancias que pueden no converger exactamente con los totales auxiliares. En este caso, es preferible escoger aquella distancia que menores valores arroje sobre $M1$, con sujeción a que los pesos tengan una interpretación adecuada.

Por otro lado, siendo $\hat{t}_{x,\pi}$ el estimador HT de la p-ésima variable de información auxiliar, el coeficiente de variación relativo promedio se define de la siguiente manera:

$$M2 = \frac{1}{P}\sum_{p=1}^{P}\frac{\sqrt{Var(\hat{t}_{x,\pi})}}{t_x}$$

El propósito de esta medida es que el investigador utilice variables de información auxiliar que estén bien representadas en la encuesta. De la misma manera, se define la proporción de ponderadores extremos menores que un límite inferior (L) predefinido, o mayores que un límite superior (U) predefinido. Estas medidas están, respectivamente, dadas por:

$$M3 = \frac{1}{n} \sum_{k \in s}^{P} I\,(g_k < L)$$

$$M4 = \frac{1}{n} \sum_{k \in s} I\,(g_k > U)$$

Como los pesos de calibración están definidos como la multiplicación de los pesos ajustados d_k con un ponderador g_k, no sería deseable que hubiese ponderadores extremos, muy alejados de la unidad. Ello sería indicio de que existe un alejamiento grave entre los pesos dados por el diseño y los nuevos pesos de calibración. Si se tiene en cuenta que el insesgamiento está garantizado al utilizar los pesos d_k, una proporción de valores g_k muy alejados sería una señal de alarma y podría indicar la existencia de sesgo.

Siendo $\sigma(g)$ la desviación estándar muestral de los ponderadores g_k y $\bar{g}$ su promedio muestral, otra medida de interés sería el coeficiente de variación de los ponderadores, que está supeditado a la siguiente expresión:

$$M5 = \frac{\sigma(g)}{\bar{g}}$$

Una dispersión muy alta de los ponderadores sería indeseable, puesto que indicaría que hay valores influyentes que alejarían los pesos calibrados de los pesos muestrales. Asimismo, se define la distancia de χ^2 entre los pesos de calibración y los pesos originales, dada por:

$$M6 = \frac{1}{n} \sum_{k \in s} \frac{(w_k - d_k)^2}{d_k} = \frac{1}{n} \sum_{k \in s} (g_k - 1)^2$$

Si la calibración fue exitosa, esta medida debe ser pequeña, lo que indicaría una cercanía de los pesos calibrados a los pesos originales y, a su vez, señalaría que el insesgamiento se mantiene. Por otro lado, la eficiencia de los estimadores de calibración puede calcularse sobre la base de la siguiente expresión:

$$M7 = \frac{1}{P} \sum_{p=1}^{P} \frac{Var\,(\hat{t}_{x_{p,cal}})}{Var(\hat{t}_{x_{p,\pi}})}$$

Nótese que esta medida es similar al efecto de diseño generalizado, que se utiliza para evaluar la eficiencia de los escenarios de estratificación. Siguiendo el mismo razonamiento, se quiere que esta medida sea pequeña y siempre menor que uno, lo que significa que la inferencia estadística al utilizar los pesos calibrados es mayor que con el estimador HT. Por último, como se indicó en los capítulos anteriores, el efecto de diseño debido a la ponderación desigual $DEFF^w$ se define como:

$$M8=1+cv^2(w_k)$$

En este caso, es deseable que esta medida sea muy cercana a uno, lo que indica que la dispersión de los pesos finales está controlada.

2. Calibración integrada para hogares y personas

Una de las preguntas recurrentes en la calibración de encuestas de hogares es a qué nivel se debería realizar este ajuste. En principio, es posible realizar la calibración al nivel de las personas o de los hogares. Cada una de estas opciones entraña algunas ventajas y consideraciones que se deben tener en cuenta:

- El hecho de calibrar al nivel de los hogares implica que el hogar tendrá unos nuevos pesos que cumplan con las restricciones de calibración, y esos pesos se aplicarán a las personas que lo habitan. De esta forma, todas las personas pertenecientes a un mismo hogar tendrán el mismo peso de muestreo, sin importar sus diferencias en cuanto a composición demográfica. Por ejemplo, hombres, mujeres, menores y mayores de 15 años tendrán el mismo peso de muestreo. Esta propiedad es atractiva porque emula el diseño de muestreo que se definió en la fase de planificación. Sin embargo, realizar la calibración a nivel de los hogares implica que, dentro de las unidades primarias de muestreo (UPM), los hogares no tendrán un peso homogéneo, lo que se distancia de las propiedades del diseño sistemático simple que se utiliza para la selección de los hogares dentro de las UPM.
- Por otro lado, la decisión de calibrar a nivel de las personas implica que los pesos de muestreo de los hogares también pueden verse alterados y que los pesos finales de muestreo de las personas serán diferentes dentro de los hogares. De esta forma, de acuerdo con las características de las personas, se tendrá un peso diferente. Por ejemplo, es posible que hombres, mujeres, menores y mayores de 15 años no tengan el mismo peso de muestreo. Por consiguiente, cuando se calibra por personas y se utiliza un filtro sobre esa base para crear una base de hogares, las características observadas de los jefes de hogares influirían en los pesos de muestreo resultantes.

Dado que la calibración puede generar factores de expansión diferentes para los miembros de un mismo hogar, es necesario analizar a qué nivel se realiza este procedimiento (persona u hogar). En principio, y debido al diseño de la encuesta, los pesos de muestreo originales son idénticos para todos los miembros de un mismo hogar. Sin embargo, cuando

en la posestratificación se trata de ajustar los totales de las restricciones de calibración, y debido a que la población no está equitativamente distribuida, también se presenta un reajuste en los factores de calibración. Podría ser conveniente revisar otras metodologías de calibración (por ejemplo, la de *raking*) y su impacto en los pesos de calibración dentro de los hogares.

Por ejemplo, si la calibración se realiza a nivel de las personas y se calibra sobre la población en edad de trabajar, esto traerá como consecuencia que los factores de expansión sean diferentes para los miembros de un mismo hogar. Ello se debe a que la metodología buscará ajustar los totales de las personas en edad de trabajar y las personas que no están en la fuerza de trabajo de manera independiente. Por esta razón, en la mayoría de los hogares, donde hay personas que son parte de la fuerza de trabajo y personas que no lo son, los pesos de muestreo no serán equivalentes.

En general, la mayoría de las encuestas de hogares en la región tienen una naturaleza multipropósito. Permiten obtener estimaciones de indicadores a nivel de persona (tasa de participación y tasa de desocupación, entre otros) y, al mismo tiempo, indicadores a nivel de hogar (pobreza monetaria, necesidades básicas insatisfechas y pobreza multidimensional). En este documento se hace hincapié en la recomendación de disponer de factores de expansión coherentes entre las diferentes unidades de análisis.

Por ejemplo, una práctica común que pone en tela de juicio las propiedades estadísticas del estimador consiste en generar factores de expansión al nivel de las personas y atribuir el factor de expansión del jefe de hogar al mismo hogar (Alexander, 1987). Esta es una elección arbitraria si los factores de expansión se han obtenido mediante una calibración que tiene en cuenta las características de las personas (por ejemplo, edad o sexo). Este acercamiento deliberado no permite sopesar las propiedades estadísticas del estimador resultante y, por ende, sus resultados no pueden ser interpretados de manera confiable, y mucho menos comparados.

Una elección más parsimoniosa puede consistir en optar por un enfoque del tipo de ponderación integrada de los hogares (*integrated household weighting*). Como expone Heldal (1992), al realizar una calibración al nivel de las personas, ya no será posible agregar a las personas de un mismo hogar para obtener un único peso del hogar, pues las características de esas personas serán, en general, diferentes y sus respectivos factores de expansión también lo serán. Por tanto, al definir $w_{k|i}$ como el factor de expansión de la persona k que pertenece al hogar i, y $w_{II,i}$ como el factor de expansión del hogar i, es necesario que el sistema de pesos satisfaga la siguiente restricción:

$w_{k|i}=w_{II,i}$ para toda persona k en el hogar i

De esta forma, sería posible obtener pesos que fueran coherentes con las restricciones de calibración a nivel de persona y que, al mismo tiempo, permitieran la integración con los hogares al cambiar de unidad de observación. En la literatura se han descrito varios métodos para lograr esta estandarización. A continuación se profundiza en algunos de ellos.

a) Enfoque de Estevao y Särndal

En principio, se debe destacar que es posible realizar el proceso de calibración de factores de expansión sobre la base de datos de las personas o de los hogares. De estos dos escenarios, el de calibrar sobre la base de personas parecería ser la opción más rápida, ya que, en la mayoría de los casos, las cifras que se utilizan para calibrar están al nivel de los individuos. Por ejemplo, en una encuesta de fuerza de trabajo, es evidente que las variables más importantes se encuentran al nivel de las personas y que la calibración de los factores de expansión se debería realizar desde la base de datos de personas.

Ahora bien, en una situación en que se desea calibrar por sexo, se debería tener acceso a las proyecciones demográficas por sexo correspondientes al período de referencia de la encuesta y se procedería a calibrar los factores de expansión, mediante un enfoque de posestratificación. En este escenario, las ecuaciones de calibración estarían dadas por la siguiente expresión:

$$\left(\sum_{k \in s} w_k x_{1k}, \sum_{k \in s} w_k x_{2i} \right) = (t_{x1}, t_{x2})$$

Donde la suma se hace sobre las personas en la muestra s; además, x_{k1} toma el valor de 1 si el individuo k es mujer y de 0 en otros casos. Por supuesto, $x_{k2}=1-x_{k1}$, t_{x1} es la proyección demográfica del total de mujeres y t_{x2} es la proyección demográfica del total de hombres. En este caso, las covariables de la calibración son variables dicotómicas. Nótese que las ecuaciones de calibración están al nivel de la muestra s que surge de una base de datos de personas. Como el muestreo ha sido en varias etapas, una posibilidad que surge al calibrar los factores de expansión es la de utilizar la muestra de hogares s_{II} que surge de una base de datos con información de los hogares y calibrar mediante un enfoque de calibración general con covariables continuas. De esta forma, las ecuaciones de calibración estarían dadas por la siguiente expresión:

$$\left(\sum_{i \in s_{II}} w_{II,i} z_{1i}, \sum_{i \in s_{II}} w_{II,i} z_{2i} \right) = (t_{z1}, t_{z2})$$

Donde la suma se realiza ahora al nivel de la muestra de hogares s_{II}. Nótese que $z_{1i}=\sum_{k \in s_i} x_{k1}$ se refiere al número de hombres en el hogar i; $z_{2i}=\sum_{k \in s_i} x_{k2}$ es el número de mujeres en el hogar i, y los totales de calibración $t_{z1}= t_{x1}$ y $t_{z2}=t_{x2}$ siguen siendo el número de hombres y mujeres en la población, respectivamente, por lo que coinciden plenamente.

Al respecto, cabe mencionar que, al calibrar con el primer escenario, se reproducen los totales auxiliares sobre la base de personas, mientras que, al calibrar sobre el segundo escenario, se reproducen los totales sobre la base de hogares. Sin embargo, teniendo en cuenta los principios del muestreo en varias etapas y que, en un hogar, la probabilidad de inclusión de las personas es de 1 (inclusión forzosa), entonces generar factores de expansión para las personas en el segundo escenario es muy sencillo, puesto que:

$$w_{k|i}=\frac{w_{II,i}}{Pr(k \in U_i \mid i \in sI)}=\frac{w_{II,i}}{1}=w_{II,i}$$

Es decir que, en este escenario de calibración, todas las personas dentro del hogar comparten los mismos pesos de muestreo y, además, estos pesos son iguales al peso del hogar. Estevao y Särndal (2006) recrean la calibración conjunta para hogares y personas. En resumen, después de haberse calibrado la base de hogares, se construyen los pesos a nivel de persona recurriendo a la siguiente expresión:

$$w_k=d_{k|i}\, w_{II,i} \quad \forall k \in s_i$$

Como todos los individuos pertenecientes a un hogar son seleccionados para que respondan la encuesta de hogares, se tiene que $d_{(k|i)}=1$, por definición. Por lo tanto, el peso del individuo (en la base de datos de la muestra de personas) será idéntico al peso calibrado del hogar; es decir, $w_k = w_{II,i}\ \forall\ k \in S_i$. Además, dado que el muestreo es de conglomerados en la última etapa y todos los individuos del hogar son seleccionados, el peso de muestreo del hogar será el promedio de los pesos individuales.

b) Enfoque de Lemaitre y Dufour

Un segundo enfoque, condensado en Lemaitre y Dufour (1987), consiste en crear nuevas variables de calibración a nivel de persona, definidas como el promedio de las variables originales en el hogar. Así pues, se definen las siguientes cantidades:

$$z_{ik}=\sum_{i \in s_{II}} x_{ik} \ \ y \ \ \bar{z}_{ik}=\frac{z_{ik}}{N_i}$$

Donde z_{ik} es la agregación a nivel de hogar de las covariables originales de calibración a nivel de persona y N_i es el tamaño del i-ésimo hogar. Al ejecutar el algoritmo de calibración utilizando las variables z en vez de las variables x, se reproducen las ecuaciones de calibración a satisfacción. Dado que todos los individuos comparten las mismas covariables en la calibración, sus pesos serán idénticos para todos los que comparten un mismo hogar. Nótese que esta calibración se realiza con la base de datos a nivel de personas.

En la literatura estadística se ha estudiado este enfoque integrado. Neethling y Galpin (2006) concluyeron que, en ambos enfoques, las estimaciones resultantes redujeron el sesgo, aumentaron la precisión y proporcionaron un único conjunto de ponderaciones para los datos de las encuestas estudiadas. Además, si se opta por el segundo enfoque, según el cual el tamaño de la base de datos sería igual al número de personas entrevistadas, se tendría margen suficiente para actualizar las restricciones de calibración con el fin de ejercer un mayor control sobre los tamaños de los subgrupos de interés.

3. Calibración sobre razones, medias y proporciones

Gutiérrez, Zhang y Rodríguez (2016) afirman que, además de utilizar ponderadores calibrados a tamaños o totales, también es posible imponer restricciones de calibración sobre razones, que a su vez son una generalización de medias y proporciones. Por ejemplo, considérense Q subgrupos de interés (dominios, estratos o posestratos). Si las razones para dichos subgrupos son conocidas, se pueden encontrar pesos w_k que satisfagan la siguiente restricción:

$$\hat{R}_{cal} = (\hat{R}_{1,cal}, \ldots, \hat{R}_{Q,cal})' = (R_1, \ldots, R_Q)' = R$$

Donde $\hat{R}_{q,cal} = \frac{\sum_{k \in s} w_k y_{qk}}{\sum_{k \in s} w_k x_{qk}}$. De esta forma, es posible imponer la siguiente restricción en las ecuaciones de calibración:

$$\hat{R}_{cal} = \hat{R}_U$$

Es decir, para $q=1,\ldots,Q$, se define la siguiente variable de información auxiliar:

$$z_{qk} = \begin{cases} y_{qk} - R_q\, x_{qk} & \text{si } k \in s_q \\ 0 & \text{en caso contrario} \end{cases}$$

Donde:

$$t_{z_q} = \sum_{k \in U} z_{qk} = \sum_{k \in U} y_{qk} - R_q\, x_{qk} = 0$$

Como caso particular, si las medias de los subgrupos son conocidas, la restricción queda como:

$$\bar{y}_{cal} = (\bar{y}_{1,c\ldots}, \ldots, \bar{y}_{Q,cal})' = (\ldots, \ldots, \bar{y}_Q)' = \bar{y}$$

Así, la restricción para las ecuaciones de calibración, $\bar{y}_{cal} = \bar{y}$, para cada $q=1,\ldots,Q$, se define a partir de la siguiente variable de calibración:

$$z_{qk} = \begin{cases} y_{qk} - \bar{y}_q & \text{si } k \in s_q \\ 0 & \text{en caso contrario} \end{cases}$$

4. Calibración con valores perdidos y totales estimados

Existen algunas condiciones que deben mantenerse al utilizar los estimadores de calibración en las encuestas de hogares. Una de ellas es que la información de las covariables de calibración esté completa en la base de datos de las encuestas. Por ejemplo, supóngase que un país está interesado en evaluar la posibilidad de actualizar las covariables de calibración en su encuesta continua sobre la fuerza de trabajo. Con el advenimiento de nuevos flujos de migración internacional en la región, es posible considerar que la nacionalidad del respondiente guarda una relación directa con su condición de actividad. Por lo tanto, incluir esta variable

en el sistema de calibración podría ser atractivo para reducir los sesgos generados por la falta de respuesta (o problemas de cobertura del marco de muestreo) de los extranjeros en la encuesta.

Al actualizar el sistema de calibración, es necesario tener en cuenta que las nuevas covariables deben tener información completa en la base de datos de la encuesta. Siguiendo con el ejemplo planteado, si la nacionalidad de los respondientes tiene observaciones incompletas, entonces habría serias dificultades para considerarla en el nuevo sistema. En particular, habría que sopesar las siguientes consideraciones:

- Los estimadores de calibración no están basados en modelos estadísticos. A pesar de que estos estimadores se conocen como asistidos por modelos (porque se basan en información auxiliar externa a la encuesta), siguen adheridos al marco inferencial basado en el diseño de muestreo, donde se supone que los valores observados a nivel de unidad para las variables de la encuesta son valores verdaderos, no aleatorios y fijos.
- Desde un punto de vista matemático, la calibración es un problema de optimización con restricciones sobre los totales auxiliares disponibles, que implicaría una definición errónea si faltaran valores de las covariables de calibración para algunas unidades.
- Los INE suelen utilizar como covariables de calibración variables estructurales, es decir, variables cuyos valores muestrales se observan con alta calidad (es decir, que no tienen valores perdidos y son fiables), o pueden reconstruirse de forma confiable y precisa a partir de fuentes externas (por ejemplo, registros estadísticos, archivos censales o encuestas pasadas).

Uno de los objetivos de actualizar el sistema de calibración (incluida la variable de nacionalidad en el ejemplo anterior) consiste en no solo mantener la coherencia con las cifras de migración oficiales, sino también utilizarla para tratar la falta de respuesta de unidad. En este caso, se calibra la muestra efectiva, es decir, el subconjunto donde las variables auxiliares se observan completamente. Si se supone que la información de la covariable no está completa en las bases de datos de la encuesta, el investigador tendría dos opciones posibles:

i) descartar la nacionalidad como variable auxiliar en la calibración, o

ii) imputar o rellenar los valores faltantes de la variable de nacionalidad antes de la calibración.

Nótese que en la imputación se asumirá implícita o explícitamente algún tipo de modelo, lo que hará que los estimadores de calibración finales ya no se basen completamente en el diseño de muestreo y se pierda todo el andamiaje inferencial en la encuesta, incluida su comparabilidad en el tiempo. Además, no se lograría una protección definitiva contra el sesgo generado por la falta de respuesta de la unidad después de la calibración. Es más, si el modelo de imputación no queda correctamente especificado, el error podría ser aún más grande.

Sin embargo, si las unidades de la muestra pudieran vincularse sin error a un registro estadístico completo, actualizado y de alta calidad en que se disponga de la nacionalidad, entonces se podrían rellenar los valores perdidos en la base de datos para esta covariable y la inferencia seguiría siendo robusta y fiel al paradigma básico de las encuestas de hogares.

En resumen, la recomendación es que se realicen todos los esfuerzos en la consecución de las covariables de calibración para los respondientes en la etapa de recopilación de información primaria. Si esto no fuese posible, sería igualmente plausible como solución apoyarse en los registros estadísticos para conseguir la nacionalidad del respondiente. De la misma forma, se debería evitar la adopción de modelos de imputación en las covariables de calibración.

Por otro lado, la calibración con totales de control estimados se utiliza cada vez más. Si bien es cierto que el procedimiento de calibración exige que los totales de control se conozcan de antemano, también es cierto que no es posible realizar los censos con mayor frecuencia. Por ende, las estructuras poblacionales y demográficas observadas en los censos pueden desactualizarse rápidamente. Una vez más, considérese el caso de la nacionalidad. Ante una explosión migratoria en un país en el período intercensal, las proyecciones censales para la variable de nacionalidad podrían quedar obsoletas rápidamente y sería necesario utilizar estimaciones de otras operaciones estadísticas para poder calibrar con totales estimados de control actualizados.

Un ejemplo de la situación anterior se da en los Estados Unidos con la Encuesta sobre la Comunidad Estadounidense, que brinda estimaciones actualizadas y oportunas con información anual detallada acerca del ingreso, la educación, el empleo, la cobertura en salud, los costos del hogar y las condiciones de los residentes del país. Esta encuesta complementa los datos poblacionales recopilados por el censo, que se realiza cada diez años.

Por lo tanto, es posible que una encuesta mediana o pequeña, cuya muestra se denota como s_A, se apoye en totales de control estimados por una encuesta más grande, cuya muestra se denota con s_B. Este caso se conoce con el nombre de calibración con totales de control estimados (Dever, 2008) y los estimadores derivados con esta técnica se denotan con un asterisco. En este caso, los estimadores de calibración buscarían nuevos ponderadores w_k*que satisficieran la siguiente restricción:

$$\sum_{k \in s_A} w_k^* \boldsymbol{x}_k = \sum_{j \in s_B} w_j \boldsymbol{x}_j = \hat{\boldsymbol{t}}_{\boldsymbol{x},cal}$$

Por lo tanto, el estimador de un total con las observaciones de la muestra pequeña tendría la siguiente forma funcional:

$$\hat{t}^*_{y,cal} = \sum_{k \in s_A} w_k^* y_k = (\hat{\boldsymbol{t}}_{\boldsymbol{x},cal} - \hat{\boldsymbol{t}}_{\boldsymbol{x},\pi}) \widehat{\boldsymbol{B}}_{s_A}$$

Donde:

$$\hat{\boldsymbol{B}}_{s_A} = \left(\sum_{s_A} w_k \, q_k \, \boldsymbol{x}_k \boldsymbol{x}_k' \right)^{-1} \sum_{s_A} w_k \, q_k \, \boldsymbol{x}_k \, y_k$$

Sin embargo, Dever y Valliant (2016) muestran que, al utilizar la metodología de calibración con totales de control estimados, existe sesgo para los estimadores de razón, definidos así:

$$\hat{R}^*_{y,cal} = \frac{\hat{t}^*_{y,cal}}{\hat{t}^*_{z,cal}}$$

Donde $\hat{t}^*_{ycal}$ y $\hat{t}^*_{zcal}$ denotan dos estimadores de calibración con totales de control estimados. Nótese que esta es la misma forma que tomaría cualquier promedio estimado:

$$\bar{y}^*_{y,cal} = \frac{\hat{t}^*_{y,cal}}{\hat{N}^*_{cal}}$$

En este caso, el denominador de $\hat{R}^*_{cal}$ sería $\hat{t}^*_{z,cal} = \hat{N}^*_{cal}$. Dever y Valliant (2016) presentan la siguiente expresión para el sesgo de un promedio $\bar{y}^*_{cal}$:

$$Bias\ \bar{y}^*_{y,cal} \approx \frac{1}{E(\hat{N}^*_{cal})} Bias\ \bar{y}_{y,cal} \approx [Bias(\hat{t}^*_{y,cal}) - \bar{y}\, Bias(\hat{N}^*_{cal})]$$

En esta ecuación, $Bias(\hat{t}^*_{cal})$ y $Bias(\hat{N}^*_{cal})$ representan los sesgos de los estimadores de calibración con totales de control estimados del total poblacional (t_y) y del tamaño poblacional (N), respectivamente. El sesgo de estos estimadores puede llegar a ser despreciable si el mecanismo que genera la falta de respuesta es aleatorio o completamente aleatorio (véase el capítulo XII), y si se incluye una columna de unos en la matriz de las variables de calibración, lo que mostraría que no hay errores de cobertura.

Asimismo, la estructura de varianza de estos estimadores es bastante compleja, como exponen Dever y Valliant (2016). Sin embargo, es posible utilizar métodos de estimación de varianza basados en réplicas, como los propuestos por Opsomer y Erciulescu (2022).

F. Recorte y redondeo

1. Recorte de pesos extremos

Un inconveniente que se genera debido a la multitud de ajustes en los factores de expansión es que, si bien el estimador resultante tendrá un sesgo cercano a cero, la distribución de los pesos puede mostrar datos extremos, sobre todo a la derecha de la distribución (valores muy grandes). Estos valores hacen que la varianza del estimador crezca y que la precisión

de la inferencia decrezca. Para hacer frente a este problema, es posible considerar un procedimiento de *trimming* o recorte de pesos, siguiendo las recomendaciones de Valliant, Dever y Kreuter (2018, sec. 14.4), que puede resumirse en los siguientes pasos:

i) Recortar cualquier peso superior a un umbral preestablecido en la distribución de pesos ajustados. Por lo general, este umbral se fija en alrededor de 3,5 veces la mediana de los pesos. Por lo tanto:

$$U=3,5\times mediana(\mathbf{w_k})$$

ii) Truncar cualquier peso con magnitud superior a U de la siguiente manera:

$$w_k^*=\begin{cases} U & \text{si } w_k \geq U \\ w_k & \text{en caso contrario} \end{cases}$$

iii) Determinar la cantidad neta perdida debido al recorte de pesos extremos, mediante la siguiente expresión:

$$K=\sum_{s_r}(w_k^*=w_k)$$

iv) Distribuir K equitativamente entre las unidades que no fueron recortadas.

v) Iterar hasta que todos los nuevos pesos calibrados estén por debajo del umbral U.

Al final del proceso se debe asegurar que los datos extremos en los factores de expansión se hayan manejado correctamente y que la distribución general de los pesos no haya sufrido cambios estructurales en los subgrupos poblacionales de interés.

2. El problema del redondeo de los factores de expansión

Cuando el factor de expansión no es entero, su interpretación se torna compleja desde el punto de vista práctico, aunque en teoría ello no tenga ninguna repercusión negativa. Sin embargo, este inconveniente puede hacer que, en la práctica, las oficinas nacionales de estadística y los usuarios de las bases de datos de encuestas de hogares tomen la decisión —bienintencionada, pero errada— de redondear estas cantidades al entero más cercano. Esta práctica es perjudicial porque añade sesgo a la inferencia y causa problemas de sobreestimación o subestimación en algunos dominios de estudio. Sartore y otros (2019) plantean que el redondeo de los factores de expansión puede ser problemático porque las estimaciones ponderadas pueden crecer o decrecer enormemente.

Los siguientes ejemplos muestran de forma directa las repercusiones perjudiciales que conlleva esta práctica y que son consecuencia directa del sesgo de redondeo:

- En encuestas de establecimientos, redondear el factor de expansión en las unidades que tienen flujos de ventas grandes trae problemas de sesgo en este dominio de estudio.

- En encuestas agropecuarias, si una unidad productiva produce un cuarto de la producción nacional, el redondeo de su factor de expansión es muy perjudicial.
- En encuestas de hogares, en que los diseños de muestreo son generalmente autoponderados (todas las viviendas comparten el mismo factor de expansión) dentro de los estratos, redondear el factor de expansión implica sesgar por completo todo el estrato.

Suponiendo que una muestra probabilística $s=(I_1,\ldots,I_k,\ldots,I_N)'$ fue seleccionada de una población finita U mediante un diseño de muestreo que conlleva probabilidades de inclusión $\pi_k=E(I_k)$ para todos los individuos $k\in U$ (donde I_k toma el valor de uno si fue seleccionado o de cero en caso contrario), entonces, desde el punto de vista teórico, los estimadores de muestreo $\hat{t}_y=\sum_s d_k\, y_k$ son insesgados cuando el factor de expansión d_k es idéntico al inverso de la probabilidad de inclusión, puesto que:

$$E(\hat{t}_y)=E\left(\sum_s \frac{y_k}{\pi_k}\right)=E\left(\sum_U I_k \frac{y_k}{\pi_k}\right)=\sum_U E(I_k)\frac{y_k}{\pi_k}=\sum_U \pi_k \frac{y_k}{\pi_k}=t_y$$

De las anteriores relaciones se desprende que, cuando el factor de expansión se redondea de forma determinística, $E(\hat{t}_y)\neq t_y$. Para evitar el sesgo de redondeo, es necesario emplear un método aleatorio que favorezca el insesgamiento en los estimadores de muestreo. En general, este problema puede abordarse desde una perspectiva probabilística. De hecho, si en primera instancia se utiliza como redondeo la parte entera (el entero máximo que sea menor o igual) del factor de expansión, bastará con añadir de forma aleatoria una unidad a algunos factores de expansión para asegurar que la suma de los factores redondeados sea idéntica a la original. Con esta simple idea se devuelve la propiedad del insesgamiento a los estimadores de muestreo. El procedimiento se describe a continuación:

i) Para $k \in s$, definir:

$$\phi_k = d_k - \lfloor d_k \rfloor$$

ii) Seleccionar una submuestra $s_a=(c_1,\ldots,c_k,\ldots,c_n)'$ de s con probabilidades de inclusión ϕ_k, para $k\in s$. Nótese que c_k tomará el valor de 1 si el elemento k está en la submuestra y de 0 si no fue seleccionado en la submuestra.

iii) Si $c_k=0$, entonces $\tilde{d}_k=\lfloor d_k \rfloor$; en caso contrario, si $c_k=1$, entonces $\tilde{d}_k=\lfloor d_k \rfloor+1$.

En primer lugar, la submuestra s_a no necesariamente será de tamaño fijo, puesto que $\sum_s \phi_k$ no será entera en todos los casos. De ahí que sea posible utilizar un algoritmo de muestreo de Poisson (Gutiérrez, 2016, sec. 4.1) para seleccionar esta submuestra. Sin embargo, si esta suma es entera, es posible utilizar un algoritmo de muestreo más eficiente que determine una submuestra de tamaño fijo, por ejemplo, el método de Brewer (Tillé, 2006). Por otro lado, la esperanza de estos factores redondeados condicionados a la submuestra s_a es igual a los factores de expansión originales, como se indica a continuación:

$$E(\tilde{d}_k \mid s_a) = \lfloor d_k \rfloor + E(c_k \mid s_a) = \lfloor d_k \rfloor + \phi_k = d_k$$

Por ello, es importante destacar que el uso de este método aleatorio de redondeo siempre produce insesgamiento en los estimadores de muestreo, ya que:

$$E\left(\sum_s \tilde{d}_k y_k\right) = E\left[E\left(\sum_s \tilde{d}_k y_k \mid s_a\right)\right] = E\left(\sum_s E(\tilde{d}_k \mid s_a) y_k\right) = E\left(\sum_s d_k y_k\right) = t_y$$

Por último, cuando los factores de expansión de la encuesta están calibrados, se presenta un problema de optimización un poco más complejo, puesto que, al utilizar el redondeo aleatorio, los factores de expansión perderán la propiedad de calibración. Sartore y otros (2019) y Tillé (2019) han presentado diferentes soluciones a este problema, siendo la última mucho más fácil de implementar en el *software* estadístico R. Según esta perspectiva, la calibración de los factores de expansión crea nuevos pesos, denominados w_k, que conservan la siguiente propiedad respecto de un conjunto de totales auxiliares $\boldsymbol{t_x}$ disponibles para toda la población:

$$\sum_s w_k \, \boldsymbol{x}_k = \boldsymbol{t_x}$$

El siguiente algoritmo hace uso del muestreo balanceado (Tillé, 2006, cap. 8), que representa una forma de calibración desde el diseño de muestreo y es una solución óptima para seleccionar la submuestra s_a y, por ende, preservar la coherencia de los pesos calibrados con los totales auxiliares.

i) Para $k \in s$, definir $\phi_k = w_k - \lfloor w_k \rfloor$ y:

$$\tilde{\boldsymbol{x}}_k = \phi_k \boldsymbol{x}_k$$

ii) Seleccionar una submuestra balanceada $s_a = (c_1, \ldots, c_k, \ldots, c_n)'$ de s con probabilidades de inclusión ϕ_k, tal que:

$$\sum_{k \in s_a} \frac{\tilde{\boldsymbol{x}}_k}{\phi_k} \cong \sum_{k \in s} \tilde{\boldsymbol{x}}_k$$

iii) Si $c_k = 0$, entonces $\tilde{w}_k = \lfloor w_k \rfloor$; en caso contrario, si $w_k = 1$, entonces $\tilde{w}_k = \lfloor w_k \rfloor + 1$.

Es importante recalcar que la restricción en la submuestra balanceada implica que los pesos redondeados deben cumplir la siguiente relación:

$$\sum_s c_k \, \boldsymbol{x}_k \cong \sum_U \boldsymbol{x}_k - \sum_U \lfloor w_k \rfloor \, \boldsymbol{x}_k$$

Ello significa que los nuevos pesos, además de redondeados, también están calibrados, es decir:

$$\sum_s \tilde{w}_k \, \boldsymbol{x}_k \cong \boldsymbol{t_x}$$

Nótese que el redondeo aleatorio depende de la selección de la submuestra s_a para completar los restos de la parte entera. En esta selección intervienen diferentes algoritmos

de muestreo que se pueden aplicar fácilmente si se utiliza la biblioteca *sampling* (Tillé y Matei, 2016). Por ejemplo, supóngase una muestra de tamaño $n=200$ que fue seleccionada de una población de tamaño $N=9.200$ con factores de expansión desiguales que no están calibrados. Considérese que el vector de probabilidades de inclusión en la muestra toma la siguiente forma:

$$\boldsymbol{\pi}_s = \left(\underbrace{15/500}_{50\ veces}, \ldots, \underbrace{15/800}_{80\ veces}, \ldots, \underbrace{15/700}_{70\ veces} \right)'$$

Por lo tanto, el vector de pesos de muestreo estará definido de la siguiente manera:

$$\boldsymbol{d}_s = \left(\underbrace{33,33333}_{50\ veces}, \ldots, \underbrace{53,33333}_{80\ veces}, \ldots, \underbrace{46,66667}_{70\ veces} \right)'$$

De la misma manera, el vector de excesos $\phi_k = d_k - \lfloor d_k \rfloor$ estará dado por la siguiente expresión:

$$\boldsymbol{\phi}_s = \left(\underbrace{0,33333}_{130\ veces}, \ldots, \underbrace{0,66667}_{70\ veces} \right)'$$

Luego del cálculo de ϕ_k, se selecciona la submuestra s_a. En particular, en este caso se utiliza el algoritmo de Brewer (Gutiérrez, 2016), puesto que $\sum_s \phi_k = 90$ y es entero. Al final del proceso de redondeo aleatorio, la suma de los nuevos factores coincidirá con la suma de los factores originales. Por último, si en una segunda instancia se considera que los pesos están calibrados mediante sendas covariables de calibración, será posible utilizar el método del cubo para que la submuestra esté balanceada y los pesos redondeados sigan las restricciones de calibración con una tolerancia predefinida.

Capítulo X
Estimación del error de muestreo

Después de seleccionar la muestra y realizar el proceso de medición, es necesario efectuar la estimación de los parámetros y de sus respectivos errores estándar. El error estándar se define como la raíz cuadrada de la varianza

Mientras que los investigadores pueden elegir libremente el diseño de muestreo y el estimador, no ocurre lo mismo con el cálculo de las medidas de confiabilidad y precisión. Dado que la base científica del muestreo es la inferencia estadística, se deben respetar las normas básicas para la asignación y el posterior cálculo del margen de error, el cual constituye una medida unificada del error total de muestreo que cuantifica la incertidumbre acerca de las estimaciones de una encuesta. La forma de estimar el error estándar depende de:

- La complejidad del diseño de muestreo: estratificación, selección proporcional al tamaño, múltiples etapas.
- La complejidad del estimador: ajuste de pesos por falta de respuesta, calibración, razón de totales, medias, percentiles, coeficientes de regresión.

Existen tres alternativas para calcular el error estándar de las estimaciones de una encuesta. Sobre la base de la estrategia de muestreo, es posible encontrar fórmulas exactas que describan la varianza del estimador. Sin embargo, cuando el estimador utilizado no es una función lineal de totales, es posible utilizar un enfoque de linealización de Taylor para aproximar la varianza del estimador a una función lineal. Por último, es posible apoyarse en los métodos computacionales modernos y aplicar los principios de los pesos replicados para aproximar la varianza de cualquier estimador en una encuesta de hogares.

Los programas estadísticos más utilizados en la actualidad incluyen procedimientos para la estimación de la varianza teniendo en cuenta diseños de muestreo complejos. Una forma sencilla de utilizarlos es seguir estos pasos en una base de datos agregada:

i) Modificar los pesos, de manera que cumplan con las restricciones poblacionales básicas.

ii) Definir los estratos de interés en los que el diseño de muestreo se realiza de forma independiente.

iii) Definir estrictamente las unidades primarias de muestreo (UPM) como aglomerados poblacionales que incluyen a hogares y personas.

A. Fórmulas exactas y linealización de Taylor

En la mayoría de los casos de interés, se pueden encontrar fórmulas exactas que corresponden a cada diseño de muestreo y a cada estimador utilizado. Cuando se emplean diseños simples, es posible implementarlas para calcular la estimación de los errores directamente. Sin embargo, cuando se utilizan diseños en múltiples etapas con estimadores simples, las fórmulas se pueden tornar extremadamente complicadas. En el caso de los diseños en múltiples etapas con estimadores complejos, simplemente no se trata de una opción viable.

La estimación de la varianza en una estrategia de muestreo no siempre es una tarea sencilla. A partir de la teoría, se establece un camino lógico basado en las probabilidades de inclusión de primer y segundo orden. En general, en el caso de los diseños de muestreo sin reemplazo, la fórmula exacta para calcular una varianza del estimador de Horvitz-Thompson está dada por:

$$Var(\hat{t}_{y,\pi}) = \sum_{U} \sum_{U} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

Donde $\Delta_{kl} = \pi_{kl} - \pi_l \pi_l$. La probabilidad de inclusión de segundo orden se denota análogamente como π_{kl} y define la probabilidad de que los elementos k y l pertenezcan a la muestra al mismo tiempo, es decir:

$$\pi_{kl} = Pr(k \in s, l \in s) = Pr(I_k I_l = 1) = \sum_{s \ni k,l} p(s)$$

Donde el subíndice $s \ni k,l$ se refiere a la suma sobre todas las muestras que contienen los elementos k-ésimo y l-ésimo. Evidentemente, debido a las limitaciones de la capacidad de cómputo de los programas informáticos y el hecho de que es imposible observar los registros relativos a toda la población finita, la realización de este cálculo para los estimadores de los indicadores de interés en las encuestas simplemente no es viable.

Dado que en la práctica de las encuestas de hogares nunca se podrá contar con la varianza exacta de un estimador, para evaluar la precisión de la estrategia de muestreo se deberá estimar dicha varianza. De acuerdo con Gutiérrez (2016), uno de los estimadores insesgados de esta varianza estará dado por la siguiente expresión:

$$\widehat{Var}_1(\hat{t}_{y,\pi}) = \sum_{S} \sum_{s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

Asimismo, si el diseño prevé un tamaño de muestra fijo, uno de los estimadores insesgados estará dado por:

$$\widehat{Var}_2(\hat{t}_{y,\pi}) = -\frac{1}{2}\sum_S \sum_s \frac{\Delta_{kl}}{\pi_{kl}}\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2$$

De esta forma, cuando el tamaño de la muestra es suficientemente grande, se puede construir un intervalo de confianza de nivel $(1-\alpha)$ para el total poblacional t_y, como se indica a continuación:

$$IC(1-\alpha) = \left[\hat{t}_{y,\pi} - z_{1-\alpha/2}\sqrt{(Var(\hat{t}_{y,\pi})}, \hat{t}_{y,\pi} + z_{1-\alpha/2}\sqrt{Var(\hat{t}_{y,\pi})}\right]$$

Donde $z_{1-\alpha/2}$ se refiere al percentil $(1-\alpha/2)$ de una variable aleatoria con distribución normal estándar. Como cada diseño de muestreo da lugar a una forma cerrada para las probabilidades de inclusión de primer y segundo orden, las fórmulas de estimación de la varianza se reducen ostensiblemente. Por ejemplo, en el caso de un diseño de muestreo aleatorio simple, la fórmula de estimación de la varianza es:

$$\hat{Var}(\hat{t}_{y,\pi}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)S_{y_s}^2$$

Donde $S_{y_s}^2$ es la varianza de los valores de la característica de interés en la muestra aleatoria s, dada por:

$$S_{y_s}^2 = \frac{1}{n-1}\sum_{k\in s}(y_k - \bar{y}_k)^2$$

Por otra parte, si se utiliza un diseño de muestreo aleatorio estratificado y el parámetro de interés es una media, la fórmula del estimador de Horvitz-Thompson es $\bar{y}_\pi = \frac{1}{N}\sum_s d_k y_k = \sum_{h=1}^H W_h\ \bar{y}_h$, donde $W_h = N_h/N$. Siendo S_{yh}^2 la varianza muestral en el estrato h de los valores de la característica de interés y definiendo $w_h = n_h/n$, la fórmula de estimación de la varianza es:

$$\hat{Var}(\bar{y}_\pi) = \sum_{h=1}^{H} w_h^2\ \frac{1-f_h}{n_h}\ S_{yh}^2$$

Cuando el diseño de muestreo se vuelve más complejo, también lo hace la estimación de la varianza. Por ejemplo, si el diseño de muestreo es estratificado y de dos etapas, de manera que dentro de cada estrato $U_h\ h=1,\ldots,H$ existen N_{Ih} UPM, entre las que se selecciona una muestra s_{Ih} de n_{Ih} unidades mediante un diseño de muestreo aleatorio simple y, además, se considera que el submuestreo dentro de cada unidad primaria seleccionada es también aleatorio simple, de manera que para cada UPM seleccionada $i \in s_{Ih}$ de tamaño N_i se selecciona una submuestra s_i de elementos de tamaño n_i, la forma final del estimador de la varianza del estimador de Horvitz-Thompson para el total poblacional se expresaría de la siguiente manera:

$$\widehat{Var}\,(\hat{t}_{y,\pi}) = \sum_{h=1}^{H} \left[\frac{N_{Ih}^2}{n_{Ih}} \left(1 - \frac{n_{Ih}}{N_{Ih}}\right) S^2_{\hat{t}_{yh}\,S_I} + \frac{N_{Ih}}{n_{Ih}} \sum_{i \in S_{Ih}} \frac{N_i^2}{n_i} \left(1 - \frac{n_i}{N_i}\right) S^2_{y_{S_i}} \right]$$

Donde $S^2_{\hat{t}_{y}S_I}$ y $S^2_{y_{S_i}}$ son, respectivamente, las varianzas muestrales de los totales estimados en las UPM seleccionadas y las varianzas muestrales de los hogares incluidos en la submuestra dentro de las UPM seleccionadas en la muestra de la primera etapa.

Existen fórmulas computacionales para estimar la varianza de estadísticas descriptivas como la media muestral para algunos diseños complejos, que incorporan elementos como la estratificación y el muestreo por conglomerados. Sin embargo, en el caso de estadísticas analíticas más complejas, como coeficientes de correlación y coeficientes de regresión, no es fácil encontrar fórmulas específicas para diseños muestrales distintos del muestreo aleatorio simple. Estas fórmulas son sumamente complicadas o, en última instancia, inadecuadas para el análisis matemático.

B. Técnica del último conglomerado

Debido a dificultades algebraicas y computacionales, la estimación de la varianza en encuestas complejas, que contemplan sistemas de conglomeración, selección en varias etapas y estratificación, puede tornarse bastante tediosa, costosa y lenta. En esta sección se explica por qué la técnica del último conglomerado es una buena opción a la hora de aproximar la varianza en una encuesta compleja.

Para estimar la varianza de los estimadores de interés en encuestas de múltiples etapas, los programas computacionales existentes utilizan un método conocido como "técnica del último conglomerado" (*ultimate cluster*). Esta técnica, que solo tiene en cuenta la varianza de los estimadores en la primera etapa, supone que el muestreo fue realizado con reemplazo. Así, se ignoran los procedimientos de muestreo en etapas posteriores de la selección, a menos que el factor de corrección para poblaciones finitas no sea despreciable a nivel de la primera etapa de muestreo.

En particular, se considera un estimador del total poblacional dado por la siguiente combinación lineal:

$$\hat{t}_{y,\pi} = \sum_{k \in s} d_k\, y_k = \sum_{k \in U} I_k\, d_k\, y_k$$

Donde I_k corresponde a variables indicadoras de la pertenencia del elemento k a la muestra s. Se asume que el factor de expansión de la encuesta d_k cumple con los supuestos básicos de un ponderador que hace que $\hat{t}_y$ sea insesgado, es decir:

$$E_p\,(I_k\, d_k) = 1$$

Se supone un diseño de muestreo en varias etapas (dos o más) en el que la primera etapa implica la selección de una muestra s_I de m_I UPM U_i $(i \in s_I)$, de manera que:

- Si la selección se realizó con reemplazo, la i-ésima UPM tiene probabilidad de selección p_{I_i}.
- Si la selección se realizó sin reemplazo, la i-ésima UPM tiene probabilidad de inclusión π_{I_i}.

En las subsiguientes etapas de muestreo, se procede a seleccionar una muestra de elementos para cada una de las UPM seleccionadas en la primera etapa de muestreo. Dentro de la i-ésima UPM se selecciona una muestra s_i de elementos. En particular, la probabilidad condicional de que el k-ésimo elemento pertenezca a la muestra porque la UPM que la contiene fue seleccionada en la muestra de la primera etapa está dada por la siguiente expresión:

$$\pi_{k|i}=Pr(\, k\in s_i \mid i\in s_I\,)$$

Por ejemplo, en el caso del muestreo sin reemplazo en todas las etapas, la probabilidad de inclusión del k-ésimo elemento en la muestra s está dada por:

$$\begin{aligned}\pi_k &= Pr(k\in s\,)\\ &= Pr(k\in s_i\,,\; i\in s_I)\\ &= Pr(k\in s_i \mid i\in s_I)\; Pr(i\in s_I)=\pi_{k|i}\times\pi_{I_i}\end{aligned}$$

Dado que el inverso de las probabilidades de inclusión es un ponderador natural, se definen las siguientes cantidades:

1. $d_{I_i}=\frac{1}{\pi_{I_i}}$, que es el factor de expansión de la i-ésima UPM.
2. $d_{k|i}=\frac{1}{\pi_{k|i}}$, que es el factor de expansión del k-ésimo elemento dentro de la i-ésima UPM.
3. $d_k=d_{I_i}\times d_{k|i}$, que es el factor de expansión final del k-ésimo elemento para toda la población U.

Según la teoría de muestreo, si el diseño de muestreo es con reemplazo, además del estimador de Horvitz-Thompson es posible utilizar otro estimador insesgado, denominado "estimador de Hansen-Hurwitz" (Gutiérrez, 2016). A diferencia del primero, este tiene una expresión de varianza muy sencilla de calcular y, por consiguiente, las expresiones de la estimación de la varianza del estimador de Hansen-Hurwitz son más manejables desde el punto de vista computacional. En efecto, en el marco de un diseño de muestreo en varias etapas, el estimador de Hansen-Hurwitz para el total poblacional está dado por la siguiente expresión:

$$\hat{t}_{y,p}=\frac{1}{m_I}\sum_{i=1}^{m_I}\frac{\hat{t}_{yi}}{p_{I_i}}$$

Donde p_{Ii} corresponde a la probabilidad de selección de la unidad i y m_I es el tamaño de la muestra (con reemplazo) del muestreo en la primera etapa. En este caso, la varianza estimada del estimador de Hansen-Hurwitz es:

$$\widehat{Var}\,(\hat{t}_{y,p})=\frac{1}{m_I\,(m_I\text{-}1)}\sum_{i=1}^{m_I}\left(\frac{\hat{t}_{yi}}{p_{I_i}}-\hat{t}_{y,p}\right)^2$$

Donde las cantidades $\hat{t}_{y_i}$ representan los totales estimados de la variable de interés en la i-ésima UPM y están dadas por:

$$\hat{t}_{y_i} = \sum_{k \in s_i} \frac{y_k}{\pi_{k|i}} = \sum_{k \in s_i} \hat{t}_{y_i} d_{k|i} y_k$$

La técnica del último conglomerado consiste en utilizar la expresión de la estimación de la varianza del estimador de Hansen-Hurwitz en lugar de la expresión exacta en diseños de muestreo complejos que no contemplan selecciones con reemplazo en la primera etapa. Para lograrlo, es necesario equiparar algunas cantidades antes de poder utilizar la aproximación. La aproximación de la varianza requiere la equiparación adecuada de los términos. En primer lugar, se observan los estimadores $\hat{t}_{y,p}$ y $\hat{t}_{y,\pi}$. Para realizar esta comparación, se debe asumir la siguiente igualdad en las probabilidades de inclusión de la primera etapa:

$$\pi_{I_i} = p_{I_i} \times m_I$$

Por lo tanto, el estimador del total poblacional quedaría definido como un estimador de tipo Hansen-Hurwitz. En efecto,

$$\hat{t}_{y,\pi} = \sum_{k \in s} d_k y_k = \sum_{i=1}^{m_I} \sum_{k \in s_i} d_k y_k = \sum_{i=1}^{m_I} \sum_{k \in s_i} \frac{1}{\pi_{I_i} \pi_{k|i}} y_k = \sum_{i=1}^{m_I} \frac{\hat{t}_{y_i}}{\pi_{I_i}} \approx \frac{1}{m_I} \sum_{i=1}^{m_I} \frac{\hat{t}_{y_i}}{p_{I_i}}$$

Dado que la forma del estimador se ha equiparado con un estimador de tipo Hansen-Hurwitz, es posible utilizar su estimación de varianza. Asimismo, después de un poco de álgebra, es posible obtener la siguiente aproximación, cuya gran ventaja es que solo utiliza los factores de expansión finales d_k, que los institutos nacionales de estadística (INE) suelen proporcionar cuando divulgan los microdatos de sus encuestas, en lugar de los factores de expansión de la primera etapa o los factores de expansión condicionales dentro de las UPM.

$$\begin{aligned}
\widehat{Var}_2(\hat{t}_{y,p}) &= \frac{1}{m_I(m_I-1)} \sum_{i=1}^{m_I} \left(\frac{\hat{t}_{y_i}}{p_{I_i}} - \hat{t}_y \right)^2 \\
&= \frac{m_I}{m_I-1} \sum_{i=1}^{m_I} \frac{1}{m_I^2} \left(\frac{\sum_{k \in s_i} d_{k|i} y_k}{p_{I_i}} - \sum_{i=1}^{m_I} \sum_{k \in s_i} d_k y_k \right)^2 \\
&= \frac{m_I}{m_I-1} \sum_{i=1}^{m_I} \left(\frac{\sum_{k \in s_i} d_{k|i} y_k}{m_I p_{I_i}} - \frac{1}{m_I} \sum_{i=1}^{m_I} \sum_{k \in s_i} d_k y_k \right)^2 \\
&= \frac{m_I}{m_I-1} \sum_{i=1}^{m_I} \left(\frac{\sum_{k \in s_i} d_{k|i} y_k}{\pi_{I_i}} - \frac{1}{m_I} \sum_{i=1}^{m_I} \sum_{k \in s_i} d_k y_k \right)^2 \\
&= \frac{m_I}{m_I-1} \sum_{i=1}^{m_I} \left(\sum_{k \in s_i} d_k y_k - \frac{1}{m_I} \sum_{i=1}^{m_I} \sum_{k \in s_i} d_k y_k \right)^2
\end{aligned}$$

En virtud de lo expuesto, al definir $\check{t}_{yi} = \sum_{k \in s_i} d_k \ y_k$ como la contribución[1] de la i-ésima UPM a la estimación del total poblacional y $\bar{\check{t}}_y = \frac{1}{m_I} \sum_{i=1}^{mI} \check{t}_{yi}$ como la contribución promedio en el muestreo de la primera etapa, el estimador de varianza toma la siguiente forma, denominada estimador de varianza del último conglomerado:

$$\widehat{Var}\,(\hat{t}_{y,p}) = \frac{m_I}{m_I - 1} \sum_{i=1}^{m_I} \left(\check{t}_{yi} - \frac{1}{m_I} \sum_{i=1}^{m_I} \check{t}_{yi} \right)^2 = \frac{m_I}{m_I - 1} \sum_{i=1}^{m_I} (\check{t}_{yi} - \bar{\check{t}}_y)^2$$

Por ejemplo, si en la encuesta se plantea un escenario de muestreo estratificado, con tres etapas de selección dentro de cada estrato, al utilizar la técnica del último conglomerado, la aproximación del estimador de la varianza estaría dada por:

$$\widehat{Var}\,(\hat{t}_{y,p}) = \sum_h \frac{n_h}{n_h - 1} \sum_{i \in s_h} (\hat{t}_{yi} - \bar{\hat{t}}_{yh})^2$$

Donde $\hat{t}_{yi} = \sum_{k \in s_{hi}} w_k \ y_k$, $\bar{\hat{t}}_{yh} = (1/n_h) \sum_{i \in s_h} \hat{t}_{yi}$ y n_h es el número de UPM seleccionadas en el estrato h. Si bien este procedimiento, propuesto por Hansen, Hurwitz y Madow (1953), tiende a sobrestimar la varianza verdadera, es una técnica valorada por los investigadores porque utiliza directamente los pesos finales de muestreo o factores de expansión publicados por los INE.

La técnica del último conglomerado es una solución práctica al problema de la estimación de la varianza, que, en la mayoría de las encuestas en las que se basan las estadísticas oficiales de los países, puede tornarse bastante complejo. Si bien la expresión del estimador de la varianza no constituye un estimador estrictamente insesgado, se considera una aproximación bastante precisa.

Por último, es importante reflexionar acerca de la definición práctica y el concepto de esta aproximación. ¿Qué es un último conglomerado? Es la primera unidad de muestreo en un diseño complejo. Por ejemplo, se considera el siguiente diseño de muestreo en cuatro etapas:

$$\underbrace{Municipio}_{UPM} \rightarrow \underbrace{Sector}_{UPM} \rightarrow \underbrace{Vivienda}_{UTM} \rightarrow \underbrace{Hogar}_{UFM}$$

En la primera etapa, las UPM son los municipios. Dentro de cada municipio se seleccionan unidades secundarias de muestreo (USM), que corresponden a sectores cartográficos. El submuestreo continúa pasando por las unidades terciarias de muestreo (UTM) hasta seleccionar las unidades finales de muestreo (UFM), que son los hogares.

En general, la primera etapa de muestreo de una encuesta obedece a dos tipos de diseño: estratificado o con probabilidad de selección proporcional al tamaño del municipio. En los dos casos se crean subgrupos de inclusión forzosa. Tanto en el muestreo estratificado como en el proporcional, estos serán las grandes ciudades, pues la medida de tamaño determinará probabilidades de inclusión superiores a uno. Para aplicar la

[1] La suma de estas contribuciones en la muestra de la primera etapa da como resultado la estimación $\hat{t}_y$.

aproximación en este caso, los municipios pertenecientes a este subgrupo de inclusión forzosa no se considerarán UPM, sino que formarán un estrato de grandes ciudades. En cada ciudad de este estrato se realizará un muestreo de la siguiente manera:

$$\underbrace{Sector}_{UPM} \rightarrow \underbrace{Vivienda}_{USM} \rightarrow \underbrace{Hogar}_{UFM}$$

Es necesario tener en cuenta esa particularidad de algunas encuestas para aplicar correctamente esta técnica de aproximación de varianzas. En resumen, para aquellas ciudades que pertenecen al estrato de inclusión forzosa, las UPM serán los sectores cartográficos y, para el resto del país, las UPM serán los municipios cuya probabilidad de inclusión en la muestra de la primera etapa es inferior a uno.

C. Linealización de Taylor

Cuando se trata de estimar parámetros que tienen una forma no lineal, es posible recurrir a herramientas de análisis matemático para aproximar sus varianzas y posteriormente publicar las cifras oficiales con sus respectivos errores estándar. Según Valliant, Dever y Kreuter (2013), esta técnica se basa en la expresión del estimador como función de estimadores lineales de totales. Por ejemplo, si se desea estimar un parámetro poblacional θ, que a su vez depende de Q estimadores de totales $(t_1,t_2,\ldots,t_Q)$, su estimador de muestreo se debe expresar como $\hat{\theta}=f(\hat{t}_1,\ldots,\hat{t}_Q)$, donde $\hat{t}_j=\sum_{k\in s} w_k\, y_{jk}$ es un estimador del j-ésimo total. Por consiguiente, si el estimador de interés no es una función lineal de totales, es necesario aproximar las propiedades estadísticas comunes como insesgamiento, eficiencia y precisión de los estimadores. La técnica de linealización de Taylor se utiliza a menudo para encontrar aproximaciones lineales de primer orden. Gutiérrez (2016, cap. 8) presenta una explicación detallada de esta técnica aplicada a diferentes escenarios de estimación y enumera los siguientes pasos para construir un estimador linealizado de la varianza de una función no lineal de totales:

i) Expresar el estimador del parámetro de interés $\hat{\theta}$ como una función de estimadores de totales insesgados. Así,

$$\hat{\theta}=f(\hat{t}_1,\hat{t}_2,\ldots,\hat{t}_Q)$$

ii) Determinar todas las derivadas parciales de f con respecto a cada total estimado t_q y evaluar el resultado en las cantidades poblacionales t_q. Así,

$$a_q=\left.\frac{\partial f(\hat{t}_1,\ldots,\hat{t}_Q)}{\partial \hat{t}_q}\right|_{\hat{t}_1=t_1,\ldots,\hat{t}_Q=t_Q}$$

iii) Aplicar el teorema de Taylor para funciones vectoriales a fin de linealizar la estimación $\hat{\theta}$ con $a=(t_1,t_2,\cdots,t_Q)'$. En el paso anterior, se vio que $\nabla\hat{\theta}=(a_1,\cdots,a_Q)$.

Por consiguiente,

$$\hat{\theta}=f(\hat{t}_1,\ldots,\hat{t}_Q)\cong\theta+\sum_{q=1}^{Q}a_q(\hat{t}_q-t_q)$$

iv) Definir una nueva variable E_k con $k\in s$ al nivel de cada elemento observado en la muestra aleatoria. Así,

$$E_k=\sum_{q=1}^{Q}a_q\,y_{qk}$$

v) Si los estimadores $\hat{t}_q$ son estimadores de Horvitz-Thompson, la expresión que aproxima la varianza de $\hat{\theta}$ está dada por:

$$Var(\hat{\theta})=Var\left(\sum_{q=1}^{Q}a_q\,\hat{t}_{q,\pi}\right)=Var\left(\sum_{S}\frac{E_k}{\pi_k}\right)=\sum_{U}\sum_{U}\Delta_{kl}\frac{E_k}{\pi_k}\frac{E_l}{\pi_l}$$

Como se señaló anteriormente, Gutiérrez (2016) afirma que, para encontrar una estimación de la varianza de $\hat{\theta}$, no es posible utilizar directamente los valores E_k, porque estos dependen de los totales poblacionales (las derivadas a_q se evalúan en los totales poblacionales que son desconocidos). Por consiguiente, los valores E_k se aproximan reemplazando los totales desconocidos por sus estimadores. Siendo e_k la aproximación de la variable linealizada dada por:

$$e_k=\sum_{q=1}^{Q}\hat{a}_q\,y_{qk}$$

Donde $\hat{a}_q$ corresponde a un estimador de $\hat{a}_q$. La aproximación de Taylor para el estimador de la varianza del estimador de Horvitz-Thompson para un total está dada por la siguiente expresión:

$$\widehat{Var}(\hat{t}_{y,\pi})=\sum\sum_{S}\frac{\Delta_{kl}}{\pi_{kl}}\frac{e_k}{\pi_k}\frac{e_l}{\pi_l}$$

Por ejemplo, en este contexto, si se quisiera estimar la tasa de desocupación (función no lineal de totales), definida como el cociente entre el total poblacional de personas en edad de trabajar que carecen de un empleo (t_y) y la cantidad de personas que pertenecen a la población económicamente activa (t_z), la estimación de la aproximación de la varianza del estimador de esta razón $\hat{\theta}=\frac{\hat{t}_{y,\pi}}{\hat{t}_{z,\pi}}$ estaría definida en términos de variables linealizadas de la siguiente manera:

$$e_k=\frac{1}{\hat{t}_{z,\pi}}(y_k-\hat{\theta}z_k)$$

Si, además, el muestreo de la encuesta es de dos etapas con selección aleatoria simple sin reemplazo en cada etapa, el estimador de la varianza tomaría la siguiente forma:

$$\widehat{Var}(\hat{\theta}) = \frac{N_I^2}{n_I}\left(1 - \frac{n_I}{N_I}\right) S^2_{\hat{t}_e S_I} + \frac{N_I}{n_I}\sum_{i \in S_I} \frac{N_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right) S^2_{e_{S_i}}$$

Donde $S^2_{\hat{t}_e S_I}$ es la varianza muestral de los totales estimados t_{ei} de las UPM seleccionadas en la primera etapa del muestreo y $S^2_{e_{S_i}}$ es la varianza muestral entre los valores e_k para los elementos incluidos en la submuestra dentro de cada UPM seleccionada en la primera etapa. De la misma manera, en el caso particular de la estimación de un promedio mediante el estimador de Hájek, las expresiones anteriores pueden adaptarse convenientemente.

Si se utiliza un estimador de calibración para el total poblacional de la característica de interés t_y, de acuerdo con los lineamientos de Gutiérrez (2016, sec. 10.6), la varianza estimada del estimador mediante la técnica de linealización de Taylor utilizaría las siguientes variables linealizadas:

$$e_k = y_k - x_k \hat{\theta}$$

Donde x_k corresponde a las variables relacionadas con el vector de totales auxiliares t_x, medidas en la misma encuesta, y $\hat{\theta}$ es el vector estimado de coeficientes de regresión entre los valores que toman la característica de interés y_k y el vector de información auxiliar x_k.

En la región, tanto en la Encuesta Nacional Permanente de Hogares (PNADC) del Brasil como en la Encuesta de Caracterización Socioeconómica Nacional (CASEN) de Chile, se utilizan sistemas de linealización de Taylor junto con el acercamiento del último conglomerado. En resumen, la linealización de Taylor supone que es posible definir una aproximación lineal de $\hat{\theta}$ de la siguiente manera:

$$\hat{\theta} - \theta \approx \sum_{j=1}^{p} \frac{\partial f(\hat{t}_1, \ldots, \hat{t}_p)}{\partial \hat{t}_j}(\hat{t}_j - t_j) = \sum_{k \in s} w_k e_k + c$$

Donde $e_k = \sum_{j=1}^{p} \frac{\partial f(\hat{t}_1, \ldots, \hat{t}_p)}{\partial \hat{t}_j} y_{jk}$ son variables linealizadas, mientras la cantidad c representa una constante determinística que no contribuye a la varianza de $\hat{\theta}$. Expresar esta aproximación de esa manera es muy conveniente, pues, al final, las cantidades que intervienen en la varianza se pueden expresar como una suma ponderada de las variables e_k y, por consiguiente, es posible aplicar todos los principios establecidos anteriormente. De esta forma, considerando el escenario de muestreo planteado en las secciones anteriores, el estimador de la varianza de la aproximación lineal de $\hat{\theta}$ está dado por:

$$\widehat{Var}(\hat{\theta}) = \sum_h \frac{n_h}{n_h - 1} \sum_{i \in s_h} (\hat{t}_{e_i} - \bar{\hat{t}}_{e_h})^2$$

Donde $\hat{t}_{e_i} = \sum_{k \in s_{hi}} w_k e_k$ y $\bar{\hat{t}}_{e_h} = (1/n_h) \sum_{i \in s_h} \hat{t}_{e_i}$. Por ejemplo, si se desea estimar una razón, las nuevas variables linealizadas son $e_k = (1/\hat{t}_{y_2})(y_{1k} - \hat{\theta} y_{2k})$.

D. Pesos replicados

La complejidad del cálculo de los errores de muestreo depende del estimador elegido y del diseño de muestreo utilizado para la recolección de la información primaria. En algunos casos, el proceso de linealización puede resultar complicado, por lo que es posible optar por una estrategia computacional aproximada que permite pasar por alto el proceso teórico de definición de las cantidades que estiman la varianza del estimador. Este conjunto de métodos supone la selección sistemática de submuestras para estimar el parámetro de interés, utilizando los mismos principios de estimación que con la muestra completa. Así, se obtienen estimaciones puntuales para cada réplica, que se utilizan para estimar la varianza del estimador de interés.

A falta de fórmulas adecuadas, en los últimos años se han propuesto diversas técnicas empíricas mediante las cuales se obtienen varianzas aproximadas que parecen satisfactorias para fines prácticos (Kish, 1965). Estos métodos utilizan una muestra de datos para construir submuestras y generar una distribución para las estimaciones de los parámetros de interés utilizando cada submuestra. Los resultados de la submuestra se analizan para obtener una estimación del parámetro e intervalos de confianza para esa estimación. El enfoque general de esta técnica computacional se basa en:

i) Dividir toda la muestra en pequeños subconjuntos (réplicas).

ii) Repetir los mismos procesos de ajuste de ponderadores en cada réplica.

iii) Hacer la estimación en cada subgrupo.

iv) Calcular la varianza del estimador de manera simple como la varianza muestral de todas las estimaciones en cada réplica.

Con esta metodología, no es necesario que las bases de datos públicas contengan la información asociada a los estratos o UPM y esto protege la anonimización de los encuestados. Tampoco es necesario conocer el diseño de muestreo utilizado en la encuesta, pues, al proporcionar los pesos replicados en las bases de datos, los investigadores pueden estimar el error de muestreo de forma automática y sin necesidad de intrincadas fórmulas matemáticas. Estos métodos han resultado ser eficientes y precisos para la mayoría de los parámetros de interés. Entre las encuestas de los Estados Unidos en las que se utilizan estos métodos, se encuentran la Encuesta sobre la Comunidad Estadounidense, la Encuesta sobre la Vivienda en los Estados Unidos y la Encuesta Continua de Población. En América Latina y el Caribe, estas técnicas para la estimación de la varianza de algunos estimadores complejos se han utilizado en la Encuesta Nacional de Hogares Continua del Brasil, la Encuesta Nacional de Empleo de Chile y la Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU) del Ecuador.

En particular, hay tres métodos que abordan este problema: la técnica de réplicas repetidas balanceadas (McCarthy, 1969; Judkins, 1990), la técnica de jackknife (Krewski y Rao, 1981) y la técnica de bootstrap (Rao y Wu, 1988). La idea general que subyace a estos métodos es que, a partir de la muestra completa, en cada réplica se selecciona un conjunto

de UPM, manteniendo todas las unidades seleccionadas dentro de esas UPM. Después, es necesario volver a ponderar los pesos de muestreo para mantener la representatividad y, de esta manera, obtener un nuevo conjunto de pesos de muestreo para cada réplica. Con estos pesos se calcula la estimación de interés y se obtienen tantas estimaciones como réplicas definidas. Wolter (2007) proporciona todos los detalles teóricos referentes al problema de la estimación de la varianza mediante los pesos replicados.

Con respecto a las técnicas de remuestreo y la utilización de los pesos replicados para el cálculo de los errores de muestreo, se recalca que la técnica de jackknife es útil para estimar parámetros lineales, pero no tanto cuando se trata de estimar percentiles o funciones de distribución. La técnica de réplicas repetidas balanceadas es útil para estimar parámetros lineales y no lineales, pero puede ser deficiente en el caso de dominios pequeños, que pueden determinar estimaciones nulas en la configuración de los pesos. Sin embargo, como se explicará más adelante, el ajuste de Fay aplicado a la técnica anterior contribuye a superar todos los inconvenientes mencionados. En este caso, es importante utilizar una matriz de Hadamard que no produzca más de 120 conjuntos de pesos replicados para que la publicación de la base de datos no se sobrecargue. Por último, la técnica de bootstrap ha de utilizarse con cautela, porque debe reproducir el diseño de muestreo exacto y esto se logra construyendo una población a partir de los pesos de muestreo.

1. Técnica de jackknife

Con este método se obtienen estimaciones eficientes para estimadores lineales y no lineales (a excepción de los percentiles). En su forma básica, los pesos replicados se crean al retirar una UPM del análisis. Por ende, el número de pesos replicados será igual al número de UPM que existan en la muestra. Además, al retirar una UPM de la réplica, se retiran también todas las unidades dentro de esa UPM. El procedimiento de jackknife se basa en un método utilizado por Quenouille (1956) para reducir el sesgo de las estimaciones. El refinamiento ulterior del método (Mosteller, 1968) llevó a su aplicación en las ciencias sociales, en las que no se dispone fácilmente de fórmulas para el cálculo de errores de muestreo.

Este procedimiento ofrece mayor flexibilidad, pues la técnica de jackknife puede aplicarse a una gran variedad de diseños muestrales, y facilidad de uso, pues no se requiere un *software* especializado. Se parte de una muestra de tamaño n, que se divide en A grupos de igual tamaño $m=n/A$. A partir de esta división, la varianza de un estimador $\hat{\theta}$ se estima a partir de la varianza observada en los A grupos.

Para cada grupo $(a=1,2,...,A)$, se calcula $\hat{\theta}_{(a)}$, una estimación para el parámetro θ, calculada de la misma forma que la estimación $\hat{\theta}$ obtenida con la muestra completa, pero solo con la información restante (tras la eliminación del grupo a). Para $a=1,2,...,A$ se define

$$\hat{\theta}_a = A\hat{\theta} - (A-1)\,\hat{\theta}_{(a)}$$

como un seudovalor de θ. El estimador obtenido mediante la técnica de jackknife se presenta como una alternativa a $\hat{\theta}$ y se define como:

$$\hat{\theta}_{JK} = \frac{1}{A}\sum_{a=1}^{A}\hat{\theta}_a$$

mientras que el estimador de la varianza obtenido mediante la técnica de jackknife se expresa como:

$$\widehat{Var}_{JK1} = \frac{1}{A(A-1)}\sum_{a=1}^{A}(\hat{\theta}_a - \hat{\theta}_{JK})^2$$

También es posible utilizar el siguiente estimador alternativo:

$$\widehat{Var}_{JK2} = \frac{1}{A(A-1)}\sum_{a=1}^{A}(\hat{\theta}_a - \hat{\theta})^2$$

En el caso de diseños estratificados y en múltiples etapas en los que las UPM se han seleccionado en el estrato h, para $h=1,\ldots,H$, el estimador de varianza de jackknife para la estimación de un parámetro poblacional está dado por:

$$\widehat{Var}_{JK}(\hat{\theta}) = \sum_{h=1}^{H}\frac{n_{Ih}-1}{n_{Ih}}\sum_{i=1}^{n_{Ih}}(\hat{\theta}_{h(i)} - \hat{\theta})^2$$

Donde $\hat{\theta}_{h(i)}$ es la estimación de θ utilizando los datos de la muestra y excluyendo las observaciones en la i-ésima UPM (Korn y Graubard, 1999, pág. 29-30). Shao y Tu (2012) garantizan la convergencia en probabilidad de este estimador hacia la varianza teórica, por lo que se puede concluir que es un estimador aproximadamente insesgado para la varianza teórica. Los pesos de la unidad k que pertenece a la UPM U_i en el estrato U_h están dados por la siguiente expresión:

$$d_{hk}^{i} = \begin{cases} 0 & \text{si } U_i \in U_h \text{ y } k \in U_i \\ d_k & \text{si } k \notin U_h \\ \dfrac{n_{Ih}}{n_{Ih}-1}\, d_k & \text{si } U_i \in U_h \text{ y } k \notin U_i \end{cases}$$

Donde n_{Ih} es el número de UPM seleccionadas en el estrato U_h. Por último, para reducir el número de pesos replicados, es posible conformar unidades de varianza, uniendo varias UPM dentro de un mismo estrato, y estratos de varianza, mediante la unificación de estratos dentro de la muestra. En el primer caso, se podrían emparejar las UPM en cada estrato de acuerdo con la medida de tamaño. En este caso, el estimador de varianza se traduce en la siguiente expresión:

$$\widehat{Var}_{JK}(\hat{\theta}) = \sum_{h}\frac{n_{Ih}-n_{Ihg}}{n_{Ih}}\sum_{i\in s_{hg}}^{n_{Ih}}(\hat{\theta}_{h(g)} - \hat{\theta})^2$$

Donde $\hat{\theta}_{hg}$ es el estimador del parámetro retirando el g-ésimo subgrupo del estrato U_h y n_{Ihg} es el tamaño del subgrupo en la muestra denotado como s_{hg}. En el cuadro X.1 se ejemplifica la estructura final de una base de datos de pesos replicados con esta técnica en un conjunto reducido de ocho unidades muestrales divididas en cuatro UPM y dos estratos. Se hace hincapié en que habrá tantos conjuntos (columnas) de pesos replicados mediante la técnica de jackknife como UPM existentes en la muestra de la primera etapa.

■ Cuadro X.1
Ejemplo reducido de creación de pesos replicados con la técnica de jackknife

k	Estrato	Unidad primaria de muestreo (UPM)	$d_k^{(1)}$	$d_k^{(2)}$	$d_k^{(3)}$	$d_k^{(4)}$
1	Estrato1	UPM1	0,00	1,03	1,03	1,03
2	Estrato1	UPM1	0,00	1,03	1,03	1,03
3	Estrato1	UPM2	1,03	0,00	1,03	1,03
4	Estrato1	UPM2	1,03	0,00	1,03	1,03
5	Estrato2	UPM3	1,03	1,03	0,00	1,03
6	Estrato2	UPM3	1,03	1,03	0,00	1,03
7	Estrato2	UPM4	1,03	1,03	1,03	0,00
8	Estrato2	UPM4	1,03	1,03	1,03	0,00

Fuente: Elaboración propia.

2. Método de réplicas repetidas balanceadas

Esta técnica se desarrolló para los diseños en los que se seleccionan dos UPM por estrato. El método funciona sistemáticamente para la estimación de parámetros lineales y no lineales (incluidos los percentiles) y, además, asegura la máxima dispersión de las UPM en las regiones geográficas (estratos) (Valliant y Dever, 2017). Cabe señalar que, si el submuestreo en cada estrato es $n_{Ih}=2$, con la técnica de jackknife se deberían definir 2^H posibles réplicas al seleccionar aleatoriamente una UPM en cada estrato, lo cual puede resultar imposible de calcular mediante programas informáticos.

Para lograr la misma eficiencia reduciendo el número de pesos replicados, es posible utilizar un enfoque ortogonal con matrices de Hadamard, que son matrices cuadradas cuyas columnas deben ser ortogonales. Por ejemplo, se considera la siguiente matriz:

$$\begin{pmatrix} +1 & +1 & +1 & +1 \\ +1 & -1 & +1 & -1 \\ +1 & +1 & -1 & -1 \\ +1 & -1 & -1 & +1 \end{pmatrix}$$

Al asumir que el valor +1 significa que la primera UPM se mantiene como parte de la réplica y la segunda UPM se retira de la réplica, el valor -1 significa que la segunda UPM se mantiene como parte de la réplica y la primera UPM se retira de la réplica. Por lo tanto,

en cada réplica se retira una UPM por estrato. Esto significa que el producto punto entre cualquier combinación de dos columnas deber ser igual a 0. Por ejemplo, al tomar las columnas 2 y 4, se obtiene:

$$(+1,-1,+1,-1)'\cdot(+1,-1,-1,+1)=1+1-1-1=0$$

De esta forma, el número de réplicas ortogonales será igual al menor múltiplo de 4 mayor o igual al número de estratos. Las UPM que se mantienen en cada réplica se conocen como semimuestras (*half-samples*). Por consiguiente, el peso de las personas en la UPM que se mantiene se multiplica por un factor de 2. Así, se obtiene:

$$d_k = \begin{cases} 0 & \text{si } k \text{ pertence a la UPM retirada} \\ 2d_k & \text{en caso contrario} \end{cases}$$

Con el método de réplicas repetidas balanceadas, el estimador de la varianza toma la siguiente forma:

$$\widehat{Var}_{BRR}(\hat{\theta}) = \frac{1}{A}\sum_{a=1}^{A}(\hat{\theta}_a - \hat{\theta})^2$$

Donde $\hat{\theta}_a$ es el estimador del parámetro de interés en la réplica a. En el cuadro X.2 se ejemplifica la estructura final de una base de datos de pesos replicados con esta técnica en el mismo conjunto reducido, considerando que hay dos estratos.

■ Cuadro X.2
Ejemplo reducido de creación de pesos replicados con la técnica de réplicas repetidas balanceadas

k	Estrato	Unidad primaria de muestreo (UPM)	$d_k^{(1)}$	$d_k^{(2)}$
1	Estrato1	UPM1	2	0
2	Estrato1	UPM1	2	0
3	Estrato1	UPM2	0	2
4	Estrato1	UPM2	0	2
5	Estrato2	UPM3	2	0
6	Estrato2	UPM3	2	0
7	Estrato2	UPM4	0	2
8	Estrato2	UPM4	0	2

Fuente: Elaboración propia.

Una desventaja del método de réplicas repetidas balanceadas es que las unidades en dominios con muestras pequeñas pueden estar ausentes en algunas combinaciones de pesos replicados por el diseño ortogonal. Esto conlleva una pérdida de precisión en el cálculo del error estándar. Una solución a este problema consiste en modificar los pesos en los pesos replicados. Para la aplicación de las réplicas repetidas balanceadas se recomienda utilizar el método de Fay, en el que se siguen los lineamientos basados en la

matriz de Hadamard, aunque las UPM no se retiran completamente, sino que su peso se modifica de la siguiente manera:

$$d_k^a = \begin{cases} \rho * d_k & \text{si } k \text{ pertence a la UPM retirada} \\ (2-\rho)\, d_k & \text{en caso contrario} \end{cases}$$

Donde $0<\rho<1$. En algunos estudios por simulación, se alcanzaron buenos niveles de eficiencia para valores de ρ iguales a 0,3, 0,5 o 0,7. Con el método de réplicas repetidas balanceadas y el ajuste de Fay, el estimador de la varianza toma la siguiente forma:

$$\widehat{Var}_{Fay}(\hat{\theta}) = \frac{1}{A(1-\rho)^2} \sum_{a=1}^{A} (\hat{\theta}_a - \hat{\theta})^2$$

Donde $\hat{\theta}_a$ es el estimador del parámetro de interés en la réplica a. En el cuadro X.3 se ejemplifica la estructura final de una base de datos de pesos replicados con el ajuste de Fay en el mismo conjunto reducido.

■ Cuadro X.3
Ejemplo reducido de creación de pesos replicados con el ajuste de Fay

k	Estrato	Unidad primaria de muestreo (UPM)	$d_k^{(1)}$	$d_k^{(2)}$
1	Estrato1	UPM1	1,5	0,5
2	Estrato1	UPM1	1,5	0,5
3	Estrato1	UPM2	0,5	1,5
4	Estrato1	UPM2	0,5	1,5
5	Estrato2	UPM3	1,5	0,5
6	Estrato2	UPM3	1,5	0,5
7	Estrato2	UPM4	0,5	1,5
8	Estrato2	UPM4	0,5	1,5

Fuente: Elaboración propia.

En general, al aplicar estos métodos, los pesos de muestreo se ajustan para generar los pesos replicados y, posteriormente, se repiten los ajustes por falta de respuesta y calibración para estos nuevos pesos. Con esta metodología se estiman los errores de muestreo y la varianza de muestreo, incluido el impacto de la falta de respuesta, que se espera sea pequeño pero relevante en el momento de calcular estimadores más precisos. De conformidad con las observaciones anteriores, cuando la encuesta presenta estratos que comprenden una sola UPM, el método de réplicas repetidas balanceadas no es aplicable, puesto que, al eliminar una unidad, algunos estratos quedarán vacíos.

3. Técnica de bootstrap

En este apartado se presenta la técnica de bootstrap (Efron y Tibshirani, 1993), ampliamente utilizada por su fácil implementación y flexibilidad en cuanto al número de pesos replicados que se crean. A partir de los pesos muestrales, se procede a crear los pesos replicados con el método de remuestreo, a fin de calcular las estimaciones de los indicadores y las varianzas. En el contexto de las encuestas de hogares, se trata de realizar un remuestreo de las UPM seleccionadas en el marco de áreas.

La técnica de bootstrap es el método basado en réplicas más versátil para el cálculo de errores estándar. Según Valliant y Dever (2017), es muy eficiente en la estimación de parámetros lineales y no lineales, a diferencia de la técnica de jackknife, que no es eficiente en la estimación de percentiles. Asimismo, a diferencia del método de réplicas repetidas balanceadas, que requiere una muestra mínima de dos UPM por estrato, también funciona para muestras pequeñas. Este método requiere una gran cantidad de pesos replicados, en general superior a 200.

Siendo s_{BS} la submuestra bootstrap, el peso replicado de la persona k perteneciente a la UPM i del estrato h sigue la siguiente expresión:

$$d_k^b = \begin{cases} 0 & \text{si la UPM } i \text{ no pertence a } s_{BS} \\ d_k\left[1 - \sqrt{\frac{n_{Ih}^*}{n_{Ih}-1}} + \sqrt{\frac{n_{Ih}^*}{n_{Ih}-1}}\,\frac{n_{Ih}}{n_{Ih}^*}\,n_{Ihi}^*\right] & \text{en caso contrario} \end{cases}$$

Donde n_{Ih} es el úmero de UPM en la muestra original del estrato h, n_{Ih}^* es el número de UPM en la muestra bootstrap y n_{Ihi}^* es el número de veces que la UPM i fue seleccionada en la muestra bootstrap. En este caso se selecciona una muestra bootstrap con $m_h^*=m_h-1$, y los pesos toman la siguiente forma:

$$d_k^a = \begin{cases} 0 & \text{si la UPM } i \text{ no pertence a } s_{BS} \\ d_k\left[1 - \sqrt{\frac{n_{Ih}^*}{n_{Ih}-1}} + \sqrt{\frac{n_{Ih}^*}{n_{Ih}-1}}\,\frac{n_{Ih}}{n_{Ih}^*}\,n_{Ihi}^*\right] & \text{en caso contrario} \end{cases}$$

Con la técnica de bootstrap, el estimador de la varianza toma la siguiente forma:

$$\widehat{Var}_{BS}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b - \hat{\theta})^2$$

Donde $\hat{\theta}_b$ es el estimador del parámetro de interés en la réplica b inducida por la muestra bootstrap. En resumen, para la b-ésima réplica con los pesos resultantes se podrán calcular las estimaciones de totales, proporciones, promedios y razones y sus respectivas varianzas o desviaciones. En general, para reflejar el ajuste de los pesos en

los pesos replicados, es necesario trabajar con la muestra originalmente seleccionada, que contendrá unidades no elegibles y unidades que no respondieron. En cada réplica se deberán realizar los mismos ajustes que se hicieron a la muestra original. En caso de calibración de pesos, también deberá incluirse este proceso en cada réplica, para asegurar que el error estándar generado por estos métodos incluya el incremento (o la disminución) de la varianza definida por estos ajustes de los pesos. En el cuadro X.4 se ejemplifica la estructura final de una base de datos de pesos replicados con la técnica de bootstrap en el mismo conjunto reducido.

■ Cuadro X.4
Ejemplo reducido de creación de pesos replicados con la técnica de bootstrap

k	Estrato	Unidad primaria de muestreo (UPM)	$d_k^{(1)}$	$d_k^{(2)}$	$d_k^{(3)}$	$d_k^{(4)}$
1	Estrato1	UPM1	2	0	1	1
2	Estrato1	UPM1	2	0	1	1
3	Estrato1	UPM2	0	2	1	1
4	Estrato1	UPM2	0	2	1	1
5	Estrato2	UPM3	1	1	2	0
6	Estrato2	UPM3	1	1	2	0
7	Estrato2	UPM4	1	1	0	2
8	Estrato2	UPM4	1	1	0	2

Fuente: Elaboración propia.

Rao y Wu (1984 y 1988) aconsejan seleccionar una muestra con reemplazo de $n_I - 1$ de las n_I UPM de la muestra, teniendo en cuenta la probabilidad de selección del diseño complejo en la primera etapa. Dado que la selección es con reemplazo, es posible que una UPM resulte seleccionada más de una vez en esta nueva muestra. Por otra parte, también es posible realizar una selección sin reemplazo. En este caso, Preston (2009) recomienda seleccionar una muestra con reemplazo de $n_I/2$ de las n_I UPM de la muestra, teniendo en cuenta la probabilidad de selección del diseño complejo en la primera etapa.

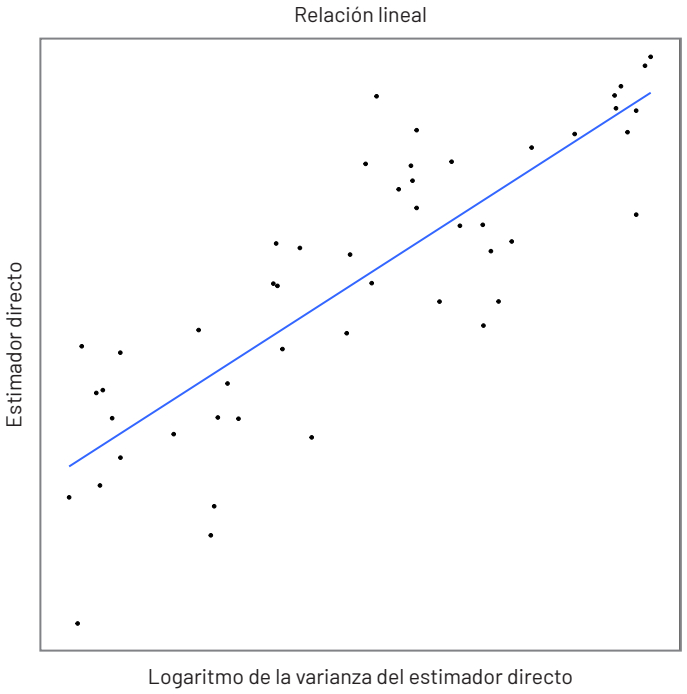
E. Función de la varianza generalizada

Existe también la posibilidad de estimar las varianzas de los estimadores de muestreo (definido sobre la medida de probabilidad determinada por el diseño de muestreo p) mediante un modelo (definido sobre una medida de probabilidad m) que simplifica el proceso computacional en la generación de las miles de estimaciones que se producen a partir de las encuestas de hogares en la región. De acuerdo con Wolter (2007), si los parámetros de este modelo pueden estimarse a partir de encuestas pasadas o de un conjunto de datos reducido, las estimaciones de la varianza (y, por consiguiente, las estimaciones del error de muestreo) pueden realizarse simplemente evaluando el modelo según los datos actuales de la encuesta.

Un caso particular de este tipo de relaciones se presenta cuando se deben obtener estimaciones de la encuesta a nivel subnacional para la publicación de cuadros de salida con la estimación puntual y el error estándar estimado. En estos casos, es muy común que la cantidad total de celdas en los cuadros de salida sea muy grande, de manera que la utilización de este tipo de modelos, denominados en la literatura función de la varianza generalizada (*generalized variance function*), puede ser una mejor opción en términos de eficiencia computacional. Otro caso especial que el personal técnico de las oficinas nacionales de estadística (ONE) debe abordar se observa cuando el tamaño de la muestra de las subpoblaciones de interés es pequeño o cuando las observaciones de la muestra en la subpoblación no son suficientemente heterogéneas. En este caso, es muy probable que las estimaciones de la varianza de los estimadores de los totales, los tamaños, las proporciones o las medias sean imprecisas. En efecto, no es insólito encontrar estimaciones de la varianza iguales a 0. En este caso, este tipo de estimaciones deben cotejarse con pericia o reemplazarse por una mejor aproximación, que puede estar basada en la función de la varianza generalizada.

Un aspecto importante en la modelación de las varianzas de los estimadores es reconocer la naturaleza de este parámetro, que será siempre positivo. Por ende, al tratar de modelarlas es útil realizar un análisis log-lineal, que permite lidiar con este tipo de estructuras. En el gráfico X.1 se muestra la relación que puede establecerse entre las estimaciones directas de las proporciones de pobreza municipal y el logaritmo de sus varianzas estimadas.

■ Gráfico X.1
Relación entre un estimador de la tasa de pobreza estimada y el logaritmo de la estimación directa de su varianza



Fuente: Elaboración propia.

En términos de notación, $Var_{GVF}\ (\hat{\theta})=E_m\ (\ \widehat{Var}(\hat{\theta}))$ será la varianza suavizada del estimador directo $\hat{\theta}$. Un aspecto importante en este tipo de modelos es que, en general, no es posible tratar $Var(\hat{\theta})$ como un valor fijo, puesto que no es estrictamente una función de las covariables auxiliares. A partir del acceso a un estimador insesgado de $Var(\hat{\theta})$, denotado por $\widehat{Var}(\hat{\theta})$, se obtiene:

$$E_{mp}\ (\widehat{Var}\ (\hat{\theta}))=E_m\ (E_p\ (\widehat{Var}\ (\hat{\theta})))=E_m\ (Var(\hat{\theta}))=Var_{GVF}\ (\hat{\theta})$$

Donde los subíndices m y p hacen referencia a la medida de probabilidad del modelo y del diseño de muestreo, respectivamente. Aunque el diseño de muestreo determina estimadores de las varianzas insesgados, estos tienden a ser inestables cuando el tamaño de la muestra es pequeño, que es precisamente el paradigma dominante en la desagregación de estimaciones. Rivest y Belmonte (2000) consideran modelos de suavización para la estimación de las varianzas directas definidos de la siguiente manera:

$$log(\widehat{Var}\ (\hat{\theta}))=\mathbf{z}_d'\boldsymbol{\alpha}+\varepsilon_d'$$

Donde z_d es un vector de covariables explicativas, α es un vector de parámetros que deben estimarse, ε_d son errores aleatorios con media 0 y varianza constante, que se asumen idénticamente distribuidos condicionalmente sobre z_d. La estimación suavizada de la varianza de muestreo del modelo anterior está dada por:

$$Var_{GVF}(\hat{\theta}) = E_{mp}(\widehat{Var}(\hat{\theta})) = exp(z_d'\boldsymbol{\alpha})\cdot\Delta$$

Donde $E_{mp}(\varepsilon_d)=\Delta$. No es necesario especificar una distribución paramétrica para los errores de este modelo. Al utilizar el método generalizado de momentos (MGM), se obtiene el siguiente estimador insesgado para Δ:

$$\hat{\Delta} = \frac{\sum_{d=1}^{D} \widehat{Var}(\hat{\theta})}{\sum_{d=1}^{D} exp(z_d'\boldsymbol{\alpha})}$$

De la misma forma, al utilizar el método de mínimos cuadrados ordinarios (MCO), la estimación del coeficiente de parámetros de regresión está dada por la siguiente expresión:

$$\hat{\boldsymbol{\alpha}} = \left(\sum_{d=1}^{D} z_d z_d'\right)^{-1} \sum_{d=1}^{D} z_d \log(\widehat{Var}(\hat{\theta}))$$

Por último, el estimador suavizado de la varianza muestral está definido por:

$$\widehat{Var}_{GVF}(\hat{\theta}) = exp(z_d'\hat{\boldsymbol{\alpha}})\hat{\Delta}$$

Rivest y Belmonte (2000) concluyeron que este estimador no sobrestima ni subestima la varianza suavizada, pues el promedio de las estimaciones suavizadas $\widehat{Var}_{GVF}(\hat{\theta})$ coincide con el promedio de las varianzas directas $\widehat{Var}(\hat{\theta})$. Por lo tanto:

$$\frac{\sum_{d=1}^{D} \widehat{Var}_{GVF}(\hat{\theta})}{D} = \frac{\sum_{d=1}^{D} \widehat{Var}(\hat{\theta})}{D}$$

La Oficina de Estadística del Canadá utiliza este tipo de modelos para elaborar las cifras oficiales del mercado de trabajo de 149 áreas censales (Beaumont y Bocci, 2016). En primer lugar, se ajusta el modelo mediante la exclusión de las áreas con menos de diez personas en la fuerza de trabajo (denominador del indicador). De la misma manera, se excluyen del modelo todas las áreas con estimador de varianza directa $\widehat{Var}(\hat{\theta})$ igual a 0, pues esto significaría que no se encontró ningún caso efectivo en el numerador del indicador. Asimismo, la estimación de la varianza directa se basa en el diseño de muestreo complejo, mientras que la estimación de la varianza suavizada está supeditada al siguiente modelo de regresión:

$$\log(\widehat{Var}(\hat{\theta})) = z_d'\boldsymbol{\alpha} + \varepsilon_d$$

Donde:

$$z_d' = \left(1, log\left(\frac{N_d^{EIB}}{N_d^{15+}}\right), log\left(1 - \frac{N_d^{EIB}}{N_d^{15+}}\right), log\left(N_d^{15+}\right)\right)'$$

N_d^{EIB} es el número de beneficiarios del seguro de desempleo en el área d y N_d^{15+}, el número de personas en la fuerza de trabajo. A fin de evitar posibles sesgos en las áreas con muestras grandes, se decidió que, en el caso de las áreas con un número de casos efectivo superior a 400, la estimación suavizada por la función de la varianza generalizada fuese igual a la estimación directa, es decir, $\widehat{Var}_{GVF}(\hat{\theta}) = \widehat{Var}(\hat{\theta})$.

En otra aplicación práctica, Fuquene y otros (2019) estiman la prevalencia de migrantes internacionales en los municipios de Colombia mediante un modelo de área en el que utilizan el enfoque de la función de la varianza generalizada, con un modelo planteado en términos de una relación log-lineal con el siguiente vector de covariables auxiliares:

$$z_d' = \left(1, \hat{\theta}_d, \sqrt{\hat{\theta}_d}, n_d, \sqrt{n_d}, \sqrt{\hat{\theta}_d \times n_d}\right)'$$

Asimismo, una de las aplicaciones más citadas se presenta en Fay y Herriot (1979). En este artículo fundacional sobre los modelos de estimación en áreas pequeñas, se relata que la Oficina del Censo de los Estados Unidos realizó un censo con una muestra cocensal del 20% en cada estado para estimar el ingreso per cápita. Para estimar este indicador a nivel desagregado, se utilizó la estimación directa acompañada con una función de la varianza generalizada como estimador suavizado. Este modelo tomó los resultados de ocho estados y los generalizó para el resto del país. Como resultado de esta modelación, el coeficiente de variación y la varianza se establecieron como una función del tamaño del área.

En la región, el Ministerio de Desarrollo y Familia de Chile y la Comisión Económica para América Latina y el Caribe (MDSF/CEPAL, 2021) utilizaron un modelo de función de la varianza generalizada para estimar las varianzas de las tasas de pobreza comunal a partir de la CASEN 2020. La variable dependiente del modelo fue el logaritmo natural de la estimación de la varianza directa de las tasas de pobreza, mientras que el intercepto, la estimación directa de la tasa de pobreza, el tamaño de la muestra comunal, la interacción entre la tasa de pobreza y el tamaño de muestra, la raíz cuadrada de la tasa de pobreza, la raíz cuadrada del tamaño de la muestra y, por último, la raíz cuadrada de la interacción entre la tasa de pobreza y el tamaño de la muestra se incluyeron como covariables. Las comunas incluidas en la modelación que tuvieron una tasa nula de pobreza y, por consiguiente, una estimación nula de la varianza del estimador directo no se incluyeron en el ajuste del modelo, pero sí se obtuvieron las predicciones de sus varianzas. En el mencionado informe se presentan modelos descriptivos que justifican la inclusión de las covariables y las relaciones establecidas en el modelo. Además, el factor de ajuste $\hat{\Delta}$ fue cercano a 1,2 en todas las series estudiadas.

F. Otras consideraciones sobre la estimación de la varianza de los estimadores de muestreo

En la práctica del muestreo, existen algunos paradigmas que determinan la planificación y el diseño de las encuestas. En esta sección se muestran ejemplos y contraejemplos que permiten ilustrar algunos mitos de la estimación del error de muestreo en las encuestas de hogares. Para ello, se considera la varianza del estimador de Horvitz-Thompson, dada a continuación:

$$\widehat{Var}(\hat{t}_{y,\pi})=\sum_{k\in U}\sum_{l\in U}\Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

Donde $\Delta_{kl}=(\pi_{kl}-\pi_k\ \pi_l)$ y π_{kl} denota la probabilidad de inclusión conjunta de los elementos k y l en la muestra s. En el caso de los diseños de muestreo de tamaño fijo, existen dos estimadores insesgados para esta varianza. El primero de ellos fue propuesto originalmente por Horvitz y Thompson (1952) y está dado por:

$$\widehat{Var}_1(\hat{t}_{y,\pi})=\sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

El segundo estimador fue propuesto por Sen (1953) y Yates y Grundy (1953) y está dado por la siguiente expresión:

$$\widehat{Var}_2(\hat{t}_{y,\pi})=\frac{1}{2}\sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2$$

1. Estimaciones negativas de varianza

La idea de que no puede haber estimaciones negativas de la varianza constituye un razonamiento bastante lógico e intuitivo: dado que la varianza es un parámetro positivo, no puede estimarse con cantidades negativas. Sin embargo, en la inferencia basada en el diseño de muestreo, es posible obtener estimaciones negativas de varianza para algunas estructuras poblacionales particulares y por ello se requiere un equipo de muestreo con mucha experiencia, que conozca las condiciones en que se podría presentar esta situación, a fin de poder evitarla.

Es necesario diferenciar los estimadores (que son funciones de variables aleatorias) de los parámetros (que son valores reales desconocidos). En efecto, para la varianza del estimador de Horvitz-Thompson (parámetro desconocido y siempre positivo), hay estimadores (funciones de variables aleatorias) que pueden arrojar estimaciones negativas. Es posible que las estimaciones de la varianza arrojen resultados negativos, que no pueden utilizarse ni interpretarse. Se considera el diseño de muestreo de tamaño fijo e igual a $n=2$ del cuadro X.5, que determina seis posibles muestras.

■ Cuadro X.5
Ejemplo reducido de un diseño de muestreo con tamaño de muestra *n*=2 para una población de *N*=4 elementos

s	I_1	I_2	I_3	I_4	$p(s)$
s_1	1	1	0	0	0,31
s_2	1	0	1	0	0,20
s_3	1	0	0	1	0,14
s_4	0	1	1	0	0,03
s_5	0	1	0	1	0,01
s_6	0	0	1	1	0,31

Fuente: Elaboración propia.

En el ejemplo del cuadro X.5, la probabilidad de obtener una muestra compuesta por los dos primeros elementos se fijó en 0,31, mientras que la probabilidad de obtener una muestra compuesta por el primer y el tercer elemento se fijó en 0,20, y así sucesivamente. En esta configuración se obtienen las estimaciones puntuales para cada una de las seis posibles muestras, así como las dos posibles estimaciones de la varianza. Teniendo en cuenta el cuadro X.6, se observa que en ambos escenarios existen estimaciones negativas.

■ Cuadro X.6
Ejemplo reducido de un diseño de muestreo con estimaciones de varianza negativas

s	I_1	I_2	I_3	I_4	$p(s)$	$\hat{t}_{y,\pi}$	$\widehat{Var}_1(\hat{t}_{y,\pi})$	$\widehat{Var}_2(\hat{t}_{y,\pi})$
s_1	1	1	0	0	0,31	9,560440	38,099984	-0,9287681
s_2	1	0	1	0	0,20	5,883191	-4,744190	2,4710422
s_3	1	0	0	1	0,14	4,933110	-3,680428	8,6463858
s_4	0	1	1	0	0,03	7,751323	-100,252974	71,6674365
s_5	0	1	0	1	0,01	6,801242	-165,715154	323,3238494
s_6	0	0	1	1	0,31	3,123994	3,426730	-0,1793659

Fuente: Elaboración propia.

A pesar de los resultados negativos para las varianzas, tanto el estimador del total como los dos estimadores de su varianza siguen siendo insesgados. En efecto, al multiplicar la estimación puntual por la probabilidad del diseño de muestreo, se obtienen los valores poblacionales. La varianza del estimador de Horvitz-Thompson para este diseño en particular es 6,744442, que corresponde a la esperanza de ambos estimadores

de varianza. Para evitar estas estimaciones negativas, Gutiérrez (2016) afirma que es necesario garantizar que la covarianza (Δ_{kl}) sea negativa para cada par de elementos en la población $(k \neq l)$. Esto no sucede con este diseño de muestreo, puesto que:

$$\Delta_{kl} = \begin{bmatrix} 0,2275 & 0,0825 & -0,1510 & -0,1590 \\ 0,0825 & 0,2275 & -0,1590 & -0,1510 \\ -0,1510 & -0,1590 & 0,2484 & 0,0616 \\ -0,1590 & -0,1510 & 0,0616 & 0,2484 \end{bmatrix}$$

2. Disminución de la varianza ante el aumento del tamaño de la muestra

Por otra parte, la idea de que al aumentar el tamaño de la muestra debería disminuir la varianza deriva de la lógica intuitiva según la cual el error de muestreo no debería existir si se realiza una medición completa de la población. Si bien esto es lo que ocurre por lo general, existen algunas excepciones. En algunas estrategias de muestreo, es posible encontrar situaciones en las que la varianza del estimador crece con el tamaño de la muestra. En esta sección se presenta un ejemplo en el que sucede exactamente eso.

Para ello se utiliza un ejemplo reducido. Se supone una población de $N=3$ elementos $U=\{1,2,3\}$ y se comparan dos diseños de muestreo, el primero con un tamaño de muestra fijo de n=1 y el segundo con un tamaño de muestra fijo de $n=2$. En ambos casos, la variable de interés es dicotómica y denota la presencia o ausencia del fenómeno en los individuos de la población. En el primer caso, el diseño de muestreo de tamaño de muestra $n=1$ es el siguiente (véase el cuadro X.7):

■ Cuadro X.7
Ejemplo reducido de un diseño de muestreo con tamaño de muestra *n*=1 para una población de *N*=3 elementos

s	I_1	I_2	I_3	$p(s)$
s_1	1	0	0	0,5
s_2	0	1	0	0,1
s_3	0	0	1	0,4

Fuente: Elaboración propia.

En este diseño de muestreo, la varianza del estimador de Horvitz-Thompson es igual a $Var(\hat{t}_{y,\pi})=1,5$. Sin embargo, en un segundo caso, se considera el siguiente diseño de muestreo de tamaño de muestra $n=2$ (véase el cuadro X.8):

■ Cuadro X.8
Ejemplo reducido de un diseño de muestreo con tamaño de muestra *n*=2 para una población de *N*=3 elementos

s	I_1	I_2	I_3	$p(s)$
s_1	1	1	0	0,7
s_2	1	0	1	0,2
s_3	0	1	1	0,1

Fuente: Elaboración propia.

En este diseño de muestreo, la varianza del estimador de Horvitz-Thompson es igual a $Var(\hat{t}_{y,\pi})=2,3$. Por lo tanto, no es exacto afirmar que siempre que un diseño de muestreo contemple un tamaño de muestra más grande se obtendrá necesariamente una reducción de la varianza.

Capítulo XI
Representatividad y falta de respuesta

El problema de la falta de respuesta total o parcial a las preguntas del cuestionario es una faceta normal, aunque no deseable, del desarrollo de una encuesta. Todas las encuestas de hogares sufren el fenómeno de la falta de respuesta, ya sea de hogares completos, de personas dentro de los hogares o con respecto a algunas de las variables de interés de los cuestionarios. En algunos casos, incluso después de un diseño cuidadoso y una planificación logística exhaustiva, este problema es tan grande que los resultados de la encuesta pueden quedar en entredicho. Por esta razón, es necesario considerarlo en la planificación y el diseño de la recopilación de información mediante encuestas, así como contemplar varios ajustes que prevean las consecuencias de este fenómeno. Por esta razón, en los capítulos anteriores se abordó el ajuste de la subcobertura, que garantiza que el tamaño efectivo de la muestra sea adecuado para realizar una inferencia precisa. Si el diseño de la encuesta no incluye estos ajustes, el tamaño de la muestra final se verá reducido, ya que muchos hogares no contestarán algunas preguntas del cuestionario y, en algunos casos, la totalidad del cuestionario.

No cabe duda de que la falta de respuesta puede disminuir considerablemente la calidad de las estadísticas calculadas sobre la base de una encuesta. De acuerdo con Lohr (2000), la mayoría de las encuestas presentan una falta de respuesta residual, incluso después de un diseño cuidadoso y un seguimiento de dicha falta de respuesta. El autor afirma que existen dos tipos de mecanismos de falta de respuesta:

i) Ignorable: cuando la probabilidad de que una persona responda no depende de la característica de interés. El mecanismo de la falta de respuesta puede explicarse mediante un modelo y dicha falta de respuesta puede ignorarse después de que el modelo la toma en cuenta.

ii) No ignorable: cuando la probabilidad de que una persona responda depende de la característica de interés. Por ejemplo, si en una encuesta de la fuerza laboral se desea estimar el número de personas empleadas o desempleadas, la falta de respuesta no puede ignorarse cuando depende de la clasificación laboral del individuo.

El fenómeno de la falta de respuesta y sus repercusiones negativas en la calidad de las estimaciones se ha estudiado ampliamente en la literatura. Por ejemplo, Lumley (2010, cap. 9) hace un análisis detallado de la falta de respuesta individual (cuando existen datos parciales de un encuestado), considerando un enfoque basado en el diseño de muestreo al ajustar los pesos muestrales. Fuller (2009, cap. 5) cita algunas técnicas de imputación para el tratamiento de la falta de respuesta y conjuga modelos probabilísticos con los pesos del diseño de muestreo para mitigar los efectos de este problema. El enfoque de Särndal (2011a) prevé el uso de modelos y conjuntos equilibrados para aumentar la representatividad de las estimaciones. De la misma forma, Särndal y Lundström (2010) proponen un conjunto de indicadores para evaluar la efectividad de la información auxiliar utilizada para controlar el sesgo generado por la falta de respuesta.

A. Concepto de representatividad

Si bien el concepto de representatividad se utiliza a menudo en la investigación en materia de encuestas, no existe una definición clara e inequívoca del término. En particular, Kruskal y Mosteller (1980) presentan una amplia descripción del supuesto significado del adjetivo "representativo". El concepto de "muestra representativa", por su parte, no está completamente estandarizado. De acuerdo con Bethlehem, Cobben y Schouten (2009), asimismo, algunos de los conceptos utilizados a este respecto son muy vagos e imprecisos, como se observa en los siguientes ejemplos:

- Muestra que reconoce la estructura general de los datos
- Muestra que está libre de fuerzas selectivas
- Muestra que tiene una cobertura suficiente de la población
- Muestra que constituye una miniatura de la población
- Muestra que contiene casos típicos o ideales
- Muestra que permite una buena estimación
- Muestra suficientemente buena para un propósito particular

En términos de notación, supóngase la selección de una muestra probabilística s de tamaño n sin reemplazo de una población finita U de tamaño N. La muestra puede verse como un vector de N indicadores $s = (I_1, I_2, \ldots, I_N)$, donde el indicador $I_k = 1$ si se selecciona el elemento k en la muestra, y $I_k = 0$ en caso contrario $(k = 1, 2, \ldots, N)$. El fenómeno de la falta de respuesta se modela por medio de las probabilidades de respuesta. Para esto, se supone

que cada elemento k de la población tiene una determinada probabilidad desconocida ϕ_k de responder cuando se selecciona en la muestra. La respuesta a la encuesta se puede representar mediante el vector de indicadores $D=(D_1, D_2,...,D_N)$, donde $D_k=1$ si el elemento k fue seleccionado en la muestra $(I_k=1)$ y respondió. De lo contrario, $D_k=0$. Por ende, se deduce que:

$$\phi_k = P(D_k=1 \mid I_k=1)$$

El concepto de representatividad más adecuado para definir un indicador de representatividad corresponde a la falta de fuerzas selectivas. Está claro que no existen fuerzas selectivas si todas las probabilidades de respuesta son uniformes. Esta observación forma la base de la primera definición de representatividad.

La respuesta a una encuesta es altamente representativa con respecto a la muestra si las probabilidades de respuesta de todos los elementos de la población son iguales y si la respuesta de un elemento es independiente de la respuesta de todos los demás. En otras palabras:

$$\phi_k = P(D_k=1 \mid I_k=1) = \phi_k \; =1,2,...,N$$

Se debe tener en cuenta que se garantiza un elevado nivel de representatividad cuando el mecanismo que genera la falta de datos no presenta ningún patrón y es completamente aleatorio (*missing completely at random* (MCAR)) para cada variable objetivo en el estudio. En este caso, la falta de respuesta no determina un sesgo en los estimadores. Aunque atractiva, esta definición no es muy útil, ya que en la práctica no es posible comparar las probabilidades de respuesta individual.

Por otra parte, supóngase que hay una variable auxiliar categórica X que tiene H categorías y divide a la población en H estratos (subpoblaciones). El número de elementos en el estrato h se denota mediante N_h, para $h=1,2,...,H$. Se asume que esta variable se ha medido en la encuesta y que su valor está disponible para cada persona encuestada y no encuestada. La probabilidad de respuesta del elemento k en el estrato h está definida por ϕ_{hk}.

La respuesta a una encuesta es poco representativa con respecto a la muestra para la variable auxiliar X si la probabilidad de respuesta promedio es la misma en cada estrato, es decir:

$$\bar{\phi}_h = \frac{1}{N_h} \sum_{k=1}^{N_h} \phi_{hk} = \phi h \; =1,2,...,H$$

Los bajos niveles de representatividad determinan que no sea posible distinguir a las personas encuestadas de las no encuestadas simplemente usando información con respecto a X. Si la respuesta es poco representativa con respecto a muchas variables auxiliares X, existirán relaciones fuertes entre las variables objetivo y las variables auxiliares. Dado que es posible estimar la media de las probabilidades de respuesta en los estratos, el supuesto de poca representatividad se puede comprobar en la práctica.

B. Indicadores de representatividad

Como ya se mencionó, el fenómeno de la falta de respuesta se observa en la mayoría de las encuestas y puede afectar considerablemente la calidad de los resultados. De hecho, las estimaciones de las características de la población estarán sesgadas si, debido a la falta de respuesta, algunos grupos de la población quedan sobrerrepresentados o subrepresentados. El problema se agrava cuando estos grupos se comportan de manera diferente con respecto a las variables de la encuesta. En relación con la falta de respuesta por unidad, en general los institutos nacionales de estadística (INE) de la región utilizan la tasa de respuesta de la encuesta como un indicador de la calidad de la operación.

Dado que una baja tasa de respuesta no necesariamente significa que las estimaciones de la encuesta sean imprecisas, el uso de la tasa de respuesta como único indicador de la calidad de la encuesta puede ser engañoso. Por ejemplo, Bethlehem, Cobben y Schouten (2009) ilustran esta situación con un ejemplo de la Encuesta Integrada de Condiciones de Vida de los Hogares (POLS) realizada en el Reino de los Países Bajos en 1998. Mientras que la tasa de respuesta tras un mes de trabajo de campo fue del 47,2%, esa proporción aumentó al 59,7% después del período completo de dos meses. En el primer mes, la recolección de datos se realizó mediante entrevistas personales asistidas por computadora (CAPI). En los casos en que se disponía de un número de teléfono fijo, las personas que no respondieron fueron contactadas en el segundo mes mediante entrevistas telefónicas asistidas por computadora (CATI). En el segundo mes de trabajo de campo, la respuesta aumentó un 12,5%. Sin embargo, esto no redundó en mejores estimaciones, pues el sesgo de los estimadores aumentó a partir del segundo mes, dado que las personas que habían proporcionado un número telefónico diferían de las que no lo habían hecho.

Además de la tasa de falta de respuesta, se necesitan otros indicadores de calidad de la encuesta que proporcionen más información sobre el posible riesgo de obtener estimadores sesgados. Shlomo, Skinner y Schouten (2012) estudian el uso de los indicadores de representatividad (indicadores R) que permiten determinar la medida en que la muestra de respondientes efectivos representa a la población y la manera en que la composición de la respuesta en la muestra diferiría de la composición de la población finita. Estos indicadores demostraron ser una guía importante para determinar la medida en que el sesgo causado por la falta de respuesta afecta la encuesta. De hecho, en Europa, el proyecto "Indicadores de la representatividad para la calidad de las encuestas" (RISQ) se basó en este enfoque para desarrollar y probar indicadores R en varias encuestas de interés. Los países que participaron en este proyecto fueron Eslovenia, Noruega y los Países Bajos (Reino de los), junto con la Universidad de Southampton (Reino Unido) y la Universidad Católica de Lovaina (Bélgica).

Los indicadores R miden hasta qué punto la composición de la respuesta a una encuesta se desvía de la muestra original. Si todas las probabilidades de respuesta son

iguales, la respuesta es altamente representativa y no habrá diferencias sistemáticas entre la composición de la respuesta y la muestra. Por el contrario, si las probabilidades de respuesta no son iguales, es importante establecer la medida en que se ve afectada la composición de la respuesta. Esto se logra mediante la definición de una función de distancia que determina la medida en que las probabilidades de respuesta individuales difieren de la probabilidad de respuesta media.

Si se conocen las probabilidades de respuesta individual $\phi_1, \phi_2, \ldots, \phi_N$ de todos los elementos de la población, la desviación estándar es:

$$S(\phi)=\sqrt{\frac{1}{N-1}\sum_{k=1}^{N}(\phi_k-\bar{\phi})^2}$$

Cabe observar que, si todas las probabilidades de respuesta son iguales, $S(\phi)=0$, y el valor de $S(\phi)$ será mayor a medida que aumente la variación de los valores de las probabilidades de respuesta. Además, el valor máximo de $S(\phi)$ es igual a 0,5. Por ende, el indicador R se define como:

$$R(\phi)=1-2\,S(\phi)$$

Este indicador asume valores en el intervalo $[0,1]$. El valor 1 supone altos niveles de representatividad. Cuanto menor es el valor, más se desvía la composición de la respuesta de la composición de la muestra. En general, los valores de las probabilidades de respuesta individuales se desconocen en la práctica. Este problema se resuelve estimando las probabilidades de respuesta. Para ello se debe disponer de información auxiliar adecuada; es decir, de variables medidas tanto para los encuestados como para los no encuestados. Para estimar estas probabilidades, es posible utilizar varias técnicas, como modelos de regresión logística y árboles de clasificación, entre otras.

Al suponer que $\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_n$ son las probabilidades de respuesta estimadas para las unidades de la muestra, la probabilidad de respuesta media se puede estimar mediante:

$$\hat{\bar{\phi}}=\frac{1}{N}\sum_{k=1}^{n}\frac{\hat{\phi}_k}{\pi_k}$$

y

$$\hat{R}(\phi)=1-2\sqrt{\frac{1}{N-1}\sum_{k=1}^{n}\frac{(\hat{\phi}_k-\hat{\bar{\phi}})^2}{\pi_k}}$$

El indicador R mide la desviación de la baja representatividad y no de la alta representatividad. Por ende, este enfoque no permite detectar y cuantificar las diferencias entre las probabilidades de respuesta individual dentro de las clases obtenidas al cruzar las variables auxiliares. Supóngase que las clases están definidas por una variable auxiliar

X que tiene H categorías, que N_h es el tamaño de la clase h y que $\bar{\phi}_h$ es la media poblacional de las probabilidades de respuesta en el estrato h. Si se utiliza un modelo estándar como el de regresión logística, el indicador R se calcula de la siguiente manera:

$$R_x(\phi)=1-2\sqrt{\frac{1}{n-1}\sum_{h=1}^{H} n_h(\bar{\phi}_h-\bar{\phi})^2}$$

En este caso, $R_x(\phi)$ mide la variación de las probabilidades de respuesta entre clases X. Si se supone que la variación dentro de la clase es 0 en todas las clases, $R_x(\phi)=R(\phi)$.

Bethlehem, Cobben y Schouten (2009) mencionan que, de julio a diciembre de 2005, Statistics Netherlands realizó un seguimiento a gran escala entre personas no encuestadas en la encuesta de población activa del Reino de los Países Bajos. En el marco del estudio, se contactó a dos muestras de personas que no respondieron a la EPA utilizando, por una parte, el enfoque de devolución de llamada con el cuestionario completo de la EPA y, por otra, el enfoque de preguntas básicas con un cuestionario muy corto. En el primer caso, se utilizaron entrevistas telefónicas asistidas por computadora y, en el segundo, se utilizó un diseño mixto de recolección de datos mediante cuestionarios en línea y entrevistas personales, tanto con el uso de papel como asistidas por computadora. Los indicadores R se estimaron utilizando modelos de regresión logística con una gran cantidad de variables explicativas que medían determinadas características demográficas, geográficas y socioeconómicas de los hogares. Los resultados de este estudio son los siguientes:

i) El valor del indicador R para la respuesta inicial de la EPA fue de 0,80, que es inferior al valor ideal de 1. En consecuencia, esta respuesta no es altamente representativa. La aplicación del enfoque de devolución de llamada aumentó la tasa de respuesta del 62,2% al 76,9% y el valor del indicador R aumentó de 0,80 a 0,85. Como los intervalos de confianza no se superponían, hubo indicios de que la respuesta adicional mejoró la composición del conjunto de datos.

ii) La aplicación del enfoque de preguntas básicas dio como resultado una conclusión diferente. Aunque la tasa de respuesta aumentó del 62,2% al 75,6%, el valor del indicador R disminuyó de 0,80 a 0,78. Dado que los intervalos para la EPA inicial y la EPA con preguntas básicas se superpusieron, el enfoque de preguntas básicas aparentemente no mejoró la composición del conjunto de datos.

Este último enfoque no es novedoso y agudiza el contraste entre las personas que respondieron al cuestionario y las que no lo hicieron, pues las probabilidades de respuesta estimadas se utilizan para calcular el indicador R y esta estimación se basa en un modelo lineal que utiliza un conjunto de variables auxiliares como variables explicativas.

En las aplicaciones prácticas, los autores también concluyen que la dependencia del indicador R del conjunto de variables auxiliares utilizadas tiene repercusiones en la comparación de diferentes conjuntos de datos (por ejemplo, en el tiempo o en dominios). Un enfoque apropiado podría ser fijar el conjunto de variables auxiliares de antemano y mantenerlas iguales para todos los conjuntos de datos. Para ello, debe elegirse el máximo

número posible de variables. Por otra parte, si bien el error estándar (estimado) puede verse afectado debido al sobreajuste, las probabilidades de respuesta estimadas no estarán sesgadas. Asimismo, otro enfoque recomendable consiste en buscar el mejor modelo para cada conjunto de datos utilizando técnicas de selección de modelos. Esto hace que los modelos dependan del tamaño de la muestra: cuanto mayor sea la muestra, más variables del modelo tendrán una contribución significativa. Las muestras pequeñas simplemente no permiten una estimación adecuada de las probabilidades de respuesta y conducen a una visión más optimista de la representatividad.

Al utilizar esta metodología, es posible determinar si la composición de la muestra de respondientes efectivos difiere o no de la de la muestra inicial. Los resultados de este proceso de seguimiento pueden ayudar a sostener la decisión de iniciar gestiones adicionales para obtener datos sobre grupos específicos de la población objetivo. Este enfoque también puede resultar útil para evaluar si volver a contactar a una muestra de personas que no respondieron sería una buena estrategia para limitar el sesgo, o si con un enfoque de preguntas básicas sería suficiente. Además, el uso de este método durante la fase de recolección de datos podría revelar que la composición de los datos observados se desvía cada vez más de la estructura poblacional esperada. Esto podría llevar a la decisión de centrar el resto del proceso de recolección en los grupos que están subrepresentados. En la literatura especializada, los cambios en el curso de la encuesta se conocen como "diseños receptivos".

Por último, Bethlehem, Cobben y Schouten (2009) afirman que otra forma de utilizar el indicador R para controlar la calidad de los datos recolectados en la encuesta es analizar la representatividad de una versión anterior de la encuesta. Los resultados de dicho análisis pueden proporcionar información para mejorar la estrategia de recopilación de datos en una nueva versión de la encuesta.

C. Clasificación de la falta de respuesta

De acuerdo con Särndal y Lundstrom (2005), existe una gran cantidad de estudios acerca de la falta de respuesta, incluidos muchos artículos recientes. En estos trabajos se examinan dos aspectos diferentes pero complementarios de la realización de una encuesta: la prevención de la falta de respuesta (antes de que ocurra) y las técnicas de estimación necesarias para tener en cuenta la falta de respuesta de manera apropiada en el proceso de inferencia. Esta segunda actividad se conoce con el nombre de ajuste para la falta de respuesta. Little y Rubin (2002) establecen tres tipos de falta de respuesta.

La falta de respuesta con un patrón completamente aleatorio (MCAR) se presenta cuando la probabilidad de que un individuo responda no depende de la característica de interés, ni de ninguna otra covariable auxiliar (por ejemplo, si la falta de respuesta en una encuesta laboral no depende de la situación laboral actual del encuestado, ni de ninguna

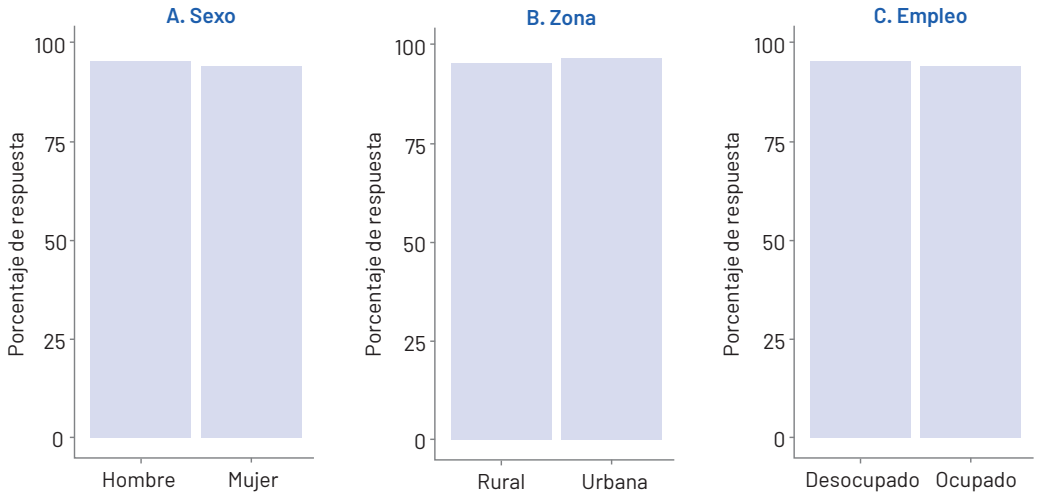
característica auxiliar). De esta forma, la falta de respuesta presenta una dispersión uniforme en toda la población.

Esto significa que, cuando el investigador produzca estadísticas descriptivas sobre las personas que respondieron a la encuesta, ese porcentaje de personas será muy similar y presentará un comportamiento uniforme en todas las posibles covariables que afecten al individuo. En el gráfico XI.1 se observan indicios de que el patrón de falta de respuesta podría ser completamente aleatorio, puesto que el porcentaje de respuesta es similar en las variables auxiliares.

■ Gráfico XI.1

Ejemplo de distribución de las personas encuestadas con un patrón de respuesta completamente aleatorio (MCAR)

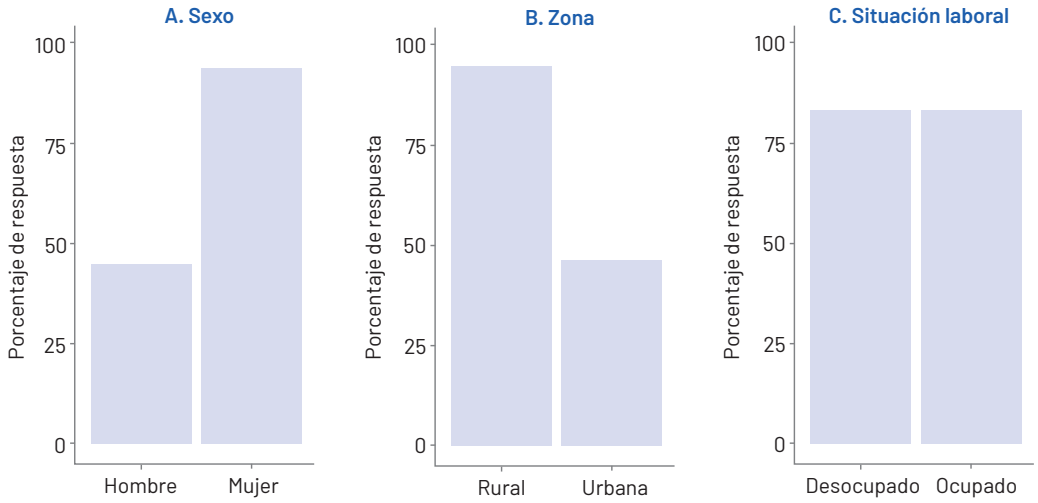
(En porcentajes)



Fuente: Elaboración propia.

La falta de respuesta aleatoria *(missing at random* (MAR)) se establece cuando la probabilidad de que una persona responda depende de algunas covariables auxiliares, pero no de la característica de interés. Por ejemplo, en una encuesta de la fuerza laboral, la falta de respuesta puede depender de la edad, el sexo o incluso el nivel económico de la persona encuestada, pero no de su clasificación laboral. En el gráfico XI.2, se muestra que el patrón de falta de respuesta podría ser aleatorio, pues el sexo y la zona de residencia de las personas encuestadas influyen en el porcentaje de respuesta, a diferencia del estado de ocupación.

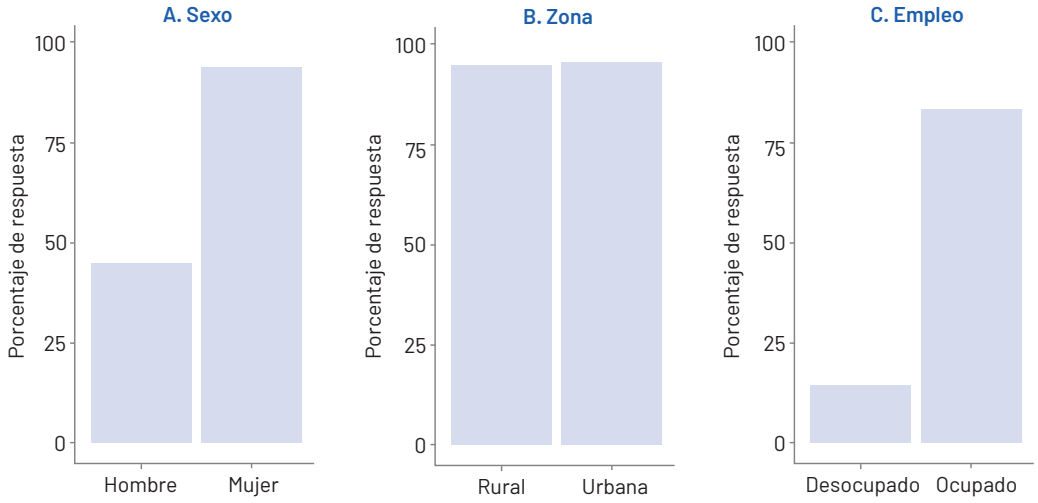
■ Gráfico XI.2
Ejemplo de distribución de las personas encuestadas con un patrón de respuesta aleatorio (MAR)
(En porcentajes)



Fuente: Elaboración propia.

Por último, la falta de respuesta no aleatoria (*not missing at random* (NMAR)) se presenta cuando la falta de respuesta depende de la característica de interés. En el gráfico XI.3, se observan indicios de este tipo de patrón de respuesta, pues la condición de ocupación influye en el porcentaje de respuesta. Esto es contraproducente, ya que no existe una forma simple de mitigar el sesgo generado por esta clase de falta de respuesta.

■ Gráfico XI.3
Ejemplo de distribución de las personas encuestadas con un patrón de respuesta no aleatorio (NMAR)
(En porcentajes)



Fuente: Elaboración propia.

D. Falta de respuesta por registro y unidad

En el cuadro XI.1 se ilustra la estructura ideal de la base de datos de una hipotética encuesta de hogares con 8 variables y 14 personas, en la que no hay problemas de falta de respuesta porque todas las personas han contestado a cada una de las variables que se indagaron en el cuestionario. De esta manera, la base de datos contiene registros válidos en cada una de sus entradas.

■ Cuadro XI.1
Ejemplo de base de datos completa (sin falta de respuesta) de una encuesta

	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5	Variable 6	Variable 7	Variable 8
Unidad 1	$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$	$y_{1,5}$	$y_{1,6}$	$y_{1,7}$	$y_{1,8}$
Unidad 2	$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	$y_{2,4}$	$y_{2,5}$	$y_{2,6}$	$y_{2,7}$	$y_{2,8}$
Unidad 3	$y_{3,1}$	$y_{3,2}$	$y_{3,3}$	$y_{3,4}$	$y_{3,5}$	$y_{3,6}$	$y_{3,7}$	$y_{3,8}$
Unidad 4	$y_{4,1}$	$y_{4,2}$	$y_{4,3}$	$y_{4,4}$	$y_{4,5}$	$y_{4,6}$	$y_{4,7}$	$y_{4,8}$
Unidad 5	$y_{5,1}$	$y_{5,2}$	$y_{5,3}$	$y_{5,4}$	$y_{5,5}$	$y_{5,6}$	$y_{5,7}$	$y_{5,8}$
Unidad 6	$y_{6,1}$	$y_{6,2}$	$y_{6,3}$	$y_{6,4}$	$y_{6,5}$	$y_{6,6}$	$y_{6,7}$	$y_{6,8}$
Unidad 7	$y_{7,1}$	$y_{7,2}$	$y_{7,3}$	$y_{7,4}$	$y_{7,5}$	$y_{7,6}$	$y_{7,7}$	$y_{7,8}$
Unidad 8	$y_{8,1}$	$y_{8,2}$	$y_{8,3}$	$y_{8,4}$	$y_{8,5}$	$y_{8,6}$	$y_{8,7}$	$y_{8,8}$
Unidad 9	$y_{9,1}$	$y_{9,2}$	$y_{9,3}$	$y_{9,4}$	$y_{9,5}$	$y_{9,6}$	$y_{9,7}$	$y_{9,8}$
Unidad 10	$y_{10,1}$	$y_{10,2}$	$y_{10,3}$	$y_{10,4}$	$y_{10,5}$	$y_{10,6}$	$y_{10,7}$	$y_{10,8}$
Unidad 11	$y_{11,1}$	$y_{11,2}$	$y_{11,3}$	$y_{11,4}$	$y_{11,5}$	$y_{11,6}$	$y_{11,7}$	$y_{11,8}$
Unidad 12	$y_{12,1}$	$y_{12,2}$	$y_{12,3}$	$y_{12,4}$	$y_{12,5}$	$y_{12,6}$	$y_{12,7}$	$y_{12,8}$
Unidad 13	$y_{13,1}$	$y_{13,2}$	$y_{13,3}$	$y_{13,4}$	$y_{13,5}$	$y_{13,6}$	$y_{13,7}$	$y_{13,8}$
Unidad 14	$y_{14,1}$	$y_{14,2}$	$y_{14,3}$	$y_{14,4}$	$y_{14,5}$	$y_{14,6}$	$y_{14,7}$	$y_{14,8}$

Fuente: Elaboración propia.

Cabe destacar que, a pesar de que se hayan tomado las medidas de ajuste necesarias en el diseño de la encuesta, una vez terminado el proceso de recolección de información, se debe lidiar con la falta de respuesta para evitar sesgos y aumentar la precisión de los estimadores. Como se mencionó anteriormente, en la literatura especializada se examinan dos enfoques complementarios para la realización de una encuesta: la prevención de la falta de respuesta (antes de que ocurra) y las técnicas de estimación necesarias para tenerla en cuenta de manera apropiada en el proceso de inferencia, después de la recolección de los datos.

Si se asume que el mecanismo de falta de respuesta presenta un patrón completamente aleatorio, es posible contemplar en el proceso de inferencia solo las unidades con registros completos y eliminar de la base de datos aquellas que no contestaron (eliminación por lista). A pesar de que este tipo de análisis es simple, para evitar la subestimación de los parámetros de interés, es necesario ajustar los factores de expansión determinados por el diseño muestral, originalmente concebido para una muestra más grande. De esta forma, es

posible suponer que la muestra de personas encuestadas corresponde a una submuestra completamente aleatoria de la población y utilizar los principios de los diseños en dos fases. Heeringa, West y Berglund (2010, cap. 11) afirman que, además de provocar posibles sesgos si el supuesto MCAR no se cumple, este tipo de análisis reduce la eficiencia de la inferencia, debido a la disminución del tamaño efectivo de la muestra. Por lo tanto, en la mayoría de las encuestas, este supuesto no se asume y se realiza un ajuste adicional después de que se produce la falta de respuesta.

En general, existen dos tipos de falta de respuesta: la de una unidad de observación (falta de respuesta por unidad) y la de una unidad con respecto a algunas variables de interés (falta de respuesta por registro). De acuerdo con Särndal, Swensson y Wretman (2003, sec. 15.5), las principales técnicas para tratar la falta de respuesta son el ajuste de los pesos de muestreo y la imputación. El ajuste por ponderación supone el aumento de los pesos aplicados en la estimación de los valores y de los encuestados para compensar los valores que se pierden debido a la falta de respuesta, mientras que la imputación implica la sustitución de los valores faltantes por valores artificiales.

La falta de respuesta tiene repercusiones evidentes en la base de datos de la encuesta. Por ejemplo, en la base de datos inicial puede faltar toda la información relativa a una unidad de observación. Esto suele suceder porque el encuestador no pudo establecer contacto con el hogar o porque la persona seleccionada no puede responder o simplemente se rehúsa a participar. En esta etapa es recomendable que el encuestador determine algunas características demográficas del hogar para poder realizar los ajustes pertinentes en el proceso de análisis. También es posible que en la base de datos falte información relativa a algunos registros de las unidades. Esto puede deberse a muchas más causas y se evidencia en la base de datos inicial porque faltan algunos registros de la unidad de observación, mientras que otros están completos. Algunas causas de este fenómeno pueden estar relacionadas con el cansancio de la persona encuestada en algún momento del cuestionario o su decisión de no responder a determinadas preguntas por considerarlas delicadas.

En general, es posible hacer frente a este fenómeno no deseado desde diversas perspectivas. A continuación se mencionan algunos puntos de vista para abordar la falta de respuesta:

- Omisión: ignorar la falta de respuesta en la encuesta y realizar inferencias con los datos recopilados de las unidades que respondieron al cuestionario, sin realizar ningún tipo de procedimiento estadístico para ajustar la inferencia. Lamentablemente, ocurre a menudo.
- Prevención: adoptar medidas para minimizar la falta de respuesta en el diseño de la encuesta. Este es el mejor método para hacer frente al problema. La capacitación del equipo encuestador, la redacción de las preguntas, la longitud del cuestionario, las revisitas y la coordinación de las entrevistas pueden contribuir a reducir las altas tasas de falta de respuesta.

- Reacción: utilizar herramientas para analizar la encuesta y corregir los sesgos causados por la falta de respuesta. En este caso, es posible ajustar los ponderadores de las unidades o establecer procedimientos de imputación en los registros.

Ignorar la falta de respuesta puede tener graves consecuencias en el análisis del constructo de interés de la encuesta. Más aún, puede llevar a errores en el momento de tomar decisiones de política pública. Por ejemplo, si se omite el efecto de la falta de respuesta en una encuesta de ingresos y gastos, se podrían subestimar el ingreso medio y el ingreso total de un país. En el caso de una encuesta de desempleo, se podría subestimar el número total de desempleados y, en el de una encuesta de victimización, el número total de víctimas.

Dado que la falta de repuesta conlleva grandes efectos de sesgo en los resultados de calidad de las estimaciones, se debe seleccionar cuidadosamente la mejor estrategia para hacer frente a sus consecuencias. Por ejemplo, si se aumenta el tamaño de la muestra general, es posible encontrar una mayor cantidad de personas de la misma clase de respondientes (homogeneidad) y seguir sin encontrar a los no respondientes. En este caso, el sesgo puede aumentar, con el agravante de que se malgastaron recursos que hubiesen servido para remediar la falta de respuesta con otras medidas.

E. Posibles soluciones

El tratamiento metodológico de la falta de respuesta varía según el enfoque aplicado. En general, se distinguen las siguientes prácticas:

- Imputación total: consiste en imputar todos los valores faltantes de los individuos que presentan al menos un valor faltante.
- Ponderación total: consiste en ponderar cada una de las variables de interés, aunque sea de manera diferenciada. No se utiliza la imputación y existirán tantos conjuntos de factores de expansión como variables con valores faltantes.
- Eliminación total: consiste en eliminar todos los registros con valores faltantes y hacer el análisis con el conjunto restante de valores disponibles.
- Enfoque combinado: consiste en imputar valores únicamente a los elementos que tienen al menos un registro faltante (no todos) y modificar los factores de expansión en aquellos casos en los que se han omitido todos los registros del cuestionario.

Según la notación de Särndal y Lundström (2005), se considera una muestra de unidades s, de la cual r denota el conjunto de encuestados que han respondido a una o más de las l variables de interés. Por tanto, una unidad que no responde a ninguna variable pertenece al conjunto $s - r$. El conjunto de unidades que han respondido a una variable del estudio en particular se denota mediante r_i. Así:

$$r_i \subseteq r \subseteq s$$

Por último, si se supone que y_k falta y se considera para la imputación, el valor imputado se denotará mediante $\hat{y}_k$. En el cuadro XI.2 se ilustra la estructura de la base de datos de una hipotética encuesta de hogares en la que algunas personas no respondieron a una o todas las variables incluidas en el cuestionario. Las unidades están representadas por las filas y las variables por las columnas. Se observa que las primeras tres personas contestaron a todas las preguntas del cuestionario, la cuarta no contestó las últimas dos preguntas, la quinta no contestó ni la primera ni la última pregunta, la sexta no contestó la tercera pregunta y así sucesivamente, hasta llegar a las últimas dos personas, que no contestaron a ninguna pregunta del cuestionario. Para este ejemplo particular, se observa que:

- El número de variables de interés en la encuesta de hogares es $I=8$.
- El número de unidades incluidas en la muestra s es $n = \#(s)=14$.
- El número de respondientes efectivos en la primera variable es $\#(r_1)=10$, en la segunda variable es $\#(r_2)=9$, y así sucesivamente hasta notar que el número de respondientes efectivos en la última variable de la base de datos es de $\#(r_8)=8$.

■ Cuadro XI.2
Ejemplo de base de datos de una encuesta con falta de respuesta

	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5	Variable 6	Variable 7	Variable 8
Unidad 1	$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$	$y_{1,5}$	$y_{1,6}$	$y_{1,7}$	$y_{1,8}$
Unidad 2	$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	$y_{2,4}$	$y_{2,5}$	$y_{2,6}$	$y_{2,7}$	$y_{2,8}$
Unidad 3	$y_{3,1}$	$y_{3,2}$	$y_{3,3}$	$y_{3,4}$	$y_{3,5}$	$y_{3,6}$	$y_{3,7}$	$y_{3,8}$
Unidad 4	$y_{4,1}$	$y_{4,2}$	$y_{4,3}$	$y_{4,4}$	$y_{4,5}$	$y_{4,6}$		
Unidad 5		$y_{5,2}$	$y_{5,3}$	$y_{5,4}$	$y_{5,5}$	$y_{5,6}$	$y_{5,7}$	
Unidad 6	$y_{6,1}$	$y_{6,2}$		$y_{6,4}$	$y_{6,5}$	$y_{6,6}$	$y_{6,7}$	$y_{6,8}$
Unidad 7	$y_{7,1}$	$y_{7,2}$	$y_{7,3}$	$y_{7,4}$	$y_{7,5}$	$y_{7,6}$	$y_{7,7}$	
Unidad 8	$y_{8,1}$	$y_{8,2}$	$y_{8,3}$	$y_{8,4}$		$y_{8,6}$	$y_{8,7}$	$y_{8,8}$
Unidad 9		$y_{9,2}$	$y_{9,3}$	$y_{9,4}$	$y_{9,5}$		$y_{9,7}$	
Unidad 10	$y_{10,1}$		$y_{10,3}$		$y_{10,5}$	$y_{10,6}$	$y_{10,7}$	$y_{10,8}$
Unidad 11	$y_{11,1}$		$y_{11,3}$		$y_{11,5}$			$y_{11,8}$
Unidad 12	$y_{12,1}$			$y_{12,4}$	$y_{12,5}$	$y_{12,6}$	$y_{12,7}$	$y_{12,8}$
Unidad 13								
Unidad 14								

Fuente: Elaboración propia.
Nota: Las celdas en gris claro representan registros respondidos y las celdas en gris oscuro representan registros no respondidos y faltantes.

1. Imputación total

En este enfoque se imputarían todos los valores y_k faltantes, sin importar si esto se debe a la falta del registro o de la persona. En este caso, tendríamos un conjunto completo de datos con los valores $\{y_{\circ k} : k \in s\}$, donde:

$$y_{\circ k} = \begin{cases} y_k & \text{cuando } k \in r_i \\ \hat{y}_k & \text{cuando } k \in s - r_i \end{cases}$$

y $\hat{y}_k$ es el valor imputado. Al utilizar este enfoque, el estimador del total estaría dado por la siguiente expresión:

$$\hat{t}_{y,\pi} = \sum_s d_k y_{\circ k} = \sum_{r_i} d_k y_k + \sum_{s-r_i} d_k \hat{y}_k$$

En el cuadro XI.3 se muestra un ejemplo de las unidades que se considerarían para el análisis después de la imputación. Se observa que las tres unidades que respondieron a todas las preguntas del cuestionario se incluyen en el análisis sin ningún ajuste, mientras que las nueve unidades que no respondieron a todo el cuestionario se incluyen en el análisis tras la imputación de las celdas correspondientes a la falta de respuesta. Además, las dos unidades que no respondieron a ninguna pregunta del cuestionario también se incluyen en el análisis, puesto que todas sus respuestas fueron imputadas. De acuerdo con este enfoque, todas las unidades del conjunto s se consideran para el análisis posterior.

■ Cuadro XI.3

Ejemplo de base de datos de una encuesta con falta de respuesta en la que se aplica el enfoque de imputación total de los valores faltantes

	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5	Variable 6	Variable 7	Variable 8
Unidad 1	$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$	$y_{1,5}$	$y_{1,6}$	$y_{1,7}$	$y_{1,8}$
Unidad 2	$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	$y_{2,4}$	$y_{2,5}$	$y_{2,6}$	$y_{2,7}$	$y_{2,8}$
Unidad 3	$y_{3,1}$	$y_{3,2}$	$y_{3,3}$	$y_{3,4}$	$y_{3,5}$	$y_{3,6}$	$y_{3,7}$	$y_{3,8}$
Unidad 4	$y_{4,1}$	$y_{4,2}$	$y_{4,3}$	$y_{4,4}$	$y_{4,5}$	$y_{4,6}$	$y_{4,7}$	$y_{4,8}$
Unidad 5	$y_{5,1}$	$y_{5,2}$	$y_{5,3}$	$y_{5,4}$	$y_{5,5}$	$y_{5,6}$	$y_{5,7}$	$y_{5,8}$
Unidad 6	$y_{6,1}$	$y_{6,2}$	$y_{6,3}$	$y_{6,4}$	$y_{6,5}$	$y_{6,6}$	$y_{6,7}$	$y_{6,8}$
Unidad 7	$y_{7,1}$	$y_{7,2}$	$y_{7,3}$	$y_{7,4}$	$y_{7,5}$	$y_{7,6}$	$y_{7,7}$	$y_{7,8}$
Unidad 8	$y_{8,1}$	$y_{8,2}$	$y_{8,3}$	$y_{8,4}$	$y_{8,5}$	$y_{8,6}$	$y_{8,7}$	$y_{8,8}$
Unidad 9	$y_{9,1}$	$y_{9,2}$	$y_{9,3}$	$y_{9,4}$	$y_{9,5}$	$y_{9,6}$	$y_{9,7}$	$y_{9,8}$
Unidad 10	$y_{10,1}$	$y_{10,2}$	$y_{10,3}$	$y_{10,4}$	$y_{10,5}$	$y_{10,6}$	$y_{10,7}$	$y_{10,8}$
Unidad 11	$y_{11,1}$	$y_{11,2}$	$y_{11,3}$	$y_{11,4}$	$y_{11,5}$	$y_{11,6}$	$y_{11,7}$	$y_{11,8}$
Unidad 12	$y_{12,1}$	$y_{12,2}$	$y_{12,3}$	$y_{12,4}$	$y_{12,5}$	$y_{12,6}$	$y_{12,7}$	$y_{12,8}$
Unidad 13	$y_{13,1}$	$y_{13,2}$	$y_{13,3}$	$y_{13,4}$	$y_{13,5}$	$y_{13,6}$	$y_{13,7}$	$y_{13,8}$
Unidad 14	$y_{14,1}$	$y_{14,2}$	$y_{14,3}$	$y_{14,4}$	$y_{14,5}$	$y_{14,6}$	$y_{14,7}$	$y_{14,8}$

Fuente: Elaboración propia.

Nota: Las celdas en gris claro corresponden a los valores observados originalmente y las celdas en gris medio corresponden a los valores imputados para los campos en los que no hubo respuesta.

2. Ponderación total

Al usar el enfoque de ponderación total, es posible usar pesos de calibración específicos $w_k=d_k\,F_{ik}$ que compensarían la falta de respuesta por unidad y registro. De esta forma, el estimador del total estaría dado por la siguiente expresión:

$$\hat{t}_{y,cal}=\sum_{r_i} w_k\,y_k=\sum_{r_i} d_k\,F_{ik}\,y_k$$

Si todos los r_i son diferentes, cada variable de estudio requerirá un conjunto de ponderadores diferentes. Al final, este enfoque genera un número no uniforme de casos por variable. En este sistema se utilizan pesos $w_k^{(i)}$ para cada variable $i \in I$, que compensan la falta de respuesta de la unidad.

Siguiendo con el ejemplo, en el cuadro XI.4 se observa que diez personas respondieron a la primera variable del cuestionario, mientras que otras cuatro no lo hicieron. Por lo tanto, en el marco de este enfoque, se crearán pesos $w_k^{(1)}$ para cada $k \in r_1$ que ponderen satisfactoriamente la información recogida con respecto a esta variable. Sin embargo, este conjunto de pesos no será único, puesto que, en particular, nueve personas respondieron a la segunda variable del cuestionario, mientras que otras cinco no lo hicieron. Por lo tanto, en el marco de este enfoque, se crearán pesos $w_k^{(2)}$ para cada $k \in r_2$ que ponderen la información recogida con respecto a esta variable. En general, $w_k^{(1)} \neq w_k^{(2)}$ y, por ende, cada una de las $I=8$ variables del estudio tendrá su propio conjunto de ponderadores.

■ Cuadro XI.4
Ejemplo de base de datos de una encuesta con falta de respuesta en la que se aplica el enfoque de ponderación total de los valores faltantes

	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5	Variable 6	Variable 7	Variable 8
Unidad 1	$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$	$y_{1,5}$	$y_{1,6}$	$y_{1,7}$	$y_{1,8}$
Unidad 2	$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	$y_{2,4}$	$y_{2,5}$	$y_{2,6}$	$y_{2,7}$	$y_{2,8}$
Unidad 3	$y_{3,1}$	$y_{3,2}$	$y_{3,3}$	$y_{3,4}$	$y_{3,5}$	$y_{3,6}$	$y_{3,7}$	$y_{3,8}$
Unidad 4	$y_{4,1}$	$y_{4,2}$	$y_{4,3}$	$y_{4,4}$	$y_{4,5}$	$y_{4,6}$		
Unidad 5		$y_{5,2}$	$y_{5,3}$	$y_{5,4}$	$y_{5,5}$	$y_{5,6}$	$y_{5,7}$	
Unidad 6	$y_{6,1}$	$y_{6,2}$		$y_{6,4}$	$y_{6,5}$	$y_{6,6}$	$y_{6,7}$	$y_{6,8}$
Unidad 7	$y_{7,1}$	$y_{7,2}$	$y_{7,3}$	$y_{7,4}$	$y_{7,5}$	$y_{7,6}$	$y_{7,7}$	
Unidad 8	$y_{8,1}$	$y_{8,2}$	$y_{8,3}$	$y_{8,4}$		$y_{8,6}$	$y_{8,7}$	$y_{8,8}$
Unidad 9		$y_{9,2}$	$y_{9,3}$	$y_{9,4}$	$y_{9,5}$		$y_{9,7}$	
Unidad 10	$y_{10,1}$		$y_{10,3}$		$y_{10,5}$	$y_{10,6}$	$y_{10,7}$	$y_{10,8}$
Unidad 11	$y_{11,1}$		$y_{11,3}$		$y_{11,5}$			$y_{11,8}$
Unidad 12	$y_{12,1}$			$y_{12,4}$	$y_{12,5}$	$y_{12,6}$	$y_{12,7}$	$y_{12,8}$
Unidad 13								
Unidad 14								

Fuente: Elaboración propia.
Nota: Las celdas en gris claro corresponden a los valores observados originalmente y las celdas en gris oscuro corresponden a valores que no fueron observados ni imputados.

3. Eliminación total

En el marco de este enfoque, se eliminarán de la base de datos todas las unidades a las que les falte al menos un registro. Se recomienda encarecidamente abstenerse de tomar este camino, puesto que, aunque supondría un solo conjunto de ponderadores, conllevaría una reducción considerable y deliberada del tamaño de la muestra (con la consiguiente pérdida de información), un incremento sustancial de los sesgos y una disminución de la precisión de los estimadores. En este enfoque, solo las unidades del conjunto de respondientes efectivos en todas las variables se consideran para el análisis posterior de la encuesta. Este subgrupo de individuos está dado por:

$$\bigcap_{i=1}^{I} r_i$$

En general, esto no es aconsejable porque conlleva problemas de sesgo (dado que las unidades que contestaron todo el cuestionario generalmente difieren de forma estructural de las unidades que no contestaron) y problemas de eficiencia estadística (puesto que el tamaño de la muestra efectiva, después de la eliminación de unidades, será insuficiente para garantizar los mínimos requeridos en la inferencia).

A partir del cuadro XI.5, en el que se representa este enfoque, resulta evidente que la reducción del tamaño de la muestra podría tener repercusiones muy negativas en la inferencia de la encuesta. En el ejemplo anterior, solo se tendrían en cuenta tres unidades para el análisis de la información, mientras que las nueve unidades que no contestaron al menos una pregunta y las dos unidades que no contestaron ninguna pregunta se eliminarían del análisis estadístico. Es decir, la mayoría de las unidades de la muestra inicial se descartarían.

■ Cuadro XI.5
Ejemplo de base de datos de una encuesta con falta de respuesta en la que se aplica el enfoque de eliminación total de las unidades con valores faltantes

	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5	Variable 6	Variable 7	Variable 8
Unidad 1	$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$	$y_{1,5}$	$y_{1,6}$	$y_{1,7}$	$y_{1,8}$
Unidad 2	$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	$y_{2,4}$	$y_{2,5}$	$y_{2,6}$	$y_{2,7}$	$y_{2,8}$
Unidad 3	$y_{3,1}$	$y_{3,2}$	$y_{3,3}$	$y_{3,4}$	$y_{3,5}$	$y_{3,6}$	$y_{3,7}$	$y_{3,8}$
Unidad 4								
Unidad 5								
Unidad 6								
Unidad 7								
Unidad 8								
Unidad 9								

	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5	Variable 6	Variable 7	Variable 8
Unidad 10								
Unidad 11								
Unidad 12								
Unidad 13								
Unidad 14								

Fuente: Elaboración propia.
Nota: Las celdas en gris claro corresponden a los valores observados originalmente en todas las unidades seleccionadas en la muestra y las celdas en gris oscuro corresponden a valores que no fueron observados o que fueron eliminados por no haber sido observados en toda la muestra.

4. Enfoque combinado

Por el contrario, es recomendable escoger un camino parsimonioso que combine estas estrategias de forma diferencial a lo largo de la encuesta. En el enfoque combinado, se utiliza la imputación para afrontar la falta de respuesta por registro para las variables (columnas de la base de datos) específicas que lo necesiten y luego se ajustan los factores de ponderación para afrontar la falta de respuesta por unidad (filas de la base de datos). En general, los pesos finales se producen utilizando un enfoque de calibración en el que se recurre a información auxiliar externa.

En caso de falta de respuesta por registro y por unidad, el enfoque combinado prevé la imputación de valores para obtener una matriz rectangular completa, seguida de un ajuste de los ponderadores. El conjunto de datos completo para la variable de interés y está determinado por el conjunto dado por $\{y_{\circ k} : k \in r\}$. En donde:

$$y_{\circ k} = \begin{cases} y_k & \text{cuando } k \in r_i \\ \hat{y}_k & \text{cuando } k \in r - r_i \end{cases}$$

Donde $\hat{y}_k$ es el valor imputado. Cabe señalar que en el enfoque de imputación total también se imputa para $k \in s - r$. En el cuadro XI.6 se representa este enfoque parsimonioso, en el que los valores imputados (en gris claro) se incluyen en la inferencia y las unidades que nunca respondieron (en gris oscuro) y de las cuales que faltan todos los registros se retiran de la base de datos final.

Se observa que las dos últimas unidades de la muestra se descartaron totalmente porque no contestaron ninguna pregunta del cuestionario. Al considerar la primera variable, se imputaron los valores faltantes de las unidades 5 y 9 y, en el caso de la segunda variable, los valores faltantes de las unidades 10, 11 y 12. El procedimiento se reiteró hasta llegar a la última variable, en cuyo caso se imputaron los valores de las unidades 4, 5, 7 y 9.

■ Cuadro XI.6

Ejemplo de base de datos de una encuesta con falta de respuesta en la que se aplica el enfoque combinado para el tratamiento de los valores faltantes

	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5	Variable 6	Variable 7	Variable 8
Unidad 1	$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$	$y_{1,5}$	$y_{1,6}$	$y_{1,7}$	$y_{1,8}$
Unidad 2	$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	$y_{2,4}$	$y_{2,5}$	$y_{2,6}$	$y_{2,7}$	$y_{2,8}$
Unidad 3	$y_{3,1}$	$y_{3,2}$	$y_{3,3}$	$y_{3,4}$	$y_{3,5}$	$y_{3,6}$	$y_{3,7}$	$y_{3,8}$
Unidad 4	$y_{4,1}$	$y_{4,2}$	$y_{4,3}$	$y_{4,4}$	$y_{4,5}$	$y_{4,6}$	$y_{4,7}$	$y_{4,8}$
Unidad 5	$y_{5,1}$	$y_{5,2}$	$y_{5,3}$	$y_{5,4}$	$y_{5,5}$	$y_{5,6}$	$y_{5,7}$	$y_{5,8}$
Unidad 6	$y_{6,1}$	$y_{6,2}$	$y_{6,3}$	$y_{6,4}$	$y_{6,5}$	$y_{6,6}$	$y_{6,7}$	$y_{6,8}$
Unidad 7	$y_{7,1}$	$y_{7,2}$	$y_{7,3}$	$y_{7,4}$	$y_{7,5}$	$y_{7,6}$	$y_{7,7}$	$y_{7,8}$
Unidad 8	$y_{8,1}$	$y_{8,2}$	$y_{8,3}$	$y_{8,4}$	$y_{8,5}$	$y_{8,6}$	$y_{8,7}$	$y_{8,8}$
Unidad 9	$y_{9,1}$	$y_{9,2}$	$y_{9,3}$	$y_{9,4}$	$y_{9,5}$	$y_{9,6}$	$y_{9,7}$	$y_{9,8}$
Unidad 10	$y_{10,1}$	$y_{10,2}$	$y_{10,3}$	$y_{10,4}$	$y_{10,5}$	$y_{10,6}$	$y_{10,7}$	$y_{10,8}$
Unidad 11	$y_{11,1}$	$y_{11,2}$	$y_{11,3}$	$y_{11,4}$	$y_{11,5}$	$y_{11,6}$	$y_{11,7}$	$y_{11,8}$
Unidad 12	$y_{12,1}$	$y_{12,2}$	$y_{12,3}$	$y_{12,4}$	$y_{12,5}$	$y_{12,6}$	$y_{12,7}$	$y_{12,8}$
Unidad 13								
Unidad 14								

Fuente: Elaboración propia.

Nota: Las celdas en gris claro corresponden a los valores observados originalmente en todas las unidades seleccionadas en la muestra, las celdas en gris medio corresponden a valores imputados y las celdas en gris oscuro, a valores que no fueron observados porque la unidad seleccionada nunca respondió.

En los capítulos anteriores se analizó la creación de factores de expansión para los individuos presentes en la base de datos final. A continuación se examinarán algunos métodos de imputación recomendables a la hora de completar una base de datos estructurada y rectangular con todas las entradas. Antes de introducir estos temas, se presentarán algunas medidas descriptivas que pueden utilizarse para generar alertas sobre la pérdida de representatividad debido a la falta de respuesta.

Capítulo XII
Falta de respuesta por unidad

En una encuesta, la información auxiliar puede utilizarse en dos etapas: en la planificación del diseño de muestreo y en la elección del estimador. En el primer caso, es posible utilizar la información auxiliar para construir estratos, definir conglomerados, asignar los tamaños de muestra dentro de los estratos, o incluso definir probabilidades de selección desiguales. En el segundo caso, la información auxiliar puede utilizarse en la estimación de los parámetros de interés, al definir nuevos ajustes de ponderación e imponer restricciones de coherencia con la información auxiliar disponible en los censos, registros o encuestas, para que la distribución de la muestra expandida coincida plenamente con algunas características poblacionales. En este capítulo se analiza el uso de la información auxiliar en el estimador para corregir los sesgos generados por la falta de respuesta.

Como ya se ha expuesto, la falta de respuesta a nivel de la unidad puede tener consecuencias muy graves en la inferencia resultante de las encuestas de hogares. Esto se debe a que, si el conjunto de respondientes tiene características distintas al conjunto de no respondientes, se introducirá un sesgo en la estimación de los parámetros de interés.

A. Sesgo sobre los estimadores

Partiendo del supuesto de que existe falta de respuesta en la muestra, considérese la siguiente forma de estimar (ingenuamente) el promedio poblacional $\bar{y}_U$ mediante el estimador de Hájek:

$$\tilde{y}_s = \frac{\sum_{sr} d_k y_k}{\sum_{sr} d_k} = \frac{\hat{t}_y}{\hat{N}}$$

Si $\bar{\phi}$ es el promedio de las probabilidades de respuesta, el sesgo introducido por la falta de respuesta puede cuantificarse de la siguiente manera:

$$B(\tilde{y}_s)= \frac{1}{N\bar{\phi}} \sum_U (y_k - \bar{y}_U)(\phi_k - \bar{\phi}) = \frac{Cov(\bar{y},\phi)}{\bar{\phi}} = \frac{Cor(Y,\phi)S(Y)S(\phi)}{\bar{\phi}}$$

Donde $Cov(Y,\phi)$ es la covarianza poblacional entre los valores de la característica de interés y las probabilidades de respuesta, $cor(Y,\phi)$ es el coeficiente de correlación poblacional y $S(Y)$ es la desviación estándar poblacional de la variable objetivo. Dado que el valor del coeficiente de correlación está restringido al intervalo $[-1,1]$, el valor máximo del sesgo absoluto será igual a:

$$|B(\tilde{y}_s)| \leq \frac{S(\phi)S(y)}{\bar{\phi}} = \frac{(1-R(\phi))S(y)}{2\bar{\phi}}$$

A pesar de que este límite superior no se puede calcular en situaciones prácticas, sí es posible estimarlo a partir de los datos de la muestra y las probabilidades de respuesta estimadas. Nótese que, si el mecanismo de falta de respuesta fuese completamente aleatorio (*missing completely at random* (MCAR)), el valor de $R(\phi)$ sería 1 y, por consiguiente, no habría sesgo. De la misma forma, en el caso extremo en que la característica de interés fuese homogénea en toda la población, tampoco habría sesgo en el estimador. Bastaría con utilizar los datos de la muestra de respondientes efectivos, sin ningún tipo de corrección.

Además de las anteriores consideraciones, es posible evaluar las propuestas de elección de variables para calibración de Graham Kalton y Flores-Cervantes (2003) y de Särndal (2011b). En particular, este último autor considera un indicador del sesgo por falta de respuesta sobre los estimadores de calibración, cuya lógica se basa en que, en el mejor de los casos —en que no hubiese errores de cobertura ni falta de respuesta—, el estimador de expansión $\hat{t}_y$ sería insesgado y la distancia que habría entre este y el estimador de calibración $\hat{t}_{y,cal}$ se podría cuantificar como $\Delta_A = \frac{(\hat{t}_{y,cal} - \hat{t}_y)}{N}$. Este indicador se sugiere como una herramienta para comparar posibles variables de calibración, de tal forma que, cuando el valor de $|\Delta_A|$ sea grande, habría un indicio para preferir un vector de calibración sobre otro. Además, al estandarizarla, esta medida puede descomponerse en los siguientes tres factores:

$$\frac{\Delta_A}{S_y} = cv_g \times R_{y,x} \times R_{D,C}$$

De esta forma, el primer factor representa el coeficiente de variación de los pesos g_k; el segundo factor al cuadrado es el coeficiente de determinación de una regresión múltiple entre la variable de estudio y las variables del vector de calibración, y el último factor al cuadrado es el coeficiente de determinación (proporción de varianza explicada) en una regresión ponderada que pasa por el origen entre las desviaciones de las covariables $D_j = \hat{t}_{x,j} - t_{x,j}$ y las covarianzas de la variable de estudio y las covariables $C_j = cov(y, x_j)$.

B. Soluciones

Como se explicó en la sección anterior, si no existe correlación entre la variable de interés y la estructura de la falta de respuesta, no hay sesgo en los estimadores. Esto quiere decir que, si la probabilidad de respuesta es homogénea (nivel alto de representatividad) o pudiera modelarse (nivel bajo de representatividad) para todos los individuos, el sesgo se podría eliminar. En esta sección se explorarán dos caminos que, al incorporar información auxiliar, eliminan el sesgo causado por el fenómeno de la falta de respuesta.

Ambas opciones, la del ajuste de los factores de expansión mediante modelos de puntaje de propensión (*propensity score*) y la de los estimadores de calibración, descansan en el paradigma de la inferencia basada en el diseño de muestreo. Por ende, se contemplan como dos posibilidades atractivas que mantienen las buenas propiedades de la estimación directa en las encuestas de hogares.

1. El puntaje de propensión

Ya se ha mencionado que uno de los ajustes que se debe realizar al obtener los ponderadores finales es la corrección por falta de respuesta. Donde:

$$d_{4k}=\frac{d_{3k}}{\hat{\phi}_k}$$

Como se explicó en los capítulos anteriores, si el patrón de falta de respuesta no es aleatorio (*not missing at random* (NMAR)), entonces $\phi_k=f(\boldsymbol{y}_k,\beta)$. En este caso, como no es posible tener acceso a los determinantes de la respuesta (porque precisamente son las mismas variables de interés en la encuesta), tampoco es posible estimar el patrón de falta de respuesta. Por ende, en este escenario siempre habrá sesgo. Por el contrario, si el patrón de falta de respuesta es completamente aleatorio o aleatorio (*missing at random* (MAR)), entonces $\phi_k=f(\boldsymbol{x}_k,\boldsymbol{\beta})$. En este caso, si fuese posible tener acceso a las covariables $\boldsymbol{x}$ que determinan el mecanismo de respuesta, sería posible estimar las probabilidades de respuesta mediante $\hat{\phi}_k=f(\boldsymbol{x}_k,\hat{\boldsymbol{\beta}})$. Efectivamente, en el caso del estimador de Horvitz-Thompson, el sesgo del estimador se anula, puesto que:

$$\begin{aligned}
E(\hat{t}_y) &= E\left(\sum_{k\in s_r} d_{3k}\, y_k\right)\\
&= E\left(\sum_{k\in s_r} \frac{y_k}{\pi_k \hat{\phi}_k}\right)\\
&= E\left(E\left(\sum_{k\in U} \frac{y_k}{\pi_k \hat{\phi}_k} I_k D_k \mid I_k\right)\right)\\
&= \sum_{k\in U} \frac{y_k}{\pi_k \hat{\phi}_k}\, E(I_k)E(D_k \mid I_k)\\
&= \sum_{k\in U} \frac{y_k}{\pi_k \hat{\phi}_k}\, \pi_k \phi_k = t_y
\end{aligned}$$

Partiendo del supuesto de que el modelo está bien establecido, se tendrá una concordancia directa entre $\widehat{\phi_k}$ y ϕ_k. Por lo tanto, se anularían en la última igualdad de la ecuación anterior. Además, el insesgamiento viene supeditado a la siguiente relación:

$$E(I_k D_k)=E\big(E(I_k D_k \mid I_k)\big)=E(I_k)E(D_k \mid I_k)=\pi_k \phi_k$$

En resumen, si se tiene acceso a información auxiliar (contenida en el marco de muestreo o en otras preguntas de la encuesta), y si se considera que el mecanismo que produce la falta de respuesta en la encuesta de hogares es aleatorio o completamente aleatorio, es posible ajustar un modelo de puntaje de propensión para la falta de respuesta (donde la variable dependiente es una variable indicadora de la respuesta del individuo, por lo general supeditada a una distribución de Bernoulli o binomial). Así, es posible definir el siguiente estimador insesgado:

$$\hat{t}_y = \sum_{k \in s_{ER}} d_{4k}\, y_k$$

Donde:

$$d_{4k} = \frac{d_{3k}}{\widehat{\phi_k}} \;\; \forall \; k \in s_{ER}$$

Siempre es muy importante realizar una validación exhaustiva de los modelos utilizados para estimar la probabilidad de respuesta. En general, es necesario que el modelo satisfaga las dos condiciones que se describen a continuación:

i) Dominio común: al igual que en un experimento aleatorizado, es necesario asegurarse de que ninguna combinación de las covariables induzca un estado (respuesta o falta de respuesta) de forma determinística. Es decir, sobre todas las combinaciones en las covariables, deben existir respondientes y no respondientes. Esta condición se puede escribir como:

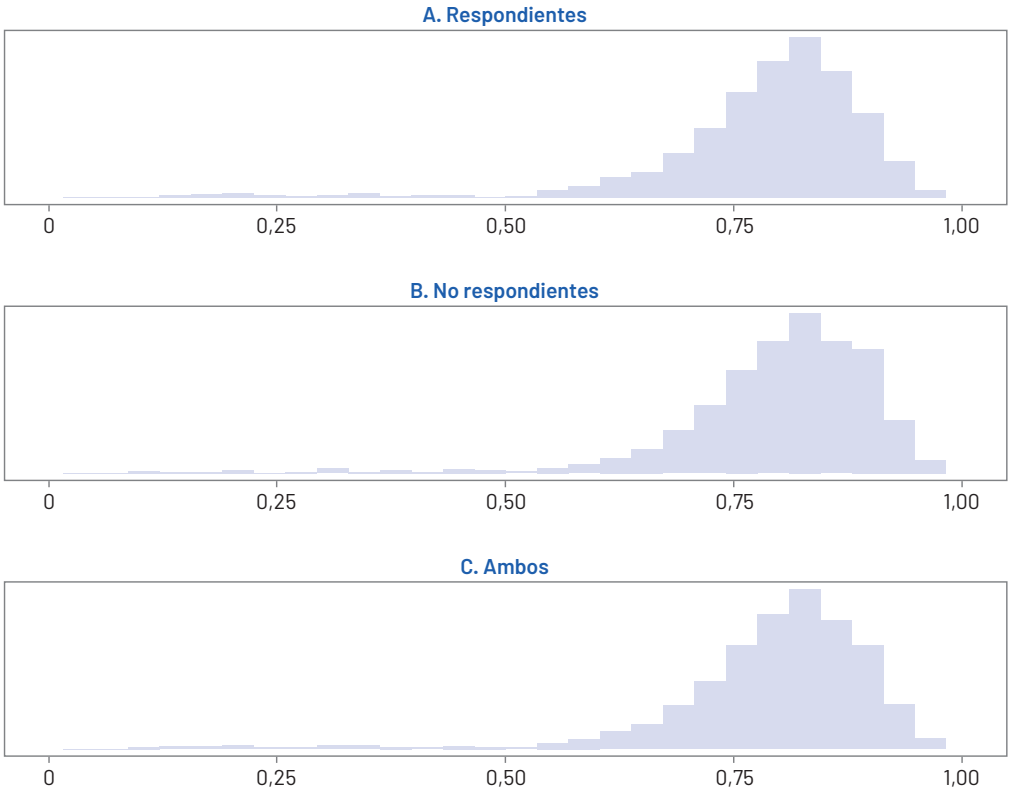
$$0 < Pr\,(D_{1,k})=1 \mid \boldsymbol{x}_1\,) < 1$$

ii) Balanceo: dado que la respuesta de las unidades de muestreo no proviene de un estudio aleatorizado, es necesario garantizar que la distribución de las $\hat{\phi}_k$ sea similar entre respondientes y no respondientes. De esta forma, es posible expandir el subconjunto de respondientes efectivos a la muestra original (que incluye a las unidades no respondientes), que a su vez se expande a toda la población de interés. Esta condición se puede escribir como:

$$\begin{aligned}\hat{\phi}_{1,k} &= Pr\,[D_{1,k} \mid I_{1,k}=1, \boldsymbol{x}] \\ &= Pr\,[D_{1,k} \mid k \in s_r, I_{1,k}=1, \boldsymbol{x}] \\ &= Pr\,[D_{1,k} \mid k \notin s_r, I_{1,k}=1, \boldsymbol{x}]\end{aligned}$$

Por último, se debe corroborar que la suma de los pesos ajustados por la falta de respuesta sea cercana al tamaño de la población que se quiere representar. El gráfico XII.1 permite ilustrar el dominio común entre respondientes y no respondientes para un modelo de puntaje de propensión. Ambas distribuciones son similares, por lo que es posible concluir que las covariables usadas representan bien la estructura estocástica en respondientes y no respondientes.

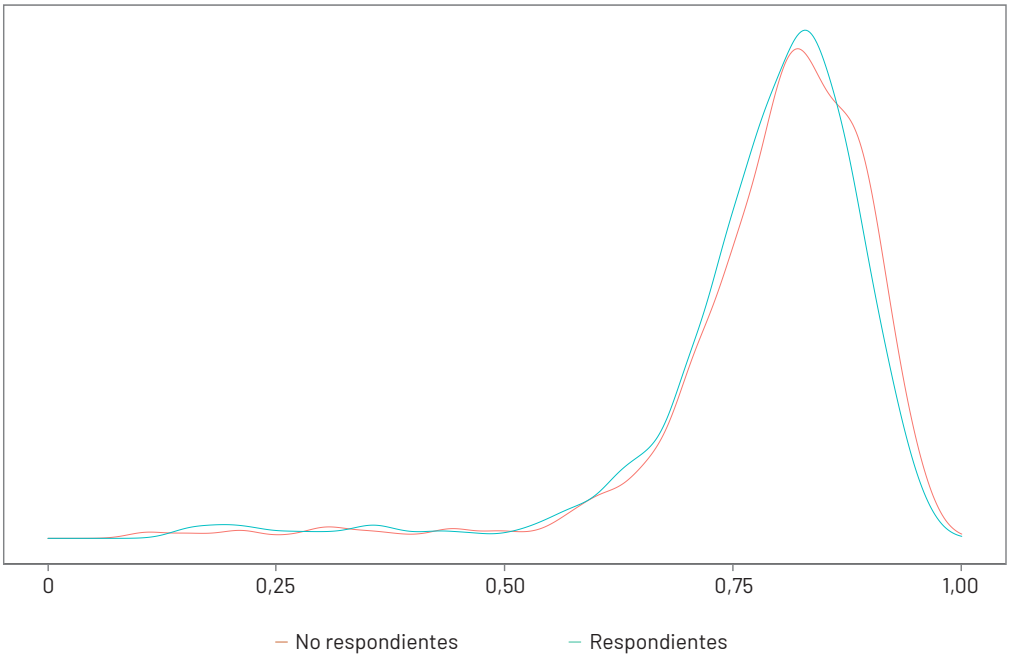
■ Gráfico XII.1
Histogramas de distribución de las probabilidades estimadas de respuesta de respondientes, no respondientes y ambos



Fuente: Elaboración propia.

Además, en el gráfico XII.2 se muestra la propiedad de balanceo en el modelo. Se observa que ambas distribuciones se alejan de los extremos (0 y 1) y presentan una caracterización similar.

■ Gráfico XII.2
Balanceo entre respondientes y no respondientes: densidades de respuesta para la población



Fuente: Elaboración propia.

2. Calibración

Como afirma Särndal (2007), la calibración proporciona una forma sistemática de utilizar la información auxiliar. En la mayoría de las aplicaciones prácticas, la calibración permite aplicar un enfoque simple para incorporar esta información dentro de la etapa de estimación. La información auxiliar se utilizaba para mejorar la precisión de los estimativos mucho antes de que el término "calibración" se popularizara. La calibración puede utilizarse efectivamente en encuestas donde la información auxiliar está disponible en diferentes niveles. Por ejemplo, al realizar un muestreo en dos etapas, puede haber información auxiliar para las unidades de la primera etapa (los conglomerados) y otro tipo de información para las unidades de la segunda etapa (elementos o conglomerados).

Como se ha detallado con amplitud en este capítulo y los anteriores, la falta de respuesta por unidad tiene consecuencias dañinas en la inferencia mediante encuestas de hogares. En este caso, es muy recomendable que se implemente el ajuste a los factores de ponderación de las unidades (hogares o personas). A pesar de que los modelos de puntaje de propensión tienen una larga trayectoria en el manejo de la falta de respuesta,

la calibración utilizada para corregir estos sesgos ofrece una perspectiva relativamente nueva. Nótese que el estimador tradicional toma la siguiente forma:

$$\hat{t}_y^* = \sum_{s_r} w_k \, y_k = \sum_{s_r} \frac{d_k}{\phi_k} \, y_k$$

La anterior expresión indica que, implícitamente, se obtiene un procedimiento en dos etapas. En primer lugar, se calculan los pesos básicos determinados por el diseño de muestreo y, posteriormente, se ajusta un modelo de puntaje de propensión para estimar las probabilidades de respuesta ϕ_k. A esta estrategia se le suele agregar una tercera etapa en la que se crean nuevos pesos calibrados con respecto a proyecciones demográficas poscensales; por ejemplo, los cruces entre edad, sexo y región. Al ajustar los pesos para que sumen exactamente la cifra de las proyecciones censales, se reduce el sesgo de subcobertura.

Särndal (2007) afirma que la práctica general consiste en suponer que el estimador $\hat{t}_y^*$ es insesgado, cuando en realidad no lo es. Esto se debe a que no es posible conocer todos los determinantes del mecanismo de respuesta para ajustar el modelo que estima las probabilidades de respuesta. Además, de la sección anterior se deduce que este supuesto implica considerar que $\pi_k \, \hat{\phi}_k$ es la verdadera probabilidad de inclusión de la unidad, cuando en realidad no es así. Por lo tanto, realizar un ajuste de los factores de expansión únicamente basado en los modelos de puntaje de propensión conllevará de manera inevitable cierta cantidad de sesgo en la estimación de los parámetros en las encuestas de hogares.

En este escenario, el enfoque de calibración doble surge como un proceso metodológico adicional que pretende corregir estos sesgos. Para poder utilizarlo, es necesario tener información auxiliar en dos niveles: la población y la muestra. Este tipo de metodología puede utilizarse en las encuestas de tipo panel o panel rotativo. En este proceso es necesario contar con dos tipos de información auxiliar:

i) Por un lado, estará la información poblacional usual que se utiliza para calibrar los factores de expansión en un proceso normal de recolección de información. Las variables que intervienen en esta calibración se denotarán como $\boldsymbol{x}_{1k}$ y, por lo general, indican la pertenencia de los individuos a regiones, o grupos de edad, sexo o área (urbana o rural).

ii) Por otro lado, se deberá tener acceso a información auxiliar en la muestra original (que incluya las unidades respondientes y no respondientes) y que se denotará como $\boldsymbol{x}_{2k}$. Por ejemplo, utilizando la información del panel en el momento de la primera medición, sería posible contar con información relativa a la condición de ocupación, ingresos o cualquier otra variable medida en la primera ronda del panel.

Por lo tanto, es posible calibrar los pesos en la muestra de respondientes (s_r) a nivel de la información auxiliar disponible en la muestra original (s), y luego a nivel nacional (U) o por los estratos de interés. Si el mecanismo que da lugar a la falta de respuesta es aleatorio o completamente aleatorio, es posible que los ponderadores de calibración eliminen el sesgo en las estimaciones finales, si es que las variables que generan este mecanismo

se han calibrado en alguno de los dos niveles mencionados. Särndal y Lundström (2005) proponen que, para lograr este objetivo, se encuentre un primer conjunto de pesos calibrados, sujetos a la siguiente restricción:

$$\sum_{s} w_{1k}\, \boldsymbol{x}_{1k} = \sum_{s} \boldsymbol{x}_{1k}$$

Luego, en una segunda etapa, se deben usar estos pesos intermedios w_{1k} para calcular los pesos finales de calibración w_k de la muestra de respondientes efectivos que están sujetos a la siguiente restricción:

$$\sum_{s_r} w_k\, \boldsymbol{x}_{2k} = \sum_{s} w_{1k}\, \boldsymbol{x}_k = \begin{pmatrix} \sum_{U} \boldsymbol{x}_{1k} \\ \sum_{s_r} w_{1k}\, \boldsymbol{x}_{2k} \end{pmatrix}$$

En este sentido, cabe señalar que la forma funcional de los pesos de calibración doble resultantes de este proceso de optimización se puede escribir de la siguiente manera:

$$w_k = d_k \times g_k \cong d_k \times \hat{\phi}_k$$

Por consiguiente, según el raciocinio de la calibración, los pesos g_k se pueden ver como una estimación de las probabilidades de respuesta ϕ_k. Por otra parte, de las expresiones sobre el sesgo de los estimadores que no contienen ningún tipo de corrección, se deduce que el sesgo se propaga a través de las variables de la encuesta y lo hace con más fuerza en el caso de las variables correlacionadas con los determinantes de la falta de respuesta.

Para mostrar cómo el ajuste a los factores de expansión con las dos metodologías antes mencionadas genera menos sesgo que los estimadores comunes, se planeó el siguiente experimento:

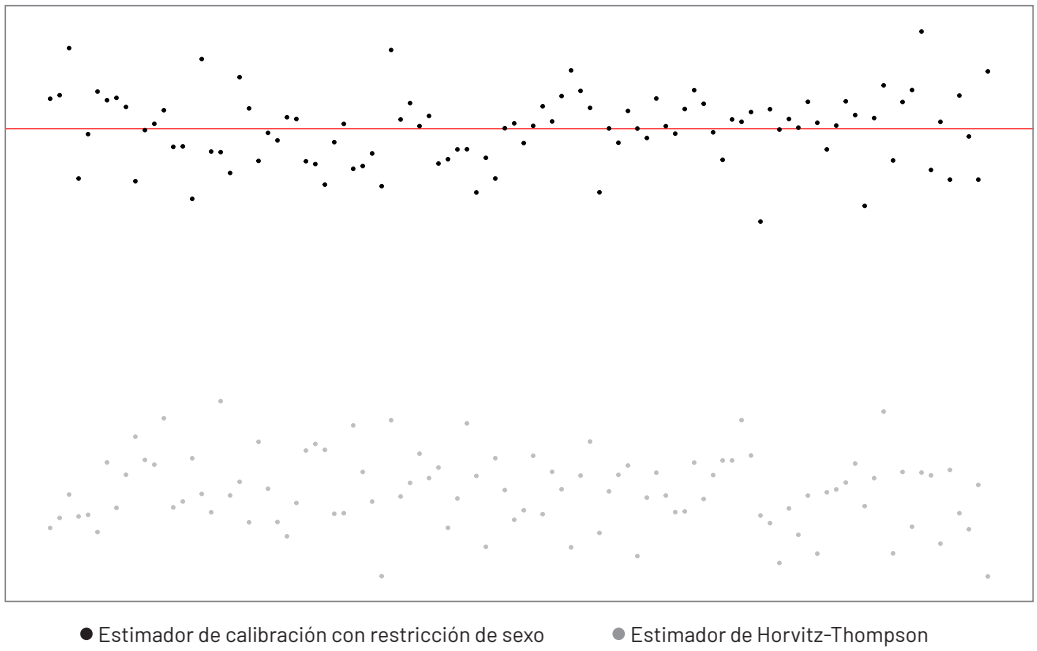
i) Se generó una población compuesta por individuos con diferente propensión de respuesta completamente aleatoria.

ii) Se utilizaron metodologías de calibración y se comparó, de forma empírica, el efecto de la falta de respuesta sobre las estimaciones finales.

En primera instancia, cabe mencionar que la población se definió a partir del ingreso del hogar y se creó a partir de variables auxiliares disponibles (sexo). De esta forma, se le dio una probabilidad de respuesta diferencial entre los grupos correspondientes al cruce de las categorías de estas dos variables. Como resultado de las simulaciones, se generaron estimaciones para el estimador de Horvitz-Thompson sin ajuste de ningún tipo y para un estimador de calibración que tuvo en cuenta los conteos poblacionales censales de las dos categorías de la variable sexo. En el gráfico XII.3 se muestra el comportamiento de ambas

estimaciones. La línea roja refleja el parámetro desconocido, los puntos negros indican las estimaciones del estimador de calibración en cada iteración de la simulación y los puntos grises muestran las estimaciones del estimador de Horvitz-Thompson en cada iteración.

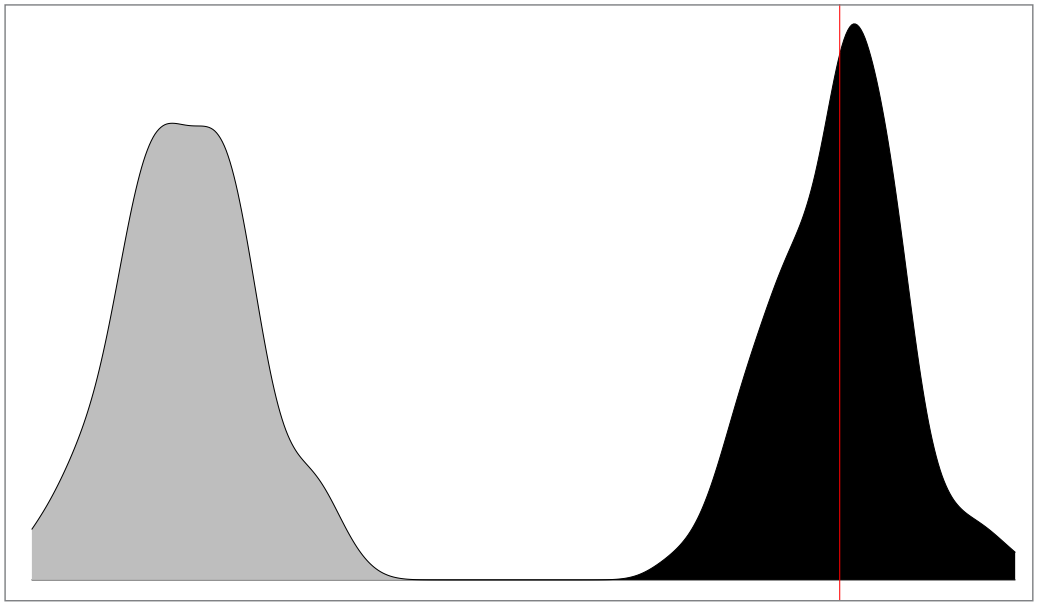
■ Gráfico XII.3
Estimaciones de Horvitz-Thompson y de calibración



Fuente: Elaboración propia.
Nota: La línea roja refleja el parámetro desconocido.

En conjunto con el gráfico XII.3, en el gráfico XII.4 se muestra la distribución sesgada del estimador de Horvitz-Thompson (puntos grises) en comparación con el insesgamiento del estimador de calibración (puntos negros). Con este modelo de respuesta, la inclusión en la calibración de las variables pertinentes corrige el sesgo generado por la falta de respuesta. En este estudió se encontró que el estimador ingenuo (de Horvitz-Thompson) produjo sesgo en la estimación de los tamaños de hombres y mujeres, en el tamaño de la población, en los ingresos de hombres y mujeres y en los ingresos de la población.

■ Gráfico XII.4
Distribuciones del estimador de Horvitz-Thompson y del estimador de calibración



Fuente: Elaboración propia.
Nota: La línea roja refleja el parámetro desconocido.

Como se mencionó anteriormente, hay mejores formas de calibrar, dado que el problema de la calibración se reduce a cómo introducir la información auxiliar en la estructura de estimación de la encuesta. Es posible que existan variables que reduzcan el sesgo, pero no todas las variables conllevarán el mismo nivel de precisión. Se deberían seleccionar las variables que reduzcan tanto el sesgo como la varianza. Por lo tanto, las variables auxiliares que se usen como insumo en los procesos de calibración deben:

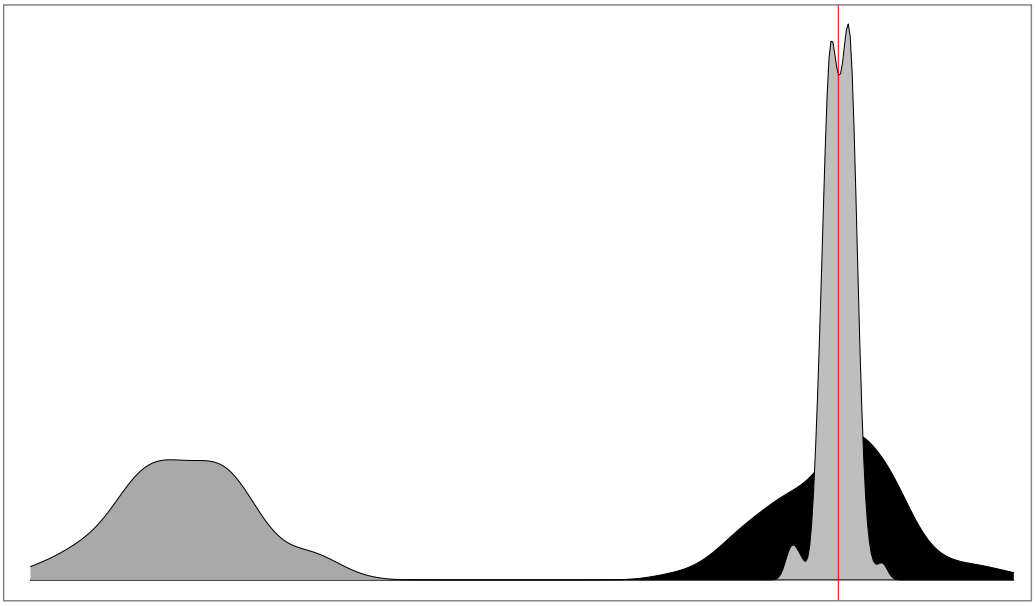
i) ser capaces de explicar la variación de la probabilidad de respuesta,

ii) estar correlacionadas con las variables de interés e

iii) identificar los dominios de estimación más importantes.

En particular, al introducir otras covariables en la calibración (grupo de edad, escolaridad, región o área), además de la corrección del sesgo, se evidencia un aumento de la precisión en las nuevas estimaciones, como muestran las distribuciones de los estimadores en el gráfico XII.5, donde se consideran tres estimadores: i) el estimador de Horvitz-Thompson (en gris claro); ii) el estimador de calibración con restricción de sexo (en negro), y iii) el estimador de calibración con todas las restricciones (en gris oscuro).

■ Gráfico XII.5
Distribuciones del estimador de Horvitz-Thompson y de dos estimadores de calibración



Fuente: Elaboración propia.
Nota: La línea roja refleja el parámetro desconocido.

C. Las consecuencias de la pandemia de COVID-19 en las encuestas de la región

Como se explica en CEPAL (2020c), en su intento por frenar la velocidad de contagio de la enfermedad por coronavirus (COVID-19), los Gobiernos de la región impusieron restricciones de movilidad que truncaron la recolección presencial de información de las encuestas de hogares. A fin de hacer frente a este inconveniente y poder seguir produciendo estadísticas oficiales pertinentes y oportunas, la mayoría de los institutos nacionales de estadística (INE) de la región decidieron realizar el seguimiento continuo de un panel seleccionado de un período reciente y, mediante contacto telefónico, continuar recolectando la información primaria. Uno de los retos más importantes que esta pandemia impuso a los INE fue la corrección del sesgo de selección en las encuestas de hogares. A pesar de los esfuerzos ingentes realizados para minimizar dicho sesgo durante la recolección, el cambio del modo presencial al modo telefónico produjo consecuencias indeseadas que se pudieron resolver gracias a algunas de las metodologías explicadas en este capítulo.

En CEPAL (2020b) se afirma que un buen punto de partida para los INE fue contar con una muestra probabilística de meses anteriores y poder conformar con ella un panel de seguimiento durante el período en que se aplicaron estas restricciones de la movilidad. En términos de notación, se puede considerar que esa es la muestra maestra. Sin embargo, se deben tener en cuenta los siguientes dos aspectos importantes:

i) no todos los hogares seleccionados de forma probabilística proporcionaron su información de contacto telefónico, y

ii) no todos los hogares con los que se puedo contactar respondieron el cuestionario de la encuesta.

Haciendo cálculos aproximados, si se supone que la cobertura de la submuestra que sí proporcionó datos de contacto asciende al 85% y que la probabilidad de que un hogar contactado responda toda la encuesta es del 80%, se contaría solamente con un 68% de la muestra original. Estas cuentas habría que ajustarlas con el efecto del desgaste en el panel, que aumenta a medida que este se sigue utilizando. En estos términos, sería un grave error y poco plausible suponer que los hogares respondientes efectivos se comportan de manera similar a los hogares no respondientes y a los hogares no cubiertos. El mejor escenario que puede plantearse es considerar que la muestra efectiva no está libre de sesgo, hacer una búsqueda exploratoria de la magnitud de dicho sesgo con los datos recolectados y tratar de minimizarlo (o incluso eliminarlo) utilizando alguna de las técnicas estadísticas que se han mencionado en este documento.

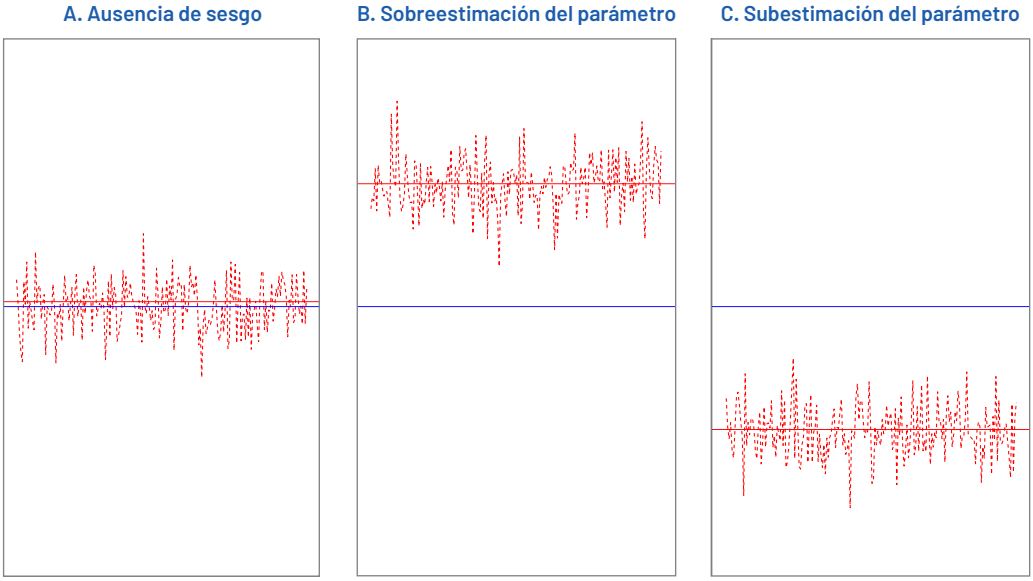
En el gráfico XII.6 se presentan tres posibles escenarios con los que se encontraron los INE en esta búsqueda. En el gráfico de la izquierda se verifica la falta de sesgo, mientras que en el del centro y en el de la derecha se confirma que la magnitud del sesgo es significativa.

En el gráfico XII.7 se muestra un escenario simulado donde se contempla el uso del estimador ajustado con la técnica de puntaje de propensión (línea verde) y el estimador de calibración en dos etapas (línea azul), comparado con el estimador sin ningún tipo de ajuste (línea negra). Lo que se esperaría es que el estimador ingenuo subestime los tamaños poblacionales y los indicadores de interés, mientras que los estimadores ajustados, siempre que el mecanismo de falta de respuesta sea aleatorio o completamente aleatorio, eliminen este sesgo.

Los caminos que se deben seguir después de corroborar la presencia (o ausencia) de sesgo dependerán de la estrategia de recolección de información que los países hayan decidido aplicar. En el escenario más optimista, ante la ausencia de sesgo, se estaría en una buena posición para reproducir los procesos usuales de inferencia. Sin embargo, ante la sospecha de que exista sesgo —posición parsimoniosa y recomendada en CEPAL (2020c)—, y dependiendo de la información auxiliar disponible, los INE pudieron disponer de las dos alternativas metodológicas que se describieron anteriormente.

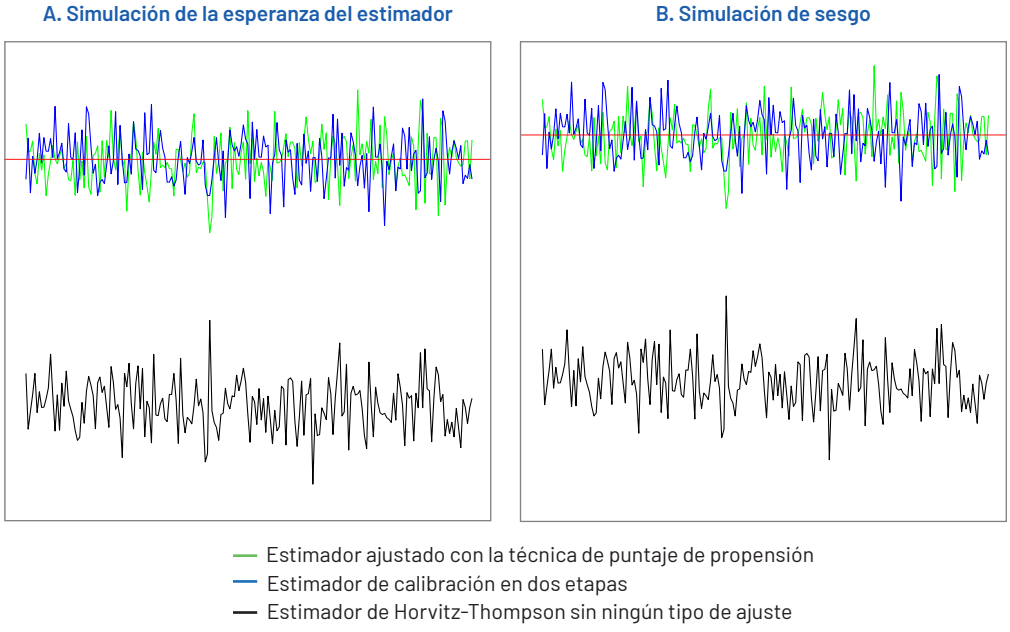
Muchos países de la región decidieron realizar un seguimiento mensual telefónico de la muestra maestra como respuesta a las restricciones de movilidad impuestas, que impidieron la recolección presencial de la información. En este caso, partiendo de una muestra probabilística, se pueden realizar ajustes a los factores de expansión de manera diferencial (CEPAL, 2020a).

■ Gráfico XII.6
Distribuciones del estimador de Horvitz-Thompson en tres escenarios de interés



Fuente: Elaboración propia.
Nota: La línea horizontal azul corresponde a la estimación publicada en el mes en que se seleccionó la muestra maestra, la línea horizontal roja representa el promedio de las simulaciones con la muestra efectiva y cada uno de los resultados de las simulaciones está representado por las fluctuaciones de las líneas punteadas.

■ Gráfico XII.7
Distribuciones del estimador de Horvitz-Thompson y de dos estimadores ajustados



Fuente: Elaboración propia.

En particular, dado que se tuvo acceso a todo un conjunto de covariables $\boldsymbol{x}$ en la muestra maestra, fue posible determinar el mejor modelo para estimar el patrón de falta de respuesta en la muestra de respondientes efectivos. En este paso se consideró que la probabilidad de respuesta dependía de alguna combinación lineal de las covariables en la muestra maestra; es decir, que se supone que el mecanismo que produce esa falta de respuesta se pudo describir mediante $\boldsymbol{x}$. Teniendo en cuenta que los pesos originales de la encuesta telefónica se denotan como d_k, y habiendo estimado $\hat{\phi}_k$ para respondientes y no respondientes de la muestra telefónica, el factor de expansión ajustado tomó la forma $w_k=\frac{d_k}{\hat{\phi}_k}$

En este sentido, la utilización del factor de expansión w_k minimizaría el sesgo de selección que se produjo por el cambio de modo en la recolección de la información. Por ejemplo, podría considerarse que un buen modelo de puntaje de propensión contempla la edad, el nivel educativo, el área de residencia (rural o urbana), el sexo del respondiente, la región geográfica, la situación de ocupación en el mes de observación de la muestra maestra y el ingreso per cápita del hogar. Nótese que todas las covariables en el modelo, salvo el área y la región geográfica, provienen necesariamente de las observaciones obtenidas en la muestra maestra.

Por otro lado, en CEPAL (2020b) se afirma que, al imponer cierta coherencia entre las cifras oficiales ya publicadas y las obtenidas a partir de la encuesta telefónica, es preferible el uso de los estimadores de calibración. Al usar este enfoque, se asegura una estructura inferencial robusta a partir de la información disponible, puesto que se reduce tanto el error de muestreo (al aumentar la precisión) como el error debido a la falta de respuesta (al eliminar el sesgo). A manera de ejemplo, podría considerarse que las dos etapas de la calibración que se describen a continuación son suficientes para eliminar el sesgo producido por el cambio en el modo de recolección:

i) En la primera etapa se calibran los pesos de la muestra maestra a partir de las variables de edad, región, área y sexo, definidas convenientemente en $\boldsymbol{x}_{1k}$. Los totales de estas variables se encuentran en los conteos censales o, en su defecto, en las proyecciones demográficas.

ii) En una segunda etapa se calibrarán los pesos de la muestra telefónica mediante las anteriores variables $\boldsymbol{x}_{1k}$ y, además, las variables de ingreso per cápita, situación de ocupación, rama de actividad y escolaridad, definidas convenientemente en $\boldsymbol{x}_{2k}$. Los totales de estas variables se estimaron en la misma publicación de la encuesta basada en la muestra maestra.

1. Ejemplo

En esta sección se revisan los pasos principales que se debieron considerar para eliminar (o al menos minimizar considerablemente) el sesgo de selección de una encuesta realizada durante la pandemia de COVID-19. Los datos que se utilizarán son artificiales, pero sirven para mostrar y ejemplificar las diferentes etapas propuestas para la evaluación y minimización del impacto del COVID-19 en la calidad de las encuestas.

Supóngase un conjunto artificial de datos que define una población finita U de tamaño $N=50.000$, y que se desea observar la situación laboral de cada persona en U. En aras de la simplicidad, supóngase que una persona solo puede estar empleada o desempleada. Se desea observar esta característica de interés en dos períodos diferentes, t_0 y t_1. Por un lado, t_0 corresponde a un período de recolección regular antes de la pandemia y, por otro, t_1 corresponde al período en que las restricciones de movimiento debido a la pandemia afectaron la recolección estándar de las encuestas por muestreo.

Si se tuviera acceso a toda la población, se encontraría que, en t_0, el 80% de las personas estaría empleada, mientras que el 20% estaría desempleada. Sin embargo, debido al impacto de la pandemia en los indicadores sociales (por ejemplo, pobreza y mercado laboral), en t_1, se observaría que muchas personas perdieron su trabajo y que la mitad de la población está desempleada.

El conjunto de datos incluido en el cuadro XII.1 muestra una versión reducida de los primeros diez individuos de esta población finita.

■ Cuadro XII.1
Ejemplo con una población total de 50.000 individuos: diez primeras filas

y_0	y_1
Ocupado	Ocupado
Ocupado	Desocupado
Ocupado	Ocupado
Ocupado	Ocupado
Ocupado	Desocupado
Ocupado	Desocupado
Ocupado	Ocupado
Desocupado	Desocupado
Ocupado	Desocupado
Ocupado	Desocupado

Fuente: Elaboración propia.
Nota: y_0 representa la característica de interés (situación laboral) en el período anterior a la pandemia de enfermedad por coronavirus (COVID-19), mientras que y_1 representa la característica de interés en el período de la pandemia de COVID-19.

En los cuadros XII.2 y XII.3 se muestran los flujos netos de la población finita en los dos períodos considerados.

■ Cuadro XII.2
Flujos netos verdaderos en la población del ejemplo antes de la pandemia de enfermedad por coronavirus (COVID-19)

y_0	Tamaño de muestra	Proporción
Desocupado	10 000	0,2
Ocupado	40 000	0,8

Fuente: Elaboración propia.
Nota: y_0 representa la característica de interés (situación laboral) en el período anterior a la pandemia de COVID-19.

■ Cuadro XII.3
Flujos netos verdaderos en la población del ejemplo en el marco de la pandemia de enfermedad por coronavirus (COVID-19)

y_0	Tamaño de muestra	Proporción
Desocupado	25 000	0,5
Ocupado	25 000	0,5

Fuente: Elaboración propia.
Nota: y_1 representa la característica de interés (situación laboral) en el período de la pandemia de COVID-19.

En el cuadro XII.4 se muestran los flujos brutos de la población finita entre los dos períodos considerados. Como se puede observar, 25.000 personas permanecieron ocupadas en los dos períodos y 15.000 personas cambiaron su situación laboral de ocupadas a desocupadas. De los desempleados en el primer período, ninguno pudo conseguir trabajo, mientras que 10.000 personas permanecieron desempleadas en ambos períodos.

■ Cuadro XII.4
Flujos brutos verdaderos del cambio en la situación laboral en la población del ejemplo

y_0	Tamaño de muestra	Proporción
Desocupado	10 000	0
Ocupado	15 000	25 000

Fuente: Elaboración propia.
Nota: y_0 representa la característica de interés (situación laboral) en el período anterior a la pandemia de enfermedad por coronavirus (COVID-19).

La medición y observación de la situación laboral se realiza mediante una encuesta por muestreo en ambos períodos. De esta forma, supóngase que se selecciona una muestra aleatoria simple sin reemplazo s_0 de tamaño $n_0=4.000$ de U. Para simplificar, supóngase que se pretende observar la misma muestra en ambos períodos (tipo panel).

Cabe tener en cuenta que la muestra anterior a la pandemia se conformó a partir del modo habitual de recolección presencial. Sin embargo, dadas las restricciones de movilidad impuestas por los Gobiernos para frenar la propagación de la pandemia, la modalidad de recolección del último período debió cambiar. Los INE utilizaron los registros de la muestra anterior a la pandemia para obtener el número de teléfono de los hogares seleccionados, tratar de ponerse en contacto con alguno de sus integrantes y administrarle un cuestionario por esa vía. Por supuesto, las tasas de muestreo en ambos períodos diferirían, puesto que no todos los hogares de la primera muestra proporcionaron un número de teléfono válido y, de los válidos, no todos contestaron la encuesta telefónica.

La muestra telefónica es más pequeña (2.305) que la muestra obtenida personalmente (4.000). Los investigadores creen que los sesgos de selección no son despreciables en la muestra telefónica. En los cuadros XII.5 y XII.6 se exponen los resultados basados en las muestras (no ponderados) de la encuesta presencial y la encuesta telefónica, respectivamente.

■ Cuadro XII.5
Resultados observados en la muestra presencial del ejemplo

	Estado	Tamaño de muestra	Proporción
Desocupado	Desocupado	820	0,205
Ocupado	Ocupado	3 180	0,795

Fuente: Elaboración propia.

■ Cuadro XII.6
Resultados observados en la muestra telefónica del ejemplo

	Estado	Tamaño de muestra	Proporción
Desocupado	Desocupado	909	0,394
Ocupado	Ocupado	1 396	0,605

Fuente: Elaboración propia.

El primer paso en la búsqueda de sesgos de selección consiste en calcular la tasa de respuesta. En este caso, de 4.000 encuestados seleccionados originalmente, solo 2.305 respondieron la entrevista telefónica. Esto equivale a una tasa de respuesta de tan solo el 58%.

Después, se debe evaluar la naturaleza de la falta de repuesta. En este paso se intenta reconocer si la falta de respuesta sigue una estructura aleatoria (MAR) o completamente aleatoria (MCAR). En este último caso, no se esperaría encontrar patrones bien definidos en las covariables. Es decir, ninguna categoría dentro de las covariables debería mostrar una tasa de respuesta diferente. Por otro lado, con el primer supuesto, se pueden encontrar patrones bien definidos en una o múltiples covariables. Para verificar cuál de los dos supuestos (MAR o MCAR) se ajusta mejor a las observaciones de la muestra seleccionada durante la pandemia de COVID-19, supóngase que se tiene acceso a los datos de la muestra del período prepandemia y se puede identificar a los individuos, encuestados y no encuestados, de la última muestra.

Como se ve en el cuadro XII.7, de los 2.305 encuestados en la encuesta telefónica, el 3,8% estaba empleado en el período anterior y aproximadamente el 96,2% estaba desempleado, lo que podría indicar un patrón de falta de respuesta aleatorio.

■ Cuadro XII.7
Proporciones relativas al estado de ocupación de los respondientes en la muestra telefónica del ejemplo

	Estado	Tamaño de muestra	Proporción
Desocupado	Desocupado	88	0,038
Ocupado	Ocupado	2 217	0,962

Fuente: Elaboración propia.

Finalmente, del cuadro XII.8 se deduce que, al examinar el estado anterior, de los 1.695 no encuestados, se observa que casi el 43,2% estaba empleado en el período anterior, mientras que el 56,8% estaba desempleado. Las proporciones no son similares en ningún aspecto, lo que apunta a un posible sesgo de selección.

■ Cuadro XII.8
Proporciones relativas al estado de ocupación de los no respondientes en la muestra telefónica del ejemplo

	Estado	Tamaño de muestra	Proporción
Desocupado	Desocupado	732	0,432
Ocupado	Ocupado	963	0,568

Fuente: Elaboración propia.

Para verificar la asociación entre la respuesta en la encuesta telefónica y la situación laboral en la encuesta presencial, se pueden utilizar herramientas de inferencia clásica, como la prueba χ^2 de Pearson y el estadístico V de Cramer. En el cuadro XII.9 se resume el comportamiento de la respuesta en la encuesta telefónica, dada la situación laboral en la encuesta presencial.

■ Cuadro XII.9
Asociación entre la respuesta telefónica y la situación laboral del período anterior en la muestra del ejemplo

Estado	Respuesta	Frecuencia
Desocupado	No	732
Ocupado	No	963
Desocupado	Sí	88
Ocupado	Sí	2 217

Fuente: Elaboración propia.

Sobre la base de la información brindada en el cuadro XII.9, es posible realizar la prueba de bondad de ajuste χ^2 de Pearson entre las dos variables de interés (respuesta en la encuesta telefónica y situación laboral en la encuesta presencial) para determinar si existe una correlación significativa. El sistema de hipótesis es el siguiente:

- H0: Las dos variables son independientes.
- H1: Las dos variables se relacionan entre sí.

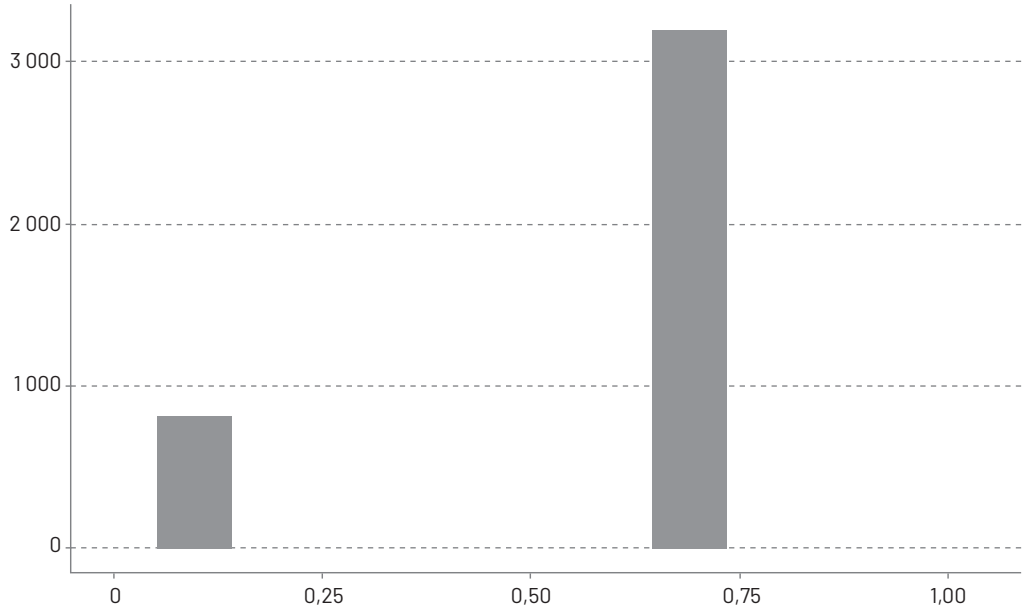
A partir de estos datos, la estadística de prueba toma el valor de 944 (muy grande) con un valor p muy pequeño y cercano a 0, lo que indica una relación bien definida. Por último, la estadística V de Cramer mide la fuerza de la asociación entre dos variables nominales y toma valores entre 0 (sin asociación) y 1 (asociación fuerte). En este caso, el valor del estadístico es cercano a 0,5, lo que indica una asociación significativa que debe tenerse en cuenta en los pasos siguientes con el fin de minimizar el posible sesgo de selección que puede afectar la inferencia de la encuesta telefónica.

Como se pudo observar anteriormente, existen indicios significativos de que el mecanismo de respuesta de la encuesta telefónica depende de la situación laboral del individuo en el período anterior. Eso significa que las personas que estaban desempleadas tienden a responder menos que las que estaban empleadas. Para simplificar el ejemplo, supóngase que la probabilidad de ser un encuestado depende de la situación laboral anterior. De esta forma, ϕ_k puede escribirse como una función de ese estado laboral anterior, incluido en las covariables z.

$$\phi_k = f(z_k, \boldsymbol{\beta})$$

Dado que ϕ_k se puede escribir como una función de las covariables disponibles, es posible afirmar que el mecanismo de respuesta sigue un enfoque aleatorio. Como el patrón de no respuesta es aleatorio, y es posible tener acceso a las covariables z que determinan el mecanismo de respuesta, también será posible estimar las probabilidades de respuesta por $\hat{\phi}_k = f(z_k, \hat{\boldsymbol{\beta}})$ para utilizarlo en la generación de nuevos pesos. Después de ajustar un modelo de puntaje de propensión, en el gráfico XII.8 se muestra el histograma de las probabilidades de respuesta estimadas, que solo toman dos valores (0,6971698 y 0,1073171), uno para cada categoría de la situación laboral en la encuesta presencial.

■ Gráfico XII.8
Histograma de los puntajes de propensión



Fuente: Elaboración propia.

Luego, utilizando los datos telefónicos y el nuevo conjunto de ponderaciones d_{4k}, ajustado por el puntaje de propensión estimado, se obtiene que el número estimado de empleados en el período del COVID-19 es $\hat{t}_y = \sum_{k \in s_{ER}} d_{4k} y_{1k} = 24.970{,}23$.

Por otra parte, también es posible calibrar los pesos en la muestra telefónica a nivel de la información auxiliar disponible en la muestra presencial, y luego a nivel nacional. Como el mecanismo que da lugar a la falta de respuesta es aleatorio, es posible que este nuevo conjunto de pesos de calibración elimine el sesgo. Para lograr este objetivo, se encuentra un primer conjunto de pesos calibrados sujetos a la siguiente restricción:

$$\sum_{s0} w_{0k}\, \boldsymbol{x}_k = \sum_{U} \boldsymbol{x}_k = \boldsymbol{t}_{\boldsymbol{x}}$$

Donde $\boldsymbol{t}_{\boldsymbol{x}}$ puede representar conteos nacionales provenientes de censos o proyecciones demográficas. En una segunda etapa, estos pesos intermedios w_{0k} deben utilizarse para calcular los pesos finales de calibración w_{1k} de la muestra de encuestados efectivos que están sujetos a la siguiente restricción:

$$\sum_{s_0} w_{1k}\, \boldsymbol{x}_k = \begin{pmatrix} \sum_{U} \boldsymbol{x}_k \\ \sum_{s_1} w_{0k}\, \boldsymbol{z}_k \end{pmatrix} = \begin{pmatrix} \boldsymbol{t}_{\boldsymbol{x}} \\ \hat{\boldsymbol{t}}_{\boldsymbol{z}} \end{pmatrix}$$

Donde $\boldsymbol{t}_{\boldsymbol{z}}$ representa las cifras estimadas provenientes de la encuesta presencial. El estimador de calibración se puede escribir de la siguiente manera:

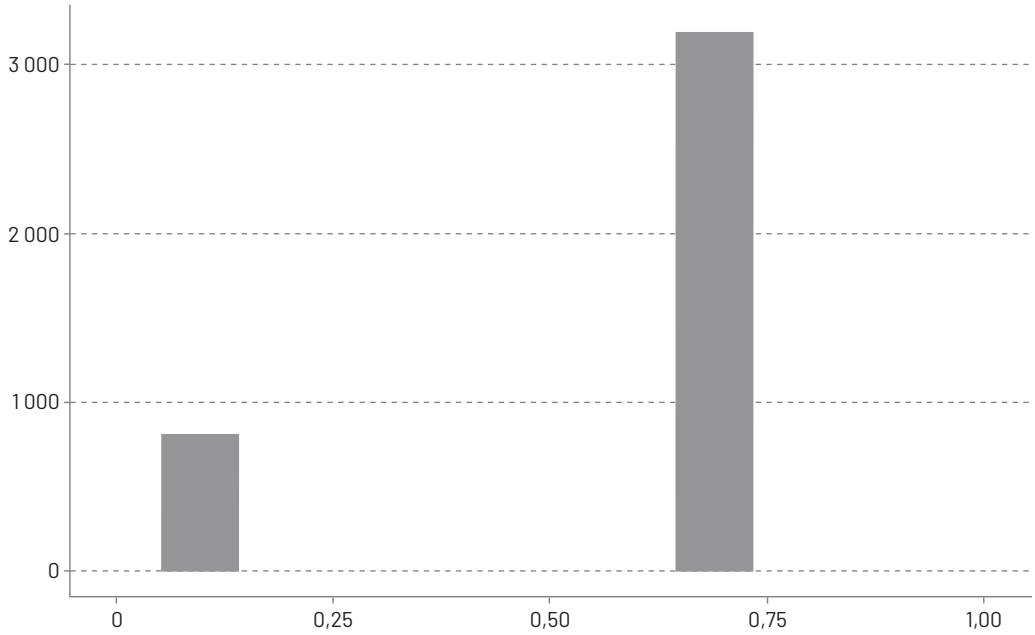
$$\hat{t}_y^{cal} = \sum_{k \in s_1} w_{1k}\, y_{1k}$$

Después de realizar la calibración en dos etapas, y utilizando los datos de la encuesta telefónica, junto con el nuevo conjunto de pesos calibrados, se obtiene que el número estimado de empleados en el período del COVID-19 es $\hat{t}_y^{cal} = \sum_{k \in s1} w_{1k}\, y_{1k} = 25.448$. Téngase en cuenta que la forma funcional de los pesos de calibración doble resultantes de este proceso de optimización se puede escribir de la siguiente manera:

$$w_{1k} = d_k \times g_{0k} \times g_{1k} \cong d_k \times \hat{\phi}_k$$

Por lo tanto, en este escenario, los pesos g_k pueden verse como una estimación de las probabilidades de respuesta ϕ_k. En el gráfico XII.9 se muestra el histograma de los puntajes de propensión pronosticados. Solo toman dos valores (0,6928125 y 0,11), uno para cada categoría de la situación laboral en la encuesta presencial.

■ Gráfico XII.9
Histograma de los pesos ajustados (puntajes de propensión con calibración)



Fuente: Elaboración propia.

Por último, cabe mencionar que, si no se considerara el mecanismo de falta de respuesta, sería fácil obtener estimaciones engañosas y sesgadas. Este enfoque ingenuo e incorrecto conduce al siguiente estimador sesgado:

$$\hat{t}_y^{exp} = \sum_{k \in s_1} d_{3k} y_{1k}$$

Donde d_{1k} se refiere a los pesos muestrales no ajustados de la encuesta telefónica. En este escenario, el número estimado de empleados en el período del COVID-19 es $\hat{t}_y^{exp} = \sum_{k \in s_1} d_{1k} y_{1k} = 19.718.$

Capítulo XIII
Falta de respuesta por registro

Es posible que, en el diseño y la realización de una encuesta, se produzcan situaciones que pueden sesgar las estimaciones finales. Este tipo de sesgos puede ocurrir antes, durante y después de la recolección de los datos. Los estadísticos deben advertir los problemas que causan los sesgos en la inferencia para procurar minimizar los errores humanos en todas las etapas de la encuesta, con la finalidad de que los resultados del estudio sean lo más confiables posible.

Como se mencionó en los capítulos anteriores, el enfoque recomendado para tratar la falta de respuesta es una combinación de los procedimientos de imputación y ponderación que se conoce como "enfoque combinado". El método consiste en imputar los valores de las celdas vacías correspondientes a los individuos de los que falta al menos un registro en la base de datos (excepto aquellos de los que falta la mayoría o la totalidad de los registros). En resumen, los individuos que no contestaron ninguna pregunta del cuestionario se eliminan del análisis, mientras que los restantes se consideran con sus respuestas originales o con la imputación de las celdas vacías.

En este capítulo, se abordan el enfoque de imputación utilizado para tratar la falta de respuesta por registro y los métodos disponibles para encontrar valores atípicos en la base de datos que, por consiguiente, también pueden imputarse. El verbo "imputar" viene del latín *imputāre*, cuya traducción al español puede ser "calcular", "estimar" o "atribuir". En el siglo XIX, se utilizó el término "ingreso imputado" para denotar el ingreso derivado de los bienes raíces. Según Van Buuren (2018), la connotación de la imputación, en el sentido de completar valores en una base de datos, apareció por primera vez en 1957 en la Oficina del Censo de los Estados Unidos. Es así como el tratamiento de los datos faltantes consiste en estimar valores razonables, en función de los observados en la base de datos.

A. Modelos de imputación

El término "imputación" se refiere al conjunto de técnicas mediante las cuales se reemplazan los valores faltantes en una o más variables con información probable, a fin de lograr valores sustitutivos que permitan el análisis posterior de la base de datos. Este proceso supone un nuevo elemento de error, conocido como "error de imputación", debido a la incertidumbre que introducen los valores no observados. En los casos de falta de respuesta por registro, es preferible utilizar técnicas de imputación en lugar de sistemas de ponderación en la muestra. De esta manera, es posible crear un conjunto completo y rectangular de datos mediante la imputación de los valores faltantes, puesto que, después de realizar la imputación, se espera que todos los valores del cuestionario de una persona contengan información y no exista ningún vacío.

Para lograr la sustitución de los valores faltantes con información probable, se pueden encontrar donantes apropiados en la misma muestra que se ha seleccionado; es decir, personas encuestadas con características demográficas similares a las de la persona que no respondió. Por lo tanto, la información del donante (o una función de estos valores) se copiará en las celdas vacías de la persona que no respondió. Para encontrar los donantes, es posible realizar un análisis estadístico basado en métodos de clasificación. Algunos de los métodos de imputación más utilizados en las encuestas de hogares son los siguientes:

- Imputación del valor medio (*mean value imputation*): se utiliza la media de la variable (de las unidades primarias de muestreo (UPM) o un subconjunto apropiado de datos). En este caso, los valores faltantes se reemplazan inmediatamente por el promedio de los datos de un subgrupo apropiado de respondientes.
- Imputación en caliente (*hot deck imputation*): se reemplazan los valores faltantes por los valores de un donante que respondió a la misma encuesta. En este caso, el valor faltante se reemplaza por la información de una persona escogida de antemano.
- Imputación en frío (*cold deck imputation*): se reemplazan los valores faltantes por los valores de un donante que respondió a la misma encuesta, pero en una edición anterior. En este caso, el valor faltante se reemplaza por la información auxiliar de una persona escogida de encuestas anteriores.
- Imputación estadística basada en modelos estadísticos: la variable dependiente es aquella que se quiere imputar y las covariables se derivan del conjunto de datos restante. En este caso, el valor faltante se reemplaza por la predicción (o una función) del modelo ajustado con la información de la muestra.

Como se mencionó anteriormente, se distinguen dos tipos de métodos de imputación. La imputación de la unidad completa, que se produce cuando se imputa toda la información de un individuo, y la imputación de registros, cuando se imputa solo una parte de la información relativa a una unidad. Mientras que la imputación de la unidad se utiliza para hacer frente a la falta total de respuesta, es decir, cuando no se dispone de ningún dato del individuo, la imputación del registro se utiliza cuando los datos proporcionados están incompletos.

El proceso de imputación se realiza a menudo en grupos no traslapados $g=1,...,G$, donde la unión de $s_1,...,s_G$ equivale a la muestra completa s. Si bien se pueden utilizar diferentes métodos de imputación para los distintos grupos, es necesario usar el mismo método dentro de cada uno de ellos. Esto se debe a que las covariables disponibles pueden diferir entre uno y otro grupo. Cuando la disponibilidad de variables auxiliares (covariables) es limitada, se puede considerar una jerarquía de métodos de imputación. Mientras que, en el caso de los grupos sobre los que se dispone de más información, es posible utilizar métodos de imputación más sofisticados, en el de los grupos con menos información auxiliar disponible, se deben utilizar métodos más simples. Särndal y Lundström (2005) examinan el uso de esta técnica en combinación con los estimadores utilizados en las encuestas de hogares que sirven de base para las estadísticas oficiales. A continuación se presenta una lista no exhaustiva de los principales métodos de imputación que se utilizan en las encuestas de hogares.

1. Imputación por regresión

En este método determinístico, el valor imputado para el valor faltante y_k se calcula utilizando una regresión lineal.

$$\hat{y}_k = x_k \hat{\beta}_i$$

Donde:

$$\hat{\beta}_i = \left(\sum_{r_i} a_k \, x_k \, x_k' \right)^{-1} \sum_{r_i} a_k \, x_k \, y_k$$

El vector de los coeficientes de regresión $\hat{\beta}_i$ se produce a partir de un ajuste de regresión múltiple utilizando los datos (y_k, x_k) disponibles para cada unidad $k \in r_i$ con pesos a_k especificados adecuadamente. En general, las predicciones del modelo de regresión no necesariamente corresponden a valores observados en otro individuo de la muestra. Por lo tanto, con este método se obtienen valores imputados que no se han observado en la encuesta. El número de modelos de regresión que se han de generar equivaldrá al número de variables con valores faltantes.

2. Imputación de razón

Un caso especial del método anterior se observa cuando solo se tiene acceso a una única covariable (positiva) $\mathrm{x}_k = x_k$, y definiendo $a_k = \frac{1}{x_k}$. En este caso, la estimación del coeficiente de regresión será:

$$\hat{\beta}_i = \frac{\sum_{ri} y_k}{\sum_{ri} x_k} = R_i$$

Por lo tanto, la imputación para el valor faltante se convierte en:

$$\hat{y}_k = x_k \hat{\beta}_i = x_k \frac{\sum_{r_i} y_k}{\sum_{r_i} x_k} = x_k R_i$$

Este método se utiliza a menudo cuando la misma variable se mide en dos momentos diferentes en la misma encuesta. Por ejemplo, si y indica la variable de estudio en el momento actual, x indica la variable en el punto de tiempo anterior y el coeficiente utilizado para la imputación es la relación entre los dos puntos en el tiempo.

3. Imputación del valor promedio

El caso más sencillo de la imputación por regresión se da cuando $a_k = x_k = 1$ para todo $k \in r_i$. En este escenario, el valor imputado se convierte en:

$$\hat{y}_k = \frac{\sum_{r_i} y_k}{\sum_{r_i} 1} = \bar{y}_{r_i}$$

Por lo tanto, todos los valores faltantes recibirán el mismo valor imputado, que es justamente el promedio de la variable en el conjunto de personas encuestadas. Cabe destacar que en este método no se requiere ninguna información adicional.

4. El vecino más cercano

Al suponer que valores similares de x producirán valores similares de y, es posible "pedir prestado" un valor de y para imputar el valor faltante de un "vecino" con valores similares en x. En este caso, el valor imputado para la unidad k está dado por:

$$\hat{y}_k = y_{l(k)}$$

Donde $l(k)$ es el "elemento donante", determinado al minimizar una ecuación de distancia. En el caso más simple, para una sola covariable de imputación x_k, la distancia entre los posibles donantes l con respecto a la unidad k es:

$$D_{lk} = | x_k - x_l |$$

El donante l del elemento k es el individuo con la menor distancia D_{lk} entre todos los posibles elementos $l \in r_i$. Cuando se contempla más de una covariable de imputación, es posible considerar la siguiente distancia:

$$D_{lk} = \left(\sum_{j=1}^{J} h_j \, (x_{jk} - x_{jl})^2 \right)$$

Donde h_j se utiliza para ponderar adecuadamente cada una de las J covariables de la matriz de imputación.

5. Imputación en caliente (*hot deck*)

Los métodos de imputación por regresión y el vecino más cercano suponen una fuerte relación entre la variable de interés y y las covariables x. Sin embargo, en algunas aplicaciones, esta relación no se puede establecer fácilmente y no es posible validar los supuestos de modelación que otros métodos requieren. Por lo tanto, en este tipo de técnica, el valor imputado para el individuo k está dado por:

$$\hat{y}_k = y_{l(k)}$$

Donde el valor imputado $y_{l(k)}$ es proporcionado por un donante seleccionado aleatoriamente del conjunto de datos de la variable de interés. Este método no se recomienda cuando existen mejores opciones, pues se carece de información auxiliar para determinar un buen sustituto.

6. Imputación múltiple

Cuando se dispone de información auxiliar que permite relacionar las covariables con la variable de interés, es posible establecer mejores modelos que no solo mantienen el insesgamiento de la inferencia, sino que también estiman el error de muestreo con bastante precisión. Con respecto a esta última categoría de imputación, se puede completar el conjunto de datos utilizando información auxiliar de los participantes en la encuesta (o en encuestas anteriores, si se trata de un diseño rotativo) y la información disponible a nivel de la población para predecir los valores faltantes mediante un modelo de regresión. Una de las técnicas más robustas es la imputación múltiple, que consiste en formular un modelo probabilístico entre la variable de interés y las covariables disponibles en la encuesta (Rubin, 1987). Se supone que este modelo presenta la siguiente forma:

$$\hat{y}_k = f(x_k, \beta) + \varepsilon_k, k \in r_i$$

Donde ε_k es un término de error aleatorio. Una vez formulado el modelo, debido a la naturaleza estocástica de ε_k, es posible generar $M > 1$ realizaciones de la variable de interés para los registros faltantes. Esto se logra de manera muy sencilla, simulando M valores del término de error. De esta forma, se generan M conjuntos de datos completos. Para cada conjunto de datos, se generarán M estimaciones de interés, que luego se promediarán para obtener una estimación puntual.

B. Ejemplo de imputación en una encuesta de ingresos y gastos

Tras examinar los propósitos de la imputación en una encuesta de hogares, es necesario escoger un método (o varios métodos) de imputación y, una vez establecido el mecanismo

de imputación, generar un conjunto de datos rectangular y completo. A la luz de las particularidades de las encuestas de ingresos y gastos de los hogares, en esta sección se analizan los pasos que se deben seguir para completar un proceso de imputación. Por sus características, este tipo de encuestas presenta tasas elevadas de falta de respuesta por registro y, en menor medida, por unidad.

En general, este tipo de encuestas se basa en un trabajo de campo masivo, que prevé varias visitas a cada hogar, durante las cuales se solicita a la persona encuestada que responda a sendos cuestionarios y registre toda la información relativa a los gastos y los ingresos del hogar durante un período de al menos dos semanas. Por supuesto, para que esto pueda realizarse, es necesario contar con la colaboración activa de todos los miembros del hogar. En el mejor de los casos, el encuestador visitará varias veces el domicilio del hogar en el período de observación y obtendrá un formulario completo. A menudo, a pesar del seguimiento exhaustivo del encuestado, no se obtendrá información sobre todas las categorías de gastos de la encuesta, sino información parcial que se traducirá en celdas vacías por falta de respuesta. En el peor de los casos, se obtendrán cuestionarios tan incompletos que deberán descartarse y considerarse faltantes.

Con el siguiente ejemplo, se trata de ilustrar de manera escueta el procedimiento de imputación en una encuesta de ingresos y gastos. Esta metodología incluye varios pasos, puesto que, antes de imputar las variables de interés, es necesario determinar las covariables que se relacionan fuertemente con ellas. Además, es necesario imputar todas las covariables en primer lugar y reemplazar los valores faltantes con información razonable que pueda utilizarse en los modelos que se ajusten. Se supone que, en el conjunto de hogares considerados con fines de imputación, se observaron al menos las siguientes variables:

- tamaño del hogar,
- número de hombres y mujeres en el hogar,
- número de niños y adultos en el hogar,
- edad del jefe del hogar,
- estado de ocupación del jefe del hogar,
- nivel educativo más alto alcanzado por el jefe del hogar y
- número de personas empleadas en el hogar.

El primer paso del camino que se seguirá en este ejemplo consiste en la imputación de los ingresos, como principal covariable del gasto y del consumo. Una vez imputadas las covariables, el segundo paso corresponde a la imputación de los filtros, que son las preguntas que se formulan para saber si un hogar ha adquirido un determinado bien o servicio. El tercer paso se refiere a la imputación de los valores de gasto anualizados en cada unidad. Esta serie de pasos metodológicos ha sido recomendada por diferentes organismos estadísticos, incluidos institutos y oficinas nacionales de estadística. Por ejemplo, Hayes y Watson (2009) y Sun (2010) siguen esta metodología en la Oficina Australiana de Estadística para la imputación en la Encuesta de Hogares, Ingresos y Dinámica Laboral de Australia (HILDA).

1. Imputación de los ingresos

En primer lugar, cabe recordar que existen múltiples fuentes de ingresos en el hogar; por ejemplo, los ingresos derivados del trabajo, la propiedad de activos, la producción de servicios para consumo propio y las transferencias gubernamentales (condicionadas o no). Asimismo, los ingresos han demostrado ser un importante predictor de los gastos, tanto teórica como empíricamente (Starick y Watson, 2011).

La imputación de los ingresos podría basarse en un enfoque de modelos predictivos y la técnica que se podría utilizar para imputar esta covariable es la del vecino más cercano con regresión. De esta forma, se define un modelo lineal para las unidades encuestadas y luego se estiman los coeficientes de regresión para obtener un valor predicho que se imputa a las unidades faltantes. Así, para cada unidad a la que falta información sobre los ingresos, se identifica un solo donante, que corresponderá al hogar cuyos ingresos disponibles son más cercanos a la predicción del modelo de regresión. Por ende, todos los componentes de los ingresos se imputarán con la información del donante. El modelo lineal se describe a continuación y la predicción de los ingresos para los hogares faltantes se calcula utilizando una regresión lineal.

$$\tilde{y}_k = x_k \hat{\beta}_i$$

Donde, $\tilde{y}_k$ es el valor predicho de los ingresos disponibles para el hogar k, x_k es el vector de las covariables del modelo y los coeficientes de regresión estimados están dados por:

$$\hat{\beta}_i = \left(\sum_{r_i} a_k \, x_k \, x_k' \right)^{-1} \sum_{r_i} a_k \, x_k \, y_k$$

Este vector de coeficiente de regresión $\hat{\beta}_i$ se produce a partir de un ajuste de regresión múltiple utilizando los datos (y_k, x_k) disponibles para cada unidad $k \in r_i$ con pesos a_k especificados adecuadamente para incluir la posible heterocedasticidad de los residuales. Por ejemplo, es recomendable que la información incluida en el vector de covariables x_k contenga los siguientes datos:

- Composición del hogar: número de adultos, número de niños, número de hombres, número de mujeres, edad media de los adultos, edad media de los niños, edad de la persona más joven, edad de la persona mayor, edad del jefe del hogar y nivel educativo más alto alcanzado por el jefe del hogar.
- Ocupación y fuerza de trabajo: situación laboral del jefe del hogar y número de personas empleadas y desempleadas en el hogar.
- Calidad de la vivienda: indicador creado a partir de la sección relativa a la calidad de la vivienda, que incluye, por ejemplo, un índice de hacinamiento (la relación entre el número de habitaciones utilizadas principalmente para dormir y el número de personas en el hogar), el material de las paredes y la principal fuente de agua potable del hogar.
- Ubicación del hogar: municipalidad y provincia, como primera y segunda desagregación cartográfica del país.

Al suponer que valores similares de las predicciones del modelo lineal $\tilde{y}$ producirán valores similares en las observaciones del ingreso y, se podría "pedir prestado" un valor real de ingreso y para imputar el valor faltante con la información de este vecino que tiene valores similares en las predicciones $\tilde{y}$ del modelo lineal. Así, el valor imputado para la unidad k está dado por:

$$\hat{y}_k = y_{l(k)}$$

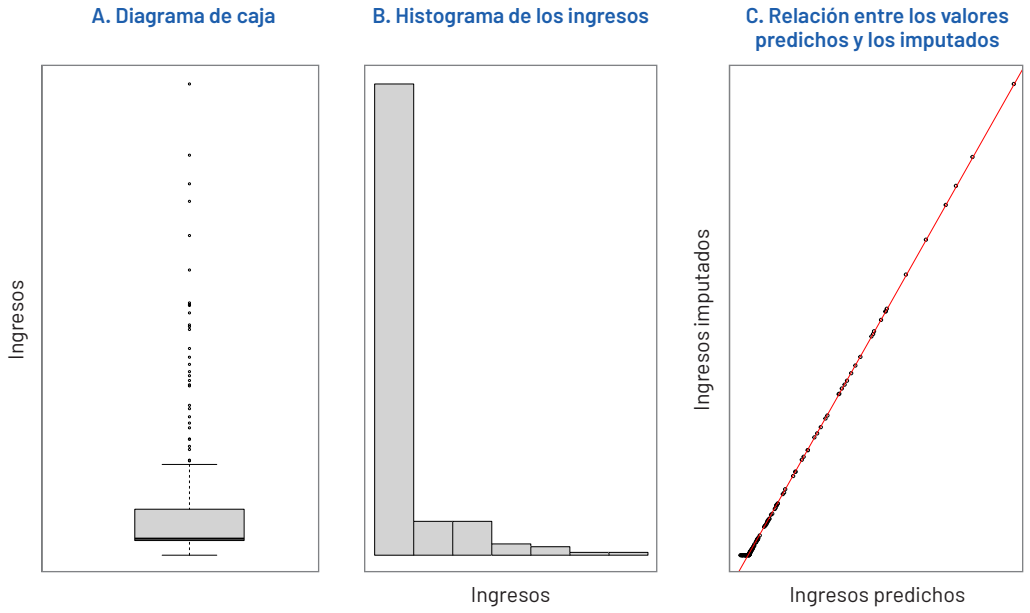
Donde $l(k)$ es el elemento donante, determinado por medio de la minimización de una medida simple de distancia entre todos los posibles donantes l y la unidad k. Esta distancia está dada por:

$$D_{lk} = | \tilde{y}_k - y_l |$$

El donante l al elemento k será el hogar del conjunto r_i con la menor distancia D_{lk}. Por regla general, todos los donantes deben estar situados en la misma provincia que la unidad faltante. En el gráfico XIII.1 se muestran un diagrama de caja, el histograma de los ingresos (antes de la imputación) y la relación lineal entre los valores predichos derivados del modelo y los valores imputados tomados de los donantes.

■ Gráfico XIII.1

Distribución de los ingresos y relación entre los valores predichos e imputados para los hogares con datos de ingresos faltantes en el marco de una encuesta



Fuente: Elaboración propia.

Por último, si la base de datos contiene hogares que declararon ingresos nulos en todo el año, es posible que esos valores se consideren como faltantes, porque se da por sentado que la probabilidad de que un hogar no genere ningún tipo de ingreso durante todo un año es bastante baja. Además, los hogares con ingresos superiores a un determinado límite también pueden considerarse como valores atípicos y ser objeto de imputación.

2. Imputación del filtro

El siguiente paso, tras imputar con éxito las covariables determinantes del gasto, consiste precisamente en utilizarlas para imputar el gasto en bienes o servicios. Por lo general, en las encuestas de ingresos y gastos se pregunta si el hogar consumió o adquirió un determinado bien o servicio. En caso de respuesta afirmativa, se pregunta por la cantidad de dinero gastado en el bien o servicio y por la cantidad de artículos adquiridos en el período de referencia. En caso de respuesta negativa, se procede a preguntar por el siguiente bien o servicio. Por supuesto, los filtros de los diferentes artículos tienen diferentes tasas de respuesta. De aquí en adelante, el valor que se ha de imputar en esta etapa es dicotómico: sí o no. Si el valor imputado es "no", significa que el hogar no tiene ningún gasto relacionado con ese registro. Debido a la naturaleza del filtro, es conveniente aplicar un modelo de regresión logística para modelar la falta de respuesta en el filtro. De esta manera, la probabilidad de consumo (o compra) de un artículo i en particular es $p_k = Pr(Filtro_i = 1)$ y puede estimarse por medio del siguiente modelo de regresión logística:

$$\tilde{p_k} = logit^{-1}(x_k \hat{\beta}_i) = \frac{exp(x_k \hat{\beta}_i)}{1+exp(x_k \hat{\beta}_i)}$$

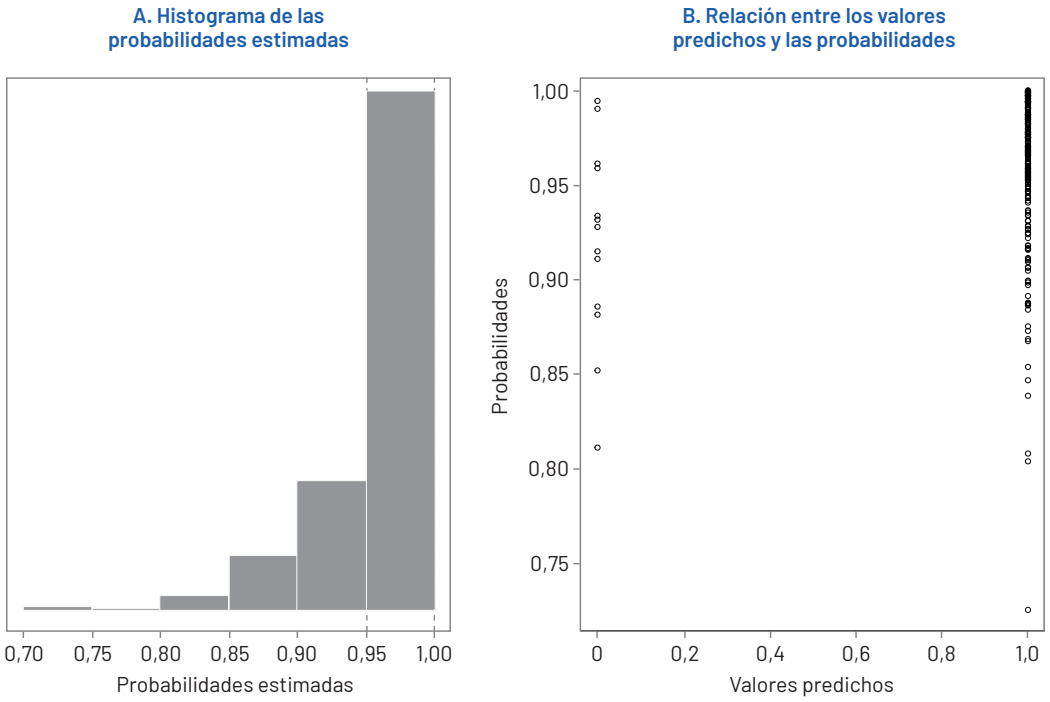
Las covariables incluidas en la matriz x podrían ser las mismas utilizadas para la imputación de los ingresos y, por supuesto, los ingresos en sí. Es decir, las covariables incluidas serían la composición del hogar, el estado de ocupación y la fuerza de trabajo de los miembros del hogar, la calidad de la vivienda, la ubicación y los ingresos del hogar. Al suponer que valores similares de p˜ producirán valores de filtro similares, se puede "pedir prestado" un valor de filtro para imputar el que falta de un vecino con un valor similar de $\tilde{p}$. Por lo tanto, el valor imputado del filtro para la unidad k está dado por $Filtro_k = Filtro_{l(k)}$, donde $l(k)$ es el elemento donante, determinado por la minimización de la distancia $D_{lk} = |\tilde{p}_k - p_l|$. Se destaca que el donante l del elemento k es el elemento del conjunto r_i que presenta el valor más pequeño de la distancia D_{lk}.

Por regla general, todos los donantes deben estar en la misma provincia que la unidad con el valor faltante. Por ejemplo, se considera la falta de respuesta de algunos hogares con respecto al filtro de compra del artículo "arroz". Dado que se trata de un artículo de consumo masivo en América Latina y el Caribe, se supone que la mayoría de los hogares responderá que efectivamente ha comprado arroz en el período de referencia. De esta manera, al utilizar la regresión logística como modelo para la falta de respuesta del filtro relativo al arroz, es posible que la distribución de las probabilidades estimadas de compra de arroz

esté sesgada hacia el valor 1 y alejada del valor 0, como se muestra en el gráfico XIII.2. Es evidente que la distribución de los valores imputados también debería estar sesgada hacia el valor 1, reflejando la realidad de la compra de un artículo esencial como el arroz.

■ Gráfico XIII.2

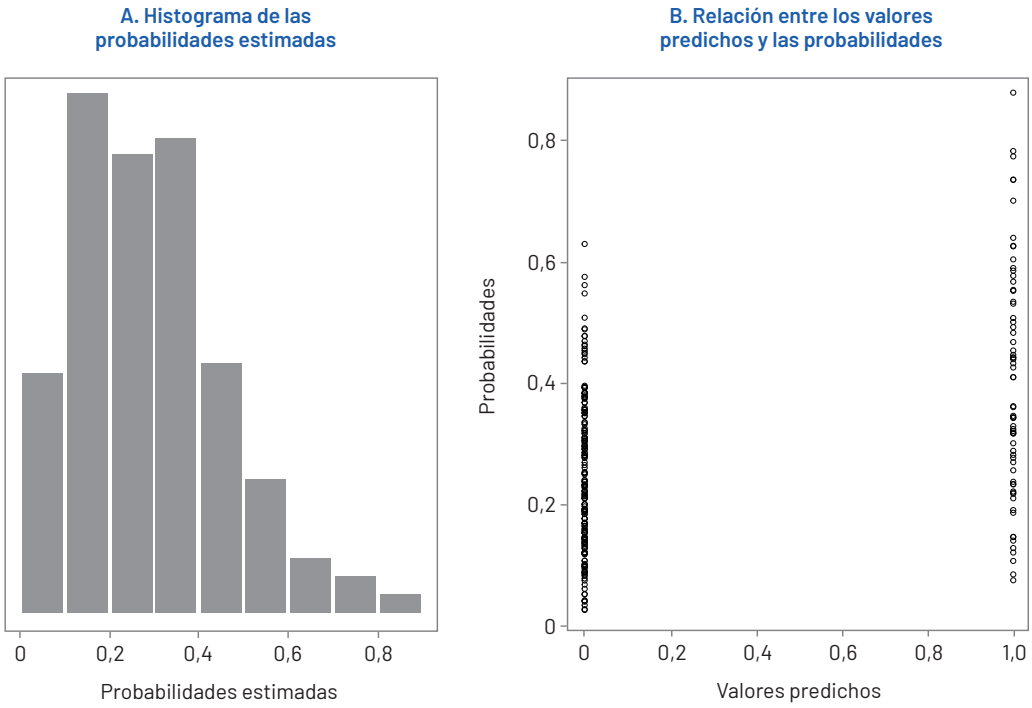
Distribución de las probabilidades estimadas de compra de arroz y valores imputados para los hogares con valores faltantes en el filtro en el marco de una encuesta



Fuente: Elaboración propia.

Por el contrario, el filtro para algunos artículos de bajo consumo estará más sesgado hacia el valor 0. En el gráfico XIII.3 se muestra la distribución de las probabilidades estimadas de compra de un artículo de bajo consumo como el salmón, así como los valores imputados.

■ Gráfico XIII.3
Distribución de las probabilidades estimadas de compra de un artículo de bajo consumo y valores imputados para los hogares con valores faltantes en el filtro en el marco de una encuesta



Fuente: Elaboración propia.

3. Imputación de los gastos

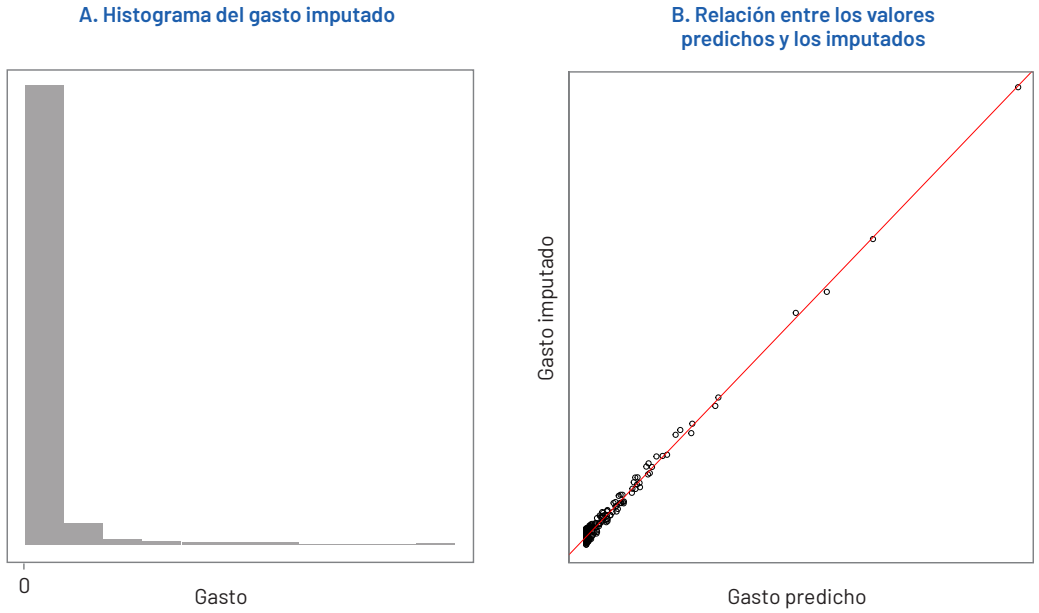
Este es el último paso del proceso de imputación y está fuertemente influenciado por los resultados de la imputación de la pregunta de filtro. En este paso, a los hogares con un valor imputado de filtro igual a 0 se imputará automáticamente un 0 como cantidad de dinero gastado en el artículo de referencia. Es decir, si el resultado de la imputación en el filtro es 0, se deduce directamente que el hogar no compró (o produjo) el artículo en el período de referencia y, por lo tanto, la frecuencia de compra, la cantidad de registros comprados y la cantidad de dinero gastado en ese artículo deben ser iguales a 0. El filtro de las unidades restantes debe tener un valor observado o imputado de 1 y, en consecuencia, se deben imputar los valores faltantes del gasto.

Cabe señalar que el grupo de donantes está restringido a las unidades con un valor de gasto distinto de 0 en el artículo específico. Es decir, en el caso de aquellas unidades con un valor de filtro distinto de 0, es necesario identificar un donante. Para la imputación de los gastos, la técnica del vecino más cercano con el método de regresión puede aplicarse de forma análoga a la imputación de los ingresos. Por lo tanto, se considera un modelo

lineal en el que las covariables incluidas en la matriz x son: la composición del hogar, el estado de ocupación y la fuerza de trabajo, la calidad de la vivienda, la ubicación del hogar y los ingresos.

A partir de los ejemplos anteriores, en el gráfico XIII.4 se muestra la distribución de los gastos imputados en el caso de un bien de bajo consumo como el salmón. Se observa que la cantidad de dinero gastado en este artículo es baja y que la relación entre los valores predichos del modelo y los valores imputados es fuertemente lineal.

■ Gráfico XIII.4
Distribución de los gastos imputados sobre el salmón y relación entre los valores predichos e imputados para los hogares con valores faltantes en el gasto en el marco de una encuesta



Fuente: Elaboración propia.

C. Consideraciones sobre la imputación múltiple

Antes de elegir un método en particular, conviene tener en cuenta los efectos de esta elección en las propiedades estadísticas de los estimadores de las encuestas de hogares. En cuanto a la imputación múltiple, las propiedades estadísticas de los estimadores deben modificarse en consecuencia. Subestimar la variación de las estimaciones puede ser un error muy grave, porque afecta la cobertura nominal de los intervalos de confianza y, a su vez, influye en las pruebas de hipótesis y en el cálculo de los valores p.

Por ejemplo, se supone que existe una muestra aleatoria s compuesta por un conjunto de n datos que relacionan dos variables X, Y, por medio del siguiente modelo de regresión simple:

$$y_k = \beta x_k + \varepsilon_k$$

Donde los errores tienen distribución normal con $E(\varepsilon_k)=0$ y $Var(\varepsilon_k)= \sigma^2$ para todo $k \in s$. Desde esta perspectiva, se supone que la variable dependiente de interés solo pudo observarse en un conjunto de personas de tamaño n_1, mientras que, en el caso de las n_0 personas restantes (es decir, $n_1 + n_0 = n$), no existen datos de la variable de interés. Además, se supone que fue posible observar los valores de la covariable X en el caso de todas las personas de la muestra.

El valor agregado de la imputación múltiple (Rubin, 1987) radica en la estimación de los errores estándar. Al no tener en cuenta la naturaleza estocástica de los valores imputados, se obtienen estimaciones de la varianza mucho menores. La idea consiste en generar $M > 1$ conjuntos de valores para los registros faltantes. Al final, el valor imputado corresponderá al promedio de esos M valores. Así, el modelo final de imputación (para los valores faltantes) toma la siguiente forma:

$$\dot{y}_i = \dot{\beta} x_{i(missing)} + \dot{\varepsilon}_i$$

En este caso, se consideran dos maneras de realizar la imputación: la primera se basa en la esperanza del modelo (sin imputación múltiple) y la segunda, en la adición del término de error del modelo (imputación múltiple).

- Ingenua: en este escenario, el valor imputado para el registro faltante toma la siguiente forma:

  $$\dot{y}_i = \hat{\beta} x_{i(missing)}$$

  Esta clase de imputación carece de aleatoriedad y, por lo tanto, la varianza de β resultará subestimada.

- Múltiple: en este caso, se tiene en cuenta el término de error en la generación de los valores imputados, de manera que:

  $$\dot{y}_i = \dot{\beta} x_{i(missing)} + \dot{\varepsilon}_i$$

La imputación múltiple puede realizarse con un enfoque frecuentista o bayesiano. Por ejemplo, es posible seleccionar M muestras de bootstrap y, para cada una, estimar los parámetros β y σ para generar $\dot{y}_i$. Al final se promedian los M valores y se imputa el valor faltante. Por otra parte, teniendo en cuenta el enfoque bayesiano, se definen las distribuciones posteriores de β y σ para generar M valores de estos parámetros y por tanto M valores de $\dot{y}_i$. De igual manera, al final se promedian los M valores y se imputa el valor faltante.

Por ejemplo, cuando se desea estimar un parámetro β, la esperanza estimada al utilizar la metodología de imputación múltiple está dada por:

$$E(\hat{\beta}|Y_{obs}) = E(E(\hat{\beta}|Y_{obs}, Y_{mis})|Y_{obs})$$

Esta expresión se estima mediante el promedio de las M estimaciones puntuales de $\hat{\beta}$ sobre las M imputaciones, dado por:

$$\bar{\hat{\beta}} = \frac{1}{M}\sum_{m=1}^{M} \hat{\beta}_m$$

Mientras tanto, la varianza estimada al utilizar la metodología de imputación múltiple está dada por la siguiente expresión:

$$E(\hat{\beta}|Y_{obs}) = E(V(\hat{\beta}|Y_{obs}, Y_{mis})|Y_{obs}) + V(E(\hat{\beta}|Y_{obs}, Y_{mis})|Y_{obs})$$

La primera parte de esta expresión se estima como el promedio de las varianzas muestrales de $\hat{\beta}$ sobre las M imputaciones, dado por:

$$\bar{U} = \frac{1}{M}\sum_{m=1}^{M} Var(\hat{\beta})$$

El segundo término se estima como la varianza muestral de las M estimaciones puntuales de $\hat{\beta}$ sobre las M imputaciones, dada por:

$$B = \frac{1}{M-1}\sum_{m=1}^{M} (\hat{\beta}_m - \bar{\hat{\beta}})^2$$

Vista la necesidad de tener en cuenta un factor de corrección (puesto que M es finito), la estimación del segundo término está dada por la siguiente expresión:

$$\left(1 + \frac{1}{M}\right) B$$

Por lo tanto, la varianza estimada es igual a:

$$\hat{V}(\hat{\beta}|Y_{obs}) = \bar{U} + \left(1 + \frac{1}{M}\right) B$$

Por ejemplo, si se quiere realizar mediciones de pobreza utilizando la imputación múltiple, en primer lugar es necesario establecer un modelo sobre los ingresos y_k y luego generar Q posibles valores y_k^q $(q=1,\ldots,Q)$ para cada persona que no respondió. A continuación, utilizando los Q conjuntos de datos completos, es necesario estimar las siguientes cantidades:

$$\hat{F}_\alpha^q = \frac{1}{N}\sum_{k\in s} w_k \left(\frac{l - y_k}{l}\right)^\alpha I(y_k < l) \; q = 1,\ldots,Q$$

El estimador final basado en la técnica de imputación múltiple será el promedio simple de las estimaciones anteriores, dado por:

$$\tilde{F}_\alpha = \frac{1}{Q}\sum_{q=1}^{Q} \hat{F}_\alpha^q$$

La varianza de esta metodología se puede dividir en dos componentes: la variación dentro de cada conjunto de datos creado y la variación entre cada estimación resultante. Por lo tanto, la varianza asociada a $\tilde{F}_\alpha$ es:

$$\hat{V}(\tilde{F}_\alpha) = \frac{1}{Q}\sum_{q=1}^{Q} \hat{V}(\hat{F}_\alpha^q) + \left(1 + \frac{1}{Q}\right)\frac{1}{Q\text{-}1}\sum_{q=1}^{Q} (\hat{F}_\alpha^q - \tilde{F}_\alpha)^2$$

Una vez obtenidos los conjuntos de datos completos, es posible estimar $\hat{V}(\hat{F}_\alpha^q)$ utilizando la técnica del último conglomerado junto con la técnica de jackknife. La característica principal del proceso de imputación es utilizar la información auxiliar para estimar con precisión los valores faltantes. De esta forma, las estimaciones poblacionales de los parámetros de interés tendrán un sesgo nulo o insignificante y la confiabilidad de la estrategia de muestreo se mantendrá como se planeó en una primera instancia. Con la siguiente simulación, se ejemplifican las ventajas de la imputación múltiple.

1. Simulación empírica

Para ejemplificar los escenarios anteriores, en esta sección se muestran los resultados empíricos de una simulación en la que se supuso un conjunto de $n=100$ datos con una pendiente $\beta=100$ y con una dispersión de $\sigma=2$. A su vez, el conjunto de datos presentará $n_0=40$ valores faltantes en la variable de respuesta. A continuación se muestran las primeras diez filas de esta base de datos simulados (véase el cuadro XIII.1). Las dos primeras columnas corresponden a los valores verdaderos de la covariable y la variable de interés, respectivamente. En la tercera columna, se señalan los valores faltantes y, en la cuarta, la información disponible en la base de datos de la muestra s.

■ Cuadro XIII.1
Ejemplo de un conjunto de datos con valores faltantes

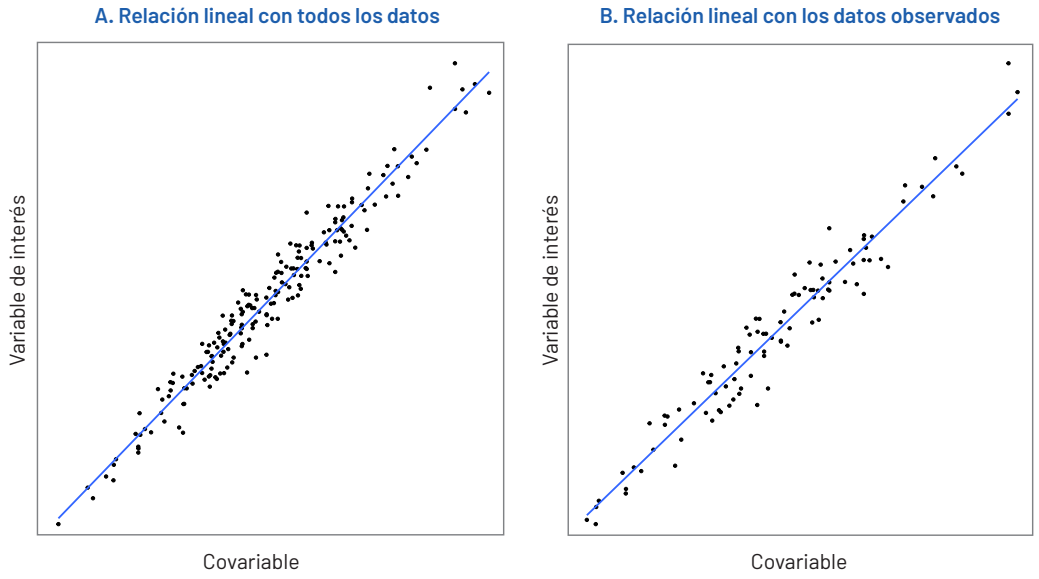
x	y	Faltantes	y (no imputado)
11	991	Sí	No disponible
12	1 282	Sí	No disponible
12	1 164	No	1 164
12	1 217	No	1 217
13	1 325	No	1 325
11	1 086	No	1 086

x	y	Faltantes	y (no imputado)
12	1 210	Sí	No disponible
13	1 272	Sí	No disponible
15	1 459	Sí	No disponible
11	1 182	No	1 182

Fuente: Elaboración propia.

Con el 40% de valores faltantes, es necesario realizar una imputación para obtener registros completos en la base de datos. En el gráfico XIII.5 se relacionan las variables con y sin valores faltantes para este ejemplo.

■ Gráfico XIII.5
Relación entre la variable de interés y la covariable en bases de datos completas y con datos faltantes



Fuente: Elaboración propia.

Al aplicar una imputación ingenua —por ejemplo, basada en un modelo de regresión simple—, se obtendría un conjunto de datos completo, ejemplificado en el cuadro XIII.2 (solo las primeras diez filas de la base de datos).

■ Cuadro XIII.2
Ejemplo de un conjunto de datos con valores imputados ingenuamente

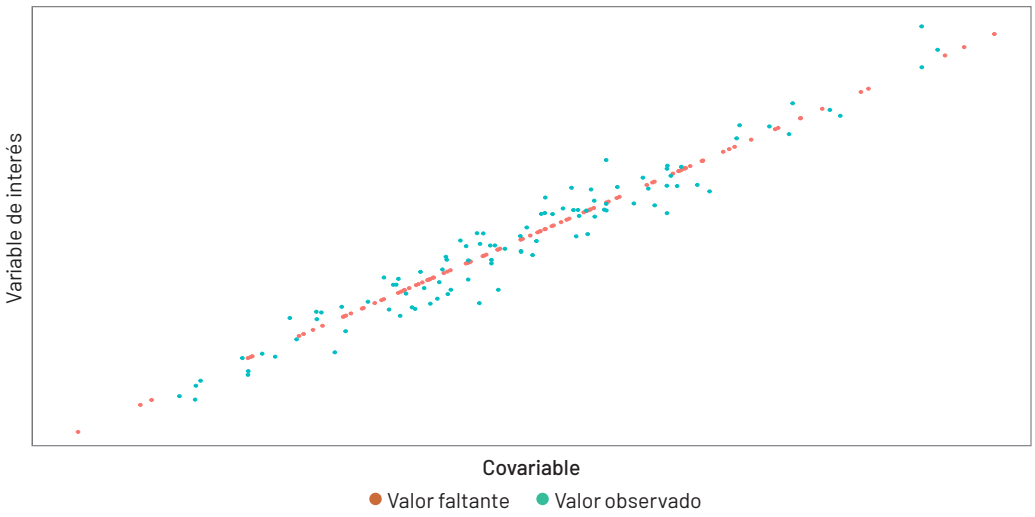
x	y (original)	Faltantes	y (imputado)
11	991	Sí	1 047
12	1 282	Sí	1 221
12	1 164	No	1 164

x	y (original)	Faltantes	y (imputado)
12	1 217	No	1 217
13	1 325	No	1 325
11	1 086	No	1 086
12	1 210	Sí	1 221
13	1 272	Sí	1 290
15	1 459	Sí	1 485
11	1 182	No	1 182

Fuente: Elaboración propia.

En general, el uso de un enfoque simple no afecta la estimación puntual del parámetro de interés, sino la estimación del error estándar, puesto que la variación natural de los datos se subestima considerablemente. Por ejemplo, en el gráfico XIII.6 se muestra que, con la imputación simple, todos los valores faltantes imputados están sobre la línea de regresión.

■ Gráfico XIII.6
Relación de la variable de interés con la covariable auxiliar para el enfoque de imputación ingenua



Fuente: Elaboración propia.

Si se considera la imputación múltiple con la técnica de bootstrap, también se obtendrá un conjunto de datos aumentado por cada una de las M realizaciones que se ejecuten. Por ejemplo, en el cuadro XIII.3 se ilustra el conjunto de datos obtenido con $M=3$ realizaciones. Cabe señalar que los valores de la variable de interés cambian en cada realización. Estas realizaciones se conocen en la literatura como "valores plausibles".

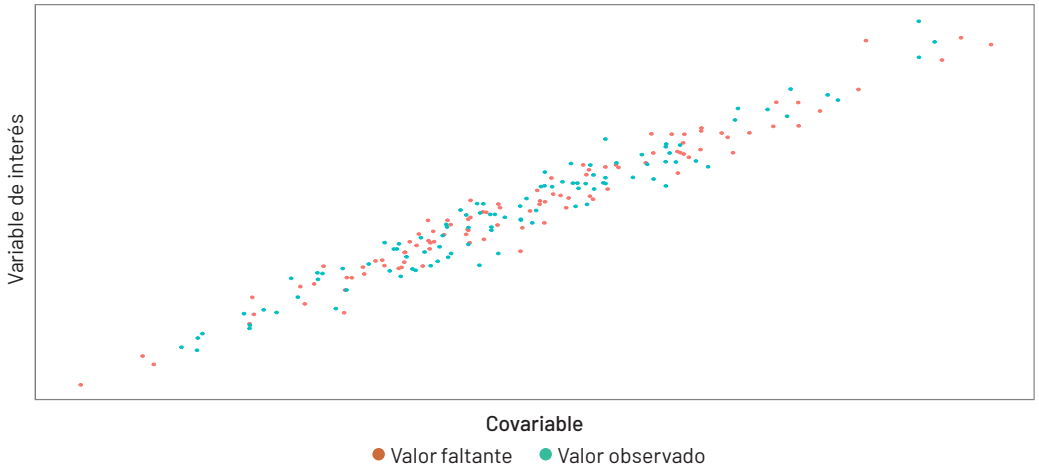
■ Cuadro XIII.3
Ejemplo de un conjunto de datos con tres valores imputados

x	y (original)	Faltantes	y1 (imputado)	y2 (imputado)	y3 (imputado)
11	991	Sí	1 047	950	1 040
12	1 282	Sí	1 221	1 254	1 198
12	1 164	No	1 164	1 164	1 164
12	1 217	No	1 217	1 217	1 217
13	1 325	No	1 325	1 325	1 325
11	1 086	No	1 086	1 086	1 086
12	1 210	Sí	1 252	1 199	1 198
13	1 272	Sí	1 304	1 302	1 292
15	1 459	Sí	1 485	1 493	1 478
11	1 182	No	1 182	1 182	1 182

Fuente: Elaboración propia.

Los valores imputados presentan una buena dispersión. En el gráfico XIII.7 se observa que este enfoque es mucho más realista, al considerar la variación natural del fenómeno de interés en los valores imputados.

■ Gráfico XIII.7
Relación de la variable de interés con la covariable auxiliar para el enfoque de imputación múltiple con la técnica de bootstrap



Fuente: Elaboración propia.

Por otra parte, en caso de distribuciones previas no informativas, es bien sabido que la distribución posterior de σ^2 es:

$$\sigma^2 \mid y,x \sim \frac{\sum_{i=1}^{n_1} (y_i - \hat{\beta}x_i)^2}{\chi^2_{n_1 - 1}}$$

con $\hat{\beta} = \frac{\sum_{i=1}^{n_1} x_i y_i}{\sum_{i=1}^{n_1} x_i^2}$. Asimismo, la distribución posterior de β es:

$$\beta|\sigma^2 y,x \sim Normal\left(\hat{\beta}, \frac{\sigma^2}{\sum_{i=1}^{n_1} x_i^2}\right)$$

Al suponer el enfoque bayesiano de imputación múltiple anterior, se llega a resultados similares. En ambos casos se observa una buena dispersión de los valores imputados, que respeta la distribución natural de la característica de interés. En resumen, a partir de esta simulación de Montecarlo, se concluye rápidamente que imputar de manera determinista puede llevar a la subestimación de la dispersión de la variable de interés. En el cuadro XIII.4 se muestra que los tres métodos de imputación arrojaron estimaciones puntuales insesgadas. Sin embargo, el error estándar de la estimación simple se subestima considerablemente. De esta forma, la amplitud de los intervalos de confianza al 95% que surgen de la estimación simple es inferior con respecto a los otros dos métodos y determina una cobertura deficiente, pues el nivel nominal de este método en realidad no es del 95%, sino del 83%.

■ Cuadro XIII.4
Comparación de tres enfoques de imputación

Propiedades	Enfoque ingenuo	Técnica de bootstrap	Enfoque bayesiano
Esperanza	100,00	100,01	100,01
Error estándar	0,24	0,41	0,42
Amplitud	0,96	1,60	1,66
Cobertura	0,83	0,97	0,95

Fuente: Elaboración propia.

Capítulo XIV
Detección de valores atípicos

En este capítulo se describen los aspectos teóricos y prácticos de la detección de valores atípicos en una base de datos completa (incluso con registros que ya han sido imputados), teniendo en cuenta los métodos que han demostrado ser eficaces para realizar inferencias en el análisis de la información recolectada a partir de las encuestas de hogares. Tras una breve introducción, se presenta un enfoque no exhaustivo del problema de la detección de valores atípicos, así como la teoría en que se basan los métodos y algunos hallazgos empíricos acerca de la imputación de estos valores.

Después de detectar los posibles valores atípicos, el investigador debe decidir qué hacer con ellos. En general, hay tres posibles soluciones: i) mantenerlos en la base de datos final (sin ningún cambio), ii) corregirlos (verificar exhaustivamente en los registros y cuestionarios, encontrar el error en la captura y reemplazarlo por el valor verdadero) o iii) imputarlos (eliminarlos y reemplazarlos por un valor adecuado que no fue proporcionado por el respondiente). El enfoque de imputación de valores atípicos sigue los mismos principios que los métodos utilizados para imputar registros en el capítulo anterior. Al final, se recomienda que, cuando se encuentren valores atípicos, se marquen para su revisión. Cuando se revisan, es posible encontrar que el valor es simplemente erróneo debido a algún problema en la recolección de datos (error de medición). También es posible que el valor sea improbable y raro, pero válido. En el primer caso, el error se corrige (si el valor atípico es erróneo) y las estimaciones se ajustan. Si el valor atípico corresponde a un dato erróneo y no se puede localizar al encuestado, se recomienda imputarlo.

Por lo tanto, el investigador está interesado en detectar esos valores y en solucionar el problema reemplazando los inverosímiles por otros más realistas. No cabe duda de que es un desafío detectar valores atípicos y distinguir aquellos que son errores de los que son

inusualmente altos (o bajos), pero correctos. Hacer estas correcciones en los microdatos (es decir, en los datos a nivel del hogar) constituye un desafío añadido. En general, un valor atípico es una observación que está distante de todas las demás observaciones o datos en la variable de interés de la base de datos.

Así como deben corregirse, eliminarse o imputarse los valores erróneos, también se deben mantener los valores improbables en el conjunto de datos, y ha de tomarse una decisión para reducir su impacto en el análisis de la encuesta. Hay que tener en cuenta que existen valores verídicos en las observaciones que, aunque tienen una incidencia baja, deben conservarse en el análisis. Los valores que se apartan de la distribución habitual pueden clasificarse como atípicos o como puntos influyentes. El tratamiento de estos valores para el análisis vendrá definido por su clasificación.

- Valores atípicos representativos: valores que se han registrado correctamente y representan otras unidades de la población con valores similares.
- Valores atípicos no representativos: valores que se han registrado incorrectamente o son únicos, lo que significa que no hay otra unidad en la muestra que se pueda representar con esta información.
- Puntos de influencia: cuando el efecto conjunto del punto de datos atípicos y su respectivo peso muestral tiene un impacto significativo en la inferencia.

A menudo, los valores atípicos pueden ser representativos de otros en la población, por lo que siguen siendo importantes y deben permanecer en el conjunto de datos. Al final, en el proceso de detección de valores atípicos, se trata de hallar un término medio entre el sesgo y la varianza. Los valores atípicos pueden tener un gran impacto en los estimadores de ubicación y escala, como la media y la varianza, así como en los estimadores de totales y tamaños de subpoblaciones. Aunque estos estimadores permanecen insesgados, su varianza crece en presencia de valores atípicos.

A. Algunos métodos de detección de valores extremos

Filzmoser, Gussenbauer y Templ (2016) afirman que se pueden cometer errores en el proceso de entrada de datos. Por ejemplo, podrían introducirse valores de gasto imposibles, es decir, demasiado altos o demasiado bajos para ser posibles. Estos valores extremos pueden tener un impacto significativo en algunos análisis particulares (como en el estudio de indicadores de desigualdad o el ajuste de modelos de regresión), que pueden verse significativamente afectados por un número reducido de valores influyentes en el conjunto de datos. En esta sección se realizará un recorrido no exhaustivo de algunos métodos para detectar valores extremos.

1. Método descendente (*top-down*)

Supóngase que $y_{(1)} \leq \cdots \leq y_{(n)}$ denota los valores ordenados de la variable de interés y en la muestra s. Si se considera el total de la variable de interés para todos los elementos en la muestra, se define el porcentaje de contribución acumulado P_j de la siguiente manera:

$$P_j = 100 \times \frac{\sum_{i=j}^{n} y_{(i)}}{\sum_{k=1}^{n} y_k} \; ; j = 1, \ldots, n$$

Grandes cambios entre los valores de P_j entrañan posibles valores atípicos. También es posible calcular esta medida si se incluye el peso de muestreo para localizar qué valores ponderados tienen efectos anormalmente grandes.

$$P_j^* = 100 \times \frac{\sum_{i=j}^{n} d_i y_{(i)}}{\hat{t}_{y,\pi}} \; ; j = 1, \ldots, n$$

2. Método de diagramas de caja (*box-plot*)

Uno de los métodos más básicos para detectar valores atípicos es construir un diagrama de caja (*box-plot*) utilizando la mediana y el rango intercuartílico *(RIC)* de la variable de interés. En primer lugar, se define su $RIC = Q_3 - Q_1$ y su mediana como $m = Q_2$. Por consiguiente, un elemento se marcará como valor atípico si cae fuera del siguiente intervalo:

$$(m - c \times RIC, \; m + c \times RIC)$$

Donde c es una constante predeterminada por el investigador, que suele fijarse entre 1,5 y 3.

3. Transformación de Box-Cox

Si la distribución de la variable es sesgada (como suelen serlo los ingresos y gastos), es útil transformarla para lograr una distribución simétrica antes de determinar los posibles valores atípicos. La transformación de Box-Cox tiene la siguiente forma:

$$y(\lambda) = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} & \text{si } \lambda \neq 0 \\ log(y) & \text{si } \lambda = 0 \end{cases}$$

Donde $\lambda \in (-5,5)$. De esta forma, una computadora iterará entre cada posible valor de λ hasta encontrar el que mejor reproduzca una distribución normal. Con esta nueva distribución, se puede utilizar el criterio de decisión basado en el método de diagramas de caja.

4. Método de distancia estandarizada

La transformación explicada en la sección anterior solo funciona para valores positivos. El método que se expone a continuación es otra forma de transformar y estandarizar los datos. Supóngase que $z_k = w_k y_k$. Si m_z es una estimación para la ubicación de z y σ_z es una estimación para la escala de z, la distancia estandarizada puede definirse como:

$$\delta_{z_k} = \frac{z_k - m_z}{\sigma_z}$$

De forma similar al método de diagramas de caja, los registros se clasificarán como valores atípicos si el valor absoluto de δ_{z_k} es mayor que un umbral predeterminado (normalmente 3). La media y la varianza de la muestra se pueden utilizar para las estimaciones de ubicación y escala para z_k. Sin embargo, no son robustas, ya que incluirán los valores atípicos potenciales, lo que a su vez reduce la probabilidad de que se detecten correctamente los registros atípicos. Por consiguiente, es posible utilizar estimadores robustos (resistentes a valores atípicos) para m_z y σ_z, como la mediana y el rango intercuartílico de z_k, respectivamente.

5. Método de Hidiroglou-Berthelot

Es posible utilizar una distancia estandarizada para detectar si la relación entre dos variables x e y en una unidad de la muestra difiere estructuralmente de las otras unidades. En este método se utiliza la idea de distancia estandarizada y también se incorpora una medida de importancia para el tamaño de la unidad, con el fin de determinar el umbral para considerar un registro como un valor atípico. El algoritmo de detección sigue los siguientes pasos:

i) Para cada elemento, calcular $r_k = y_k / x_k$ para $k \in s$.

ii) Transformar los datos para poder detectar valores atípicos en cualquier extremo de la distribución. Los datos transformados están dados por:

$$s_k = \begin{cases} 1 - \dfrac{med(r_k)}{r_k} & \text{si } 0 \leq r_k \leq med(r_k) \\ \dfrac{med(r_k)}{r_k} - 1 & \text{en caso contrario} \end{cases}$$

Donde $med(r_k)$ corresponde a la mediana de los cocientes definidos en el paso anterior.

iii) Incorporar la magnitud de los datos calculando los efectos E_k dados por:

$$E_k = s_k \left(max(x_k, y_k)\right)^{\phi}$$

El parámetro ϕ proporciona una medida de control para el impacto del tamaño en el efecto.

iv) A continuación, calcular el primer, segundo y tercer cuartil de los efectos dados por E_{Q_1}, E_{Q_2}, E_{Q_3}, respectivamente.

v) Los rangos intercuartílicos se calculan entonces como:

$$d_{Q_1} = max\,(E_{Q_2} - E_{Q_1}, |\,0{,}5 \times E_{Q_2}|\,)$$
$$d_{Q_3} = max\,(E_{Q_3} - E_{Q_2}, |\,0{,}5 \times E_{Q_2}|\,)$$

Nótese que la cantidad $|0{,}5 \times E_{Q_2}|$ se utiliza para reducir la tendencia a declarar falsos valores atípicos. Por ejemplo, esto ayudaría si la mayoría de los valores estuvieran agrupados alrededor de un valor particular, con unos pocos registros que se desvíen de ese valor. Por último, los registros se declaran como valores atípicos si el valor de su efecto E_k queda fuera del intervalo $(E_{Q_2} - c \times d_{Q_1},\ E_{Q_2} + c \times d_{Q_3})$, donde, al igual que en el método de diagramas de caja, c controla el ancho de la región de aceptación.

6. Método de la distancia de Mahalanobis

En este método se tiene en cuenta la estructura multidimensional de los datos observados en todos los registros comunes de un mismo módulo; por ejemplo, en el módulo de ingresos del hogar en una encuesta de hogares, o en el módulo de gastos de una encuesta de presupuestos familiares. En primer lugar, se supone que $y_k = (y_{k1}, y_{k2}, \ldots, y_{kQ})'$ define el vector de valores observados del individuo k en todas las Q variables del módulo de interés. Por lo tanto, la distancia de Mahalanobis para una unidad se puede definir como:

$$MD_k^2 = (\boldsymbol{y}_k - \bar{\boldsymbol{y}})' \boldsymbol{S}^{-1} (\boldsymbol{y}_k - \bar{\boldsymbol{y}})$$

Donde $\bar{\boldsymbol{y}}$ y $\boldsymbol{S}$ son, respectivamente, el vector de medias muestrales y la matriz de covarianzas de las Q variables en el módulo. Si los datos siguen una distribución normal multivariante, se puede demostrar que la distribución de esta distancia es ji al cuadrado (χ^2) con Q grados de libertad, $MD_k^2 \sim \chi_Q^2$. A continuación, las unidades se declaran como potencialmente atípicas si superan el umbral del percentil 0,95 de la distribución χ_Q^2.

7. La distancia de Cook

Los registros influyentes son valores atípicos que afectan significativamente a los modelos de regresión. Para ubicarlos, es posible utilizar la distancia de Cook, que mide cuánto impacta la unidad i-ésima en la estimación de la unidad j-ésima, en un modelo de regresión con p variables explicativas. Esta medida está dada por la siguiente expresión:

$$DC_{(i)} = \frac{\sum_{\substack{j=1 \\ j \neq i}}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{(p+1)\hat{\sigma}^2}$$

Donde $\hat{\sigma}^2=\frac{\sum_k(\varepsilon_k-\bar{e})^2}{n-p}$ es la varianza de los residuales del modelo (ε_k); además, $\hat{y}_j$ es la estimación de la j-ésima unidad en el modelo de regresión ajustado con todos los datos observados, mientras que $y_{j(i)}$ es la estimación de la j-ésima unidad cuando se excluye la i-ésima unidad. Cuanto más grande sea el valor de la estadística, más probable es que la observación de la unidad i se considere un valor influyente. Algunos autores afirman que cualquier valor $DC_{(i)}$ mayor que uno debe considerarse influyente, pero otros afirman que el umbral debe ser $4/n$ o $4/(n-p-1)$.

8. El criterio DFBETAS

Por otra parte, el estadístico DFBETAS mide cuánto influye la observación i-ésima en los estimadores de los coeficientes de regresión en un modelo lineal. La estadística se puede escribir como sigue:

$$DFBETAS_{j(i)}=\frac{b_j - b_{j(i)}}{\sqrt{S^2_{(i)}C_{jj}}}$$

Donde b_j es la estimación para el j-ésimo coeficiente de regresión y $b_{j(i)}$ es la estimación calculada sin la observación i-ésima. Además, $S^2_{(i)}$ es la varianza muestral de la variable de interés sin la observación i-ésima y C_{jj} es el j-ésimo elemento de la diagonal de la matriz $\mathbf{(x'x)}$, de dimensión $n\times n$. Se puede considerar que cualquier cifra cuyo valor absoluto sea mayor o igual que $2/\sqrt{n}$ determina que el valor atípico sea influyente.

B. Ejemplo de detección de valores atípicos en una encuesta de presupuestos familiares y gastos

En general, en cualquier tipo de encuesta, podría considerarse la estructura multivariante de los datos (relación con otras variables) en la búsqueda de valores atípicos. Para ello se debería elegir un conjunto de covariables, de acuerdo con el juicio de los expertos y la desagregación necesaria. Aunque en la base de datos existan registros y valores para una determinada observación, es posible que, después de la detección de valores atípicos, toda la información de la unidad sea declarada sospechosa. Por lo tanto, una vez que se ha detectado una unidad con valores atípicos potenciales, podría decidirse eliminar todos sus registros (creando deliberadamente una falta de respuesta por unidad) si no es posible comprobar la fiabilidad de la información. Por consiguiente, si se mantiene la unidad, toda su información se declara confiable y vale la pena analizarla. De lo contrario, el registro se eliminará de los datos de la muestra que afectan la estructura del modelo de ponderación y la unidad se declarará como una unidad elegible no respondiente (ENR) (véase el capítulo IX).

Después de decidir acerca de los valores atípicos de la unidad, es necesario detectar los valores atípicos de los registros para las variables específicas de la encuesta. Por ejemplo, en una encuesta de presupuestos familiares, las variables de interés serán los rubros asociados a los ingresos y a los gastos del hogar. En este caso, se sugiere que la variable de interés de una categoría particular se transforme utilizando el enfoque de Box-Cox, dado que las distribuciones de ingresos y gastos siempre están sesgadas. Una vez transformadas, es posible utilizar las medidas antes mencionadas para decidir acerca de la eliminación del registro. Por ejemplo, si con dos o tres de los métodos se detecta un registro como posible valor atípico, entonces se verifica la información sobre ese registro. Si la información del elemento es sospechosa, se debe eliminar y utilizar un enfoque de imputación sobre ese registro.

Siguiendo con el modelo de la encuesta de presupuestos familiares, a nivel de gasto podría ser conveniente que la detección de los valores atípicos se realizara no sobre cada artículo, sino de manera agregada para cada nivel de la clasificación de consumo (Clasificación del Consumo Individual por Finalidades (CCIF)). Además, se debe tener en cuenta que, en este tipo de encuestas, los valores nulos en el gasto o consumo de artículos particulares son frecuentes, ya que no se puede esperar que todos los hogares consuman todos los artículos posibles. Estos valores nulos se denominan ceros estructurales. Resulta pertinente decidir si en la metodología de detección de valores atípicos se tendrán o no en cuenta los ceros. En general, cuando la incidencia de los ceros es baja, no debería existir ningún inconveniente en analizar el conjunto de datos incluyendo estos ceros. Por lo tanto, esta decisión debería ser independiente para cada división.

En algunos componentes concretos, el número de ceros puede ser bastante alto y los algoritmos de detección de valores atípicos pueden fallar si ese número supera determinado umbral. Cuando se aplican métodos de detección, las observaciones podrían incluso convertirse en valores atípicos debido a los ceros. En este sentido, es necesario contemplar umbrales flexibles para cada división. Por ejemplo, un valor de cero en el gasto en alimentos sería poco realista, pero podría ser adecuado para el gasto en ropa de bebé o muebles. Por ello, es plausible recomendar la agregación de los componentes del consumo en grandes categorías a nivel de producto o servicio o grupos agregados de productos y servicios.

Un indicador fiable, no solo para medir la desigualdad en el consumo, sino también para realizar un seguimiento de los cambios en el proceso de detección de valores atípicos y la imputación posterior, es el coeficiente de Gini. Por ejemplo, en el cuadro XIV.1 se muestra la presencia de ceros en cada división de la CCIF, junto con el coeficiente de Gini, respecto de algunas de las categorías anteriormente mencionadas. Se observa que la incidencia de ceros es mucho menor en la categoría de vivienda que en las de educación o recreación.

Esta metodología también se puede aplicar a grupos. En el cuadro XIV.2 se muestra la presencia de ceros estructurales respecto de algunos artículos de la sección de alimentos. Una vez más, dependiendo del país, es concebible encontrar una mayor incidencia de ceros en algunos artículos. En este caso particular, hay una mayor cantidad de ceros en artículos como el té o el café que en otros como los cereales, el azúcar o la leche.

■ Cuadro XIV.1
Conteo de ceros y estimación del coeficiente de Gini en algunas categorías de consumo

Categoría	Ceros	Coeficiente de Gini
Alimentos	27	36
Alcohol	4 333	90
Ropa	2 558	78
Vivienda	5	48
Muebles	85	53
Salud	2 746	78
Transporte	616	69
Tecnologías de la información y las comunicaciones	551	62
Recreación	3 538	92
Educación	4 802	90
Restaurantes	1 421	65
Seguros	4 837	90
Cuidado personal	129	51

Fuente: Elaboración propia.

■ Cuadro XIV.2
Conteo de ceros y estimación del coeficiente de Gini en algunos artículos de la categoría de alimentos

Artículos	Ceros	Coeficiente de Gini
Cereales	87	37
Carne	481	47
Pescado	305	56
Leche	290	47
Aceites	482	51
Frutas	981	67
Verduras	188	47
Azúcares	290	56
Comida procesada	253	43
Jugos	3 650	79
Café	5 286	86
Té	4 709	86
Cacao	4 421	78
Agua	5 287	86
Refrescos	4 859	86
Otras bebidas	3 353	79

Fuente: Elaboración propia.

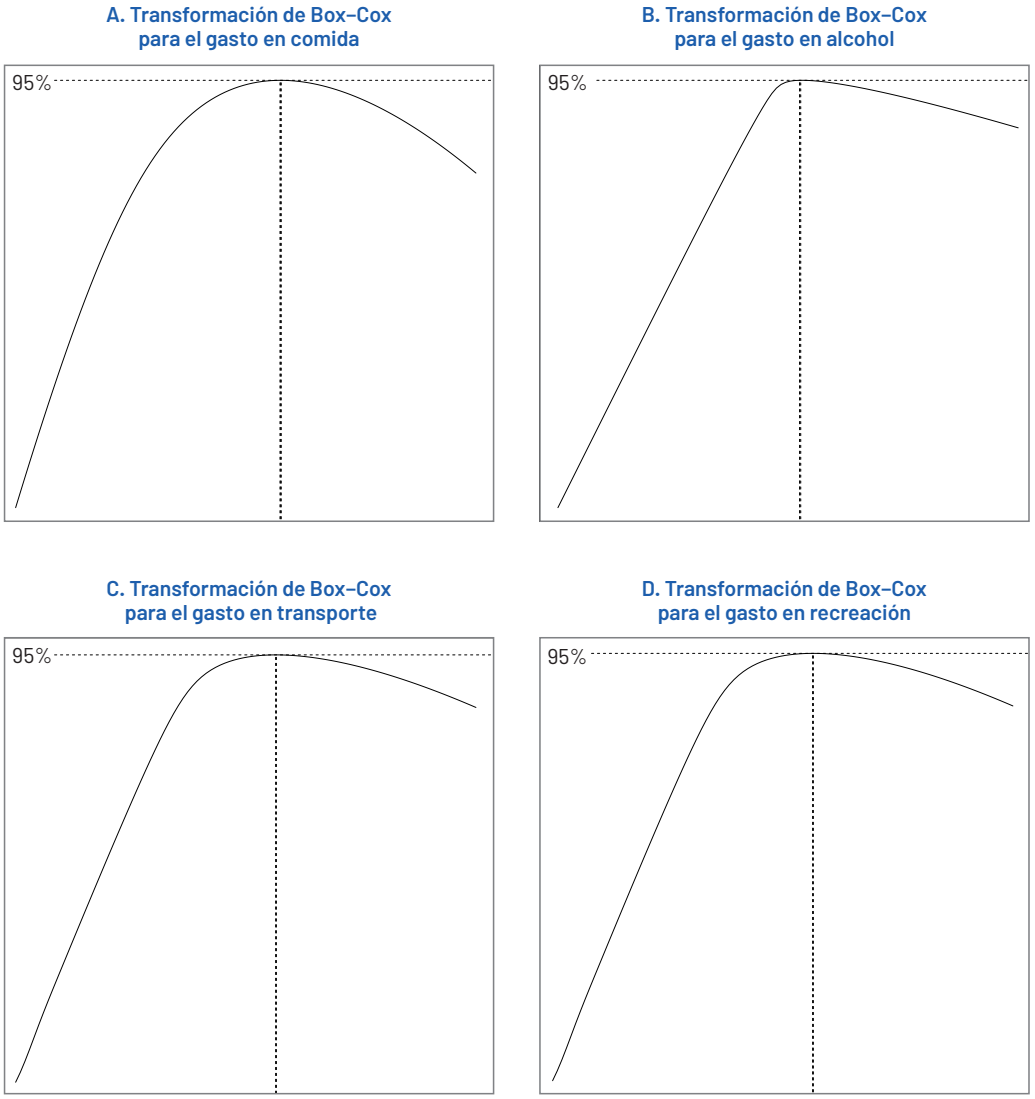
Como se mencionó anteriormente, cuando hay datos muy sesgados, los métodos para la detección de valores atípicos podrían resultar problemáticos, ya que el intervalo en que los puntos de datos no se consideran valores atípicos es simétrico alrededor de la mediana. Por ejemplo, en el gráfico XIV.1 se muestra el comportamiento estructural de algunas divisiones. Es notable que todas las distribuciones de gasto y consumo en estos conceptos están extremadamente sesgadas.

■ Gráfico XIV.1
Distribución del consumo en algunas categorías de gasto

Fuente: Elaboración propia.

Para hacer frente a este problema, es posible utilizar la transformación de Box-Cox con el fin de obtener una distribución simétrica de los datos antes de determinar los posibles valores atípicos. En el gráfico XIV.1 se muestra el proceso de iteración de esta metodología en algunas divisiones. La línea vertical en cada subgráfico corresponde al mejor valor que podría tomar λ para que los datos se ajusten a una distribución normal.

■ Gráfico XIV.2
Valores óptimos de las transformaciones de Box-Cox en algunas categorías de gasto



Fuente: Elaboración propia.
Nota: La línea vertical en cada subgráfico corresponde al mejor valor que podría tomar λ para que los datos se ajusten a una distribución normal.

Después de haber transformado apropiadamente los datos, es posible utilizar el método de diagramas de caja, uno de los más básicos —aunque muy poderoso— para detectar valores atípicos. Como se mencionó en la sección anterior, el gráfico mostrará el mínimo de la muestra, el primer cuartil, la mediana, el tercer cuartil y el máximo. La caja va del primer al tercer cuartil (que contiene por definición el 50% de los datos más internos), y la mediana suele estar marcada por una línea media. Para la aplicación específica de la detección de valores atípicos dentro de las divisiones de la CCIF, es posible que la constante predeterminada c varíe entre una división y otra. Por ejemplo, en el cuadro XIV.3 se muestra el número de valores atípicos detectados en cada división por este método.

■ Cuadro XIV.3

Conteo de valores atípicos en algunas categorías de consumo usando el método de diagramas de caja (*box-plot*)

División	Valores atípicos
Alimentos	222
Alcohol	0
Ropa	0
Vivienda	87
Muebles	330
Salud	0
Transporte	743
Tecnologías de la información y las comunicaciones	668
Recreación	0
Educación	0
Restaurantes	31
Seguros	0
Cuidado personal	400

Fuente: Elaboración propia.

Por otro lado, también es posible tener en cuenta la relación entre el gasto en cada división y el ingreso informado por el hogar en la encuesta. En general, no se puede suponer que esta relación es homogénea entre todos los encuestados, sobre todo si se tiene en cuenta que la selección de las unidades muestrales se hace en todos los grupos socioeconómicos del país. Sin embargo, sí es posible hacer este supuesto dentro de clases homogéneas, como en el cruce entre los quintiles (o deciles) del ingreso y las regiones del país. De esta forma, dentro de cada grupo se supondría que la relación entre el gasto y el ingreso es uniforme. Por ejemplo, en el cuadro XIV.4 se muestra el número de valores atípicos detectados en cada división mediante el método de Hidiroglou-Berthelot.

■ Cuadro XIV.4
Conteo de valores atípicos en algunas categorías de consumo con el método de Hidiroglou-Berthelot

División	Valores atípicos
Alimentos	74
Alcohol	141
Ropa	73
Vivienda	53
Muebles	177
Salud	71
Transporte	89
Tecnologías de la información y las comunicaciones	128
Recreación	168
Educación	117
Restaurantes	48
Seguros	100
Cuidado personal	247

Fuente: Elaboración propia.

Es necesario tener en cuenta que, como la lógica en que se basan estos dos métodos difiere, cada uno detectará un número distinto de valores atípicos. Esto representa una ventaja, porque los métodos son complementarios. Por ejemplo, en divisiones como ropa, vivienda, salud, recreación y educación, donde no se encontró ningún valor atípico posible con el método de diagramas de caja, sí se encontraron con el método de Hidiroglou-Berthelot. Si se tienen en cuenta los resultados de estos dos métodos, se puede especificar una regla lógica para asignar una marca a los registros de la base de datos que deban ser revisados por considerarse sospechosos. Por ejemplo, tal vez haya categorías en que la regla lógica sea una conjunción de los resultados de los métodos, mientras que podría haber otras en las que la regla lógica sea una disyunción entre los resultados.

Al final, se debe imputar cualquier valor que se considere atípico. Como se vio en los capítulos anteriores, la imputación puede apoyarse en un enfoque basado en modelos. Por ejemplo, para imputar el gasto per cápita anualizado, es posible utilizar el método de regresión con el vecino más cercano, donde se define un modelo lineal para las unidades

encuestadas (sin incluir los valores atípicos). Una vez estimados los coeficientes de regresión, se calcula un valor previsto para esas unidades de valores atípicos y se identifica a un solo donante como el hogar cuyo gasto total en esa división está más cerca de la predicción. En el cuadro XIV.5 se presentan algunos resúmenes de la distribución de los gastos a nivel de división antes de imputar los valores atípicos.

■ Cuadro XIV.5
Estadísticas descriptivas de algunas categorías de consumo antes de la imputación de los valores atípicos

División	Mínimo	Mediana	Máximo
Alimentos	0	164 587	1 819 370
Alcohol	0	0	5 475 960
Ropa	0	8 180	3 474 000
Vivienda	0	89 040	1 835 500
Muebles	0	10 503	3 871 476
Salud	0	2 400	735 180
Transporte	0	24 700	9 038 783
Tecnologías de la información y las comunicaciones	0	16 125	1 642 500
Recreación	0	860	6 307 200
Educación	0	0	1 800 000
Restaurantes	0	22 100	2 184 000
Seguros	0	0	2 932 400
Cuidado personal	0	18 540	1 223 734

Fuente: Elaboración propia.

Por último, en el cuadro XIV.6 se presentan algunos resúmenes de la distribución de los gastos a nivel de división después de imputar los valores atípicos de forma diferencial y con independencia en cada división. Se puede apreciar cómo la imputación realmente cambia la perspectiva del consumo mínimo y máximo. Esto significa que la detección de valores atípicos se centró en ambos lados de la distribución del gasto.

■ Cuadro XIV.6
Estadísticas descriptivas de algunas categorías de consumo después de la imputación de los valores atípicos

División	Mínimo	Mediana	Máximo
Alimentos	4 560	166 280	1 616 504
Alcohol	0	0	930 400
Ropa	0	8 250	1 446 532
Vivienda	3 000	89 280	726 000
Muebles	243	10 730	279 382
Salud	0	2 400	654 000
Transporte	60	29 900	365 000
Tecnologías de la información y las comunicaciones	0	16 627	591 000
Recreación	0	867	1 054 100
Educación	0	0	494 000
Restaurantes	0	22 133	520 000
Seguros	0	0	723 000
Cuidado personal	400	18 855	658 960

Fuente: Elaboración propia.

Capítulo XV
Agregación de encuestas

Para producir indicadores sociales de forma agregada (anual, semestral o trimestral), es común recurrir a la agregación de las bases de datos provenientes de las encuestas de hogares, cuya periodicidad suele ser mensual o trimestral. En esta sección, se exploran algunas estrategias de estimación ligadas al tratamiento de los pesos determinados por el diseño de muestreo complejo y al tratamiento de las unidades que se repiten en algún período debido al carácter rotativo de la medición.

Uno de los primeros acercamientos al problema de la estimación conjunta de indicadores sociales a partir de varios períodos de recopilación se presenta en Gurney y Daly (1965), donde se examina cómo mejorar el estimador puntual por medio de la correlación natural con períodos anteriores, siguiendo un enfoque inferencial basado en modelos estocásticos. En este orden de ideas, Lent, Miller y Duff (1999) definen una aproximación a un estimador para las distintas clasificaciones de la fuerza de trabajo, que se basa en la optimización de los coeficientes de un estimador compuesto.

Por su parte, Fuller (1990) examina los sesgos que se pueden generar en el análisis de encuestas repetidas debido a errores de medición y analiza en detalle algunos modelos estimados con mínimos cuadrados. Además, Bell (2001) describe varios acercamientos al problema de la estimación de indicadores sociales específicamente relacionados con la fuerza de trabajo, provenientes de encuestas de hogares que tienen definido un diseño de rotación y traslape entre distintos períodos de tiempo.

Asimismo, Steel y McLaren (2008) enumeran las principales dificultades al diseñar y analizar encuestas repetidas. Teniendo en cuenta los patrones de rotación en la estimación de los indicadores de nivel y de cambio, examinan su efecto en la estrategia de estimación de las varianzas de los estimadores de interés. Lewis (2017), por último, define algunos procedimientos que se deben seguir al combinar dos o más conjuntos de datos con el propósito de implementar de manera eficiente las pruebas de significación estadística

sobre indicadores de cambio a lo largo del tiempo, además de incrementar el tamaño de muestra para realizar inferencias de subgrupos poblacionales que están insuficientemente representados en una sola medición.

A. Métodos de acumulación de muestras

Antes de entrar en los detalles técnicos de este tipo de procedimientos, se analizará una situación que puede servir de ejemplo para ilustrar el problema que se quiere abordar. Con este objetivo, supóngase que un instituto nacional de estadística de América Latina ha previsto una nueva forma de analizar su encuesta de empleo. Con el fin de tener representatividad a un nivel más desagregado (por ejemplo, provincial), y para poder realizar una estimación más precisa, ha decidido realizar una agregación anual de todas las etapas de recolección de datos de su encuesta de empleo. Por ejemplo, supóngase que se planean operativos de recolección de datos trimestrales en los meses de marzo, junio, septiembre y diciembre y que, en este diseño, se considera una representatividad nacional, en el área urbana y rural, pero no provincial, ni de las ciudades principales del país. Con la metodología de agregación de muestras, se podría asegurar la representatividad en las provincias desagregadas por área (urbana o rural).

Los procesos de acumulación de muestras se realizan con frecuencia en las encuestas continuas con publicación trimestral. Por ejemplo, se pueden planear mediciones mensuales y acumular tres meses para realizar la publicación trimestral de la cifra de desempleo. De hecho, algunos países han decidido publicar cifras mensuales de desempleo teniendo en cuenta la acumulación de los últimos tres operativos de recolección de datos, lo que se conoce como trimestres móviles. Teniendo en cuenta el diseño rotativo de encuestas en América Latina, una de las bondades de estos métodos de agregación de muestras en los trimestres móviles es que el panel original se mantiene y, además, por diseño, la misma vivienda no es entrevistada dos veces en el trimestre móvil. En este tipo de diseños, es posible incluso que, al final de cada año, en diciembre, se contemple la publicación de un gran agregado anual que contenga la agregación de los 12 meses anteriores. En este escenario, sí existen viviendas que han sido entrevistadas dos o más veces y este porcentaje, dependiendo del diseño rotativo, puede no ser bajo. Por ejemplo, en un panel 2(2)2, el diseño rotativo genera un traslape natural del 50% entre un trimestre y otro.

Korn y Graubard (1999, caps. 7 y 8) describen de manera exhaustiva las opciones de ponderación y otros temas que deben considerarse cuando se combinan datos a lo largo del tiempo en encuestas complejas. En el caso de la agregación de muestras, se resalta que todas las viviendas que han sido entrevistadas en más de una ocasión deben pertenecer a la misma unidad primaria de muestreo (UPM) por diseño. Es muy importante que la identificación de las UPM y de los estratos de muestreo se realice de manera inequívoca, y se debe asegurar que se cumplan a cabalidad los siguientes principios:

i) Cuando se combinan dos o más rondas del mismo panel, es importante asegurarse de que las UPM se emparejen correctamente, de forma que el *software* las reconozca como iguales.

ii) Cuando se combinan dos o más muestras independientes, es importante asegurarse de que las UPM estén codificadas de forma que el *software* las reconozca como distintas.

Cuando se trata de estimar las varianzas de este tipo de estimadores, los cálculos analíticos se tornan mucho más complejos. Train, Cahoon y Makens (1978) muestran lo complicado que puede ser calcular las variaciones de los promedios de las estimaciones de múltiples períodos de tiempo en una encuesta repetida y cómo estos cálculos dependen en gran medida del patrón de traslape definido en el diseño de la encuesta. Para las encuestas de población activa, a menudo se utiliza un enfoque computacional basado en métodos de remuestreo, como los de jackknife, bootstrap o réplicas repetidas balanceadas. Cabe mencionar que el uso apropiado de estos métodos también dependerá del origen de la encuesta y de sus objetivos. Por ejemplo, los insumos para la aplicación de los métodos, si la encuesta está orientada a medir el desempleo, serían distintos de los que se utilizarían si está diseñada para estimar los cambios brutos entre dos períodos de tiempo.

B. Factores de expansión y estimadores de muestreo

Si el investigador está interesado en estimar la tasa de desempleo anual sobre una encuesta rotativa que se lleva a cabo durante los cuatro trimestres del año, es posible usar los cuatro conjuntos de datos y unir los trimestres para estimar la tasa de desempleo anual. Una solución inicial a este problema consiste en agregar las cuatro bases de datos y dividir los pesos de muestreo de cada período por un factor de cuatro. Este procedimiento brinda estimadores puntuales aproximadamente insesgados, aunque las estimaciones de los errores estándar se tornan un poco más complicadas, puesto que se deben concatenar exhaustivamente las UPM (o incluso crear unidades de varianza).

Por supuesto, las encuestas que utilizan diseños rotativos, en los que un hogar es entrevistado en varias ocasiones, deben adjuntar dos clases de pesos de muestreo: los transversales y los agregados. Los pesos transversales, discutidos en las secciones anteriores, son aquellos determinados por el diseño de muestreo de la encuesta en cada aplicación y permiten obtener estimaciones de los parámetros de interés de forma periódica (mensual, trimestral o semestral). De esta manera, por ejemplo, en una encuesta de fuerza de trabajo, los datos transversales se utilizarán para producir estimaciones periódicas de la participación en la fuerza de trabajo, o de la tasa de pobreza, o de la tasa

de desempleo, entre otras cosas. Por ejemplo, en la estimación de la tasa de desempleo, se utiliza un estimador de razón, definido de la siguiente forma:

$$\hat{\theta} = \frac{\sum_s d_k y_k}{\sum_s d_k z_k}$$

Donde, para la persona k-ésima, d_k representa su peso de muestreo, y_k representa su estado de ocupación (específicamente, $y_k=1$ si la persona está desempleada) y z_k es su estado en la fuerza de trabajo (específicamente, $z_k=1$ si la persona pertenece a la población económicamente activa). Esta estrategia de estimación parte del supuesto de que cada persona se representa a sí misma y a otras más en la población. Los pesos transversales asignados estarán determinados por la probabilidad de selección de las UPM, la probabilidad de selección del hogar dentro de la UPM, el ajuste por la falta de respuesta en ese mismo mes y los ajustes por elegibilidad o calibración, entre otros. Por estos motivos, así como por la incorporación de la nueva muestra en un diseño rotativo, además de por la falta de respuesta y los cambios en el tamaño de la población de interés, el peso de un individuo puede cambiar de un período a otro. De esta forma, si d_k^{t-1} y d_k^t representan el peso de muestreo del individuo k en los períodos $t-1$ y t, respectivamente, es casi seguro que:

$$d_k^{t-1} \neq d_k^t$$

Es necesario crear un nuevo conjunto de factores de expansión (pesos agregados) sobre los cuales se base la inferencia agregada del nuevo conjunto de datos. Cabe mencionar que cada factor de expansión en las encuestas mensuales se define como la cantidad de hogares que el hogar seleccionado representa en ese período de referencia. Por lo tanto, para mantener esta coherencia, es posible iniciar la construcción de los factores de expansión agregados mediante una modificación proporcional a los pesos originales de las mediciones mensuales. Por ejemplo, si se quisieran agregar tres meses, para formar una base de datos trimestral, sería necesario definir un factor de expansión trimestral d_k^+ que tenga en cuenta la siguiente relación:

$$\hat{t}_y = \sum_{s1 \cup s2 \cup s3} d_k^+ y_k \propto \sum_{s_1} d_{1k} y_k + \sum_{s_2} d_{2k} y_k + \sum_{s_3} d_{3k} y_k$$

Donde d_{ik} es el factor de expansión del mes i-ésimo $(i=1,2,3)$. En particular, para esta agregación trimestral, el factor de expansión mensual de cada individuo y hogar debe multiplicarse por el siguiente ponderador:

$$a_i = \frac{\sum_{k \in s_i} d_{ik}}{\sum_{i=1}^{3} \sum_{k \in s_i} d_{ik}} \; ; \; i = 1,2,3$$

Donde s_i representa la muestra de respondientes efectivos en el mes i-ésimo. De esta forma, los pesos iniciales agregados estarían dados por la siguiente expresión:

$$d_{ik}^+ = a_i \times d_{ik} \; ; \; k \in s_i$$

De la misma manera, para una agregación anual, el factor de expansión debe modificarse de manera proporcional a los pesos originales de las mediciones mensuales (o trimestrales), teniendo en cuenta la siguiente relación:

$$\hat{t}_y = \sum_{s_1 \cup \ldots \cup s_{12}} d_k^+ y_k \propto \sum_{s_1} d_{1k} y_k + \sum_{s_2} d_{2k} y_k + \cdots + \sum_{s_{12}} d_{12k} y_k$$

Por lo tanto, en la agregación anual, el factor de expansión de cada individuo y hogar debe multiplicarse por el siguiente ponderador:

$$b_i = \frac{\sum_{k \in s_i} d_{ik}}{\sum_{i=1}^{12} \sum_{k \in s_i} d_{ik}} \; ; \; i = 1, \ldots, 12$$

Por consiguiente, los pesos iniciales agregados estarían dados por la siguiente expresión:

$$d_{ik}^+ = b_i \times d_{ik} \; ; \; k \in s_i$$

La nueva estructura de los factores de expansión debe garantizar que la suma de los pesos en las bases agregadas esté acorde con la población que se quiere representar. En términos matemáticos, siempre se debe verificar que las siguientes relaciones se mantengan en las bases agregadas:

$$\sum_{k \in s^3} d_{ik}^+ = \sum_{i=1}^{3} \sum_{s_i} a_i \, d_{ik} \approx N$$

Donde $s^3 = s1 \cup s2 \cup s3$ corresponde a la muestra agregada de los tres primeros meses. De la misma manera, en el caso de la agregación anual, también conviene verificar la misma relación; es decir:

$$\sum_{k \in s^{12}} d_{ik}^+ = \sum_{i=1}^{12} \sum_{s_i} b_i \, d_{ik} \approx N$$

Donde $s^{12} = s_1 \cup \ldots \cup s_{12}$ corresponde a la muestra agregada anual. Además de las verificaciones sobre los tamaños nacionales, también es recomendable realizar este mismo proceso en dominios más específicos con el fin de verificar que la ponderación sea correcta. Entre esos dominios, cabe mencionar las principales ciudades del país, las áreas rural y urbana, las provincias y los grupos de sexo o edad. Una vez que se ha llevado a cabo el proceso de cómputo de los nuevos pesos agregados en las bases de datos (trimestrales o anuales), es necesario que se realice nuevamente un proceso de calibración sobre las variables que intervienen en la calibración mensual de los factores de expansión.

Ante la ausencia de proyecciones poblacionales trimestrales o anuales, es posible escoger el mes intermedio o el promedio de los meses que intervienen en la agregación. Se espera que este ajuste final de los pesos sea minúsculo y no afecte la estructura de la distribución de los pesos mensuales, puesto que se trata de calibrar unos pesos que ya se habían calibrado en las publicaciones mensuales. Por otro lado, debido a que este último

paso se realiza con propósitos de mantener la coherencia con las publicaciones, es posible que la calibración se vea reducida al considerar menos restricciones sobre los totales auxiliares más relevantes.

Se recalca que las agregaciones deberían incluir todas las viviendas que formaron parte de las muestras mensuales en el trimestre móvil. De la misma forma, las agregaciones anuales deben incluir las viviendas que han sido seleccionadas más de una vez (debido al diseño de rotación del panel) y, por ende, todas sus mediciones deben aparecer en la base de datos tantas veces como hayan sido visitadas.

Para ilustrar el procedimiento, considérese una encuesta de hogares continua en la que mes a mes se recopila información. Supóngase que esta encuesta sigue un diseño rotativo trimestral 2(2)2 y que las muestras mensuales son independientes. Es decir, la rotación de los paneles se planeó de manera trimestral y, a su vez, esta muestra está repartida de forma balanceada e independiente en los tres meses que conforman el trimestre. En este caso, las agregaciones trimestrales no deberían incluir ninguna vivienda con mediciones repetidas si el diseño de panel no las incluye. Nótese que es necesario realizar el correspondiente ajuste a los pesos de muestreo sin diferenciar si la vivienda apareció una vez o fue medida en más de una ocasión.

En el escenario utilizado como ejemplo, en la estimación del error de muestreo para las agregaciones trimestrales, se debe considerar que el muestreo es independiente en los tres meses que componen el trimestre móvil y, por ende, la posibilidad de tener viviendas repetidas es casi nula. Nótese que el estimador de un total en la agregación trimestral tomará la siguiente forma de sumas mensuales parciales:

$$\hat{t}_y = \sum_{s_1} d_{1k}^{+}\, y_k + \sum_{s_2} d_{2k}^{+} y_k + \sum_{s_3} d_{3k}^{+} y_k = \hat{t}_y^{\,1} + \hat{t}_y^{\,2} + \hat{t}_y^{\,3}$$

Donde $d_{ik}^{+} = a_i \times d_{ik}$. En este caso, la varianza del estimador está dada por:

$$Var(\hat{t}_y) = Var(\hat{t}_y^{\,1}) + Var(\hat{t}_y^{\,2}) + Var(\hat{t}_y^{\,3})$$

Sin embargo, en la estimación del error de muestreo para las agregaciones anuales, se debe considerar que el muestreo no es independiente en los 12 meses. En este caso, el estimador de interés sigue tomando la forma de sumas parciales mensuales:

$$\hat{t}_y = \sum_{i=1}^{12} \sum_{s_i} d_{ik}^{+}\, y_k = \sum_{i=1}^{12} \hat{t}_y^{\,i}$$

Donde $d_{ik}^{+} = b_i \times d_{ik}$. A diferencia de la agregación trimestral, la varianza de este estimador está supeditada a las covarianzas que se puedan crear al visitar las mismas UPM debido al diseño rotativo. Es decir:

$$Var(\hat{t}_y) = \sum_{i=1}^{12} Var(\hat{t}_y^{\,i}) + 2 \sum_{i,j=1}^{12} \sum_{j<i} Cov(\hat{t}_y^{\,i}, \hat{t}_y^{\,j})$$

C. Agregación de encuestas con diferentes tamaños de muestra

En algunos casos, el tamaño de las encuestas puede variar significativamente entre dos meses consecutivos. La pandemia de enfermedad por coronavirus (COVID-19) mostró cómo esta clase de eventos adversos puede afectar gravemente el tamaño de muestra de las encuestas con recopilación de datos regular. Por ejemplo, considérese el diseño trimestral del cuadro XV.1.

■ Cuadro XV.1
Encuestas con recopilación de datos regular: diseño trimestral

Panel y mes	M1	M2	M3
Panel	P1	P2	P3
Viviendas	5 000	4 500	2 500
Panel	P4	P5	P6
Viviendas	5 500	5 100	3 000

Fuente: Elaboración propia.
Nota: M: mes; P: panel.

En este caso, después de haber evaluado, analizado y ejecutado exhaustivamente los ajustes al factor de expansión, el principio detrás de esta agregación trimestral es intuitivo y simple: cada elemento de la base de datos agregada se representa a sí mismo y representa también a una porción de los habitantes del país en diferentes períodos de tiempo. Teniendo esto en cuenta, es necesario señalar que, en el primer mes, los paneles que se utilizan para producir las cifras oficiales son únicamente el P1 y el P4. De la misma manera, en el segundo mes se utilizan únicamente los paneles P2 y P5. Supóngase también que las cifras oficiales en estos dos primeros meses son representativas de más desagregaciones que las del tercer mes, en el que participan los paneles P3 y P6, pero con un decremento sustancial del tamaño de la muestra.

Como se puede apreciar, este cuadro contiene varios elementos que hay que tratar con precisión. En particular, el hecho de que la encuesta recopile datos mensuales en todos los dominios de interés no implica que mensualmente se tenga el mismo nivel de representatividad que en el trimestre. En realidad, como se ha mencionado, dada la baja incidencia de entrevistas en el último mes, es muy posible que no exista el mismo nivel de representatividad en comparación con los dos primeros meses. Todo ello implica que el tratamiento de los factores de expansión iniciales debe hacerse de forma diferencial.

Heeringa, West y Berglund (2017) afirman que, para evitar el sesgo que generan los tamaños de muestra pequeños, como el que evidentemente presenta el tercer mes del ejemplo, es posible ajustar los pesos de muestreo. En particular, cuando existe este tipo

de diferenciación en los tamaños de muestra de las diferentes instancias de recolección de datos, se sugiere utilizar la siguiente expresión para normalizar los pesos de muestreo:

$$d^{+}_{kth} = \delta_{th} \times d_{kth}$$

Donde d_{kth} hace referencia al peso de muestreo del individuo k del estrato h en el mes t $(t=1,2,3)$ y δ_{th} es un factor de ajuste, dependiente del tamaño de muestra, que representa el porcentaje de individuos observados en el mes t para el estrato h. Este factor, propuesto por Kish (1999, pág. 131) en el contexto de acumulación de muestras, está dado por la siguiente expresión:

$$\delta_{th} = \frac{n_{th}}{\sum_{t=1}^{3} n_{th}}$$

En general, h podría ser el estrato de muestreo o, de manera más amplia, el dominio de representatividad. Utilizando esta metodología, los factores trimestrales tendrían las siguientes propiedades, bastante favorables en un sistema de ponderación agregado:

i) definen una combinación lineal convexa;

ii) mantienen la coherencia con los tamaños por estrato o dominio;

iii) su aporte es proporcional al tamaño de muestra mensual, y

iv) se pueden expresar como un promedio equivalente en todos los estratos o dominios.

Para esta formulación en particular, la primera propiedad se mantiene puesto que $\delta_{th} > 0 \ \forall t, \forall h$ y, además, $\sum_{t=1}^{3} \delta_{th} = 1$. La segunda propiedad se verifica dado que, si s_h es la muestra del estrato h a lo largo de todos los meses, entonces:

$$\sum_{t=1}^{3} \sum_{k \in s_h} d^{+}_{kth} = \sum_{t=1}^{3} \sum_{k \in s_h} \delta_{th} \, d_{kth} = \sum_{t=1}^{3} \delta_{th} \sum_{k \in s_{th}} d_{kth} = \sum_{t=1}^{3} \delta_{th} \hat{N}_h^t \cong \sum_{t=1}^{3} \delta_{th} \hat{N}_h = \hat{N}_h$$

La tercera propiedad se verifica puesto que la suma de los factores de expansión trimestrales restringida a un mes y un dominio particular está ponderada por el factor de ajuste δ, como se demuestra a continuación:

$$\sum_{k \in s_{th}} d^{+}_{kth} = \sum_{k \in s_{th}} \delta_{th} \, d_{kth} = \delta_{th} \sum_{k \in s_{th}} d_{kth} = \delta_{th} \ \hat{N}_h^t$$

La última propiedad se puede comprobar en la encuesta, verificando que el aporte de los factores trimestrales sea proporcional al tamaño de muestra en cada dominio y en cada mes. Por último, la media de los factores de expansión es casi invariante con respecto a los meses, restringidos a un dominio específico. En efecto, nótese que:

$$\frac{\sum_{k \in s_{th}} d^{+}_{kth}}{n_h} = \frac{\sum_{k \in s_{th}} \delta_{th} \, d_{kth}}{n_h} = \frac{\sum_{k \in s_{th}} d_{kth}}{\sum_h n_h} = \frac{\hat{N}_{th}}{\sum_h n_h} \cong \frac{\hat{N}_h}{\sum_h n_h}$$

Este comportamiento se observaría en la agregación, verificando que, sin importar el mes, la media de los factores trimestrales sea similar para cada dominio

de interés. Las anteriores cuatro propiedades hacen que se cree un mejor sistema de ponderación agregado, puesto que cada individuo en un dominio de interés tendrá un factor trimestral similar, lo que dará fuerza al mes que mayor tamaño de muestra tenga. Por ello, para las cinco ciudades principales, es de esperar que la agregación defina estimadores que colinden con los valores promedio de los estimadores puntuales de los tres meses considerados.

La agregación anual consistiría en la extensión de esta metodología, considerando un período más extenso de 12 meses. En particular, se sugiere utilizar la siguiente expresión para normalizar los pesos de muestreo:

$$d^{+}_{kth} = \delta_{th} * d_{kth}$$

Donde d_{kth} hace referencia al peso de muestreo del individuo k del estrato h en el mes t $(t=1,\ldots,12)$ y δ_{th} representa el porcentaje de individuos observados en el mes t para el estrato h, dado por:

$$\delta_{th} = \frac{n_{th}}{\sum_{i=1}^{12} n_{th}}$$

D. Efecto del tipo de encuesta en la eficiencia de los indicadores

Lograr una estimación adecuada del error de muestreo en las comparaciones de múltiples períodos de tiempo, ya sea agregando los datos o no, debe ser una de las principales tareas del investigador. Además, dependiendo del parámetro, la naturaleza del error de muestreo cambia, así como el tamaño de muestra requerido para satisfacer las necesidades de precisión de las estimaciones. A continuación se ilustran diferentes tipos de parámetros.

1. Cambios netos

Considérese el cambio neto de la media de la variable de interés y en dos períodos de tiempo $(t_2 \text{ y } t_1)$:

$$\Delta = \bar{y}_2 - \bar{y}_1$$

Este parámetro de cambio en los dos períodos de tiempo se estima de forma aproximadamente insesgada mediante la siguiente expresión:

$$\hat{\Delta} = \hat{\bar{y}}_2 - \hat{\bar{y}}_1 = \frac{\sum_{k \in s_2} \frac{y_k}{\pi_k}}{\sum_{k \in s_2} \frac{1}{\pi_k}} - \frac{\sum_{k \in s_1} \frac{y_k}{\pi_k}}{\sum_{k \in s_1} \frac{1}{\pi_k}}$$

Donde s_2 y s_1 representan las muestras seleccionadas en los períodos de interés y π_k es la probabilidad de inclusión del elemento k. La varianza del estimador de cambio se calcula mediante la siguiente expresión:

$$Var(\hat{\Delta})=Var(\hat{y}_2)+Var(\hat{y}_1)-2Cov(\hat{y}_2,\hat{y}_1)$$

En general, el último término se puede expresar como:

$$2Cov(\hat{y}_2,\hat{y}_1) = 2\sqrt{Var(\hat{y}_2)}\ \sqrt{Var(\hat{y}_1)}\sqrt{T_2}\sqrt{T_1}R_{12}$$

Donde T_2 y T_1 representan el porcentaje de muestra común que se traslapa en ambas instancias de recopilación de datos y R_{12} representa la correlación de la variable de interés x en los períodos observados. Suponiendo que la variación de la variable de interés es homogénea en ambos períodos $Var(\hat{y}_1)=Var(\hat{y}_2)=Var(\hat{y})$ y que el traslape es común por diseño $T_2=T_1=T$, entonces la expresión de la varianza se reduce de la siguiente manera:

$$Var(\hat{\Delta})=2Var(\hat{y}) - 2Var(\hat{y})TR_{12}=2Var(\hat{y})(1 - TR_{12})$$

Kish (2004) comenta que la varianza de este indicador cambiará de acuerdo con el tipo de encuesta que se elija. En efecto:

i) Encuesta repetida: donde $T=0$ y

$$Var(\hat{\Delta})=2Var(\hat{y})$$

ii) Encuesta de panel: donde $T=1$, $R_{12} > 0$ y

$$Var(\hat{\Delta})=2Var(\hat{y})\ (1 - R_{12})$$

iii) Encuesta rotativa: donde $T\neq 0$, $R_{12} > 0$ y

$$Var(\hat{\Delta})=2Var(\hat{y})\ (1 - TR_{12})$$

Además, si se supone que la correlación es positiva para la variable de interés en los dos períodos de tiempo, se obtiene la siguiente conclusión:

$$2Var(\hat{y})(1 - R_{12}) < 2Var(\hat{y})(1 - TR_{12}) < 2Var(\hat{y})$$

Es decir, cuando se utiliza un diseño de panel, para medir los cambios netos se necesita un tamaño de muestra menor que cuando se aplica un diseño sin traslape en una encuesta repetida. Una solución intermedia es el diseño rotativo.

2. Promedio trimestral

Considérese una encuesta continua y mensual en la que se quiere estimar el promedio trimestral de la variable de interés x en tres períodos de tiempo $(t_3, t_2 \text{ y } t_1)$:

$$\Theta = \frac{\bar{y}_3 + \bar{y}_2 + \bar{y}_1}{3}$$

Un estimador del promedio trimestral que es aproximadamente insesgado está dado por la siguiente expresión:

$$\Theta = \frac{1}{3}(\hat{\bar{y}}_3+\hat{\bar{y}}_2-\hat{\bar{y}}_1)=\frac{1}{3}\left(\frac{\sum_{k\in s_3}\frac{y_k}{\pi_k}}{\sum_{k\in s_3}\frac{1}{\pi_k}}+\frac{\sum_{k\in s_2}\frac{y_k}{\pi_k}}{\sum_{k\in s_2}\frac{1}{\pi_k}}+\frac{\sum_{k\in s_1}\frac{y_k}{\pi_k}}{\sum_{k\in s_1}\frac{1}{\pi_k}}\right)$$

Donde s_3, s_2 y s_1 representan las muestras seleccionadas en los períodos de interés y π_k es la probabilidad de inclusión del elemento k. La varianza del estimador del promedio trimestral se calcula mediante la siguiente expresión:

$$Var(\hat{\Theta})=\frac{1}{9}\,[Var(\hat{\bar{y}}_3)+Var(\hat{\bar{y}}_2)+Var(\hat{\bar{y}}_2)+ 2Cov(\hat{\bar{y}}_3,\hat{\bar{y}}_2))+2Cov(\hat{\bar{y}}_3,\hat{\bar{y}}_1))+2Cov(\hat{\bar{y}}_2,\hat{\bar{y}}_1)]$$

Suponiendo que la variación de la variable de interés es homogénea en los tres períodos, que el traslape es común por diseño y que los errores de muestreo son débilmente estacionarios (media y correlación constante) entre dos y tres meses, entonces la expresión de la varianza se reduce de la siguiente manera:

$$Var(\hat{\Theta})=\frac{1}{9}\,Var(\hat{\bar{y}})[3+6TR]$$

Donde $R = R_{12}= R_{23}= R_{13}$ es la correlación constante de la variable de interés en dos y tres meses (que se supone homogénea). La varianza de este indicador cambiará de acuerdo con el tipo de encuesta que se elija:

iv) Encuesta repetida: donde $T= 0$ y

$$Var(\hat{\Theta})=\frac{1}{3}\,Var(\hat{\bar{y}})$$

v) Encuesta de panel: donde $T=1$, $R > 0$ y

$$Var(\hat{\Theta})=\frac{1}{9}\,Var(\hat{\bar{y}})[3+6R]$$

vi) Encuesta rotativa: donde $T\neq 0$, $R > 0$ y

$$Var(\hat{\Theta})=\frac{1}{9}\,Var(\hat{\bar{y}})[3+6TR]$$

De esta forma, si se supone que la correlación es positiva para la variable en los tres períodos de tiempo, se llega a la siguiente conclusión:

$$\frac{1}{9}Var(\hat{\bar{y}})[3+6R] > \frac{1}{9}Var(\hat{\bar{y}})[3+6TR] > \frac{1}{3}Var(\hat{\bar{y}})$$

Es decir, se necesita un tamaño de muestra mayor para estimar un promedio trimestral con un diseño de panel sin traslape. De la misma forma, una solución intermedia es el diseño de panel rotativo.

E. Pruebas de hipótesis sobre indicadores agregados

A fin de determinar si un cambio en la dinámica de los parámetros de interés entre dos períodos de tiempo es significativo, es necesario llevar a cabo una prueba de hipótesis. Por ejemplo, tomando en cuenta la dinámica del mercado de trabajo, es posible realizar comparaciones entre dos trimestres seguidos o entre dos años consecutivos para saber si hay un cambio significativo e importante en la reducción de la desocupación (entre diferentes grupos y en distintos períodos de tiempo).

Para realizar comparaciones entre grupos de un mismo corte transversal —por ejemplo, comparar la situación laboral de hombres y mujeres en un mes específico—, es necesario tener en cuenta que el muestreo de la primera etapa es de UPM y que el tamaño de muestra de hombres y mujeres es aleatorio. Para realizar comparaciones nacionales o regionales en dos períodos de tiempo —por ejemplo, comparar la situación laboral de un país entre dos trimestres—, es necesario tener en cuenta que el muestreo puede no ser independiente entre trimestres ni entre años, siendo este el caso de las encuestas con diseños de panel rotativo. Considérese el siguiente sistema de hipótesis:

$$H_0 : \theta_2 - \theta_1 = 0 \quad vs. \quad H_1 : \theta_2 - \theta_1 \neq 0$$

Para llevar a cabo la prueba de hipótesis, se trabaja con el siguiente estimador de diferencias:

$$\hat{\Delta} = \hat{\theta}_2 - \hat{\theta}_1$$

La varianza asociada a este estimador está dada por:

$$Var(\hat{\Delta}) = Var(\hat{\theta}_2) + Var(\hat{\theta}_1) - 2Cov(\hat{\theta}_1, \hat{\theta}_2)$$

Por último, el término de covarianza se puede escribir como:

$$Cov(\hat{\theta}_2, \hat{\theta}_1) = \sqrt{Var(\hat{\theta}_1)} \sqrt{Var(\hat{\theta}_2)} \sqrt{T_2} \sqrt{T_1} R_{12}$$

Existen muchos escenarios de comparación que son de interés cuando se analizan datos de una encuesta de empleo. Estas comparaciones se hacen más complejas cuando se incluye en el análisis el diseño de panel de la encuesta. Sin embargo, cuando se cumple el siguiente principio, no habrá lugar a confusión: a no ser que los dos estimadores puntuales estén compuestos de observaciones provenientes de un conjunto disyunto de UPM, el término de covarianza no será nulo.

Normalmente no es posible generalizar la estructura de varianza en una base de datos agregada. No obstante, si se toman como punto de partida los ejemplos expuestos en el capítulo sobre el tamaño de la muestra, se pueden definir tres escenarios de interés. En primer lugar, se puede suponer que existe independencia en el muestreo de dos meses consecutivos. En este caso, $T_1 = T_2 = 0$; luego, el término de la covarianza se anularía.

En segundo lugar, en un diseño de panel 2(2)2, si se quiere comparar estimadores nacionales entre trimestres consecutivos o entre el mismo mes de dos años consecutivos, entonces $T_1 = T_2 \approx 0.5$ y $R_{12} \neq 0$. En este caso, el término de covarianza sería igual a: $Cov(\hat{\theta}_1,\hat{\theta}_2) = \frac{1}{2}\sqrt{Var(\hat{\theta}_1)}\sqrt{Var(\hat{\theta}_2)}R_{12}$. Por último, si se quiere comparar estimadores entre subgrupos en un mismo mes, se pueden distinguir dos casos de interés.

i) Si no existe independencia en el muestreo de los subgrupos (por ejemplo, hombres y mujeres): por no ser estratos de muestreo, $T_1 \neq T_2$ y $R_{12} \neq 0$, y el término de covarianza en este caso sería igual a $Cov(\hat{\theta}_1,\hat{\theta}_2) = \sqrt{Var(\hat{\theta}_1)}\sqrt{Var(\hat{\theta}_2)}\sqrt{T_1}\sqrt{T_2}R_{12}$.

ii) Si existe independencia en el muestreo de los subgrupos (por ejemplo, dos ciudades principales o dos regiones): por ser estratos de muestreo, $R_{12} = 0$, y el término de covarianza será nulo.

Una vez que se haya concluido la estructura de varianza del estimador de interés, el siguiente paso es definir el estadístico de prueba para determinar si el parámetro ha cambiado entre grupos o a lo largo del tiempo, mediante la siguiente expresión:

$$t = \frac{\hat{\Delta}}{\sqrt{Var(\hat{\Delta})}}$$

Este estadístico de prueba sigue una distribución t de Student con gl grados de libertad, que surgen de restar al número de UPM seleccionadas el número de estratos de muestreo considerados en la agregación. De esta forma, se tiene que:

$$gl = \sum_{h=1}^{H}(n_{Ih} - 1) = \sum_{h=1}^{H} n_{Ih} - H = \#UPM - \#Estratos$$

Los grados de libertad permiten realizar una inferencia precisa a medida que crecen. Por ejemplo, considérese el percentil 0,975, para el cual los valores críticos de la distribución varían con respecto a sus grados de libertad: $t_{0,975,1} = 12,7$, $t_{0,975,20} = 2,08$, $t_{0,975,40} = 2,02$, $t_{0,975,\infty} = 1,96$. Los grados de libertad son determinantes a la hora de hacer inferencias dentro de subpoblaciones de interés. En este caso, los grados de libertad no se consideran fijos, sino variables. Korn y Graubard (1999) proponen el siguiente método de cálculo sobre los grados de libertad en subpoblaciones:

$$gl_{subpoblación} = \sum_{h=1}^{H} v_h (n_{Ih} - 1)$$

Donde v_h es una variable indicadora que toma el valor 1 si el estrato h contiene uno o más casos de las subpoblaciones de interés y el valor 0 en caso contrario.

Capítulo XVI
Procesamiento longitudinal de las encuestas rotativas

Algunos institutos nacionales de estadística (INE) pueden necesitar una estructura de ponderación longitudinal que permita a los analistas de las distintas áreas producir estadísticas basadas en el seguimiento continuo de los hogares, afianzándose en el sistema de rotación de las encuestas. Antes de establecer los pasos para la creación de un sistema de pesos longitudinales, es necesario definir qué es una encuesta longitudinal y, en particular, la manera en que las encuestas continuas con instancias de recopilación de datos transversales pueden convertirse en encuestas longitudinales.

De acuerdo con Lynn (2009), una encuesta longitudinal es aquella que recolecta los datos de los mismos elementos muestrales en múltiples ocasiones a lo largo del tiempo. Por ejemplo, una encuesta trimestral con un diseño rotativo 4(0)1 permitiría realizar observaciones continuas en el 25% de las viviendas durante todo un año. Para crear un sistema de ponderación longitudinal, es necesario concentrarse en la estimación del cambio del indicador en dos períodos de tiempo consecutivos y en la correspondiente estimación de la varianza. Cabe resaltar que este proceso debe tener en cuenta que las muestras no son independientes y, por lo tanto, se deben calcular la varianza de la primera ronda, la varianza de la segunda ronda y la correlación entre las dos rondas de interés. Estos tres componentes servirán para el cálculo de los coeficientes de variación y la determinación del tamaño de la muestra en cada ronda.

Asimismo, con el análisis de los datos longitudinales, es posible realizar otros tipos de análisis, como los siguientes:

- Inferencia sobre la caracterización de las unidades de observación que han pasado de un estado a otro: a partir de las bases de datos longitudinales, es posible determinar las características de los hogares o las personas que han

sufrido cambios en las variables de interés. Por ejemplo, se pueden determinar las características de los hogares que han salido de la pobreza extrema o han caído en esa situación, sin importar si se registró un cambio neto significativo en el período de estudio.

- Inferencia acerca de la estabilidad (o inestabilidad) de determinadas características de interés: al combinar varios períodos de seguimiento, es posible detectar que algunas unidades de observación experimentan períodos de estabilidad (o fluctuación) con respecto al fenómeno de interés. Por ejemplo, el análisis de este tipo de problemas puede mejorar la comprensión de las situaciones que confluyen para que un hogar caiga en la pobreza extrema y se mantenga en ese estado durante un período determinado.
- Caracterización de acontecimientos y fenómenos: con las encuestas longitudinales, es posible determinar cabalmente la duración de los períodos en los que una unidad de observación cambia de un estado a otro y persiste en este último (por ejemplo, caer en la pobreza, pasar a la inactividad económica, quedar desocupado o abandonar la educación, entre otros).
- Análisis del impacto, los efectos y las relaciones causales: los datos longitudinales pueden utilizarse efectivamente a la hora de establecer relaciones causales entre una intervención y un fenómeno de interés. Por ejemplo, es posible evaluar la magnitud del impacto de la pandemia de enfermedad por coronavirus (COVID-19) en la tasa de desocupación y sus efectos a lo largo del año.

A. Diseño de paneles rotativos en las encuestas de la región

En algunas encuestas de hogares de América Latina, se prevé más de una visita a los hogares a fin de obtener estimaciones precisas acerca de los cambios de estado que los hogares o las personas que los conforman puedan sufrir. Por ejemplo, en las encuestas de la fuerza laboral, una persona puede pasar de estar ocupada en un período a estar inactiva en el período siguiente. Estos cambios y la dinámica propia que los caracteriza son de interés para los investigadores y deben contemplarse desde una perspectiva más amplia en el diseño de las encuestas. Este tipo de variaciones en los individuos se capta mediante el componente longitudinal de la encuesta, constituido por los respondientes efectivos observados de forma sistemática en los períodos de interés.

Por ejemplo, se considera una encuesta continua de hogares con un diseño rotativo 4(0)1, en el que un hogar es entrevistado durante cuatro trimestres consecutivos y luego sale del panel definitivamente. En el cuadro XVI.1 se presenta un ejemplo de este tipo de diseño.

■ Cuadro XVI.1
Ejemplo de rotación de paneles en una encuesta de hogares con un diseño rotativo 4(0)1

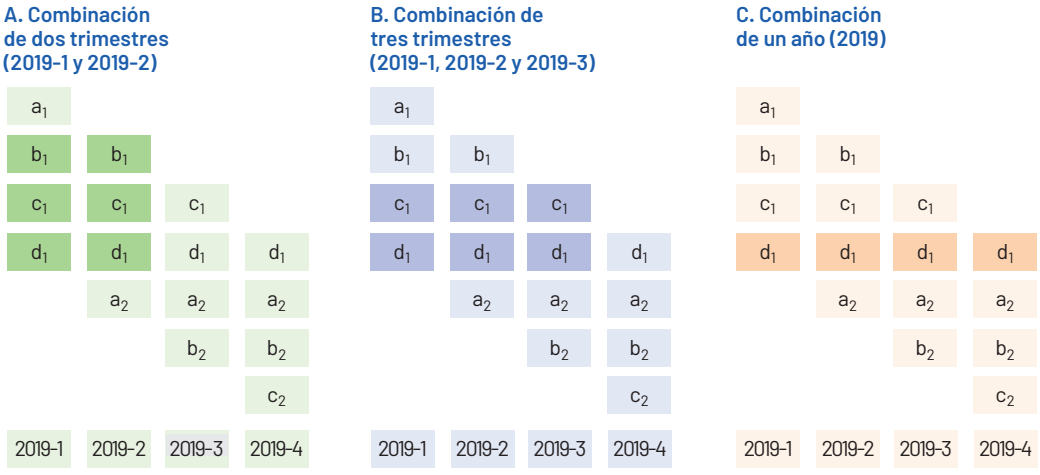
Trimestre	Panel 1	Panel 2	Panel 3	Panel 4
T1	a_1	b_1	c_1	d_1
T2	b_1	c_1	d_1	a_2
T3	c_1	d_1	a_2	b_2
T4	d_1	a_2	b_2	c_2
T5	a_2	b_2	c_2	d_2
T6	b_2	c_2	d_2	a_2

Fuente: Elaboración propia.

Entre el primer y el segundo período de medición, hay un traslape del 75% de los hogares. En particular, se conservan tres cuartas partes de la muestra, puesto que b_1, c_1 y d_1 se repiten. Esto mismo sucede en cada trimestre del diseño rotativo. Por otra parte, entre el primer y el tercer período habrá un traslape del 50%. En este caso, se conserva la mitad de la muestra, puesto que c_1 y d_1 se repiten. Este mismo patrón se encuentra a lo largo de todo el diseño rotativo. Entre el primer y el cuarto período, el traslape será del 25%, con la repetición de d_1. Por último, entre el primer y el quinto trimestre, no habrá ningún tipo de traslape.

En el diagrama XVI.1, se ejemplifican tres diseños longitudinales que pueden crearse para el año 2019. El primero de ellos representa la combinación del primer y el segundo trimestre, que se define mediante la agregación de las dos mediciones en el primer y segundo trimestre de los paneles b_1, c_1 y d_1. El segundo muestra la combinación de los tres primeros trimestres del año, definidos por los paneles c_1 y d_1. Por último, la base longitudinal anual parte de la combinación de las cuatro mediciones del panel d_1.

■ Diagrama XVI.1
Escenarios longitudinales en una encuesta de hogares con un diseño rotativo 4(0)1, 2019



Fuente: Elaboración propia.

Como se mencionó en capítulos anteriores, la pandemia de COVID-19 hizo que 2020 fuese un año atípico para la recopilación de datos de las encuestas de hogares de los INE de la región, pues la crisis de salud planteó muchos retos en lo que se refiere a la consecución de la información primaria. Debido a las restricciones de movilidad que los Gobiernos tuvieron que imponer para hacer frente a la pandemia, en algunos trimestres se optó por reproducir el mismo diseño de los trimestres anteriores. En el ejemplo anterior, se supone que no se incluyó el 25% adicional que se tenía planeado, sino que se utilizó exactamente la misma muestra que en el primer trimestre (véase el cuadro XVI.2). Asimismo, cabe recordar que, en casi toda la región, la muestra de hogares no se contactó de manera presencial sino telefónica, por lo que disminuyeron las tasas de cobertura y respuesta efectiva.

■ Cuadro XVI.2
Ejemplo de rotación de paneles en una encuesta de hogares con un diseño rotativo 4(0)1, 2020

Año	Trimestre	Panel 1	Panel 2	Panel 3	Panel 4
2020	T1	a_1	b_1	c_1	d_1
	T2	b_1	c_1	d_1	a_2
	T3	b_1	c_1	d_1	a_2
	T6	c_1	d_1	a_2	b_2

Fuente: Elaboración propia.

En este ejemplo, el traslape de la muestra de hogares entre el segundo y el tercer trimestre de 2020 fue completo. Mientras que entre el primer y el tercer trimestre de 2020 hubo un 75% de traslape, este disminuyó al 50% entre el primer y el último trimestre del mismo año.

B. Creación de bases de datos longitudinales para dos períodos consecutivos

El análisis longitudinal de las encuestas de hogares es un valioso instrumento para la toma de decisiones, puesto que ofrece una visión complementaria de los fenómenos sociales que no se puede obtener por otros medios. El seguimiento continuo de las unidades de observación no puede llevarse a cabo en todas las encuestas continuas de América Latina, sino solamente en aquellas que hayan sido planificadas con diseños rotativos. Dado que en algunas encuestas se contempla la asignación de la muestra a diferentes grupos de rotación, es posible analizar el comportamiento de los flujos brutos con respecto a indicadores tan importantes como los relacionados con la ocupación y la pobreza, entre otros.

En un contexto de estimación de cambios brutos con la definición de tablas de contingencia, Feinberg y Stasny (1983) dan por sentado que las diferencias entre los pesos de muestreo en dos períodos de tiempo distintos ocurren solamente como resultado de los

flujos naturales de entrada y salida de la población de interés. Por ejemplo, si la persona se clasifica como empleada en ambos tiempos y $w_k^{t-1}=300$ y $w_k^t=305$, el peso mínimo (300) se añade a la celda (Empleado - Empleado) de la tabla de cambios brutos y la diferencia entre los pesos (5) se añade a la celda (Fuera - Empleado). Si, por el contrario, $w_k^{t-1}=305$ y $w_k^t=300$, el peso mínimo (300) se añade a la celda (Empleado - Empleado) de la tabla de cambios brutos y la diferencia entre los pesos (5) se añade a la celda (Empleado - Fuera). Este enfoque supone que las diferencias entre los pesos están supeditadas a las fluctuaciones que se puedan presentar en la fuerza de trabajo.

El objetivo de esta sección es generar pesos longitudinales para todas las personas pertenecientes a los paneles correspondientes de la muestra original en los dos primeros trimestres de 2020. Según la metodología de Verma, Betti y Ghellini (2006), es necesario seguir un procedimiento secuencial para la creación de los factores de expansión en el panel. En primer lugar, se debe crear el conjunto de pesos iniciales (transversales en el primer período), para luego definir los pesos finales (longitudinales en el primer y el segundo período). A continuación, se resume el procedimiento:

i) Creación de pesos iniciales: se tienen en cuenta los pormenores del diseño de muestreo de la encuesta a partir de la cual se selecciona una muestra de hogares y de personas que son miembros de esos hogares. Las ponderaciones iniciales se definen a partir de los factores de expansión transversales. Este proceso comprende al menos los siguientes pasos:
   - Determinación de los pesos básicos con el ajuste de selección de paneles rotativos.
   - Ajuste por falta de respuesta y cobertura.

ii) Generación de pesos longitudinales: la muestra debe modificarse y ajustarse para reflejar los cambios en la duración del panel para la población objetivo en los dos períodos de interés. En este caso se plantean al menos tres tipos de ajuste:
   - Definición de la población longitudinal (supeditada a los hogares que salen y entran en el período de referencia).
   - Falta de respuesta y pérdidas en la muestra debido al desgaste de los respondientes (falta de respuesta en el panel).
   - Calibración de los pesos longitudinales.

El primer paso en la generación de los pesos longitudinales consiste en realizar una consolidación (combinación) de bases de datos, en la que se integren únicamente los períodos de interés. Este proceso producirá bases de datos de diferentes tamaños para dos, tres o cuatro períodos. En general, se esperaría contar con un mayor número de unidades de observación en el primer caso (dos períodos) y un número menor de unidades de observación en el último caso (cuatro períodos). Cabe señalar que, en el caso particular del ejemplo —una encuesta con un diseño rotativo 4(0)1—, no es posible realizar la integración de cinco períodos consecutivos porque el diseño solo define el traslape de hasta cuatro períodos consecutivos.

En general, es necesario suponer que, al combinar los paneles y crear una sola base de datos, se está agregando información (porque se repiten las mediciones de las individuos pertenecientes a los paneles incluidos), pero, al mismo tiempo, se reduce el número de unidades de observación (porque el número de individuos de la muestra que coinciden en los períodos de interés es necesariamente inferior al número de individuos en la muestra de un corte transversal).

1. Creación de los pesos longitudinales iniciales

Este primer paso empieza con la definición de los períodos consecutivos que se utilizarán en la combinación de las bases de datos. Si la combinación se refiere a 2020, se debe tener en cuenta que hubo un cambio abrupto debido a las restricciones de movilidad impuestas para hacer frente a la pandemia de COVID-19, que a su vez determinó un cambio en la modalidad de recolección de los datos (de presencial a telefónica) a partir del segundo trimestre de ese año.

Una vez definidos los períodos de interés, se realiza la combinación de las correspondientes bases de datos transversales. Este procedimiento debe tener en cuenta únicamente las unidades muestrales que respondieron sistemáticamente en cada uno de los períodos de interés. En el caso de la combinación de dos períodos, solo se incluirán en la base de datos combinada las unidades que hayan respondido en ambos períodos. De lo contrario (si respondieron en el primer período, pero no en el segundo o viceversa), se las excluirá.

a) Pesos básicos

La determinación de los pesos iniciales está supeditada a los pesos básicos ajustados por cobertura $d_{1,k}$ del procesamiento transversal del primer trimestre que se quiere combinar. Por ejemplo, en el primer escenario del diagrama XVI.1, se desea combinar el primer trimestre con el segundo trimestre de 2020. En este caso, se partiría de los pesos básicos ajustados por cobertura del primer trimestre de 2020. En el segundo escenario, se combinan el segundo y el tercer trimestre de 2020, de manera que se partiría de los pesos básicos ajustados por cobertura y falta de respuesta del segundo trimestre de 2020.

En general, dado que cada panel es representativo del país y se supone que debe tener las mismas características en el momento de la selección, LaRoche (2003) plantea que los pesos básicos se crean a partir del inverso de la probabilidad de inclusión de los paneles. Así:

$$d_{1,k}^{básico} = \frac{d_{1,k}}{Pr(\text{selección de paneles})}$$

Al realizar la combinación de los dos primeros trimestres de 2020 en el ejemplo, es evidente que hay tres paneles coincidentes y, como la muestra transversal contiene cuatro paneles, $Pr(\text{selección de paneles}) = \frac{3}{4}$. En cambio, por las condiciones generadas por las medidas para hacer frente a la pandemia de COVID-19, cuando se realiza la combinación del segundo y el tercer trimestre en el ejemplo, se obtiene $Pr(\text{selección de paneles}) = \frac{4}{4} = 1$.

Cabe señalar que, al combinar paneles, la inferencia que se realiza está supeditada al período del primer panel. Además, en este paso es indispensable corroborar que la suma de los pesos básicos sea cercana al tamaño de la población que se quiere representar. Es decir, $\sum_{s(1)} d_{1,k}^{básico} \cong N$; donde $s^{(1)}$ se define como el conjunto de respondientes en el primer período que pertenecen a los paneles coincidentes en la muestra para los períodos combinados.

De la misma forma, conforme a la metodología de la Encuesta sobre la Dinámica del Trabajo y los Ingresos del Canadá (Naud, 2002; LaRoche, 2003), el primer paso para crear los pesos longitudinales es el ajuste por el inverso de la probabilidad de traslape.

b) Ajuste por falta de respuesta

A continuación, es necesario realizar un ajuste por falta de respuesta sobre los pesos básicos, que debería estar supeditado a las covariables disponibles en el marco de muestreo, los registros administrativos o, teniendo en cuenta el diseño de muestreo rotativo, en rondas anteriores de la misma encuesta. En general, es recomendable tener en cuenta el paradigma principal en el tratamiento de la falta de respuesta, según el cual las personas que responden y las que no lo hacen difieren en la mayoría de los casos. Por supuesto, las unidades que no respondieron deberán excluirse de la base de datos, puesto que su peso de muestreo es nulo; es decir, $d_{1,k}^{básico}=0,\ \forall k \notin s_r^{(1)}$, donde el conjunto $s_r^{(1)}$ corresponde a las unidades que respondieron a la encuesta en el primer período de la combinación.

En este diseño, es posible utilizar un enfoque basado en la estimación de las probabilidades de respuesta de cada individuo para ajustar los pesos básicos, para lo que se necesita establecer una relación entre las unidades que respondieron y no respondieron con las covariables auxiliares x_1. En otras palabras, es necesario asegurarse de que las covariables estén disponibles para todas las unidades seleccionadas en el primer período de interés, independientemente de su respuesta final. Para el tratamiento efectivo de la falta de respuesta, se consideran las variables dicotómicas $I_{1,k}$ y $D_{1,k}$, que indican si el hogar pertenece a la muestra del primer período y si respondió a la encuesta, respectivamente. La probabilidad de respuesta estará supeditada al siguiente modelo:

$$\phi_{1,k} = Pr(D_{1,k}=1 \mid I_{1,k}=1) = f(x_1, \beta)$$

En la notación anterior, el conjunto de respondientes efectivos está conformado por las unidades muestrales que respondieron en el primer período de interés. Además, la función de enlace f es, por lo general, no lineal y su elección depende del investigador. Por otra parte, si se decide utilizar un modelo de regresión logística, la estimación de las probabilidades de respuesta tendrá la siguiente forma:

$$\hat{\phi}_{1,k} = \frac{exp(x_1'\hat{\beta})}{1+exp(x_1'\hat{\beta})}$$

Una vez modelada la falta de respuesta, se ajustan los pesos básicos utilizando el inverso de la probabilidad de respuesta sobre los respondientes efectivos en el primer período de interés, a fin de conformar el primer conjunto de pesos iniciales de las bases de datos longitudinales:

$$d_{l,k}^{inicial} = \frac{d_{l,k}^{básico}}{\hat{\phi}_{l,k}}$$

Es posible que, al construir la matriz de covariables para ajustar el modelo de respuesta, se encuentren unidades que no respondieron en el primer período y sobre las que, además, se carece de información auxiliar, porque su panel rotativo no se traslapa. En este caso, es posible calcular la tasa de respuesta efectiva y utilizarla como valor imputado a la probabilidad de respuesta $\hat{\phi}_{l,k}$. También habrá unidades recién incorporadas al panel rotativo que, por lo tanto, no habrán respondido, y para las que no se dispondrá de información auxiliar. En este caso, es necesario imputarles el factor de expansión ajustado del hogar al que pertenecen.

Como se mencionó en el capítulo XII, es necesario verificar las propiedades de balanceo y dominio común en el modelo de puntaje de propensión. Se esperaría que la distribución de las probabilidades de respuesta para las combinaciones de los dos trimestres combinados mostrara un buen balance entre las personas que respondieron y las que no respondieron (distribuciones similares) y que el dominio común de la probabilidad de respuesta excluya el cero y el uno.

2. Creación de los pesos longitudinales finales

En este paso, después de crear los pesos longitudinales iniciales, se hacen algunos ajustes concernientes al período de combinación de las bases longitudinales y a la falta de respuesta entre estos períodos y, por último, se realiza la calibración final para generar los pesos definitivos de la base de datos longitudinal.

a) Definición de la población longitudinal

La población longitudinal está conformada por todas las unidades que permanecieron en la población de interés entre el primer y el segundo período. Por ejemplo, en el caso de la encuesta ejemplificada en este capítulo, la población longitudinal del primer semestre de 2020 incluiría a todas las personas que formaron parte de la población objetivo del primer período y permanecieron en la población hasta el segundo período, inclusive.

Por supuesto, es necesario tener en cuenta que entre ambos períodos pueden haber ocurrido cambios en la población; por ejemplo, que algunas personas hayan dejado de pertenecer a la población objetivo (por diversos motivos como la muerte, el reclutamiento militar, el internamiento en una institución o la migración, entre otros). Mientras que la población de interés en el segundo período incluye a las personas que han entrado a

formar parte de la población de interés desde el primer período (por motivos como el nacimiento, la migración o el alta de una institución, entre otros), este no es el caso de la población longitudinal.

Es en esta segunda instancia en la que se generan los pesos definitivos y se construye la base longitudinal que se utilizará para realizar los análisis de interés. En primer lugar, se define la muestra longitudinal $s^{(2)}$ como aquella constituida por las unidades seleccionadas en ambos períodos de interés para los paneles coincidentes, es decir, por la intersección de las muestras transversales del primer período s^1 y el segundo período s^2:

$$s^{(2)}=s^1 \cap s^2$$

La muestra $s^{(2)}$ es representativa de la población longitudinal en los dos períodos combinados. En esta etapa, el factor de expansión longitudinal se define como idéntico al peso resultante de la sección anterior, es decir $d_{2,k}^{básico}= d_{1,k}^{básico}$.

b) Falta de respuesta y desgaste de los respondientes

La creación de la base de datos longitudinal parte de los pesos iniciales establecidos en la sección anterior. Sin embargo, hay que tener en cuenta la falta de respuesta en los períodos de la combinación. En general, se distinguen tres subconjuntos de no respondientes: i) unidades que respondieron en el primer período, pero no en el segundo, ii) unidades que no respondieron en el primer período, pero sí en el segundo, iii) unidades que no respondieron en ninguno de los períodos. En cualquiera de estos casos, es necesario identificar estas unidades, a las cuales se asignará un peso longitudinal nulo, es decir:

$$d_{2,k}^{inicial} = \begin{cases} d_{1,k}^{inicial} & \forall k \in s_r^{(2)} \\ 0 & \forall\ k \notin s_r^{(2)} \end{cases}$$

Donde el conjunto $s_r^{(2)}= s_r^1 \cap s_r^2$ corresponde a las unidades que respondieron a la encuesta en los dos períodos de la combinación, es decir a todas las unidades que respondieron en s^1 y que también respondieron en s^2. El motivo de esta determinación es que, a los efectos prácticos de comparación entre períodos, los diseños de muestreo de las encuestas rotativas en la región permiten relativamente pocas combinaciones.

De la misma forma en que se realizó el ajuste en la sección anterior, es posible utilizar un enfoque basado en la estimación de las probabilidades de respuesta de cada persona para ajustar los pesos iniciales. Para ello, se requieren covariables auxiliares x_2 en el segundo período. Es así como se consideran las variables dicotómicas $I_{2,k}$ y $D_{2,k}$, que indican si la unidad pertenece a la muestra del segundo período y si respondió a la encuesta en el segundo período, respectivamente. La probabilidad de respuesta estará supeditada al siguiente modelo:

$$\phi_{2,k} = Pr\,(D_{2,k}=1 \mid I_{2,k}=1) = f(\boldsymbol{x}_2, \boldsymbol{\beta})$$

Una vez modelada la falta de respuesta, se ajustan los pesos longitudinales utilizando el inverso de la probabilidad de respuesta sobre los respondientes efectivos en el primer período de interés:

$$d_{2,k}^{longitudinal} = \frac{d_{2,k}^{inicial}}{\hat{\phi}_{2,k}}$$

c) Calibración de los pesos longitudinales

Después de realizar el ajuste por falta de respuesta, es aconsejable imponer algunas restricciones sobre los factores de expansión. En particular, se busca que la suma de los pesos reproduzca con exactitud los conteos poblacionales o las proyecciones demográficas del país, las regiones o los departamentos y de los subgrupos clasificados según la edad, el sexo y el lugar de residencia (zonas urbana y rural), entre otros. En general, en las restricciones de la calibración pueden considerarse variables a nivel tanto de la persona como del hogar. Es importante destacar que los totales auxiliares utilizados en la calibración deben representar la población del primer período de interés, puesto que, al conformar un panel que no suma elementos a lo largo de los períodos de medición, la muestra será representativa únicamente del período en el que fue seleccionada. Teniendo en cuenta que las variables de calibración están representadas por el vector z_k y que sus totales poblacionales están disponibles en forma de proyecciones poblacionales, este conjunto de restricciones sobre los nuevos pesos longitudinales calibrados $w_{2,k}^{calibrado}$ se puede escribir como:

$$\sum_{s_r^{(2)}} w_{2,k}^{calibrado} z_k = \sum_{U} z_k$$

Por lo tanto, los pesos finales que deberían incluirse en la base de datos longitudinal de los dos trimestres combinados estarían dados por $w_{2,k}^{calibrado}$, y pueden escribirse de la siguiente manera:

$$w_{2,k}^{calibrado} = g_k * d_{2,k}^{longitudinal}$$

Donde los ponderadores g_k dependen de la muestra traslapada y representan la cercanía de los pesos longitudinales finales calibrados con respecto a los pesos longitudinales sin calibrar. En general, se esperaría que estos valores estuvieran cercanos a la unidad.

C. Creación de bases de datos longitudinales anuales

En esta sección se describen los pasos necesarios para combinar bases de datos longitudinales que permitan el seguimiento de la situación de los hogares a lo largo de todo un año. La metodología recomendada es una generalización de los pasos descritos por Verma, Betti y Ghellini (2006), que definen un procedimiento secuencial para la creación de los factores de expansión en el panel. Este procedimiento sigue exactamente los mismos pasos mencionados con respecto a la creación de bases longitudinales para dos períodos consecutivos. Es decir, en primer lugar es necesario crear el conjunto de pesos iniciales (transversales en el primer período), y luego definir los pesos finales (longitudinales en los cuatro trimestres de todo un año).

El primer paso en la generación de los pesos longitudinales consiste en realizar una consolidación de bases de datos, en la que se combinen únicamente los períodos de interés. Siguiendo con el ejemplo del diseño 4(0)1, correspondería a los cuatro trimestres del año. Para ello, es necesario filtrar cada una de las bases transversales con el identificador del panel de interés. De esta forma se obtendrán cuatro bases de datos que contendrán solamente la información de los paneles comunes.

La determinación de los pesos iniciales está supeditada a los pesos básicos ajustados por cobertura $d_{1,k}$ del procesamiento transversal del primer trimestre que se quiere combinar. En general, los pesos básicos se crean a partir del inverso de la probabilidad de inclusión de los paneles, puesto que, al realizar la combinación de los cuatro trimestres en un diseño 4(0)1, es evidente que habrá solo un panel coincidente y, como la muestra transversal contiene cuatro paneles, $Pr(\text{selección de paneles}) = 1/4$. En resumen:

$$d_{1,k}^{básico} = \frac{d_{(1,k)}}{Pr(\text{selección de paneles})} = 4 \times d_{(1,k)}$$

Como se señaló anteriormente, se debe seguir un riguroso proceso de identificación secuencial de personas que respondieron y no respondieron para poder realizar la combinación de las bases de datos transversales correspondientes. En este procedimiento, solo se deben tener en cuenta las unidades muestrales que respondieron sistemáticamente en cada uno de los períodos de interés. Por lo tanto, se recomienda seguir los siguientes pasos:

i) Trimestres T1 y T2:
   - Identificación de las unidades que respondieron en T1 y T2.
   - Identificación de las unidades que respondieron en T1 pero no en T2.

ii) Trimestres T1, T2 y T3:
   - Identificación de las unidades que respondieron en T1, T2 y T3.
   - Identificación de las unidades que respondieron en T1 y T2, pero no en T3.

iii) Trimestres T1, T2, T3 y T4:

- Identificación de las unidades que respondieron en T1, T2, T3 y T4.
- Identificación de las unidades que respondieron en T1, T2 y T3, pero no en T4.

En esta instancia se construye la base longitudinal que se utilizará para realizar los análisis de interés. En primer lugar, se define la muestra longitudinal $s^{(1234)}$ como aquella constituida por las unidades seleccionadas en ambos períodos de interés para los paneles coincidentes:

$$s^{(1234)} = s^1 \cap s^2 \cap s^3 \cap s^4$$

La muestra $s^{(1234)}$ es representativa de la población longitudinal en los períodos combinados. En esta etapa, el factor de expansión longitudinal inicial se define como idéntico al peso resultante de la sección anterior, es decir:

$$d^{inicial}_{(1234,k)} = d^{inicial}_{(1,k)}$$

Es necesario identificar las unidades que no respondieron en ninguna ocasión para asignarles un peso longitudinal nulo; es decir, $d^{inicial}_{(1234,k)} = 0$ para aquellas unidades $k \notin s_r^{(1234)}$, donde el conjunto $s_r^{(1234)}$ corresponde a las unidades que respondieron a la encuesta en los cuatro trimestres de la combinación. Sin embargo, las unidades que no respondieron en ninguna ocasión se utilizarán para ajustar el modelo de puntaje de propensión, antes de excluirlas totalmente de la base de datos (puesto que su peso de muestreo es nulo).

El paso siguiente, tras la identificación de respondientes y no respondientes a lo largo del año, consiste en realizar el ajuste por falta de respuesta, que debería estar supeditado a las covariables disponibles en el marco de muestreo o en rondas anteriores de la misma encuesta. Como se ha indicado a lo largo de este documento, se recomienda utilizar un enfoque basado en la estimación de la propensión de respuesta de cada individuo para ajustar los pesos básicos. La probabilidad de respuesta estará determinada por el siguiente modelo:

$$\phi_{(1234,k)} = Pr\,(D_{(1234,k)} = 1 \mid I_{(1234,k)} = 1) = f(x, \beta)$$

En la notación anterior, el conjunto de respondientes efectivos está conformado por las unidades muestrales que respondieron en todos los períodos de interés. Los pesos básicos se ajustan utilizando el inverso de la probabilidad estimada de respuesta sobre los respondientes efectivos en el primer período de interés, y así se conforma el primer conjunto de pesos iniciales de las bases de datos longitudinales. Por último, se debe corroborar que la suma de los pesos ajustados por la falta de respuesta sea cercana al tamaño de la población que se quiere representar.

Tras el ajuste por falta de respuesta, el proceso termina con la calibración final, en la que se imponen algunas restricciones sobre los factores de expansión finales. En particular, se busca que la suma de los pesos reproduzca con exactitud los conteos poblacionales o las proyecciones demográficas.

Cabe destacar que la base de datos longitudinal representa con exactitud el conjunto de individuos comunes en los períodos de interés. Por consiguiente, las estimaciones transversales que se hagan a partir de esta base (por ejemplo, de la pobreza en un determinado trimestre) solo deben tomarse como referencia, porque no reemplazarán a las estimaciones transversales ya publicadas. La población de interés de la encuesta transversal no es la misma que la de la combinación (panel) y, por ende, estas estimaciones no coincidirán y no deberían coincidir.

La utilidad de crear bases de datos longitudinales radica sobre todo en la necesidad de estimar parámetros de cambio como flujos brutos, el error de muestreo asociado y los correspondientes intervalos de confianza. La ganancia en el análisis es muy grande cuando se conocen las estimaciones entre estados de una variable de interés. Por ejemplo, los usuarios de la encuesta pueden estar interesados en analizar el cambio bruto (flujos) entre diferentes estados de la fuerza de trabajo, como la situación, en el trimestre T2, de las personas que estaban ocupadas en el trimestre T1. Este tipo de análisis se traduce comúnmente en la estimación de matrices de transición, que se tratan en el capítulo XVII.

Capítulo XVII
Análisis de flujos brutos y matrices de transición

Existen ventajas significativas a la hora de analizar datos que provienen de encuestas longitudinales. Lynn (2009) menciona que la más conocida es la posibilidad de realizar análisis de los flujos brutos. A partir de bases de datos longitudinales, es posible conocer el estado de una unidad de observación en varios períodos consecutivos, lo que permite descomponer los cambios netos que se podrían encontrar en las encuestas transversales. Por ejemplo, a partir de recopilaciones de datos transversales (puntuales en un trimestre), en una encuesta con diseño rotativo es posible estimar el cambio en las diferentes clasificaciones del mercado de trabajo de un período a otro. Si se utiliza la parte longitudinal de la encuesta en ambos períodos, es posible descomponer estas clasificaciones. A partir de este análisis se puede determinar si son exactamente los mismos ocupados los que entran al mercado de trabajo en dos ciclos económicos de interés.

Un problema habitual en este tipo de encuesta es la falta de respuesta, que rara vez se puede considerar aleatoria o ignorable. Si la falta de respuesta depende de la situación laboral (por ejemplo, las tasas de falta de respuesta son más elevadas entre los desempleados), las estimaciones ingenuas de los cambios brutos podrían estar sesgadas. Si el enfoque corrige la falta de respuesta, pero no tiene en cuenta las complejas características del diseño de muestreo, también habrá problemas de sesgo.

A. Matrices de transición

La inferencia realizada en la estimación de parámetros mediante encuestas de hogares depende del diseño de muestreo utilizado para seleccionar una muestra de elementos de la población. Por lo general, el diseño de muestreo que se usa para la obtención de estadísticas oficiales no es simple, ya que, por la complejidad de la estructura poblacional, es necesario recurrir a técnicas especializadas que permitan conseguir de manera adecuada la información de la muestra. En muchas ocasiones, el diseño de muestreo complejo de las encuestas de hogares está supeditado a la estratificación de las unidades cartográficas del marco de muestreo y a la posterior selección en múltiples etapas de las unidades. Por lo general, las probabilidades de inclusión son desiguales y proporcionales al número de habitantes de las áreas en que se divide el marco de muestreo.

Después de haber seleccionado una muestra probabilística, es común recurrir a la clasificación de los elementos de la población en diferentes categorías de una o más variables nominales. Dicha clasificación se puede convertir en una tabla de contingencia si se resumen simultáneamente dos variables o los cambios en el comportamiento de una misma variable en dos períodos de tiempo. Tal es el caso de la estimación de los cambios brutos entre diferentes estados de clasificación para dos períodos. En particular, este problema de estimación es fundamental para la implementación de políticas públicas en el mercado laboral.

Como ya se ha discutido en profundidad, en muchas encuestas por muestreo a gran escala de la región se utilizan diseños de panel rotativo, en que los individuos son entrevistados varias veces antes de quedar excluidos de la muestra. Estas encuestas a gran escala se utilizan para producir estimaciones puntuales en el tiempo de manera continua, mensual y trimestral. Su estructura de seguimiento en panel nace de la necesidad de suavizar los cambios coyunturales que podría experimentar el fenómeno de interés. Además, contribuye a reducir los costos, al mantenerse los mismos entrevistados y objetivos en más de una entrevista.

Debido a los efectos de la pandemia de enfermedad por coronavirus (COVID-19) y, en particular, a las diferentes restricciones de movilidad que aplicaron los Gobiernos de los países de la región (como cuarentenas obligatorias y toques de queda), la forma de recopilar la información de las encuestas de hogares tuvo que ser modificada. Se cambiaron abruptamente los operativos de recolección presenciales por operativos telefónicos que conllevaron tasas de respuesta bajas. Estas, a su vez, dieron lugar a nuevos procesos estadísticos de ajuste de los factores de expansión con el fin de detectar y eliminar los posibles sesgos que hubieran podido introducirse.

Por lo tanto, tiene sentido considerar la posibilidad de que, en estos tipos de operativos de recolección de datos, la falta de respuesta no sea completamente aleatoria. De hecho, es muy probable que la falta de respuesta dependa de la situación laboral de los encuestados (por ejemplo, que los desempleados conformen la mayor parte de los no respondientes) y que las estimaciones de los cambios brutos deban corregirse para evitar posibles sesgos.

B. Modelos de Markov

Al considerar el problema de la estimación de los cambios brutos entre dos períodos de tiempo sobre la población de interés, y teniendo en cuenta que existe una falta de respuesta no ignorable, cabe suponer que el resultado de cada entrevista corresponde a la clasificación del respondiente en una de G posibles categorías excluyentes. Así pues, uno de los objetivos de la investigación podría ser la estimación del cambio bruto entre estas categorías, utilizando la información de los individuos que fueron entrevistados en dos períodos de tiempo consecutivos. En el cuadro XVII.1 se ejemplifica la distribución (no observable) de los flujos brutos en una población.

■ Cuadro XVII.1
Distribución no observable de los flujos brutos en una población

Estado (T1/T2)	Estado 1	Estado 2	⋯	Estado G
Estado 1	X_{11}	X_{12}	⋯	X_{1G}
Estado 2	X_{21}	X_{22}	⋯	X_{2G}
⋮	⋮	⋮	⋱	⋮
Estado G	X_{G1}	X_{G2}	⋯	X_{GG}

Fuente: Elaboración propia.

Donde X_{ij} es el número de unidades en la población finita clasificadas como i en el período de tiempo $t\text{-}1$ y j en el período de tiempo $t\ (i,j=1,\ldots,G)$. Siguiendo las consideraciones de Feinberg y Stasny (1983), se supone que los datos son el resultado de un proceso de dos etapas. En la primera etapa (proceso no observable), los individuos son ubicados dentro de las celdas de una matriz $G \times G$ de acuerdo con las probabilidades de una cadena de Markov, con los siguientes parámetros:

i) η_i, la probabilidad inicial de que un individuo esté en el estado i en el período de tiempo $t\text{-}1$.

ii) p_{ij}, la probabilidad de transición desde el estado i al estado j.

En esta primera etapa, los parámetros deben cumplir con las siguientes restricciones: $\sum_i \eta_i = 1$ y $\sum_j p_{ij} = 1$ para todo i.

Está claro que, una vez que se realice la encuesta y se obtengan los datos recolectados en ambos períodos, los individuos que fueron no respondientes en uno o ambos períodos o fueron incluidos o excluidos de la muestra entre estos dos momentos no tendrán una clasificación establecida entre las distintas categorías o estados de la variable de interés, puesto que para ser clasificado es necesario proporcionar una respuesta en ambos tiempos. De esta forma, se tiene un grupo de individuos clasificado en ambos períodos, un grupo de individuos que solo está clasificado en uno de los dos períodos y un grupo de individuos que no respondieron la encuesta en ningún momento y, por ende, nunca fueron clasificados.

En el caso del primer grupo de individuos, que respondieron en los períodos de tiempo $t\text{-}1$ y t, los datos de clasificación pueden resumirse en una matriz de tamaño $G \times G$. La información disponible sobre los individuos que fueron no respondientes en la encuesta del período de tiempo $t\text{-}1$, pero sí respondieron en el período de tiempo t, puede resumirse en un complemento columna. Por su parte, la información sobre los individuos que no respondieron en el período de tiempo t, pero sí respondieron en el período de tiempo t-1, puede resumirse en un complemento fila. Finalmente, los individuos que no respondieron en ningún momento son incluidos en una única celda de faltantes.

Las relaciones anteriores se ilustran en el cuadro XVII.2, donde N_{ij} $(i,j=1,\ldots,G;\ G=4)$ denota el número de individuos respondientes que tienen clasificación i en el período de tiempo $t\text{-}1$ y j en el período de tiempo t, R_i denota el número de individuos que fueron no respondientes en el período de tiempo t y tienen clasificación i en el período de tiempo $t\text{-}1$, C_j denota el número de individuos que fueron no respondientes en el período de tiempo $t\text{-}1$ y tuvieron clasificación j en el período de tiempo t, y M denota el número de individuos seleccionados que no respondieron en ningún momento. En este estudio particular, se consideran cuatro estados de clasificación (ocupados formales, ocupados informales, desocupados e inactivos (por ende $G=4$) en dos períodos consecutivos de tiempo $t\text{-}1$ (primer trimestre de 2020) y t (segundo trimestre de 2020).

■ Cuadro XVII.2

Distribución observable de los flujos brutos sobre la situación laboral en la población con falta de respuesta en ambos períodos

Estado (T1/T2)	Formal	Informal	Desocupado	Inactivo	Complemento fila
Formal	N_{11}	N_{12}	N_{13}	N_{14}	R_1
Informal	N_{21}	N_{22}	N_{23}	N_{24}	R_2
Desocupado	N_{31}	N_{32}	N_{33}	N_{34}	R_3
Inactivo	N_{41}	N_{42}	N_{43}	N_{44}	R_4
Complemento columna	C_1	C_2	C_3	C_4	M

Fuente: Elaboración propia.

Nota: N: número de individuos respondientes en ambos períodos; R: número de individuos no respondientes en el primer período; C: número de individuos no respondientes en el segundo período; M: número de individuos no respondientes en ningún período.

En la segunda etapa (proceso observable), cada individuo en la celda ij de la matriz puede ser no respondiente en el período $t\text{-}1$ y perder la clasificación por fila, o ser no respondiente en el período t y perder la clasificación por columna, o bien ser no respondiente en ambos períodos y perder ambas clasificaciones. En consecuencia, se genera una estructura probabilística con los siguientes parámetros:

i) $\psi(i,j)$, la probabilidad inicial de que un individuo en la celda ij responda en el período $t\text{-}1$.

ii) $\rho_{RR}(i,j)$, la probabilidad de transición de que un individuo en la celda ij pase de ser respondiente en el período $t\text{-}1$ a ser respondiente en el período t.

iii) $\rho_{MM}(i,j)$, la probabilidad de transición de que un individuo en la celda ij pase de ser no respondiente en el período $t\text{-}1$ a ser no respondiente en el período t.

Como se puede deducir, las probabilidades del proceso observables dependen del estado de clasificación del individuo. Para estimar todos los parámetros pertinentes, se consideraron los modelos reducidos que se explican a continuación:

- Modelo A: en este modelo se considera que la probabilidad inicial de que un individuo sea respondiente en el período $t\text{-}1$ es la misma respecto de todas las clasificaciones de la encuesta, es decir, $\psi(i,j) = \psi$. Las probabilidades de transición entre respondientes y entre no respondientes no dependen de la clasificación del individuo en la encuesta, es decir, $\rho_{MM}(i,j) = \rho_{MM}$ y $\rho_{RR}(i,j) = \rho_{RR}$. Esto significa que la probabilidad de transición entre respondientes es la misma para formales, informales, inactivos y desocupados. Asimismo, las probabilidades de respuesta se consideran idénticas para las diferentes clasificaciones.
- Modelo B: en este modelo se considera que la probabilidad inicial de que un individuo sea respondiente en el período $t\text{-}1$ depende de su clasificación en el período $t\text{-}1$, es decir, $\psi(i,j) = \psi(i)$. De la misma manera que en el modelo A, las probabilidades de transición entre respondientes y entre no respondientes no dependerán de la clasificación del individuo en la encuesta, es decir, $\rho_{MM}(i,j) = \rho_{MM}$ y $\rho_{RR}(i,j) = \rho_{RR}$. Esto significa que la probabilidad de respuesta difiere entre formales, informales, inactivos y desocupados, mientras que la probabilidad de tránsito entre respondientes es la misma.
- Modelo C: en este modelo se parte del supuesto de que la probabilidad inicial de que un individuo sea respondiente en el período $t\text{-}1$ es la misma para todas las clasificaciones incluidas en la encuesta, es decir, $\psi(i,j) = \psi$. Sin embargo, las probabilidades de transición entre respondientes y entre no respondientes dependerán de la clasificación del individuo en el período $t\text{-}1$; es decir, $\rho_{MM}(i,j) = \rho_{MM}(i)$ y $\rho_{RR}(i,j) = \rho_{RR}(i)$. Es decir, la probabilidad de transición entre respondientes es la misma para formales, informales, inactivos y desocupados. Asimismo, las probabilidades de transición difieren para las diferentes clasificaciones en el período de tiempo inicial.
- Modelo D: en este modelo se parte del supuesto de que la probabilidad inicial de que un individuo sea respondiente en el período $t\text{-}1$ es la misma para todas las clasificaciones incluidas en la encuesta, es decir, $\psi(i,j) = \psi$. Sin embargo, las probabilidades de transición entre respondientes y entre no respondientes dependerán de la clasificación del individuo en el período t; es decir, $\rho_{MM}(i,j) = \rho_{MM}(j)$ y $\rho_{RR}(i,j) = \rho_{RR}(j)$. Esto significa que la probabilidad de respuesta es la misma para formales, informales, inactivos y desocupados. Asimismo, la probabilidad de tránsito entre respondientes difiere para las diferentes clasificaciones en el período de tiempo final.

Como ampliación de las ideas de Feinberg y Stasny (1983), Gutiérrez (2014) utilizó una metodología de estimación basada en el enfoque de máxima seudoverosimilitud para estimar los parámetros anteriores. El objetivo final del proceso es estimar el número de individuos en las celdas de una tabla de contingencia poblacional, en la que estos se clasifican según la situación laboral, medida en dos puntos de tiempo diferentes con un diseño de muestreo complejo. En resumen, el ajuste de los modelos de falta de respuesta puede realizarse siguiendo los algoritmos de estimación propuestos en Gutiérrez (2014) y utilizando el paquete computacional *SuRF* del *software* estadístico R (Jacob, 2020).

C. Estimación de las matrices de transición

Considérese que se dispone de información sobre la situación laboral de 41.274 personas en los dos primeros trimestres de 2020 (año en que la pandemia de COVID-19 produjo alteraciones importantes en los procesos de recolección regular de datos de las encuestas y en el mercado de trabajo). La clasificación de la situación laboral en la muestra puede observarse en el cuadro XVII.3, que contiene los complementos de columna y de fila para las personas que no respondieron en alguno de los dos períodos. Nótese que los valores del complemento fila son mucho mayores que los del complemento columna, lo que puede deberse a los efectos de la pandemia. Además, la suma de todas las entradas de la clasificación reflejada en el cuadro XVII.3 da como resultado el número de personas en la muestra traslapada, es decir, 41.274 individuos.

■ Cuadro XVII.3
Distribución observada de los flujos brutos sobre la situación laboral en la muestra no ponderada con falta de respuesta en dos períodos de tiempo

Estado (T1/T2)	Formal	Informal	Desocupado	Inactivo	Complemento fila
Formal	11 483	718	592	1 828	451
Informal	703	2 513	495	2 769	198
Desocupado	191	181	503	794	81
Inactivo	364	641	388	15 386	382
Complemento columna	160	65	48	257	83

Fuente: Elaboración propia.

Como se supone que la muestra proviene de un muestreo complejo sobre la población del país, para utilizar el método de estimación propuesto en esta sección se necesita primero estimar el tamaño de la población en cada celda del cuadro XVII.3. Estas estimaciones se proporcionan en el cuadro XVII.4, donde el tamaño de población estimado es de 15.597.572.

■ Cuadro XVII.4
Distribución poblacional estimada de los flujos brutos sobre la situación laboral con falta de respuesta en dos períodos de tiempo

Estado (T1/T2)	Formal	Informal	Desocupado	Inactivo	Complemento fila
Formal	3 269 673	201 639	175 719	503 740	155 902
Informal	232 095	641 565	146 416	725 006	58 649
Desocupado	55 243	50 337	157 642	233 597	26 695
Inactivo	102 490	161 363	98 898	4 299 066	118 393
Complemento columna	47 104	26 276	19 746	100 775	25 545

Fuente: Elaboración propia.

La elección del mejor modelo se llevó a cabo utilizando la prueba estadística de ajuste χ^2, calculada sobre la distancia entre los valores observados y los valores predichos por el modelo. En estos términos, el mejor modelo fue el C, puesto que minimizaba esta distancia con un valor de $\chi^2 = 0{,}706$. La distribución nula del estadístico es *ji* al cuadrado con G^2-D grados de libertad, donde D indica el número de parámetros estimados. Respecto del conjunto de datos considerado en este documento, en el cuadro XVII.5 se presentan los valores críticos de la distribución nula y los valores del estadístico de prueba para cada modelo. Se concluye que el modelo más apropiado para el conjunto de datos es el modelo C.

■ Cuadro XVII.5
Medidas de resumen sobre el ajuste de los cuatro modelos considerados en la estimación de los flujos brutos sobre la situación laboral

	Modelo A	Modelo B	Modelo C	Modelo D
Grados de libertad	7	4	1	1
Valor crítico	14,07	9,49	3,84	3,84
Valor χ^2_{RS}	15,6706	18,3659	0,2418	3,9137

Fuente: Elaboración propia.
Nota: El valor χ^2_{RS} corresponde al percentil teórico de la distribución ji al cuadrado.

Recuérdese que en este modelo se considera que la probabilidad inicial de que un individuo sea respondiente en el primer trimestre de 2020 es la misma para todas las clasificaciones incluidas en la encuesta, es decir, $\psi(i,j) = \psi$. Sin embargo, las probabilidades de transición entre respondientes y no respondientes dependerán de la clasificación del individuo en el primer trimestre de 2020; es decir, $\rho_{MM}(i,j) = \rho_{MM}(i)$ y $\rho_{RR}(i,j) = \rho_{RR}(i)$. Partiendo de estos supuestos, en el cuadro XVII.6 se ilustra la estimación poblacional, insesgada con respecto al diseño de muestreo complejo de la encuesta, de los cambios brutos en la situación laboral.

■ Cuadro XVII.6
Distribución poblacional estimada de los flujos brutos sobre la situación laboral para el proceso no observable (sin falta de respuesta) en dos períodos con el modelo C

Estado (T1/T2)	Formal	Informal	Desocupado	Inactivo
Formal	4 627 632	287 124	252 084	713 284
	(102 470)	(18 979)	(20 399)	(30 358)
Informal	327 066	911 996	210 372	1 022 351
	(33 292)	(45 645)	(16 508)	(50 332)
Desocupado	79 346	72 944	230 949	335 746
	(10 592)	(8 858)	(21 180)	(28 023)
Inactivo	143 303	227 545	140 923	6 014 907
	(11 192)	(15 849)	(10 550)	(123 559)

Fuente: Elaboración propia.
Nota: Los errores estándar se muestran entre paréntesis.

En el modelo C se considera que los parámetros de la primera etapa del proceso (no observable) se definen como las probabilidades de transición de una clasificación a otra en los períodos de observación. Estas estimaciones definen las matrices de transición laboral, que, en los períodos estudiados, corresponden a las entradas del cuadro XVII.7. En particular, se resalta el hecho de que el 12,1% de los trabajadores formales pasaron directamente a la inactividad. Mientras tanto, el cambio fue mayor entre los trabajadores informales y los desocupados, de los cuales el 41,3% y el 46,6%, respectivamente, pasaron a la inactividad. Además, en los períodos estudiados, el 92,1% de los individuos inactivos siguió en este estado.

■ Cuadro XVII.7
Estimación de las matrices de transición sobre la situación laboral en dos períodos con el modelo C

Estado (T1/T2)	Formal	Informal	Desocupado	Inactivo
1,058 mm	0,787 (0,010)	0,048 (0,003)	0,042 (0,003)	0,121 (0,004)
Informal	0,132 (0,012)	0,368 (0,013)	0,085 (0,005)	0,413 (0,013)
Desocupado	0,110 (0,012)	0,101 (0,010)	0,321 (0,021)	0,466 (0,025)
Inactivo	0,021 (0,001)	0,034 (0,002)	0,021 (0,001)	0,921 (0,010)

Fuente: Elaboración propia.
Nota: Los errores estándar se muestran entre paréntesis.

Por otro lado, las probabilidades iniciales de clasificación en el primer período de interés se encuentran en el cuadro XVII.8, donde también se observan las probabilidades de transición de los no respondientes y de los respondientes, diferenciadas por situación laboral en el primer trimestre de 2020. Nótese que la probabilidad inicial de respuesta se estimó en $\hat{\psi} = 0,981(0,002)$ respecto de todas las clasificaciones de la situación laboral. Se puede apreciar que, si $1\text{-}\hat{\rho}_{MM}$ indica la probabilidad de que un individuo responda en

el segundo trimestre de 2020 dado que no respondió en el primer trimestre de 2020, en función de cada situación laboral, ello significa que las personas en situación de informalidad e inactivas son más propensas a no responder en ambos períodos. En consecuencia, habría indicios de un patrón de falta de respuesta no ignorable que el modelo ha logrado identificar correctamente.

■ Cuadro XVII.8
Estimación de los demás parámetros del modelo C

Situación laboral (primer trimestre de 2020)	$\hat{\eta}$	$\hat{\rho}_{RR}$	$\hat{\rho}_{MM}$
Formal	0,787 (0,010)	0,048 (0,003)	0,042 (0,003)
Informal	0,132 (0,012)	0,368 (0,013)	0,085 (0,005)
Desocupado	0,110 (0,012)	0,101 (0,010)	0,321 (0,021)
Inactivo	0,021 (0,001)	0,034 (0,002)	0,021 (0,001)

Fuente: Elaboración propia.
Nota: Los errores estándar se muestran entre paréntesis; $\hat{\eta}$ corresponde a la probabilidad inicial de que un individuo esté en un determinado estado de la situación laboral; $\hat{\rho}_{RR}$ es la probabilidad de transición de ser respondiente en ambos períodos; $\hat{\rho}_{MM}$ es la probabilidad de transición de ser no respondiente en ambos períodos.

Por último, sería deseable establecer si existen diferencias importantes en el impacto que la pandemia causó en el mercado laboral entre hombres y mujeres. Para realizar estas comparaciones, se ajustó el modelo C en cada una de las subpoblaciones y se encontraron ajustes precisos y satisfactorios con $\chi^2_{hombres}=0,350$ y $\chi^2_{hombres}=0,470$. La estimación de la probabilidad inicial de respuesta en ambos casos se calculó como $\hat{\psi}=0,981(0,002)$. Las estimaciones de las probabilidades $\hat{\eta}$ y $\hat{\rho}_{RR}$ no presentaron cambios significativos entre un grupo y otro. Sin embargo, la estimación de las probabilidades $\hat{\rho}_{MM}$ mostró diferencias entre los hombres y las mujeres que se clasificaron como trabajadores formales e inactivos en el primer trimestre de 2020. En particular, en el caso del grupo de trabajadores formales, $\hat{\rho}_{MM}^{hombre}=0,253$, mientras que $\hat{\rho}_{MM}^{mujer}=0,331$, lo que indica que las mujeres en situación de formalidad fueron más propensas a no responder en el segundo trimestre de 2020, en comparación con los hombres. Por otro lado, en el grupo de inactivos, $\hat{\rho}_{MM}^{hombre}=0,112$, mientras que $\hat{\rho}_{MM}^{mujer}=0,000$, lo que indica que las mujeres desocupadas definitivamente presentaron una menor probabilidad que los hombres de no responder en el segundo trimestre de 2020.

La estimación de las matrices de transición en ambos subgrupos poblacionales se muestra en los cuadros XVII.9 y XVII.10. Es importante resaltar que, según las estimaciones resultantes de este análisis, la peor parte se la llevaron las mujeres. Se observa que, en el caso de los trabajadores formales, existe una diferencia importante entre hombres y mujeres, siendo los primeros los menos afectados al pasar a la inactividad (un 10,2% en el caso de los hombres y un 14,5% en el de las mujeres). Este mismo fenómeno se presenta de manera más notoria en el grupo de los trabajadores informales, donde el 36,2% de los hombres y el 46,9% de las mujeres pasaron a la inactividad. De la misma forma, entre los individuos desocupados, un mayor porcentaje de mujeres pasó de la desocupación a la

inactividad (un 39,2% de hombres frente a un 53,4% de mujeres). Por último, el porcentaje de personas que siguió en la inactividad es mayor entre las mujeres (un 93,1% frente al 90,2% en el caso de los hombres).

■ Cuadro XVII.9
Estimación de las matrices de transición laboral en el caso de los hombres con el modelo C

p_{ij}	Formal	Informal	Desocupado	Inactivo
Formal	0,791 (0,013)	0,056 (0,004)	0,049 (0,005)	0,102 (0,005)
Informal	0,138 (0,011)	0,403 (0,018)	0,095 (0,008)	0,362 (0,017)
Desocupado	0,146 (0,025)	0,118 (0,017)	0,343 (0,031)	0,392 (0,037)
Inactivo	0,031 (0,004)	0,036 (0,003)	0,030 (0,003)	0,902 (0,021)

Fuente: Elaboración propia.
Nota: Los errores estándar se muestran entre paréntesis; p_{ij} corresponde a la probabilidad de transición entre los cuatro estados de interés sobre la situación laboral.

■ Cuadro XVII.10
Estimación de las matrices de transición laboral en el caso de las mujeres con el modelo C

p_{ij}	Formal	Informal	Desocupado	Inactivo
Formal	0,781 (0,010)	0,039 (0,003)	0,033 (0,003)	0,145 (0,007)
Informal	0,125 (0,012)	0,331 (0,017)	0,073 (0,007)	0,469 (0,018)
Desocupado	0,078 (0,012)	0,086 (0,012)	0,301 (0,028)	0,534 (0,036)
Inactivo	0,017 (0,001)	0,034 (0,002)	0,017 (0,001)	0,931 (0,013)

Fuente: Elaboración propia.
Nota: Los errores estándar se muestran entre paréntesis; p_{ij} corresponde a la probabilidad de transición entre los cuatro estados de interés sobre la situación laboral.

Capítulo XVIII
Criterios de calidad y difusión

Como se ha explicado a lo largo de este documento, las encuestas de hogares de la región presentan un diseño complejo, probabilístico, estratificado, en múltiples etapas y con probabilidades de inclusión no uniformes. Por consiguiente, las estimaciones elaboradas a partir de estas operaciones estadísticas están sujetas al error muestral, y es preciso evaluar su validez estadística mediante diversos indicadores de calidad que describan su precisión y confiabilidad y que, a su vez, alerten al usuario cuando la precisión de la estimación no sea confiable. Una vez obtenido el indicador de interés (por ejemplo, la proporción de personas en situación de pobreza y de pobreza extrema), se estiman los intervalos de confianza y otros indicadores de calidad, sobre la base de la información del diseño muestral complejo, resumida en el factor de expansión, los estratos y las unidades primarias de muestreo (UPM).

A partir de las encuestas de hogares se estiman una gran cantidad de indicadores sociales. En esta sección, se plantea la necesidad de definir una metodología de procesamiento de las bases de datos de las encuestas de hogares en los institutos nacionales de estadística (INE) de la región que permita a los analistas, investigadores y usuarios de dichas bases decidir acerca de la pertinencia de una estimación sobre la base de algunos criterios de calidad que resumen la precisión de las estimaciones. En cualquier encuesta, existirán estimaciones para interacciones entre variables categóricas derivadas del procesamiento de las bases de datos que, por su baja calidad estadística, no se deben tener en cuenta a la hora de tomar decisiones de política pública. Esta afirmación se realiza simplemente porque las encuestas no se elaboran teniendo en cuenta todas las maneras posibles que existen de procesarlas, sino que son el resultado de una planeación rigurosa sobre los indicadores más relevantes del estudio.

Cabe subrayar que la relación entre precisión y diseño va en una sola dirección: si la precisión de un estimador es deficiente, significa que el indicador no se consideró como relevante en el momento de diseñar la encuesta. Sin embargo, en muchas ocasiones, a pesar

de no haberse tenido en cuenta desde un inicio en el diseño de la encuesta, una estimación puede ser considerada precisa y confiable. En este documento se definen y establecen algunos criterios que aplican actualmente algunas oficinas nacionales de estadística y entidades dedicadas a la investigación social para evaluar dicha precisión y confiabilidad.

A. Medidas de calidad

Los criterios que aparecen en esta sección pueden tenerse en cuenta para determinar si una estadística se debe considerar precisa y confiable.

1. Intervalos de confianza

En general, la precisión de una estadística se debe estudiar a la luz del intervalo de confianza generado por la medida de probabilidad asociada al diseño de muestreo de la encuesta. Por ejemplo, si el parámetro de interés sobre el cual se busca realizar la inferencia es θ, y se ha definido una subpoblación de interés, entonces un intervalo del 95% de confianza sobre esa subpoblación estará dado por la siguiente expresión (Heeringa, West y Berglund, 2010):

$$\left(\hat{\theta} - t_{0,975,gl} * se(\hat{\theta}),\ \hat{\theta} + t_{0,975,gl} * se(\hat{\theta})\right)$$

Donde $\hat{\theta}$ es un estimador por muestreo para el parámetro de interés θ y $t_{0,975,gl}$ es el percentil 0,975 de una distribución t de Student con gl grados de libertad, que se determinan al restar del número de UPM seleccionadas el número de estratos de muestreo considerados. A su vez, $se(\hat{\theta})$ es el error estándar de la estimación, definido por la raíz cuadrada de la varianza estimada del estimador; es decir:

$$se(\hat{\theta}) = \sqrt{\widehat{Var}(\hat{\theta})}$$

En el caso particular de las proporciones, los intervalos de confianza deben estar contenidos dentro del intervalo $(0,1)$. Sin embargo, en algunas ocasiones puede ocurrir que el error estándar de una estimación cercana al 0 o al 1 sea demasiado grande y que el límite inferior o superior del intervalo de confianza sea inferior a 0 o superior a 1, respectivamente. En este caso, es necesario estimar el intervalo de confianza con una variante que permita considerar estas restricciones. Una solución a este problema es considerar una transformación del estimador. De esta manera, si $\hat{P}$ es una estimación de la proporción, se define la transformación logit de la proporción.

$$\hat{L} = log\left(\frac{\hat{P}}{1-\hat{P}}\right) = logit\,(\hat{P})$$

Nótese que la aproximación de Taylor de primer orden para $\hat{L}$ es:

$$\hat{L} \cong L(P) + \frac{\partial \hat{L}}{\partial \hat{P}}|_{(\hat{P}=P)} (\hat{P} - P) = L(P) + \left(\frac{1}{P(1-P)}\right)(\hat{P} - P)$$

Luego, la varianza de $\hat{L}$ se puede escribir como:

$$Var(\hat{L}) = AVar(\hat{L}) = \frac{Var(\hat{P})}{P^2 (1-P)^2}$$

De esta forma, es posible definir un intervalo de $(1-\alpha)100\%$ de confianza para L como:

$$\left(\hat{L} - t_{0,975,gl}\sqrt{Var(\hat{L})}\ \hat{L} + t_{0,975,gl}\sqrt{Var(\hat{L})}\right)(\hat{L}_1, \hat{L}_2)$$

Finalmente, se tiene que:

$$\hat{P} = logit^{-1}(\hat{L}) = \frac{exp(\hat{L})}{1+exp(\hat{L})}$$

Por tanto, un intervalo de confianza para $\hat{P}$ está determinado por:

$$(logit(\hat{L}_1), logit(\hat{L}_2)) = \left(\frac{exp(\hat{L}_1)}{1+exp(\hat{L}_1)}, \frac{exp(\hat{L}_2)}{1+exp(\hat{L}_2)}\right) \subseteq (0,1)$$

En los casos en los que el intervalo de confianza clásico se sale de los límites naturales de la proporción, es recomendable utilizar este último enfoque.

2. Coeficiente de variación

Esta medida representa un acercamiento al error de muestreo que permite verificar si la inferencia es válida. Se define como sigue:

$$CV(\hat{\theta}) = \frac{se(\hat{\theta})}{\hat{\theta}} = \frac{\sqrt{\widehat{Var}(\hat{\theta})}}{\hat{\theta}}$$

Esta medida de precisión de las estimaciones se ha consolidado como un estándar de calidad que ha permeado la práctica de los INE en lo que se refiere a la publicación de estadísticas oficiales. Su uso es transversal, puesto que, por su definición, tiene una naturaleza relativa, al liberar al usuario de la unidad de medida determinada por la variable de interés. Además, es posible reformular los intervalos de confianza en términos del coeficiente de variación, de la siguiente manera:

$$\hat{\theta} \pm t_{0,975,gl} * se(\hat{\theta}) = \hat{\theta}\left(1 \pm t_{0,975,gl} * CV(\hat{\theta})\right)$$

Como afirman Singh, Westlake y Feder (2004), esta es una medida de fácil interpretación, proporcional a la amplitud del intervalo de confianza, que brinda una manera

estandarizada y relativa de medir la precisión alrededor de la estimación puntual, la cual permite comparar dos estimaciones del mismo indicador en diferentes subpoblaciones y, además, se utiliza en el diseño y rediseño de las encuestas, entre otras cualidades. Por ejemplo, desde el punto de vista teórico, Särndal, Swensson y Wretman (2003) señalan que un estadístico puede expresar su opinión de que un valor del coeficiente de variación del 2% es bueno, considerando las limitaciones de la encuesta, mientras que un valor del coeficiente de variación del 9% puede ser considerado inaceptable. De esta forma, muchos INE en todo el mundo han considerado que las precisiones de las estadísticas resultantes de una encuesta están supeditadas al comportamiento de su coeficiente de variación. En el contexto de la calidad de las estimaciones provenientes de encuestas de hogares, mucho se ha discutido acerca del uso del coeficiente de variación para la validación de la confiabilidad y precisión de las cifras que provienen de estudios por muestreo.

Cabe subrayar que, cuando se están estimando proporciones, hay que tener en cuenta algunos aspectos importantes relativos a esta medida. En primer lugar, el hecho de fijar un umbral para el coeficiente de variación requiere una interpretación directa sobre la amplitud relativa del intervalo de confianza. Por ejemplo, si la oficina nacional de estadística decide fijar como umbral para el coeficiente de variación un 30%, esto implica que la amplitud relativa (AR) del intervalo de confianza se fija de forma automática en alrededor del 118%, puesto que:

$$CV(\hat{\theta}) = 30\% \Rightarrow AR = \frac{2 * t_{0,975,gl} * se(\hat{\theta})}{\hat{\theta}} \approx 118\%$$

Por otro lado, como en todo fenómeno dicotómico resumido en una proporción, la varianza y el error estándar de la proporción obtienen su valor máximo en $P = 0,5$. Por lo tanto, en este valor es necesario aumentar el tamaño de muestra para asegurar la precisión definida. A partir de $P = 0,5$, a derecha e izquierda, los fenómenos son simétricos. Por ejemplo, según este paradigma, la precisión de una proporción $P = 0,9$ es la misma que la de una proporción $P = 0,1$. Asimismo, la precisión de una proporción $P = 0,7$ es la misma que la de una proporción $P = 0,3$. Sin embargo, el coeficiente de variación no es una medida simétrica alrededor de $P = 0,5$, como sí lo son la varianza y el error estándar y, por su definición, cuando la proporción es pequeña, el coeficiente de variación tiende a ser muy grande, lo que indicaría, erróneamente, que la precisión es baja.

3. Coeficiente de variación logarítmico

El coeficiente de variación es una medida que define la precisión de un indicador, pero, en el caso de las proporciones, no constituye una medida simétrica, como sí lo son el error estándar o la varianza. Por ejemplo, supóngase que se desea estimar una proporción P. Si la estimación del parámetro de interés es muy cercana a cero, sin importar lo pequeña que sea su varianza, el coeficiente de variación será muy grande y no representará la calidad de la estrategia de muestreo. Sin embargo, el coeficiente de variación del complemento

de la proporción $(1-P)$ será muy pequeño y confiable. Esto constituye una paradoja, puesto que, si bien se mide el mismo fenómeno, los coeficientes de variación son contradictorios. Debido a ello, las estimaciones que tienen una magnitud pequeña (muy cercana a cero) son automáticamente castigadas por este indicador, incluso si la variabilidad de la cifra es pequeña.

Algunos autores han propuesto la posibilidad de realizar una transformación logarítmica sobre la proporción y utilizar su coeficiente de variación como una medida robusta del error de muestreo en las proporciones cercanas a cero y a uno, que además sea simétrica alrededor de $P = 0,5$, que es donde se maximiza la variabilidad de la proporción (Barnett-Walker y otros, 2003). Por lo tanto, si $P \leq 0,5$, se define $\hat{L} = -log\,(\hat{P})$. En este caso, la aproximación de Taylor de primer orden es:

$$\hat{L} \cong L + \frac{\partial \hat{L}}{\partial \hat{P}}\Big|_{(\hat{P}=P)} (\hat{P} - P) = L + \left(\frac{-1}{P}\right)(\hat{P} - P)$$

Luego, la varianza de $\hat{L}$ será $Var(\hat{L}) \cong AV(\hat{L}) = \frac{Var(\hat{P})}{P^2}$ y, por consiguiente, el error estándar de la transformación equivaldrá al coeficiente de variación de la proporción, dado por:

$$SE\,(\hat{L}) = \sqrt{AVar\,(\hat{L})} = \frac{\sqrt{Var(\hat{P})}}{\hat{P}} = CV\,(\hat{P})$$

De esta manera, se puede definir una medida de suavización como el coeficiente de variación asociado a la transformación:

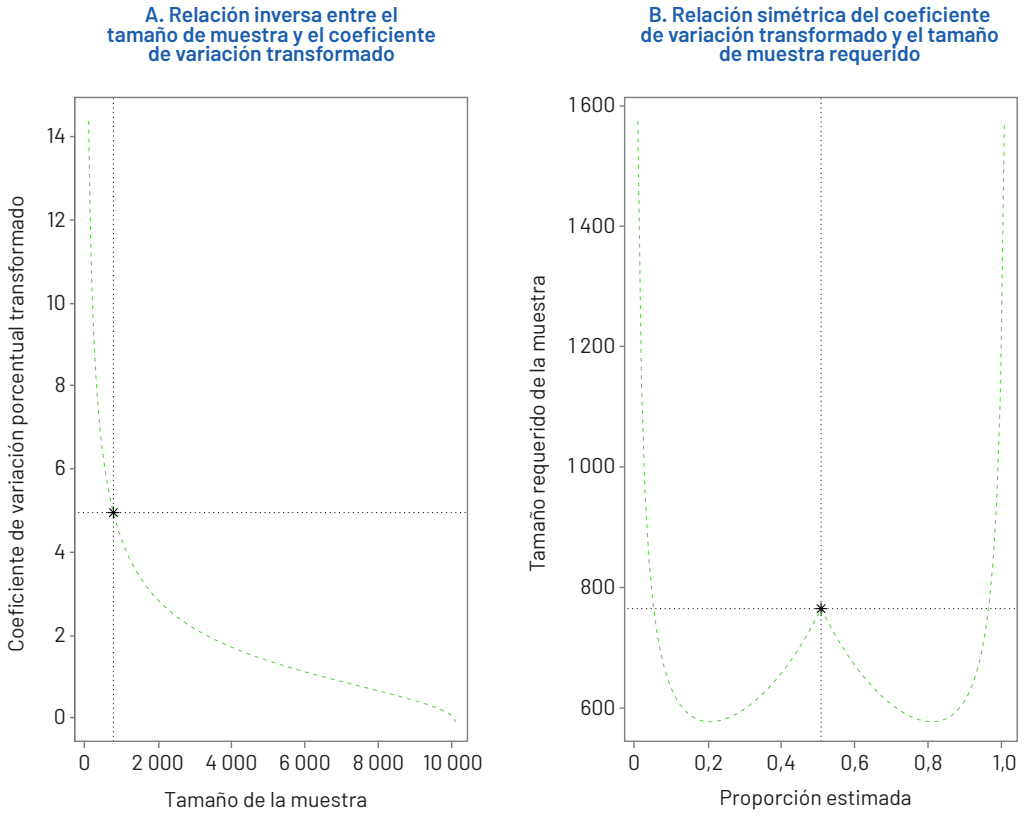
$$CV\,(\hat{L}) = \frac{SE\,(\hat{L})}{\hat{L}} = \frac{CV\,(\hat{P})}{\hat{L}}$$

De manera similar, para mantener la simetría, cuando $P > 0,5$, se realiza un ajuste definiendo $\hat{L} = -log(1-\hat{P})$. Por lo tanto, para proporciones centrales, los coeficientes de variación de $\hat{P}$ y $\hat{L}$ serán comparables, puesto que $\hat{L}$ toma valores cercanos a 1 cuando $P \in (0,2;\ 0,8)$. En este caso, el $CV(\hat{L})$ será similar a $CV(\hat{P})$.

A continuación se presenta un ejemplo sencillo. Considérese una proporción estimada $\hat{P} = 0,1\%$, con un error estándar $SE(\hat{P}) = 0,2\%$. En este caso, el intervalo de confianza clásico está dado por $(-0,10\%; 0,30\%)$, junto con un coeficiente de variación $CV(\hat{P}) = 99,70\%$, por lo que la cifra no sería publicable en primera instancia. Sin embargo, a partir de la amplitud del intervalo de confianza, es fácil observar que esta estimación es buena, informativa y precisa. Por lo tanto, utilizando la transformación logit, el intervalo de confianza de la transformación estaría dado por $(0,01\%; 0,71\%)$ y el $CV(\hat{L}) = \frac{CV(\hat{P})}{\hat{L}} = \frac{99,7}{(-log(0,001)} \cong 14,5\%$. De este modo, se concluye que la cifra sí podría publicarse.

Además, este enfoque representa una excelente aproximación al enfoque clásico cuando las proporciones estimadas no son pequeñas. Por ejemplo, considérese una proporción estimada del $\hat{P} = 30\%$, con un $CV(\hat{P}) = 4,83\%$ y un intervalo de confianza clásico dado por $(27,16\%; 32,84\%)$. Utilizando la transformación logit, el intervalo de confianza estaría dado por $(27,24\%; 32,91\%)$ y el coeficiente de variación logit sería de $CV(\hat{L}) = 4,01\%$ (véase el gráfico XVIII.1).

■ Gráfico XVIII.1
Relación entre el tamaño de muestra y la precisión de un indicador utilizando la transformación logit



Fuente: Elaboración propia.

En el gráfico XVIII.1 se aprecia que, al igual que sucede con el coeficiente de variación original, el tamaño de muestra aumentará a medida que se requiera mayor precisión en la estimación. Sin embargo, a diferencia del coeficiente de variación original, el tamaño de muestra será idéntico para los fenómenos que induzcan proporciones simétricas. Además, el tamaño de muestra necesario para estimar de manera eficiente una proporción $P \leq 0{,}5$ con una precisión superior a un determinado umbral del coeficiente de variación CVE es:

$$n \geq \frac{P\,(1\text{-}P)\,DEFF}{\frac{P\,(1\text{-}P)\,DEFF}{N} + log^2\,(P)\,P^2\,CVE^2}$$

La expresión anterior se obtiene teniendo en cuenta el siguiente desarrollo algebraico. En particular, cuando $P > 0,5$, se desea que el coeficiente de variación logarítmico sea inferior a un umbral δ. Por lo tanto, habiendo definido $S^2=P\,(1\text{-}P)\,DEFF$, se tiene la siguiente implicación:

$$CV(\hat{L}) \leq \delta \Rightarrow n \geq \frac{S^2}{\delta^2\,(1\text{-}\hat{P})^2\,log^2\,(1\text{-}\hat{P}) + \frac{S^2}{N}}$$

De forma análoga, cuando $P > 0,5$, se tiene que:

$$CV(\hat{L}) \leq \delta \Rightarrow n \geq \frac{S^2}{\frac{S^2}{N} + log^2\,(\hat{P})\,\hat{P}^2\,\delta^2}$$

4. El efecto de diseño

Cuando se selecciona una muestra a partir de un diseño de muestreo complejo, es muy improbable que exista independencia entre las observaciones. Además, como el muestreo de las encuestas de hogares es complejo, la distribución de la variable de interés no es la misma para todos los individuos. Por ello, al analizar datos que provienen de encuestas de hogares, para realizar la inferencia correcta se deben tener en cuenta estas grandes desviaciones con respecto al análisis estadístico clásico, en el que se consideran muestras aleatorias simples. En la mayoría de las ocasiones, esto supone aumentar el tamaño de muestra para obtener la precisión deseada.

Lumley (2010) afirma que el efecto de diseño compara la varianza de una media o total con la varianza de un estudio del mismo tamaño mediante un muestreo aleatorio simple sin reemplazo, y que su cálculo será incorrecto si los pesos de muestreo se han reescalado o no son recíprocos con respecto a las probabilidades de inclusión. Además, en R se compara la varianza de la estimación con la varianza de una estimación basada en una muestra aleatoria simple del mismo tamaño que el de la subpoblación. Entonces, por ejemplo, en el muestreo aleatorio estratificado, el efecto de diseño calculado en un estrato será igual a uno.

5. Tamaño de muestra

El tamaño de la muestra afecta de manera indirecta la amplitud del intervalo de confianza, a través del error estándar, que generalmente decrece a medida que el tamaño de la muestra se hace más grande. Un tamaño de muestra adecuado garantiza la convergencia en la distribución de los estimadores a la distribución teórica de donde se obtienen los percentiles

en el cálculo del intervalo de confianza. En la fase de diseño, el tamaño de muestra requerido para estimar el promedio de una variable de interés en una encuesta de hogares, con un error de muestreo relativo inferior a $\delta \in (0,1)$ y una confianza estadística superior a $1\text{-}\alpha$, está dado por la siguiente expresión:

$$n \geq \frac{S_y^2 DEFF}{\frac{\delta^2 \bar{y}^2}{z_{1-\alpha/2}^2} + \frac{S_y^2 DEFF}{N}}$$

Donde $z_{1-\alpha/2}$ es el percentil $(1-\alpha/2)$ asociado a una distribución normal estándar. Por ejemplo, en un diseño de muestreo en varias etapas, si el valor del coeficiente de correlación intraclase es grande, el valor del efecto de diseño $(DEFF)$ también lo será y, por consiguiente, el tamaño de la muestra deberá ser más grande. Por ejemplo, al medir ingresos en la región, debido a la realidad económica de los países, es común encontrar que las condiciones de la vivienda están muy asociadas con el ingreso de los individuos. Esto quiere decir que los ingresos no están uniformemente dispersos entre todos los hogares, por lo que el coeficiente de correlación intraclase será alto. Por otro lado, si lo que se quiere estimar es una proporción P, la expresión apropiada para calcular el tamaño de muestra estará dada por:

$$n \geq \frac{P\,(1\text{-}P)\,DEFF}{\frac{\delta^2}{z_{1-\alpha/2}^2} + \frac{P\,(1\text{-}P)\,DEFF}{N}}$$

Como se puede observar, el tamaño de la muestra es un indicador de la calidad de la encuesta, muy importante en la etapa de planificación y diseño. Sin embargo, se tiene que considerar que:

- Si el parámetro de interés fue tenido en cuenta en la planificación de la encuesta con el propósito de tener representatividad con respecto a una subpoblación, entonces el tamaño de la muestra será apropiado y, por ende, el error de muestreo estará controlado, al igual que el coeficiente de variación y el intervalo de confianza. La precisión de la inferencia será óptima.
- Si el parámetro de interés fue tenido en cuenta en la planificación de la encuesta, pero la tasa de falta de respuesta fue elevada, el tamaño de la muestra será mucho menor que el planeado inicialmente y, por ende, el error de muestreo será más alto, al igual que el coeficiente de variación, y el intervalo de confianza será muy amplio, lo que hará que la precisión de la inferencia no sea apropiada.
- Si el parámetro de interés no fue tenido en cuenta en la planificación y el diseño de la encuesta de hogares, entonces es posible que el tamaño de la muestra sea inferior al necesario, por lo que el error de muestreo será mayor, junto con el coeficiente de variación. Por lo tanto, el intervalo de confianza será más amplio y la precisión de la inferencia será deficiente.

6. Tamaño de muestra efectivo

Esta medida se basa en el principio general de que, en la inferencia propia de las encuestas de hogares con diseños de muestreo complejos, no existe una sucesión de variables que sean independientes y estén idénticamente distribuidas. Por lo tanto, si se piensa en la muestra $(y_1,\ldots,y_n)$ como un vector en el espacio n-dimensional, según la teoría estadística clásica, se supondría que cada componente del vector puede variar por sí mismo. Sin embargo, debido a la forma jerárquica de la selección de los hogares y a la interrelación de la variable de interés con las UPM, la variabilidad de la inferencia en las encuestas complejas tiene un fuerte componente asociado al mismo conglomerado, por lo que la dimensión final del vector $(y_1,\ldots,y_n)$ es mucho menor que n. De esta forma, se ha definido el tamaño de muestra efectivo (Naciones Unidas, 2007, cap. VI) como sigue:

$$n_{eff} = \frac{n}{DEFF}$$

En resumen, el diseño clásico de las encuestas de hogares consiste en seleccionar un conjunto de hogares dentro de una misma UPM y repetir esta estrategia de selección sistemáticamente en todo el país. Por lo tanto, se puede pensar que, si la variable de interés tiene una alta correlación intraclase, la realidad de las personas y de los hogares dentro de una misma UPM será muy homogénea. Así, se podría interpretar que la información está repetida y que los individuos u hogares de una misma UPM no contribuyen de manera diferenciada. Por lo tanto, debido a los efectos de diseño del muestreo complejo, la cantidad de individuos que están aportando a la inferencia del indicador no equivale al número de personas, ni al número de hogares de la muestra, sino al tamaño de muestra efectivo n_{eff}, que contiene los efectos de la aglomeración.

7. Grados de libertad

La amplitud del intervalo de confianza de un indicador no solo está supeditada al error estándar, sino también al percentil de la distribución *t de Student*, con sus correspondientes grados de libertad. De esta manera, cuantos más grados de libertad se consideren, menor será la amplitud del intervalo y mayor será la precisión de la inferencia. En el caso más general en que la subpoblación sea toda la población objetivo, los grados de libertad se reducirán a la siguiente expresión:

$$gl = \#UPM - \#Estratos$$

Los grados de libertad constituyen una medida de cuántas unidades independientes de información se tienen en la inferencia. Nótese que, en el caso extremo de realizar un censo en cada UPM, sin importar el número de individuos que compongan el conglomerado, el número de unidades independientes será únicamente el número de UPM seleccionadas en la primera etapa de muestreo. Esto se debe a que la UPM es la unidad de muestreo que contribuye en mayor medida a la variabilidad de las estimaciones. En las aplicaciones

reales de encuestas de hogares, en que se realiza un submuestreo dentro de la UPM, la variabilidad de la estimación puede verse como la contribución del conglomerado a la gran media, más una contribución (considerada insignificante) de la segunda etapa de muestreo. Cabe subrayar la importancia de utilizar la distribución *t de Student* como base inferencial para la construcción de los intervalos de confianza. Es necesario recordar, además, que el percentil 0,975 para la distribución *t de Student* varía con respecto a sus grados de libertad.

A nivel desagregado, los grados de libertad son determinantes a la hora de hacer inferencias dentro de subpoblaciones de interés. En este caso, los grados de libertad no se consideran fijos, sino variables. Korn y Graubard (1999, pág. 209) proponen el siguiente método de cálculo sobre los grados de libertad en una subpoblación U_g:

$$gl_g = \sum_{h=1}^{H} v_h * (n_{lh}^g - 1)$$

Donde v_h es una variable indicadora que toma el valor 1 si el estrato h contiene uno o más casos de la subpoblación de interés y el valor 0 en caso contrario, mientras que n_{lh}^g es el número de UPM en el estrato h $(h=1,\ldots,H)$ con uno o más casos de la subpoblación.

8. Conteo de casos no ponderado

El número de casos no ponderados en una muestra es simplemente el conteo de los individuos dentro de la muestra que se ven afectados por un fenómeno de interés en estudio. Esta cifra está supeditada únicamente a razones y proporciones, tiene un efecto indirecto en la determinación de la precisión del estimador de interés y está determinada por la siguiente expresión:

$$n_y = \sum_s \delta_k^y$$

Donde δ_k^y es una variable indicadora sobre cada individuo k de la muestra s que toma el valor de 1 si el individuo está afectado por el fenómeno inducido por la variable de interés y. Se trata de una cantidad aleatoria por definición, que también puede ser calculada en la muestra de un subgrupo poblacional específico U_g, de la siguiente manera:

$$n_y^g = \sum_s z_{g_k} \delta_k^y = \sum_{s_g} \delta_k^y$$

Si la incidencia del fenómeno es muy baja (cuando la proporción P es cercana a 0), tanto el coeficiente de variación original como su transformación logarítmica tendrán magnitudes altas, puesto que:

$$\lim_{(n_y \to 0)} CV(\hat{\theta}) = \lim_{(n_y \to 0)} CV(\hat{L}) = \infty$$

En muchos países, las encuestas de hogares son utilizadas por las autoridades gubernamentales para asignar recursos a una población potencial. En estos casos, es de particular interés conocer el número de personas que serán susceptibles de participar en la repartición de recursos. Por ende, si la estimación de la incidencia total del fenómeno en la población no es precisa, difícilmente se podrá establecer un rubro presupuestario para atender a esta población. Por ejemplo, si la estimación del total de personas afectadas por el fenómeno es del orden del 5% y su margen de error es del 5%, el coeficiente de variación será del 100% y el intervalo de confianza de la proporción será (0%,10%). Este intervalo es demasiado amplio para tomar algún tipo de decisión sobre los recursos públicos de un país. Nótese que esta amplitud se magnifica cuando el número de casos no ponderado no es suficiente.

B. Criterios de calidad en subpoblaciones

En esta sección, se abordan cuestiones referentes al procesamiento apropiado de los indicadores de interés en las subpoblaciones. Después de realizarse las estimaciones y calcularse los respectivos criterios de calidad, se plantea la utilización de umbrales apropiados para la supresión, la revisión o la publicación de cifras.

El análisis apropiado de las estadísticas generadas a partir de las encuestas de hogares debe basarse en una definición clara tanto de las subpoblaciones sobre las que se quiere realizar la inferencia como de las variables que generan el indicador de interés. De hecho, como se mostrará más adelante, algunas variables pueden definir una subpoblación, y es posible que esto dé lugar a confusiones. Para aclararlo, se proponen a continuación algunos ejemplos que permiten dilucidar la forma de calcular las medidas de calidad sobre un conjunto no exhaustivo de indicadores de interés.

1. Promedio del ingreso per cápita en el país

En este caso, la variable de interés es una característica continua $y_k \geq 0 \ (\forall \ k \in U)$ definida sobre toda la población del país, mientras que el indicador se escribe como una razón:

$$\hat{\bar{y}}_{nacional} = \frac{\hat{t}_y}{N} = \frac{\sum_h \sum_i \sum_k w_k y_k}{\sum_h \sum_i \sum_k w_k}$$

Donde los subíndices h, i y k se refieren, respectivamente, a los estratos, las UPM y los individuos. Nótese que la variable que define la población es siempre determinista, puesto que $z_{g_k} = 1$ para todos los individuos que residen en el país, es decir, para todos los de la muestra. En este caso, los grados de libertad corresponden a todas las UPM menos todos los estratos de la encuesta en el país.

2. Promedio del ingreso per cápita en una ciudad

En este caso, la variable de interés está definida sobre un subgrupo poblacional U_g, correspondiente a la ciudad de interés. El estimador del indicador se escribe como una razón:

$$\hat{\bar{y}}_{ciudad} = \frac{\hat{t}_{y_g}}{\hat{N}_g} = \frac{\sum_h \sum_i \sum_k w_k z_{g_k} y_k}{\sum_h \sum_i \sum_k w_k z_{g_k}}$$

La variable que define la subpoblación es dicotómica y está dada por:

$$z_{g_k} = \begin{cases} 1 & \text{si } k \text{ reside en la ciudad } U_g \\ 0 & \text{en caso contrario} \end{cases}$$

En este caso, el tamaño de muestra es $n_g = \sum_s z_{g_k}$, es decir, el tamaño de muestra de la ciudad. Los grados de libertad corresponden a todas las UPM de la ciudad, menos todos los estratos de la ciudad.

3. Proporción de personas pobres en el área urbana

En este ejemplo, el estimador del indicador se escribe como una razón sobre el área urbana U_g:

$$\hat{P}_{urbano} = \frac{\hat{t}_{y_g}}{\hat{N}_g} = \frac{\sum_h \sum_i \sum_k w_k z_{g_k} y_k}{\sum_h \sum_i \sum_k w_k z_{g_k}}$$

Donde y_k es la variable de interés que define una característica dicotómica de la siguiente manera:

$$y_k = \begin{cases} 1 & \text{si el ingreso per cápita de la persona está por debajo de la línea de pobreza} \\ 0 & \text{en caso contrario} \end{cases}$$

Las mediciones se realizan sobre la subpoblación definida por la siguiente variable:

$$z_{g_k} = \begin{cases} 1 & \text{si la persona reside en el área urbana } U_g \\ 0 & \text{en caso contrario} \end{cases}$$

En este caso, el tamaño de muestra es $n_g = \sum_s z_{g_k}$, es decir, el tamaño de muestra del área urbana. Los grados de libertad corresponden a todas las UPM del área urbana menos todos los estratos del área urbana.

4. Tasa de desocupación nacional

Este indicador está definido como la división entre el total de personas desocupadas y el total de personas activas en la fuerza de trabajo. El estimador del indicador está definido como una razón de dos estimadores de totales poblacionales:

$$\widehat{TD}_{nacional} = \frac{\hat{t}_y}{\hat{t}_z} = \frac{\sum_h \sum_i \sum_k w_k z_k y_k}{\sum_h \sum_i \sum_k w_k z_k}$$

Donde las variables de interés toman la siguiente forma:

$$y_k = \begin{cases} 1 & \text{si el individuo está desocupado} \\ 0 & \text{si el individuo no está desocupado} \\ NA & \text{si el individuo no está en edad de trabajar} \end{cases}$$

La variable que define la subpoblación es:

$$z_k = \begin{cases} 1 & \text{si el individuo está activo} \\ 0 & \text{si el individuo está inactivo} \\ NA & \text{si el individuo no está en edad de trabajar} \end{cases}$$

En este caso, el tamaño de muestra es $n = \sum_s z_k$, es decir, el número de personas de la muestra que están en edad de trabajar y activas. Los grados de libertad corresponden a todas las UPM menos todos los estratos de la encuesta en el país en los que se encontraron hogares con individuos en edad de trabajar y activos. Nótese que estos son los mismos grados de libertad definidos por la tasa de ocupación. Además, el conteo de casos no ponderado corresponde al número de individuos desocupados en la muestra.

5. Tasa de desocupación masculina en migrantes

Este indicador está definido como la división del total de hombres migrantes desocupados entre el total de hombres migrantes activos. El estimador del indicador se define como una razón de dos estimadores de totales poblacionales:

$$\widehat{TD}_{hombre\text{-}migrante} = \frac{\hat{t}_{y_g}}{\hat{t}_{z_g}} = \frac{\sum_h \sum_i \sum_k w_k z_{g_k} y_k}{\sum_h \sum_i \sum_k w_k z_{g_k}}$$

Donde las variables de interés toman la siguiente forma:

$$y_k = \begin{cases} 1 & \text{si el individuo está desocupado} \\ 0 & \text{si el individuo no está desocupado} \\ NA & \text{si el individuo no está en edad de trabajar} \end{cases}$$

La variable que define la subpoblación es:

$$z_{g_k} = \begin{cases} 1 & \text{si el individuo está activo y es hombre y migrante} \\ 0 & \text{si el individuo está inactivo y es hombre y migrante,} \\ NA & \text{si el individuo no está en edad de trabajar o es mujer o no es migrante} \end{cases}$$

En este caso, el tamaño de muestra es $n=\sum_s z_{g_k}$, es decir, el número de personas de la muestra que están en edad de trabajar, son hombres migrantes y están activos. El conteo no ponderado de casos corresponde al número de individuos de la muestra que son hombres migrantes y están desocupados. Además, los grados de libertad corresponden a todas las UPM menos todos los estratos de la encuesta en el país en los que se encontraron hogares con hombres migrantes y activos en la fuerza de trabajo.

C. Secuencia lógica para crear reglas de supresión

En esta sección se ha querido subrayar el hecho de que la precisión de una estimación recae directamente sobre los intervalos de confianza, que pueden descomponerse en elementos fundamentales que permiten crear una secuencia lógica de revisión, publicación o supresión de cifras. Esta afirmación se basa en el hecho de que la longitud de los intervalos de confianza indica con seguridad si un estimador es o no preciso. Considérense los siguientes ejemplos prácticos:

- La incidencia de la pobreza en un departamento de un país se estimó en un 5,2%, con un intervalo de confianza de (5,15%, 5,25%).
- La tasa de desocupación en el país en el caso de los hombres se ubicó en un 7,5%, con un intervalo de confianza de (7,1%, 7,9%), mientras que, en el caso de las mujeres, se ubicó en un 9,2%, con un intervalo de confianza de (8,8%, 9,6%).
- La tasa de asistencia neta estudiantil en primaria para el último quintil de ingresos se estimó en un 85%, con un intervalo de confianza de (48,2%, 100,0%).

Está claro que, en la última situación ejemplificada, el intervalo de confianza no brinda la precisión adecuada para que una oficina nacional de estadística publique esta cifra con suficiente confianza, o para que un gobierno pueda implementar algún tipo de política pública educativa, y mucho menos para estimar los recursos que requiere una intervención estatal sobre la población de interés. Como se ha descrito a lo largo de este documento, al utilizar únicamente el coeficiente de variación como criterio para la supresión de cifras, no se tienen en cuenta todas las variantes asociadas a la inferencia en un muestreo complejo. A continuación se enumeran algunas recomendaciones internacionales que incorporan otros criterios adicionales.

- **Coeficiente de variación:** En CEPAL (2018b), se realizó un análisis de distintas experiencias internacionales, sobre la base de la información publicada en las páginas web de los INE respectivos, con el fin de determinar cómo se utilizan los criterios de supresión de información y los umbrales que las oficinas nacionales de estadística definen para validar las cifras. En las encuestas de hogares, se encontró que los Estados Unidos y los países del MERCOSUR utilizan un umbral de $CV > 30\%$, el Canadá y México utilizan como referencia un umbral de $CV > 25\%$, Chile y Costa Rica utilizan un umbral de $CV > 20\%$, el Ecuador y el Perú utilizan un umbral de $CV > 15\%$ y Colombia usa un umbral de $CV > 10\%$. De esta forma, cualquier cifra estimada cuyo coeficiente de variación sea superior al umbral predefinido es suprimida o marcada como una cifra poco confiable.
- **Tamaño de muestra:** este criterio se debe considerar uno de los más importantes a la hora de decidir la trayectoria de publicación de una cifra, puesto que los desarrollos teóricos en términos de inferencia estadística para encuestas dependen de este término. La cobertura de los intervalos de confianza y la distribución de los estimadores dependen de que el tamaño de la subpoblación y su tamaño de muestra asociado no sean pequeños. En este sentido, Barnett-Walker y otros (2003) proponen que todas las estimaciones basadas en un tamaño de muestra inferior a 100 unidades se supriman o se marquen como no confiables.
- **Tamaño de muestra efectivo:** al igual que ocurre con el criterio anterior, el tamaño de muestra efectivo permite que se cumplan las aproximaciones teóricas, en términos de convergencia de las distribuciones de los estimadores y la cobertura de los intervalos de confianza. Hornik y otros (2002) consideran que, si el tamaño de muestra efectivo no es superior a 140, no se debería considerar la publicación de la cifra. Por otro lado, teniendo en cuenta el tamaño de muestra determinado por la transformación logarítmica, Barnett-Walker y otros (2003) afirman que, si la proporción se encuentra entre 0,05 y 0,95, el tamaño de muestra efectivo es máximo cuando $P = 0{,}5$, siendo su valor $n_{eff}=68$.
- **Conteo de casos no ponderado:** cuando la incidencia de un fenómeno es muy baja y el diseño de la encuesta no lo tuvo en cuenta, es posible que las estimaciones asociadas a tamaños, totales y proporciones sobre este fenómeno no sean confiables. En particular, en el caso de las proporciones, es posible restringir las estimaciones de modo que $\hat{P} < 0{,}001$, pero resulta más rápido crear una regla a partir del conteo de casos en la muestra. Por ejemplo, en National Research Council (2015), se afirma que, si el número de casos no ponderados es inferior a 50 unidades, la estimación no se publica.

- **Grados de libertad:** este criterio apunta a aislar el efecto del tamaño de la muestra en una encuesta compleja y plantea una aproximación al número de unidades independientes en la inferencia. Además, a medida que crece, la amplitud del intervalo de confianza se estabiliza. Parker, Talih y Malec (2017) consideran que, si los grados de libertad determinados por la subpoblación son menos de ocho, la cifra debería suprimirse.
- **Coeficiente de variación logarítmico:** esta medida de suavización adopta valores elevados cuando las proporciones estimadas están demasiado cercanas a 0 o a 1. Barnett-Walker y otros (2003) proponen que la cifra se suprima si el coeficiente de variación logarítmico es superior al 17,5%.

Los criterios mencionados en este documento no deberían aplicarse de manera independiente, sino que tendrían que seguir cierta lógica. Es posible, por ejemplo, para una variable con poca homogeneidad en las UPM, que, con un tamaño de muestra de $n=90$, se haya estimado un efecto de diseño de $DEFF=0,5$, lo que implicaría un tamaño de muestra efectivo de $n_{eff}=180$. En este caso, si los criterios de supresión se aplicaran de manera independiente, se concluiría que la cifra se debería suprimir por tener un tamaño de muestra insuficiente, pero que, a la vez, la cifra se debería publicar, por tener un tamaño de muestra efectivo suficiente. Ello podría dar lugar a contradicciones por parte de los INE y malas interpretaciones por parte de los usuarios finales de los datos.

De manera general, se recomienda que los INE estudien en profundidad sus políticas de supresión, revisión y publicación de cifras en cada una de las encuestas que realizan. Asimismo, se propone que, de manera independiente, definan las reglas apropiadas para cada caso y que los criterios de supresión queden plasmados en forma de diagrama de flujo en la documentación de las encuestas. Además, en cada encuesta se debería considerar un algoritmo de forma particular. Es decir, los criterios de supresión no necesariamente deben coincidir para cada operación estadística, pero sí deberían establecerse unos mínimos en cada oficina nacional de estadística con el fin de garantizar la calidad de las estimaciones publicadas provenientes de las encuestas de hogares.

Teniendo en cuenta las particularidades de cada encuesta, se podría mantener un umbral del coeficiente de variación de $CV>20\%$ y del coeficiente de variación logarítmico de $CV(\hat{L})>17,5$ a la hora de generar alertas sobre la calidad de la estimación. Además, se recomienda tener como mínimo 14 grados de libertad, cifra que implica que hay al menos 15 UPM que inducirían una convergencia en la distribución del estimador (Gutiérrez, 2016, fig. 8.1). A partir de esta cifra, es posible continuar el análisis de las recomendaciones con

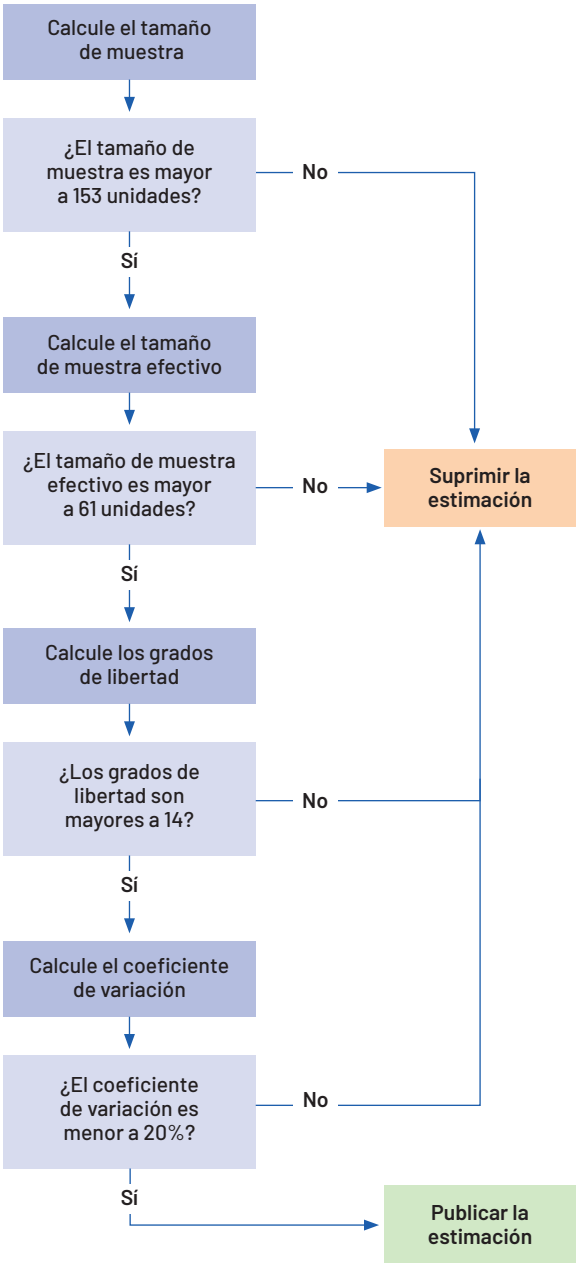
las particularidades de cada encuesta en cada país. Por ejemplo, si se espera un mínimo de 15 UPM, y en cada una se seleccionan 12 hogares, entonces la muestra esperada es de 180 hogares. Tomando una tasa de efectividad del 85%, esta se reduce a $n < 153$ como generador de alertas. Asimismo, tomando en cuenta un efecto de diseño promedio de 2,5 (calculado según el indicador de interés), se obtendría un tamaño de muestra efectivo de $n_{eff} = 61$ como generador de alertas. A la vez, se recomienda mantener el conteo de casos no ponderado en $n_y < 40$ e inmediatamente generar una alerta sobre todas las cifras que indiquen un efecto de diseño $DEFF < 1$.

Por ejemplo, en el diagrama XVIII.1 se muestra un ejemplo de cómo podrían usarse los criterios de supresión de cifras para la estimación de proporciones o razones. En primera instancia, se realiza la estimación clásica de los parámetros de interés y se genera una tabla que contenga el cálculo de todos los criterios descritos anteriormente. Luego, en función de la naturaleza del fenómeno investigado, se deben establecer los criterios que se van a tener en cuenta y los umbrales en cada caso. El próximo paso será decidir, sobre cada cifra de la tabla generada, si se va a publicar o suprimir y, en algunos casos, si se revisará la cifra con mayor detenimiento. En el diagrama XVIII.1, se proponen seis criterios como condiciones necesarias para la publicación inmediata de una cifra. Los primeros cuatro son condiciones necesarias para la revisión temática. Si alguno de estos cuatro criterios no se satisface, se suprime la cifra.

En la región, son cada vez más las oficinas e institutos nacionales de estadística que han establecido criterios de calidad adicionales al coeficiente de variación para la publicación de cifras oficiales provenientes de encuestas de hogares. Por ejemplo, el Instituto Nacional de Estadísticas de Chile (INE, 2020, imagen 1) ha dispuesto la supresión de cifras basada en un estándar de publicación sujeto a la adecuación de los criterios de calidad mencionados en este capítulo. Asimismo, el Grupo de Trabajo sobre encuestas de hogares de la Conferencia Estadística de las Américas ha elaborado una serie de recomendaciones metodológicas regionales sobre la evaluación de la calidad de las cifras estimadas a partir de encuestas de hogares (CEPAL, 2023, figura 5.1.).

■ Diagrama XVIII.1

Ejemplo de un diagrama de flujo para la publicación, supresión y revisión de estimaciones de proporciones o razones en encuestas de hogares



Fuente: Elaboración propia.

Capítulo XIX

Comparabilidad: actualización del diseño de las encuestas, impacto de las actualizaciones y empalme de series de tiempo

Entre los objetivos más importantes de las oficinas nacionales de estadística (ONE) se encuentra garantizar la comparabilidad de las estadísticas oficiales publicadas regularmente. Como se ha visto a lo largo de este documento, la estrategia de muestreo está compuesta por un diseño de muestreo y un estimador. Desde el punto de vista estadístico, esta dupla crea una medida de probabilidad discreta que permite la estimación de estadísticas oficiales basadas en las encuestas de hogares, que serán publicadas y divulgadas posteriormente por las ONE. Esta medida de probabilidad no solo posibilita la estimación puntual del parámetro de interés, sino también la estimación de su varianza y error de muestreo, lo que redunda en una inferencia completa y correcta.

La estabilidad de la medida de probabilidad de las encuestas de hogares a lo largo del tiempo presenta grandes ventajas, pues, además de permitir la comparación transversal de subgrupos poblacionales de interés (clasificados por región, zona de residencia, sexo, edad, nivel de educación, condición de discapacidad y etnia, entre muchos otros criterios), permitirá la comparación temporal de dichos subgrupos. Los parámetros del mercado de trabajo (tasa de desocupación, tasa de participación o tasa de informalidad, entre otros) constituyen uno de los ejemplos más conocidos, pues, dada su importancia, las ONE recopilan datos de manera continua para estimarlos y divulgarlos regularmente (de forma trimestral o incluso mensual). La comparación temporal de estas cifras permite la adopción de políticas públicas oportunas.

En virtud de lo expuesto, el cambio de alguno de los componentes de la estrategia de muestreo afectará la comparabilidad de las estadísticas oficiales en determinados momentos y, si el efecto es significativo, podrá incluso poner en tela de juicio la idoneidad de la inferencia para la estimación de los parámetros de interés. Por ejemplo, en el caso de que, debido a la coyuntura socioeconómica de un país, efectivamente haya un cambio negativo en el mercado de trabajo en un período de interés, este cambio no podrá captarse debidamente si existe un cambio simultáneo en el diseño de muestreo (a su vez determinado por la forma en que se recoge la información) o en el estimador de muestreo (debido a un cambio en el ajuste de los factores de expansión, incluidos los modelos de falta de respuesta y la calibración).

Debido a la evolución natural de las encuestas, la adopción de nuevos métodos de medición, los cambios en la recolección de los datos o incluso la elaboración de nuevas proyecciones censales debido a la realización decenal de los censos de población y vivienda, es casi imposible no realizar ningún cambio en las encuestas de hogares. En este capítulo se describe la mejor manera de realizar estos cambios para minimizar su impacto, medir el efecto de estos a lo largo del tiempo y, llegado el caso, hacer comparables las series de tiempo interrumpidas en un determinado momento debido a un cambio en la medida de probabilidad.

A. Actualización del diseño de las encuestas

Aunque la comparabilidad de las estimaciones es un aspecto fundamental de la divulgación de las estadísticas oficiales basadas en las encuestas de hogares, es casi imposible mantener intacta la medida de probabilidad de la inferencia a lo largo del tiempo. El diseño de las encuestas de hogares se actualiza frecuentemente para reflejar los cambios relacionados con la población de interés. Estas actualizaciones son necesarias para mantener la eficiencia de la estrategia de muestreo y optimizar la metodología. A continuación, se enumeran seis casos en los que los cambios pueden afectar la comparabilidad:

i) Cambios en la forma de medición de los constructos. Es posible que la forma en que se miden los constructos cambie a través del tiempo. Por ejemplo, la concepción y definición de las fuentes de ingresos (CEPAL, 2018a) para la medición de la pobreza monetaria o la actualización de los estándares de la Organización Internacional del Trabajo (OIT, 2013a) para la definición de los indicadores del mercado laboral son algunos casos en los que los cuestionarios deben modificarse para permitir una inferencia oportuna.

ii) Cambios en la definición de las subpoblaciones. El establecimiento de nuevos estándares para la clasificación de las subpoblaciones puede repercutir significativamente en la comparabilidad de las series. Por ejemplo, los nuevos estándares pueden afectar la definición de razas, etnias o pueblos originarios, de migrantes y extranjeros, o de preferencias sexuales y de género, así como la clasificación de las ocupaciones, entre otras categorías.

iii) Cambios en la forma de recolección de la información primaria. Con el pasar de los años y la adopción de nuevos procesos tecnológicos en las ONE de la región, es muy probable que las entrevistas presenciales mediante cuestionarios en papel se sustituyan gradualmente por una forma de recolección presencial con dispositivos digitales o que incluso se implementen cambios más drásticos mediante operativos telefónicos o mixtos. Estas nuevas formas de recoger la información pueden conllevar cambios en la comparabilidad de las series.

iv) Cambios en la división territorial del país. Aunque este tipo de cambio no es tan frecuente, es posible encontrar nuevas definiciones territoriales dentro de los países; por ejemplo, la creación de nuevas divisiones administrativas mayores (regiones, estados o departamentos) o menores (municipios, distritos, cantones o provincias). Estos cambios en la división administrativa y territorial de los países pueden causar discontinuidad, sobre todo en las estadísticas generadas a nivel subnacional.

v) Cambios debidos a la realización de nuevos censos en el país. La mera realización de los censos puede tener consecuencias inesperadas en la comparabilidad de las series. Por ejemplo:

- La forma en que se desarrollan los censos tiene repercusiones directas en la construcción de los marcos de muestreo, que determinan el diseño de muestreo de la encuesta. Si el censo anterior fue un censo de hecho y el actual es un censo de derecho (o viceversa), la definición y la construcción cartográfica de las áreas de enumeración o empadronamiento determinarán diferencias en su tamaño y composición. Dado que estas áreas son el principal insumo para la creación de las unidades primarias de muestreo (UPM), los cambios en la cartografía de los marcos de muestreo serán significativos y redundarán en la interrupción de las series temporales.
- La actualización de las proyecciones demográficas trae consigo un cambio en los totales de control utilizados en los estimadores de calibración. Ante cambios significativos en la población proyectada y la observada en el censo, es muy probable que las estadísticas oficiales de nivel (totales y tamaños, en particular) se vean afectadas y ya no sean comparables, pues se observará un incremento (o una disminución) de las proyecciones de la población civil no institucionalizada (Oficina de Estadísticas Laborales, 2014).

vi) Mejoras deliberadas. Cualquier otro cambio en la estrategia de muestreo puede causar discrepancias en las series de tiempo. Por ejemplo, una nueva forma de administración del marco de muestreo con modificaciones en el tamaño de las UPM o la adopción de una nueva estrategia de estratificación socioeconómica de las UPM del marco tendrán efectos en las cifras. Asimismo, la implementación de cambios y mejoras en el diseño de muestreo, la adopción de un nuevo estimador o la modificación de los ajustes de los pesos de muestreo y los factores de expansión pueden tener consecuencias inesperadas en la comparabilidad de las cifras.

B. Impacto de las actualizaciones

Dado que las actualizaciones en el diseño de las encuestas son inevitables, es recomendable definir de antemano los cambios que se implementarán y planear un experimento controlado a lo largo de un período suficiente de tiempo —por ejemplo, un año para acontecimientos con estacionalidad, como las estadísticas del trabajo—, en el que la operación estadística se realice con dos enfoques simultáneos: el original (sin cambios) y el nuevo (con los cambios de la actualización en el diseño). Esta opción implica que la ONE debe disponer de una cantidad suficiente de recursos presupuestarios, logísticos y humanos durante el tiempo de realización de ambos procesos. En consecuencia, no todas las ONE de la región podrán asumir esa carga y, para algunas, esta opción resultará inviable. Sin embargo, las autoridades de las ONE deberían realizar todas las gestiones posibles para conseguir recursos suficientes y garantizar la medición del impacto de los cambios propuestos.

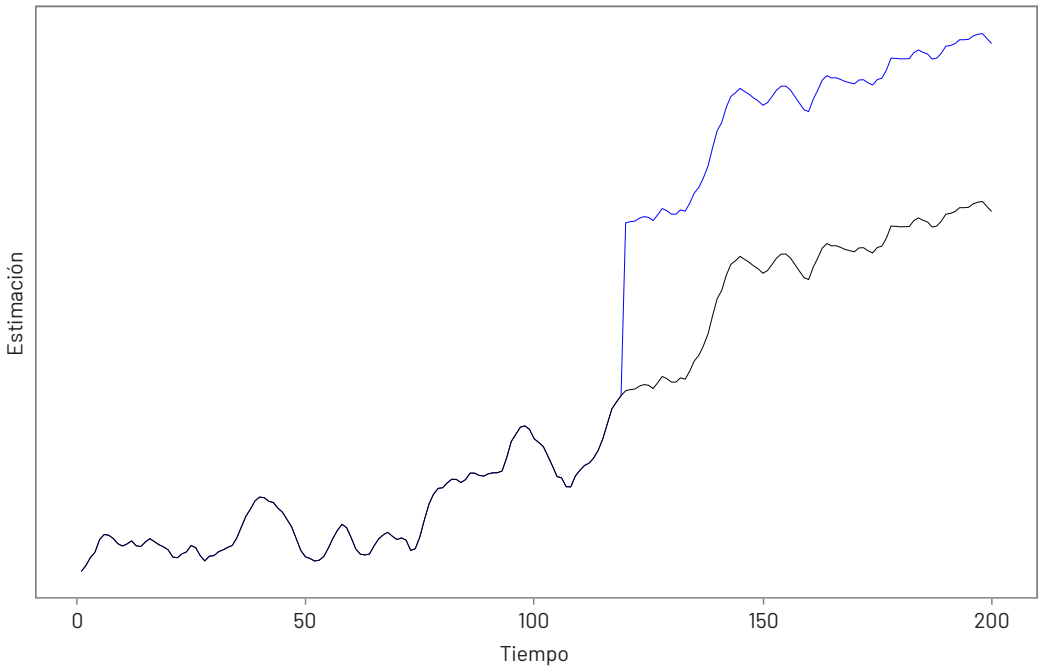
En virtud de lo expuesto, es necesario tener en cuenta que, sin este tipo de experimentos paralelos, será muy difícil medir el verdadero cambio, encontrar la fuente de la discontinuidad en la serie y corregir el sesgo generado. Según Imbens y Rubin (2015), la aleatorización es la única forma de evitar los sesgos de selección en los experimentos controlados y el único supuesto científicamente aceptado para medir este tipo de efectos. De acuerdo con esta perspectiva, en el marco de los experimentos controlados, se deberán seleccionar aleatoriamente las UPM que participarían en las dos operaciones de recolección de datos. Esto no supone una carga adicional para las ONE, garantes de la aleatorización en las encuestas de hogares.

Brakel, Smith y Compton (2008) señalan que existen varias posibilidades para llevar a cabo este tipo de experimentos paralelos: es posible que las dos operaciones estadísticas de campo tengan muestras del mismo tamaño, que la nueva operación tenga una muestra más pequeña o que incluso esté limitada a una determinada subpoblación de interés. En cualquier caso, constituyen formas correctas para evitar efectos de confusión. De lo contrario, incluso ante cambios nulos, no se podrá discernir si esto es el resultado de la coyuntura de interés o de la actualización de la encuesta. En el gráfico XIX.1 se muestra un ejemplo simulado del resultado esperado de un experimento paralelo. La línea negra representa la serie original, la línea azul representa la nueva serie con los cambios de la actualización del diseño y la distancia entre ambas representa el impacto de la actualización en cada punto del tiempo.

Sin embargo, en algunos casos es imposible realizar dos operativos de recolección de datos paralelos. La pandemia de enfermedad por coronavirus (COVID-19) y sus efectos en las condiciones socioeconómicas de los hogares y en el modo de implementar las encuestas constituyen un claro ejemplo de este escenario. De acuerdo con la CEPAL (2020a), desde el comienzo de la emergencia sanitaria derivada de la pandemia, las ONE tuvieron que interrumpir abruptamente la recopilación de información primaria de muchas de sus operaciones estadísticas, incluidas las encuestas de hogares. A pesar de esto, pudieron seguir recolectando datos al sustituir las entrevistas presenciales por

encuestas telefónicas. Esta actualización repentina del diseño —cambio en la metodología de recolección de datos— fue necesaria para seguir produciendo indicadores de empleo y pobreza, particularmente importantes en el contexto de la pandemia, dado el profundo impacto que las restricciones de movimiento y las cuarentenas tuvieron en la ocupación de las personas de la región y, por ende, en sus ingresos. En este caso, no fue posible que las ONE realizaran experimentos paralelos.

■ Gráfico XIX.1
Series de tiempo para la actualización del diseño de una encuesta



Fuente: Elaboración propia.
Nota: La línea negra representa la serie original y la línea azul representa la serie nueva.

Según la CEPAL (2020a), en términos generales, la pandemia obligó a que los países efectuaran cambios en varios aspectos de la metodología de recolección y análisis de la información, que se resumen a continuación:

- Cambió el modo de recolección de datos de presencial a telefónico (o mixto, en algunos casos), así como las definiciones de la estructura de elegibilidad de las viviendas seleccionadas y sus correspondientes códigos de disposición.
- Cambió el mecanismo de supervisión de los encuestadores y, en algunos casos, se suprimieron las actualizaciones cartográficas del número de hogares particulares en las UPM seleccionadas.

- Se introdujo un nuevo proceso de ajuste de factores de expansión, buscando eliminar el sesgo de cobertura (no todos los hogares de los operativos de recolección de datos anteriores contaban con números telefónicos de contacto) y de falta de respuesta (algunos hogares contactados telefónicamente no contestaron el cuestionario).
- Se revisaron los sistemas de calibración de los factores de expansión y, en aras de la flexibilidad de la metodología de estimación, se restringió el número de restricciones de calibración.

En algunos casos especiales, ante la imposibilidad de realizar dos encuestas paralelas, es posible obtener dos series paralelas. Un ejemplo podría ser un cambio en la forma de medición de las estadísticas del mercado de trabajo en un país, en particular, la adopción del estándar de la 19a Conferencia Internacional de Estadísticos del Trabajo (OIT, 2013b). En algunos países resultaría posible adoptar este estándar mediante la adición de nuevas preguntas al cuestionario original basado en la resolución de la 13a Conferencia Internacional de Estadígrafos del Trabajo (OIT, 1982). La actualización de las proyecciones de población y los totales de control en los estimadores de calibración constituye otro caso especial. Dado que el cambio solo afecta los procesos computacionales, es posible tener dos series paralelas, sin necesidad de realizar dos procesos de recolección de datos.

Independientemente de la posibilidad de contar con dos series en paralelo, existen diferentes métodos para establecer la magnitud del impacto debido a un cambio en la encuesta. En general, se enumeran las siguientes posibilidades: i) cuando se dispone de las dos series en paralelo, es posible cuantificar el impacto mediante estudios de causalidad basados en modelos econométricos; ii) cuando se dispone solamente de una serie, es posible estimar el efecto del cambio utilizando modelos de series temporales en los que se incluyan parámetros que indiquen el momento a partir del cual se inició el cambio y sus efectos en la serie (análisis de intervenciones).

En ambos casos, es necesario realizar este tipo de análisis, en primer lugar, para cuantificar el efecto del cambio. Luego, si el efecto resulta estadísticamente significativo, es necesario realizar un empalme de las series de tiempo para obtener una serie ajustada comparable con ambas series: la original y la nueva. Esta se conoce como "serie empalmada".

Por ejemplo, si el indicador de interés es un total, Gbur y Alexander (1984) proponen la utilización de un modelo lineal para determinar los efectos de la actualización del diseño. Este modelo puede escribirse de la siguiente manera:

$$\hat{\theta}_{tdg}=\hat{N}_{tdg}\,\theta_{td}+\hat{N}_{tdg}\,\beta_{t}+\varepsilon_{tdg}$$

Donde $\hat{\theta}_{tdg}=\sum_{k\in s_t} w_{ktg}\, y_{ktg}$ representa la estimación del indicador de interés en el tiempo t para el dominio d. El subíndice $g=1,2$, solo toma dos valores e indica si la variable de interés se observó bajo las condiciones de la actualización o no (tratamiento/control).

Además, $\hat{N}_{tdg}=\sum_{k\in s_t} w_{ktg}$ es la suma de los factores de expansión en el tiempo t, del dominio d en el tratamiento g. Este modelo relaciona el estimador directo $\hat{\theta}_{tdg}$ con el indicador verdadero θ_{tdg} y el efecto de la actualización en el tiempo y, denotado por β_t. Por supuesto, ε_{tdg} denota los errores aleatorios con vector de medias nulo y matriz de varianzas V, cuyas entradas (varianzas y covarianzas) se estiman a partir de los principios de la estimación directa. Cabe señalar que se supone independencia en la selección de los hogares o personas en cada grupo del tratamiento.

Evidentemente, si β_t es estadísticamente igual a 0, se afirma que no existe un efecto de la actualización del diseño en la serie original y, por ende, se garantiza la comparabilidad entre las estimaciones de la serie original y la serie nueva. Sin embargo, en caso contrario, es necesario realizar un proceso de empalme de series como los que se especifican en la siguiente sección.

C. Empalme de series de tiempo

A continuación, se describen brevemente algunas técnicas para empalmar dos series. Todas ellas deberán adaptarse a las necesidades de cada ONE y de cada encuesta. En términos de notación, $\hat{\theta}_t^R$ representa la estimación original en el tiempo t, $\hat{\theta}_t^N$ denota la nueva estimación en el tiempo t y $\hat{\theta}_t^E$ corresponde a la estimación empalmada en el tiempo t.

1. Factor de suavizado

DiNatale (2003) presenta el siguiente ajuste, que suaviza sistemáticamente el cambio entre la nueva serie en el momento de la actualización del diseño (t_b) y la serie original en el punto inmediatamente anterior $t_b - 1$. Es necesario determinar el punto donde ocurrió el cambio t_b, así como el punto que indicará el comienzo del empalme t_1. Luego, se debe calcular el factor de ajuste que representa el cambio (porcentual).

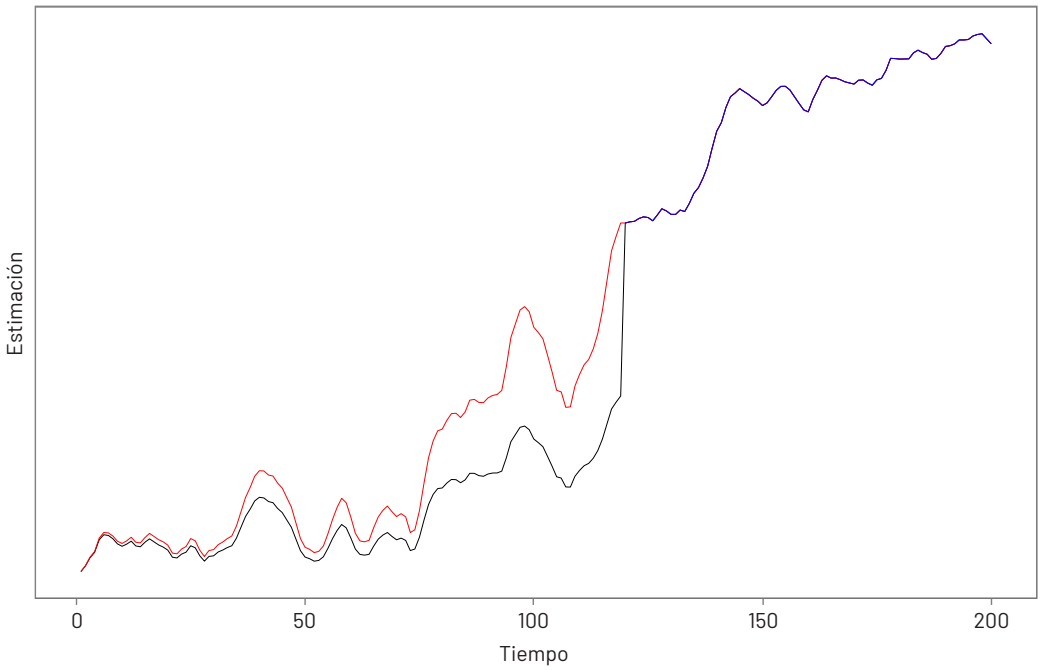
$$\gamma_t = \left(\frac{\hat{\theta}_{t_b}^N}{\hat{\theta}_{t_b - t}^R} - 1 \right) \times \psi_t$$

Así, al suponer que $\psi_t = \frac{t}{t_b - 1} \in (0,1)$ es un factor que aumenta a medida que el tiempo se acerca al punto de quiebre t_b, la serie empalmada se define como:

$$\hat{\theta}_{t_b}^N = \hat{\theta}_{t_b}^R * (1 + \gamma_t), \text{ siendo } t = 1,2,\cdots,t_b - 1$$

En el gráfico XIX.2 es posible observar la manera en que se empalman las series original y nueva a partir del ajuste proporcional. La estructura de la serie se mantiene integralmente.

■ Gráfico XIX.2
Empalme de series de tiempo con el método del factor de suavizado



Fuente: Elaboración propia.
Nota: La línea negra corresponde a la serie original, la línea azul representa la serie nueva y la línea roja denota la serie empalmada.

2. Ajuste sintético aditivo y multiplicativo

En este caso, se supone que la serie original y la nueva se observan desde t_b y que la serie empalmada sigue un ajuste aditivo dado por la siguiente expresión:

$$\hat{\theta}_t^E = \hat{\theta}_t^R + (\hat{\theta}_{t_b}^N - \hat{\theta}_{t_b}^R) \times \psi_t, \text{siendo } t = 1,2,\cdots,t_b - 1$$

Este método tiene la desventaja de que la serie empalmada podría producir valores fuera del rango del indicador de interés (por ejemplo, valores negativos). Por lo tanto, para evitar estos inconvenientes, es posible recurrir a un ajuste multiplicativo, que se ilustra a continuación:

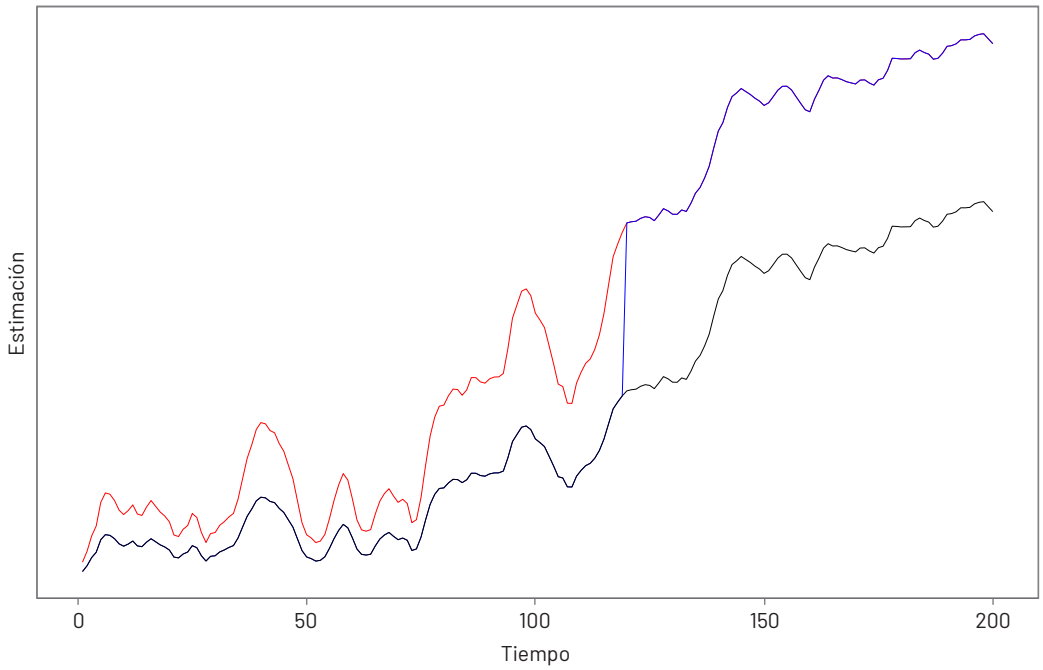
$$\hat{\theta}_t^E = \hat{\theta}_t^R \left(\frac{\hat{\theta}_{t_b}^N}{\hat{\theta}_{t_b}^R} \right) \times \psi_t, \text{siendo } t = 1,2,\cdots,t_b - 1$$

Según Brakel, Smith y Compton (2008), con los métodos anteriores la serie empalmada puede ser mayor que 1 o menor que 0, un resultado especialmente contraproducente en el caso de las proporciones y tasas. Para hacer frente a esta situación, estos autores proponen el siguiente ajuste sintético multiplicativo:

$$\hat{\theta}_t^E = \hat{\theta}_t^R + (\hat{\theta}_{t_b}^N - \hat{\theta}_{t_b}^R) \times \left(\frac{\hat{\theta}_t^R (1 - \hat{\theta}_t^R)}{\hat{\theta}_{t_b}^R (1 - \hat{\theta}_{t_b}^R)} \right), \text{siendo } t = 1,2,\cdots, t_b - 1$$

En el gráfico XIX.3 se muestra el empalme de las series mediante el ajuste sintético multiplicativo.

■ Gráfico XIX.3
Empalme de series de tiempo mediante ajuste sintético multiplicativo



Fuente: Elaboración propia.
Nota: La línea negra corresponde a la serie original, la línea azul representa la serie nueva y la línea roja denota la serie empalmada.

3. Modelos estructurales

Cuando no se puede realizar un experimento paralelo y, por lo tanto, se carece de dos series paralelas, es posible ajustar un modelo estructural de series de tiempo con una intervención justo en el momento de la actualización del diseño. Para simplificar la notación, en esta

sección se supone que la serie no tiene estacionalidad ni ciclos. Si los tuviera, el espíritu del ajuste se mantendría de todos modos. Por ende, de acuerdo con Brakel, Smith y Compton (2008), el modelo estructural para la serie está dado por el nivel L_t más el impacto β en el momento de la actualización t_b y se escribe de la siguiente manera:

$$\hat{\theta}_t = L_t + \beta\delta_t + e_t$$

Donde δ_t es una variable indicadora del momento en que se implementó la actualización de la encuesta:

$$\delta_t = \begin{cases} 1 & \text{si } t \geq t_b \\ 0 & \text{si } t < t_b \end{cases}$$

Además, L_t es una tendencia estocástica autorregresiva que depende de una pendiente:

$$L_t = L_{t-1} + R_{t-1} + w_t$$
$$R_t = R_{t-1} + \eta_t$$

Donde R_t es la pendiente y e_t, w_t y η_t son ruidos de las diferentes ecuaciones. Este modelo debe escribirse en la forma de estado-espacio para que, mediante la aplicación del filtro de Kalman, se puedan estimar los parámetros y extraer los diferentes componentes de la serie (nivel, pendiente y efecto). El modelo de estado-espacio está conformado por las ecuaciones de observación (medición) y estado (transición) que, respectivamente, están dadas por las siguientes expresiones:

$$\hat{\theta}_t = Z_t\,\alpha_t + \epsilon_t$$
$$R_t = \alpha_{t-1} + \omega_t$$

Donde α_t se conoce como el vector de estado. Para el modelo estructural de referencia, α_t está dado por $\alpha_t = (L_t, R_t, \beta)'$. De esta forma, la ecuación de transición está definida por:

$$\alpha_t = \begin{bmatrix} L_t \\ R_t \\ \beta \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} L_{t-1} \\ R_{t-1} \\ \beta \end{bmatrix} + \begin{bmatrix} W_t \\ \eta_t \\ 0 \end{bmatrix}$$

Por su parte, la ecuación de medición se expresa de la siguiente manera:

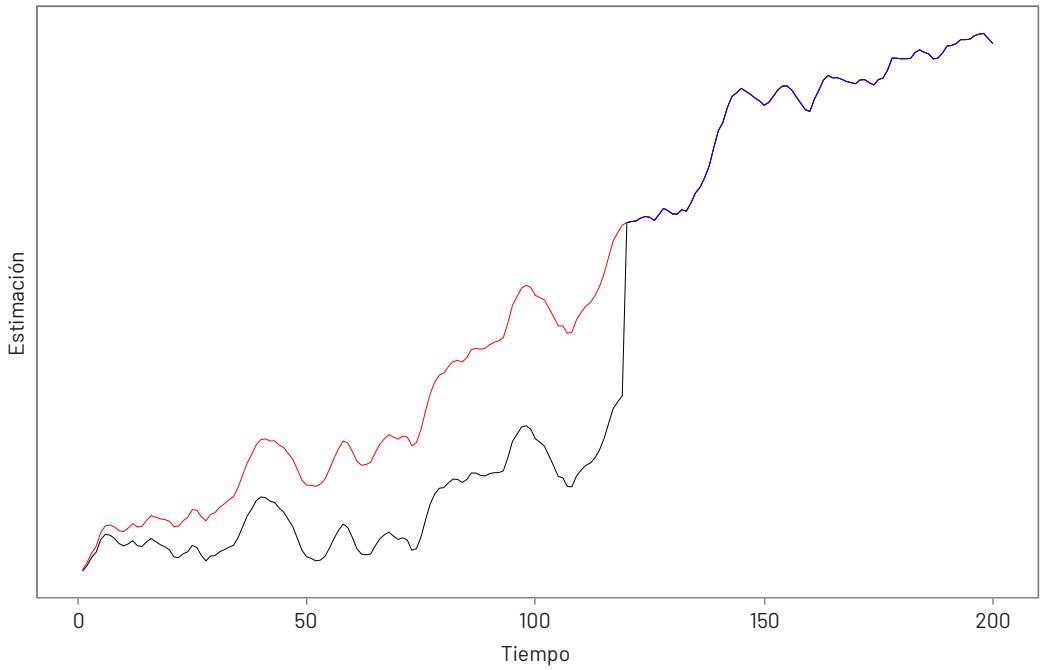
$$\hat{\theta}_t = [1 \;\; 0 \;\; \delta_t] \begin{bmatrix} L_t \\ R_t \\ \beta \end{bmatrix} + e_t$$

Para empalmar la serie antes del punto t_b, se toma la serie original y se añade gradualmente el efecto $\hat{\beta}$. Por ende, la serie empalmada será:

$$x_t^{adj,2} = \begin{cases} x_t + \hat{\beta} * \psi_t & \text{si } t < t_0 \\ x_t & \text{en caso contrario} \end{cases}$$

Los resultados de la aplicación del modelo son bastante satisfactorios, puesto que, además de extraer la estructura de la serie, con el modelo estructural se puede estimar correctamente el impacto de la intervención, sin necesidad de tener las dos series en paralelo. En el gráfico XIX.4 se muestra el empalme de las series utilizando este enfoque. Este tipo de modelos tiene diversas ventajas metodológicas, que incluyen la posibilidad de ajustar más de un punto de intervención e incluso incluir intervenciones de todo tipo (efecto en un solo tiempo, el mismo efecto o efecto creciente a partir de un tiempo). También es posible extraer la tendencia para suavizar la serie, que puede incluir componentes de estacionalidad o ciclos, e incluir otras series como covariables (con relaciones cambiantes en el tiempo).

■ Gráfico XIX.4
Empalme de series de tiempo mediante un modelo estructural simple



Fuente: Elaboración propia.
Nota: La línea negra corresponde a la serie original, la línea azul representa la serie nueva y la línea roja denota la serie empalmada.

Por último, cuando se dispone de ambas series en paralelo, también es posible proponer un modelo estructural bivariado. Si se da por sentado que ninguna de las dos series tiene un componente estacional, es posible formular el modelo estructural bivariado para el vector $(\hat{\theta}_t^R, \hat{\theta}_t^N)'$ de la siguiente manera:

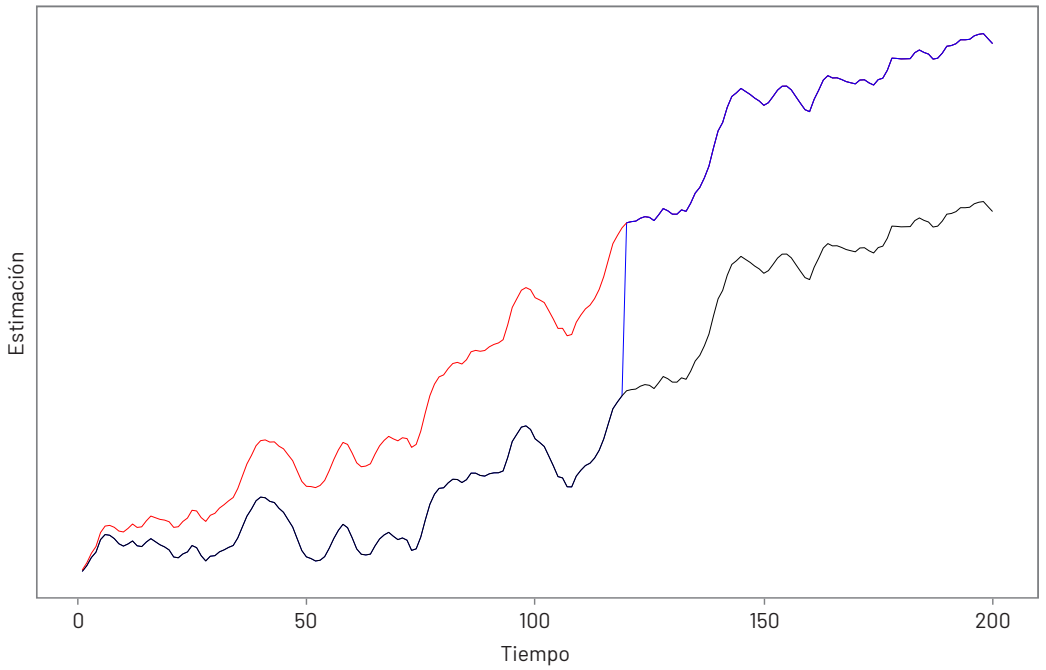
$$\hat{\theta}_t^R = L_t + e_{1,t}$$
$$\hat{\theta}_t^N = L_t + \beta\delta_t + e_{2,t}$$

Para este modelo estructural, la ecuación de transición toma la misma forma que en el modelo univariado, mientras que la ecuación de medición se expresa de la siguiente manera:

$$\begin{bmatrix} \hat{\theta}_t^R \\ \hat{\theta}_t^N \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & \delta_t \end{bmatrix} \begin{bmatrix} L_t \\ R_t \\ \beta \end{bmatrix} + e_t$$

Al igual que en el caso del modelo estructural univariado, los resultados de la aplicación permiten extraer la estructura de la serie y estimar correctamente el impacto de la intervención. En el gráfico XIX.5 se muestra el empalme de las series utilizando este enfoque.

■ Gráfico XIX.5
Empalme de series de tiempo mediante un modelo estructural bivariado



Fuente: Elaboración propia.
Nota: La línea negra corresponde a la serie original, la línea azul representa la serie nueva y la línea roja denota la serie empalmada.

Bibliografía

Alexander, Ch. (1987), "A class of methods for using person controls in household weighting", *Survey Methodology*, vol. 13, N° 1.

Araujo, M. C. (2007), "The 1990 and 2001 Ecuador poverty maps", *More Than a Pretty Picture: Using Poverty Maps to Design Better Policies and Interventions*, T. Bedi, A. Coudouel y K. Simler (eds.), Washington, D.C., Banco Mundial.

Arias, O. y M. Robles (2007), "The geography of monetary poverty in Bolivia: the lessons of poverty maps", *More Than a Pretty Picture: Using Poverty Maps to Design Better Policies and Interventions*, T. Bedi, A. Coudouel y K. Simler (eds.), Washington, D.C., Banco Mundial.

Baillargeon, S. y L.-P. Rivest (2011), "The construction of stratified designs in R with the package stratification", *Survey Methodology*, vol. 37, N° 1.

Baillargeon, S., L.-P. Rivest y M. Ferland (2007), "Stratification en enquêtes entreprises: une revue et quelques avancées", Asamblea anual de la SSC [en línea] https://www.mat.ulaval.ca/fileadmin/mat/documents/lrivest/Publications/86-BaillargeonRivestFerland07.pdf.

Ballin, M. y G. Barcaroli (2013), "Joint determination of optimal stratification and sample allocation using genetic algorithm", *Survey Methodology*, vol. 39, N° 2.

Barcaroli, G. (2014), "SamplingStrata: an R package for the optimization of stratified sampling", *Journal of Statistical Software*, vol. 61, N° 1 [en línea] https://doi.org/10.18637/jss.v061.i04.

Barnett-Walker, K. C. y otros (2003), *2001 National Household Survey on Drug Abuse. Imputation Report* [en línea] https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=51437d6cddc4269adba9dee8ace045bca15a81b6.

Beaumont, J. F. y C. Bocci (2016), "Small area estimation in the Labour Force Survey", documento presentado al Comité Asesor sobre Métodos Estadísticos.

Béland, Y. y otros (2005), The Canadian Community Health Survey: Building on the success from the past", *Proceedings of the American Statistical Association Joint Statistical Meetings*, Minneapolis, American Statistical Association.

Bell, P. (2001), "Comparison of alternative labour force survey estimators", *Survey Methodology*, vol. 27, N° 1.

Bethlehem, J., F. Cobben y B. Schouten (2009), "Indicators for the representativeness of survey response", *Statistics Canada's International Symposium Series: Proceedings* [en línea] https://www150.statcan.gc.ca/n1/en/pub/11-522-x/2008000/article/10976-eng.pdf?st=7EdCIHii.

Biemer, P. P. y L. E. Lyberg (2003), *Introduction to Survey Quality*, Wiley-Interscience.

Brakel, J., P. Smith y S. Compton (2008), "Quality procedures for survey transitions - experiments, time series and discontinuities", *Survey Research Methods*, vol. 2, N° 3.

Casas-Cordero Valencia, C., J. Encina y P. Lahiri (2016), "Poverty mapping for the Chilean comunas", *Analysis of Poverty Data by Small Area Estimation*, M. Pratesi (ed.), Wiley [en línea] https://doi.org/10.1002/9781118814963.ch20.

CEPAL (Comisión Económica para América Latina y el Caribe) (2023), "Recomendaciones metodológicas sobre la medición de la calidad de las cifras provenientes de encuestas de hogares. Documento del Grupo de Trabajo en Encuestas de Hogares de la Conferencia Estadística de las Américas", inédito.

_____ (2022), "BADEHOG: Banco de Datos de Encuestas de Hogares", Santiago.

_____ (2021), "Recomendaciones para los censos de población y vivienda en América Latina. Revisión 2020", *Documentos de Proyectos* (LC/TS.2021/150), Santiago [en línea] https://repositorio.cepal.org/bitstream/handle/11362/47562/S2100743_es.pdf1.

_____ (2020a), "Continuidad del levantamiento de las encuestas de hogares tras la coyuntura de la enfermedad por coronavirus (COVID-19)", *Informes COVID-19*, Santiago, octubre.

_____ (2020b), "Recomendaciones para la publicación de estadísticas oficiales a partir de encuestas de hogares frente a la coyuntura de la enfermedad por coronavirus (COVID-19)", *Informes COVID-19*, Santiago.

_____ (2020c), "Recomendaciones para eliminar el sesgo de selección en las encuestas de hogares en la coyuntura de la enfermedad por coronavirus (COVID-19)", *Informes COVID-19*, Santiago [en línea] https://doi.org/10.18356/9789210054263.

_____ (2018a), *Medición de la pobreza por ingresos: actualización metodológica y resultados*, Metodologías de la CEPAL, N° 2, Santiago [en línea] http://repositorio.cepal.org/bitstream/handle/11362/44314/1/S1800852_es.pdf.

_____ (2018b), "Taller regional sobre desagregación de estadísticas sociales mediante metodologías de estimación en áreas pequeñas" [en línea] https://www.cepal.org/es/cursos/taller-regional-desagregacion-estadisticas-sociales-mediante-metodologias-estimacion-areas.

_____ (1983), *Las encuestas de hogares en América Latina*, Cuadernos de la CEPAL (E/CEPAL/G.1244), Santiago.

Clark, R. G. y D. G. Steel (2007), "Sampling within households in household surveys", *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 170, N° 1.

Cochran, W. G. (1977), *Sampling Techniques*, Wiley.

Costa, G. (2007), "Coordenação de amostras PPT em pesquisas repetidas, utilizando o método de amostragem de Pareto", Tesis doctoral, Instituto Brasileño de Geografía y Estadística (IBGE)-Escuela Nacional de Ciencias Estadísticas (ENCE).

Dalenius, T. y J. L. Hodges (1959), "Minimum variance stratification", *Journal of the American Statistical Association*, vol. 54, N° 285.

DANE (Departamento Administrativo Nacional de Estadística) (2018), "Encuesta Nacional de Presupuestos de los Hogares (ENPH)" [en línea] https://www.dane.gov.co/index.php/estadisticas-por-tema/pobreza-y-condiciones-de-vida/encuesta-nacional-de-presupuestos-de-los-hogares-enph.

_____ (2017), "Gran Encuesta Integrada de Hogares" [en línea] http://formularios.dane.gov.co/Anda_4_1/index.php/catalog/458.

Dever, J. (2008), *Sampling Weight Calibration with Estimated Control Totals* [en línea] https://drum.lib.umd.edu/bitstream/handle/1903/8815/umi-umd-5841.pdf?sequence=1&isAllowed=y.

Dever, J. y R. Valliant (2016), "General regression estimation adjusted for undercoverage and estimated control totals", *Journal of Survey Statistics and Methodology*, vol. 4, N° 3 [en línea] https://doi.org/10.1093/jssam/smw001.

Deville, J.-C. y C.-E. Särndal (1992), "Calibration estimators in survey sampling", *Journal of the American Statistical Association*, vol. 87, N° 418, Taylor & Francis [en línea] https://doi.org/10.1080/01621459.1992.10475217.

DGEEC (Dirección General de Estadística, Encuestas y Censos) (2018a), "Aspectos metodológicos de la Encuesta de Ingresos y Gastos y de Condiciones de Vida (EIGyCV)" [en línea] http://www.dgeec.gov.py/microdatos/register/eig/Metodologia%20EIG%20y%20CV.pdf.

_____ (2018b), "Encuesta Permanente de Hogares" [en línea] http://www.dgeec.gov.py/Publicaciones/Biblioteca/eph2016/Boletin-de-pobreza-2016.pdf.

DIGESTYC (Dirección General de Estadística y Censos) (2018a), "Encuesta de Hogares de Propósitos Múltiples" [en línea] http://www.digestyc.gob.sv/index.php/temas/des/ehpm.html.

_____ (2018b), "Encuesta de Ingresos y Gastos de los Hogares" [en línea] http://www.censos.gob.sv/enigh/descargas/ENIGH_Publicacion.pdf.

DiNatale, M. L. (2003), "Creating Comparability in CPS Employment Series" [en línea] https://www.bls.gov/cps/cpscomp.pdf.

Duncan, G. J. y G. Kalton (1987), "Issues of design and analysis of surveys across time", *International Statistical Review*, vol. 55, N° 1 [en línea] https://doi.org/10.2307/1403273.

Efron, B. y R. J. Tibshirani (1993), *An Introduction to the Bootstrap*, Monographs on Statistics and Applied Probability, N° 57, Boca Raton, Florida, Chapman & Hall/CRC.

Estevao, V. y C.-E. Särndal (2006), "Survey estimates by calibration on complex auxiliary information", *International Statistical Review*, vol. 74, N° 2.

Fay, R. E. y R. A. Herriot (1979), "Estimates of income for small places: An application of James-Stein procedures to census data", *Journal of the American Statistical Association*, vol. 74, N° 366a [en línea] https://doi.org/10.1080/01621459.1979.10482505.

Feinberg, S. y E. Stasny (1983), "Estimating monthly gross flows in labour force participation", *Survey Methodology*, vol. 9, N° 1.

Filzmoser, P., J. Gussenbauer y M. Templ (2016), *Detecting Outliers in Household Consumption Survey Data*, Viena, Universidad Tecnológica de Viena.

Foster, J., J. Greer y E. Thorbecke (1984), "A class of decomposable poverty measures", *Econometrica*, vol. 52, N° 3 [en línea] https://doi.org/10.2307/1913475.

Fuller, W. A. (2009), *Sampling Statistics*, Wiley.
_____ (1990), "Analysis of repeated surveys", *Survey Methodology*, vol. 16, N° 2.
Fuller, W. A. y J. N. K. Rao (2001), "A regression composite estimator with application to the Canadian Labour Force Survey", *Survey Methodology*, vol. 27, Nº 1.
Fuquene, J. y otros (2019), "Prevalence of international migration: an alternative for small area estimation" [en línea] https://arxiv.org/pdf/1905.00353.pdf.
Gambino, J. G. (2009), "Design effect caveats", *The American Statistician*, vol. 63, N° 2 [online] https://doi.org/10.1198/tast.2009.0028.
Gambino, J. G. y P. L. do N. Silva (2009), "Chapter 16 - Sampling and estimation in household surveys", *Handbook of Statistics*, Elsevier [online] https://doi.org/10.1016/S0169-7161(08)00016-3.
Gambino, J. G., B. Kennedy y M. P. Singh (2001), "Regression composite estimation for the Canadian Labour Force Survey: evaluation and implementation", *Survey Methodology*, vol. 27, N° 1.
Gbur, E. y Ch. Alexander (1984), "A linear model approach to the estimation of survey redesign effects", *SRD Research Report*, N° CENSUS/SRD/RR-84/24, Washington, D.C.
Grafstrom, A. y A. Matei (2015), "Coordination of conditional Poisson samples", *Journal of Official Statistics*, vol. 31, N° 4 [en línea] https://doi.org/10.1515/jos-2015-0039.
Groves, R. y otros (2009), *Survey Methodology*, Hoboken, John Wiley & Sons.
Gunning, P. y J. M. Horgan (2004), "A new algorithm for the construction of stratum boundaries in skewed populations", *Survey Methodology*, vol, 30, N° 2.
Gurney, M. y J. Daly (1965), "A multivariate approach to estimation in periodic sample surveys", *Proceedings of the Social Statistics Section*, Washington, D.C., American Statistical Association.
Gutiérrez, H. A. (2020), samplesize4surveys: Sample Size Calculations for Complex Surveys [en línea] https://rdrr.io/cran/samplesize4surveys/.
_____ (2016), *Estrategias de muestreo: diseño de encuestas y estimación de parámetros*, Ediciones de la U.
_____ (2015), "TeachingSampling: selection of samples and parameter estimation in finite population" [en línea] https://CRAN.R-project.org/package=TeachingSampling.
_____ (2014), "The estimation of gross flows in complex surveys with random nonresponse", *Survey Methodology*, vol. 40, N° 2.
Gutiérrez, H. A., H. Zhang y C. Montaño (2016), "Cálculo del tamaño de muestra para la estimación de una varianza en poblaciones finitas con funciones en R", *Comunicaciones en estadística*, vol. 9, N° 1 [en línea] https://doi.org/10.15332/s2027-3355.2016.0001.06.
Gutiérrez, H. A., H. Zhang y N. Rodríguez (2016), "The performance of multivariate calibration on ratios, means and proportions", *Revista Colombiana de Estadística*, vol. 39, N° 2 [en línea] https://doi.org/10.15446/rce.v39n2.55424.
Hansen, M. H, W. N. Hurwitz y W. G. Madow (1953), *Sample Survey Methods and Theory. Volume 1: Methods and Applications*, Nueva York, Wiley.
Hayes, C. y N. Watson (2009), "HILDA imputation methods", *HILDA Project Technical Paper Series*, N° 2/09, Universidad de Melbourne [en línea] https://melbourneinstitute.unimelb.edu.au/assets/documents/hilda-bibliography/hilda-technical-papers/htec209.pdf.

Heeringa, S. G., B. T. West y P. A. Berglund (2017), *Applied Survey Data Analysis*, Chapman & Hall/ CRC Statistics in the Social and Behavioral Sciences Series, CRC Press.

_____ (2010), *Applied Survey Data Analysis*, Chapman & Hall/ CRC Statistics in the Social and Behavioral Sciences Series, CRC Press.

Heldal, J. (1992), "A method for calibration of weights in sample surveys" [en línea] https://www.ssb.no/a/histstat/aap/aap_metode_199203.pdf .

Hornik, R. y otros (2002), *Evaluation of the National Youth Anti-Drug Media Campaign: Fourth Semi-Annual Report of Findings* [en línea] https://archives.drugabuse.gov/sites/default/files/fullreport.pdf.

Horvitz, D. G. y D. J. Thompson (1952), "A generalization of sampling without replacement from a finite universe", *Journal of the American Statistical Association*, vol. 47, N° 260.

IBGE (Instituto Brasileño de Geografía y Estadística) (2018a), "Pesquisa de Orçamentos Familiares" [en línea] https://ww2.ibge.gov.br/home/estatistica/pesquisas/pesquisa_resultados.php?id_pesquisa=25.

_____ (2018b), "Pesquisa Nacional por Amostra de Domicílios Contínua" [en línea] https://www.ibge.gov.br/estatisticas-novoportal/sociais/trabalho/9173-pesquisa-nacional-por-amostra-de-domicilios-continua-trimestral.html?redirect=1.

_____ (2014), "Pesquisa Nacional por Amostra de Domicilios Continua: notas metodológicas".

IBM (2017), IBM SPSS Complex Samples [en línea] ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/23.0/en/client/Manuals/IBM_SPSS_Complex_Samples.pdf.

Imbens, G. W. y D. B. Rubin (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, vol. 84, Cambridge University Press.

INDEC (Instituto Nacional de Estadística y Censos) (2018a), "Encuesta Nacional de Gastos de los Hogares" [en línea] https://www.indec.gob.ar/engho/.

_____ (2018b), "Encuesta Permanente de Hogares" [en línea] https://www.indec.gov.ar/bases-de-datos.asp.

INE (Instituto Nacional de Estadística) (2018a), "Encuesta Nacional de Empleo e Ingresos 2-2018" [en línea] https://www.ine.gob.gt/sistema/uploads/2019/07/05/publicacion_ENEI_2_2018.pdf.

_____ (2018b), "Encuesta de Hogares" [en línea] http://www.ine.gob.bo/sitio_EH/Encuesta_Hogares.html.

_____ (2018c), "Encuesta de Presupuestos Familiares (EPF)" [en línea] https://www.ine.cl/estadisticas/ingresos-y-gastos/epf.

_____ (2018d), "Encuesta Nacional de Condiciones de Vida" [en línea] https://www.ine.gob.gt/index.php/encuestas-de-hogares-y-personas/condiciones-de-vida.

_____ (2018e), "Ficha técnica de Encuesta de Hogares por Muestreo" [en línea] http://www.ine.gov.ve/index.php?option=com_contentandid=333andItemid=103.

_____ (2018f), *Encuesta Permanente de Hogares de Propósitos Múltiples 2018: metodología* [en línea] https://ine.gob.hn/v4/wp-content/uploads/2023/07/03-EPHPM-Metodologia-2018.pdf.

_____ (2016a), "Encuesta Continua de Hogares (ECH)" [en línea] http://ine.gub.uy/encuesta-continua-de-hogares1.

_____ (2016b), "Encuesta de Gastos e Ingresos de los Hogares – ENGIH 2016/2017" [en línea] http://www.ine.gub.uy/engih2016.
_____ (2016c), "Encuesta Nacional de Condiciones de Vida 2014", tomo I [en línea] https://www.ine.gob.gt/sistema/uploads/2016/02/03/bWC7f6t7aSbEI4wmuExoNR0oScpSHKyB.pdf
_____ (2013), *Encuesta de Hogares por Muestreo*, documento metodológico, Caracas.
_____ (2004), *ENCOVI: encuesta nacional de condiciones de vida 2004*, Tegucigalpa.
INE (Instituto Nacional de Estadísticas de Chile) (2020), *Estándar para la evaluación de la calidad de las estimaciones en encuestas de hogares*, Santiago, febrero.
INEC (Instituto Nacional de Estadística y Censos) (2019), "Comentario: situación del mercado laboral: marzo 2019" [en línea] https://www.inec.gob.pa/archivos/P9471Comentarios.pdf.
_____ (2018a), "Encuesta de Condiciones de Vida" [en línea] http://www.ecuadorencifras.gob.ec/encuesta-de-condiciones-de-vida-ecv/.
_____ (2018b), "Encuesta de Ingresos y Gastos de los Hogares" [en línea] http://www.contraloria.gob.pa/inec/Aplicaciones/EIGH2008/intro.html.
_____ (2018c), "Encuesta Nacional de Ingresos y Gastos de los Hogares" [en línea] http://www.inec.go.cr/encuestas/encuesta-nacional-de-ingresos-y-gastos-de-los-hogares.
_____ (2018d), "Instituto Nacional de Estadística y Censo – Panamá" [en línea] https://www.contraloria.gob.pa/inec/Publicaciones/Publicaciones.aspx?ID_SUBCATEGORIA=38andID_PUBLICACION=91andID_IDIOMA=1andID_CATEGORIA=5.
_____ (2018e), Instituto Nacional de Estadística y Censos [en línea] http://www.ilo.org/surveydata/index.php/catalog/1393/study-description.
_____ (2017), "Encuesta Nacional de Hogares" [en línea] http://www.inec.go.cr/encuestas/encuesta-nacional-de-hogares.
INEGI (Instituto Nacional de Estadística y Geografía) (2020), *Cómo se hace la ENOE: métodos y procedimientos*, Ciudad de México.
_____ (2016), "Encuesta Nacional de Ingresos y Gastos de los Hogares. 2016 Nueva serie" [en línea] https://www.inegi.org.mx/programas/enigh/nc/2016/.
_____ (2012), "Metodología de la construcción del marco maestro de muestreo 2012 y del diseño de la muestra maestra 2012".
INEI (Instituto Nacional de Estadística e Informática) (2016), "Encuesta Nacional de Hogares sobre Condiciones de Vida y Pobreza" [en línea] http://webinei.inei.gob.pe/anda_inei/index.php/catalog/543.
INIDE (Instituto Nacional de Información de Desarrollo) (2018) [en línea] http://www.inide.gob.ni/.
Jacob, G. (2020), *Surf: Survey-based Gross-Flow Estimation.*
Jarque, C. M. (1981), "A solution to the problem of optimum stratification in multivariate sampling", *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 30, N° 2 [en línea] https://doi.org/10.2307/2346387.
Judkins, D. R. (1990), "Fay's method for variance estimation", *Journal of Official Statistics*, vol. 6, N° 3.
Kalton, G. (2013), "Some issues in the design and analysis of longitudinal surveys", *Proceedings of the 59th World Statistics Congress of the International Statistical Institute*, La Haya, Instituto Internacional de Estadística.

Kalton, G. y C. F. Citro (1993), "Panel surveys: adding the fourth dimension", *Survey Methodology*, vol. 19, N° 2.
Kalton, G. e I. Flores-Cervantes (2003), "Weighting methods", *Journal of Official Statistics*, vol. 19, N° 2.
Kim, J. K. y M. K. Riddles (2012), "Some theory for propensity-score-adjustment estimators in survey sampling", *Survey Methodology*, vol. 38, N° 2.
Kish, L. (2004), *Statistical Design for Research*, Wiley [en línea] https://www.wiley.com/en-us/Statistical+Design+for+Research-p-9780471691204.
_____ (1999), "Cumulating/combining population surveys", *Survey Methodology*, vol. 25, N° 2.
_____ (1965), *Survey Sampling*, John Wiley & Sons.
Korn, E. L. y B. I. Graubard (1999), *Analysis of Health Surveys*, Wiley.
Kozak, M. (2004), "Optimal stratification using random search method in agricultural surveys", *Statistic in Transition*, vol. 6, N° 5.
Krewski, D. y J. N. K. Rao (1981), "Inference from stratified samples: Properties of the linearization, Jackknife and balanced repeated replication methods", *The Annals of Statistics*, vol. 9, N° 5.
Kruskal, W. y F. Mosteller (1980), "Representative sampling, IV: The history of the concept in statistics, 1895-1939", *International Statistical Review*, vol. 48, N° 2.
_____ (1979a), "Representative sampling, I: Non-scientific literature", International Statistical Review, vol. 47, N° 1.
_____ (1979b), "Representative sampling, II: Scientific literature, excluding statistics", *International Statistical Review*, vol. 47, N° 2.
_____ (1979c), "Representative sampling, III: The current statistical literature", International Statistical Review, vol. 47, N° 3.
LaRoche, S. (2003), *Longitudinal and Cross-Sectional Weighting of the Survey of Labour and Income Dynamics*, Statistics Canada.
Lavallée, P. y M. A. Hidiroglou (1988), "On the stratification of skewed populations", *Survey Methodology*, vol. 14, N° 1.
Lemaitre, G. y J. Dufour (1987), "An integrated method for weighting persons and families", *Survey Methodology*, vol. 13, N° 2.
Lent, J., S. M. Miller y M. Duff (1999), "Effects of composite weights on some estimates from the current population survey", *Journal of Official Statistics*, vol. 15, N° 3.
Lewis, T. (2017), "Estimation strategies involving pooled survey data", *SAS Global Forum 2017* [en línea] https://support.sas.com/resources/papers/proceedings17/0767-2017.pdf.
Likert, R. (1932), "A technique for the measurement of attitudes", *Archives of Psychology*, vol. 22, N° 40.
Little, R. y D. B. Rubin (2002), *Statistical Analysis with Missing Data*, Wiley.
Lohr, S. (2000), *Sampling: Design and Analysis*, Duxbury Press.
López-Calva, L. F., L. Rodríguez-Chamussy y M. Székely (2007), "Poverty maps and public policy in Mexico", *More Than A Pretty Picture: Using Poverty Maps to Design Better Policies and Interventions*, T. Bedi, A. Coudouel y K. Simler (eds.), Washington, D.C., Banco Mundial.
Lumley, T. (2016), "Survey: analysis of complex survey samples".
_____ (2010), *Complex Surveys: a Guide to Analysis Using R*, Wiley Series in Survey Methodology, Wiley.

Lynn, P. (2009), *Methodology of Longitudinal Survey*, Wiley Series in Survey Methodology, Wiley.
Macari, A. y J. P. Ferreira (2020), "Encuesta Continua de Hogares (ECH) en Uruguay: nueva metodología 2021", presentación realizada en el seminario web COVID-19: Evaluación del Efecto del Modo de Recolección sobre las Estadísticas Oficiales, 7 de diciembre [en línea] https://rtc-cea.cepal.org/sites/default/files/2020-12/Presentación%20Uruguay.pdf.
Macqueen, J. (1967), "Some methods for classification and analysis of multivariate observations", *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press.
McCarthy, P. J. (1969), "Pseudo-replication: half samples", *Review of the International Statistical Institute*, vol. 37, N° 3 [en línea] https://doi.org/10.2307/1402116.
McLaren, C. y D. G. Steel (2001), "Rotation patterns and trend estimation for repeated surveys using rotation group estimates", *Statistica Neerlandica*, vol. 55, N° 2, Wiley.
MDSF/CEPAL (Ministerio de Desarrollo Social y Familia/Comisión Económica para América Latina y el Caribe) (2021), *Estimaciones comunales de pobreza por ingresos en Chile mediante métodos de estimación en áreas pequeñas*, Santiago, diciembre.
Ministerio de Desarrollo Social y Familia (2015), "Observatorio Social" [en línea] http://observatorio.ministeriodesarrollosocial.gob.cl/casen/casen_obj.php.
Mosteller, F. (1968), *Data Analysis, Including Statistics* [en línea] https://books.google.cl/books?id=6ptDHQAACAAJ.
Naciones Unidas (2018), *Progresos realizados para lograr los Objetivos de Desarrollo Sostenible. Informe del Secretario General* (E/2018/64), Consejo Económico y Social [en línea] https://digitallibrary.un.org/record/1627573/files/E_2018_64-ES.pdf.
_____ (2016), *Global Sustainable Development Report 2016* [en línea] https://sustainabledevelopment.un.org/globalsdreport/2016.
_____ (2015), "Transformar nuestro mundo: la Agenda 2030 para el Desarrollo Sostenible", Resolución aprobada por la Asamblea General el 25 de septiembre de 2015 [en línea] http://unctad.org/meetings/es/SessionalDocuments/ares70d1_es.pdf.
_____ (2011), *Canberra Group Handbook on Household Income Statistics. Second Edition 2011* (ECE/CES/11), Nueva York [en línea] https://www.unece.org/fileadmin/DAM/stats/groups/cgh/Canbera_Handbook_2011_WEB.pdf.
_____ (2008), "Diseño de muestras para encuestas de hogares: directrices prácticas", *Estudios de Métodos*, serie F, N° 98 (ST/ESA/STAT/SER.F/98), Nueva York.
_____ (2007), "Encuestas de hogares en los países en desarrollo y en transición", *Estudios de Métodos*, serie F, N° 96 (ST/ESA/STAT/SER.F/96), Nueva York.
National Research Council (2015), *Realizing the Potential of the American Community Survey: Challenges, Tradeoffs, and Opportunities*, Washington, D.C., National Academies Press [en línea] https://doi.org/10.17226/21653.
Naud, J. F. (2002), "Combined-panel longitudinal weighting. Survey of Labour and Income Dynamics", Statistics Canada [en línea] https://publications.gc.ca/collections/Collection/Statcan/75F0002MIE/75F0002MIE2004008.pdf.
Neethling, A. y J. S. Galpin (2006), "Weighting of household survey data: a comparison of various calibration, integrated and cosmetic estimators", *South African Statistical Journal*, vol. 40, N° 2.

Oficina de Estadísticas Laborales (2014), *Redesign of the Sample for the Current Population Survey* [en línea] https://www.bls.gov/cps/sample_redesign_2014.pdf.
Oficina del Censo de los Estados Unidos (1965), *Atlantida: A Case Study in Household Sample Surveys*, Washington, D.C.
Ohlsson, E. (1995), "Coordination of samples using permanent random numbers", *Business Survey Methods*, Wiley.
OIT (Organización Internacional del Trabajo) (2013a), "Estadísticas del trabajo, el empleo y la subutilización de la fuerza de trabajo" (ICLS/19/2013/2), 19 Conferencia Internacional de Estadísticos del Trabajo, Ginebra [en línea] http://www.ilo.org/wcmsp5/groups/public/—-dgreports/—-stat/documents/publication/wcms_220537.pdf.
_____ (2013b), "Resolución sobre las estadísticas del trabajo, la ocupación y la subutilización de la fuerza de trabajo", 19ª Conferencia Internacional de Estadísticos del Trabajo, Ginebra [en línea] http://www.ilo.org/wcmsp5/groups/public/—-dgreports/—-stat/documents/normativeinstrument/wcms_234036.pdf.
_____ (1982), "Resolución sobre estadísticas de la población económicamente activa, del empleo, del desempleo y del subempleo adoptada por la decimotercera Conferencia Internacional de Estadísticos del Trabajo" [en línea] http://www.ilo.org/wcmsp5/groups/public/—-dgreports/—-stat/documents/normativeinstrument/wcms_087483.pdf.
ONE (Oficina Nacional de Estadística) (2018a), "Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH)" [en línea] https://www.one.gob.do/enhogar.
_____ (2018b), "Oficina Nacional de Estadística (ONE)" [en línea] https://www.one.gob.do/encuestas/enigh.
ONEI (Oficina Nacional de Estadística e Información) (2018), "Oficina Nacional de Estadísticas" [en línea] http://www.one.cu/sitioone2006.asp.
Opsomer, J. D. y A. L. Erciulescu (2022), "Replication variance estimation after sample-based calibration", *Survey Methodology*, vol. 47, N° 2.
Park, I. y otros (2003), "Design effects and survey planning" [en línea] https://www.researchgate.net/publication/242534565_DESIGN_EFFECTS_AND_SURVEY_PLANNING.
Parker, J. D., M. Talih y D. J. Malec (2017), "National Center for Health Statistics data presentation standards for proportions", *Vital and Health Statistics*, Serie 2, N° 175.
Presser, S. y otros (2004), *Methods for Testing and Evaluating Survey Questionnaires*, John Wiley & Sons.
Preston, J. (2009), "Rescaled bootstrap for stratified multistage sampling", *Survey Methodology*, vol. 35, N° 2.
Quenouille, M. H. (1956), "Notes on bias in estimation", *Biometrika*, vol. 43, N° 3/4.
R Core Team (2020), *R: A Language and Environment for Statistical Computing*, Viena, R Foundation for Statistical Computing [en línea] https://www.R-project.org/.
Rao, J. N. K. e I. Molina (2015), *Small-Area Estimation*, John Wiley & Sons [en línea] https://onlinelibrary.wiley.com/doi/book/10.1002/9781118735855.

Rao, J. N. K. y C. F. J. Wu (1988), "Resampling inference with complex survey data", *Journal of the American Statistical Association*, vol. 83, N° 401 [en línea] https://doi.org/10.1080/01621459.1988.10478591.

_____ (1984), "Bootstrap inference for sample surveys", *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Rivest, L.-P. y E. Belmonte (2000), "A conditional mean squared error of small area estimators", *Survey Methodology*, vol. 26, N° 1.

Rosén, B. (1997), "On sampling with probability proportional to size", *Journal of Statistical Planning and Inference*, vol. 62, N° 2 [en línea] https://doi.org/10.1016/S0378-3758(96)00186-3.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, Wiley Series in Probability and Mathematical Statistics. Applied probability and statistics, John Wiley & Sons.

Särndal, C.-E. (2011a), "Three factors to signal non-response bias with applications to categorical auxiliary variables", *International Statistical Review*, vol. 79, N° 2.

_____ (2011b), "The 2010 Morris Hansen lecture: Dealing with survey nonresponse in data collection, in estimation", *Journal of Official Statistics*, vol. 27, N° 1.

_____ (2007), "The calibration approach in survey theory and practice", *Survey Methodology*, vol. 33, N° 2.

Särndal, C.-E. y S. Lundström (2010), "Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias", *Survey Methodology*, vol. 36, N° 2.

_____ (2005), *Estimation in Surveys with Nonresponse*, Wiley.

Särndal, C.-E., B. Swensson y J. Wretman (2003), *Model Assisted Survey Sampling*, Springer Science + Business Media.

Sartore, L. y otros (2019), "Developing integer calibration weights for census of agriculture", *Journal of Agricultural, Biological and Environmental Statistics*, vol. 24, N° 1 [en línea] https://doi.org/10.1007/s13253-018-00340-4.

SAS (2010), *SAS/STAT 9.22 User's Guide. Introduction to Survey Sampling and Analysis Procedures* [en línea] https://support.sas.com/documentation/cdl/en/statugsurveysamp/63778/PDF/default/statugsurveysamp.pdf.

Schwarz, N. y otros (1991), "Rating scales: numeric values may change the meaning of scale labels", *Public Opinion Quarterly*, vol. 55, N° 4.

Sen, A. R. (1953), "On the estimate of the variance in sampling with varying probabilities", *Journal of the Indian Society of Agricultural Statistics*, vol. 5.

Sethi, V. K. (1963), "A note on the optimum stratification of populations for estimating the population means", *Australian & New Zealand Journal of Statistics*, vol. 5, N° 1, Wiley.

Shao, J. y D. Tu (2012), *The Jackknife and Bootstrap*, Springer Series in Statistics, Nueva York, Springer.

Shlomo, N., Ch. Skinner y B. Schouten (2012), "Estimation of an indicator of the representativeness of survey response", *Journal of Statistical Planning and Inference*, vol. 142, N° 1 [en línea] https://doi.org/10.1016/j.jspi.2011.07.008.

Silva, P. L. do N. (2004), "Calibration estimation: when and why, how much and how", *Textos para Discussão*, N° 15, Río de Janeiro, Instituto Brasileño de Geografía y Estadística.

Singh, M. P., J. G. Gambino y H. J. Mantel (1994), "Issues and strategies for small area data", *Survey Methodology*, vol. 20, N° 1, Statistics Canada.

Singh, A. C., M. Westlake y M. Feder (2004), "A generalization of the coefficient of variation with application to suppression of imprecise estimates" [en línea] http://www.asasrms.org/Proceedings/y2004/files/Jsm2004-000884.pdf.

Starick, R. y N. Watson (2011), "Evaluation of alternative income imputation methods for the HILDA survey" [en línea] https://melbourneinstitute.unimelb.edu.au/assets/documents/hilda-bibliography/hilda-conference-papers/2007/Watson,-Nicole_final-paper.pdf.

STATA (2013), *STATA Survey Data* [en línea] https://www.stata.com/manuals13/svy.pdf.

Steel, D. y C. McLaren (2008), "Design and Analysis of Repeated Surveys" [en línea] https://www.researchgate.net/publication/44843092_Design_and_Analysis_of_Repeated_Surveys.

Sun, C. (2010), "HILDA expenditure imputation", *HILDA Project Technical Paper Series*, N° 1/10, Universidad de Melbourne [en línea] https://melbourneinstitute.unimelb.edu.au/assets/documents/hilda-bibliography/hilda-technical-papers/htec110.pdf.

Tillé, Y. (2019), "A simple and efficient way of rounding calibration weights" [en línea] https://libra.unine.ch/export/DL/INSTITUTDESTATISTIQUE/39140.pdf.

_____ (2006), *Sampling Algorithms*, Springer Series in Statistics, Springer-Verlag [en línea] https://doi.org/10.1007/0-387-34240-0.

Tillé, Y. y A. Matei (2016), *Sampling: Survey Sampling* [en línea] https://CRAN.R-project.org/package=sampling.

Train, G., L. Cahoon y P. Makens (1978), "The current population survey variances, inter-relationships and design effects" [en línea] http://www.asasrms.org/Proceedings/papers/1978_090.pdf.

Valliant, R. y J. A. Dever (2017), *Survey Weights: A Step-by-step Guide to Calculation*, Texas, Stata Press.

Valliant, R., J. A. Dever y F. Kreuter (2018), *Practical Tools for Designing and Weighting Survey Samples*, Statistics for Social and Behavioral Sciences, Springer International Publishing [en línea] https://doi.org/10.1007/978-3-319-93632-1.

_____ (2013), *Practical Tools for Designing and Weighting Survey Samples*, Nueva York, Springer [en línea] https://doi.org/10.1007/978-1-4614-6449-5.

Van Buuren, S. (2018), *Flexible Imputation of Missing Data*, Nueva York, Chapman & Hall/CRC.

Vehovar, V. (1999), "Field substitution and unit nonresponse", *Journal of Official Statistics*, vol. 15, N° 2.

Verma, V., G. Betti y G. Ghellini (2006), "Cross-sectional and longitudinal weighting in a rotational household panel: applications to EU-SILC", *Documento de Trabajo*, N° 67 [en línea] http://repec.deps.unisi.it/quaderni/67DMQ.pdf.

Westat (1997), WesVar 4.3. Users Guide [en línea] http://users.nber.org/~jroth/chap1.pdf.

Wolter, K. M. (2007), *Introduction to Variance Estimation*, Nueva York, Springer.

Yates, F. y P. M. Grundy (1953), "Selection without replacement from within strata with probability proportional to size", *Journal of the Royal Statistical Society*, vol. 15, N° 2.

Anexos

Anexo 1

Una perspectiva regional de las encuestas de hogares

En esta sección se presenta una breve descripción de la situación de las encuestas de hogares en América Latina. Aunque no se pretende hacer un resumen exhaustivo de cada encuesta y de sus componentes metodológicos, el lector podrá conocer las características principales de las encuestas de hogares y sus condiciones de aplicación.

En esta sección se presenta una breve descripción de la situación de las encuestas de hogares en América Latina. Aunque no se pretende hacer un resumen exhaustivo de cada encuesta y de sus componentes metodológicos, el lector podrá conocer las características principales de las encuestas de hogares y sus condiciones de aplicación.

Algunas de estas encuestas son estandarizadas por la División de Estadísticas de la Comisión Económica para América Latina y el Caribe (CEPAL) en su Banco de Datos de Encuestas de Hogares (BADEHOG), con el que se generan estimaciones comparables de indicadores de pobreza y desigualdad año tras año para América Latina (CEPAL, 2022). Esta iniciativa tuvo su desarrollo principal en la década de 1990, y su objetivo era recopilar anualmente las encuestas de hogares de 18 países de la región. Estas encuestas de hogares tienen la particularidad de que sirven para medir el ingreso y se utilizan para estimar la pobreza y la desigualdad distributiva. Actualmente, el BADEHOG tiene en su haber encuestas desde 1980 en adelante y constituye un insumo esencial para la producción de indicadores sociales armonizados en América Latina.

A. Descripción de algunas encuestas por países

1. Argentina

El Instituto Nacional de Estadística y Censos (INDEC) lleva a cabo de forma trimestral la Encuesta Permanente de Hogares, que permite caracterizar la situación social de los individuos y las familias, teniendo en cuenta su inserción en la estructura social y económica (INDEC, 2018b). Esta encuesta brinda información sobre las características demográficas básicas de los miembros del hogar, su situación laboral y sus ingresos, así como sus características educacionales y de migración. También permite caracterizar las viviendas.

Por otro lado, la Encuesta Nacional de Gastos de los Hogares proporciona información sobre los hogares argentinos mediante la recopilación de datos sobre sus gastos e ingresos. Sus resultados contribuyen a la elaboración de la canasta de bienes y servicios que se utiliza para medir el índice de precios al consumidor (IPC), y aportan información para la estimación de la pobreza y la producción de indicadores de la economía nacional (INDEC, 2018a).

2. Estado Plurinacional de Bolivia

El objetivo principal de la Encuesta Continua de Empleo aplicada cada año por el Instituto Nacional de Estadística (INE) es suministrar información sobre las condiciones de vida de los hogares, a partir de la recopilación de información de variables económicas y demográficas. Entre los ejes temáticos que aborda la encuesta, se encuentran la estimación de las necesidades básicas insatisfechas, el acceso a los servicios públicos, la caracterización demográfica de los individuos, los desplazamientos de la población en los últimos cinco años, el estado de salud de los miembros del hogar, las características educativas, las condiciones de ocupación, los ingresos percibidos y los gastos del hogar. Esta encuesta permite medir oportunamente los indicadores de pobreza de la población boliviana, así como el acceso a la vivienda y a los servicios básicos y el nivel de educación, entre otras cosas. Mediante la encuesta de hogares, el INE obtiene estadísticas e indicadores socioeconómicos y demográficos de la población que son necesarios para la formulación, evaluación, monitoreo y seguimiento de las políticas del Estado (INE, 2018b).

3. Brasil

La Encuesta Nacional de Hogares es implementada anualmente por el Instituto Brasileño de Geografía y Estadística (IBGE), con el objetivo de producir información básica para el estudio de la evolución económica del Brasil y la publicación continua de indicadores demográficos. Los constructos de ingreso, gastos y empleo son evaluados de forma continua, mientras que cada año se abordan otros módulos de interés. Otros temas investigados en la encuesta

están relacionados con las características de la vivienda, la migración de los individuos del hogar, el trabajo infantil, la fecundidad, la salud y la seguridad alimentaria, el uso de las tecnologías de la información y las comunicaciones (TIC), las transferencias de renta y el uso del tiempo (IBGE, 2018b).

La Encuesta de Presupuestos Familiares tiene como propósito obtener información general sobre domicilios, familias y personas, hábitos de consumo, gastos y recibos de las familias encuestadas, teniendo como unidad de recopilación los domicilios. Permite actualizar la canasta básica de consumo y obtener nuevas estructuras de ponderación para los índices de precios que componen el Sistema Nacional de Índices de Precios al Consumidor del IBGE y otras instituciones (IBGE, 2018a).

4. Chile

El Ministerio de Desarrollo Social y Familia lleva a cabo la Encuesta de Caracterización Socioeconómica Nacional de forma bianual. Su objetivo es conocer periódicamente la situación de los hogares y de la población, sobre todo de la que se encuentra en situación de pobreza, con relación a aspectos demográficos, de educación, salud, vivienda, trabajo e ingresos. De esta forma, la encuesta permite estimar la magnitud de la pobreza y la distribución del ingreso, detectar carencias y demandas de la población en las áreas señaladas, y evaluar las distintas brechas que separan a los diferentes segmentos sociales y ámbitos territoriales. Esta encuesta también permite medir la eficacia de los programas sociales que ha implementado el Gobierno para la toma de decisiones de política pública. Entre otros, la encuesta se compone del módulo de registro, que incluye información de identificación de los hogares; el módulo de educación, que indaga sobre la situación educacional de los miembros del hogar y la cobertura del sistema educativo; el módulo de trabajo, que permite conocer la evolución de la situación laboral y ocupacional para formular y evaluar políticas públicas; el módulo de ingresos, que permite investigar las condiciones de vida de los miembros del hogar, y el módulo de salud, donde se indaga sobre la cobertura de los programas públicos (Ministerio de Desarrollo Social y Familia, 2015).

La Encuesta de Presupuestos Familiares es una encuesta socioeconómica aplicada a hogares, cuyo propósito es recopilar información sobre los gastos en que estos incurren y los ingresos que perciben en un período de tiempo determinado. La información que recoge constituye la base para elaborar la canasta de bienes y servicios con que se calcula el IPC. También se utiliza para actualizar las líneas de pobreza extrema y de pobreza empleadas en las estadísticas oficiales de Chile (INE, 2018c).

5. Colombia

El Departamento Administrativo Nacional de Estadística (DANE) realiza la Gran Encuesta Integrada de Hogares de forma mensual. Esta encuesta tiene como objetivo general proporcionar información económica básica enfocada en las características de la fuerza de trabajo.

Además, se indaga sobre constructos sociales y económicos. Dentro de la esfera social, se pregunta por el acceso a la educación formal, las condiciones de calidad de vida, los ingresos y gastos, el trabajo infantil y aspectos de seguridad y convivencia ciudadana. En la esfera económica, se indaga sobre aspectos relacionados con la industria, el comercio, los servicios y el transporte. El instrumento de recopilación de la encuesta se divide en capítulos que abordan la información relacionada con la vivienda y el hogar, además de hacer un registro de las personas que conforman el hogar y su relación con el jefe de hogar, a fin de establecer una caracterización general de la población. Por lo demás, también se indaga sobre el acceso a la seguridad social en el ámbito de la salud y las características educativas de la población mayor de 3 años, y se clasifica a las personas mayores de 10 años en las categorías establecidas para la fuerza de trabajo (DANE, 2017).

La Encuesta Nacional de Presupuestos de los Hogares es una investigación dirigida a los hogares, en la cual se indaga en forma detallada sobre todos los ingresos de los miembros del hogar de 10 años y más (ingresos por trabajo, ingresos de capital, subsidios, transferencias e ingresos ocasionales, entre otros), así como todos los posibles gastos en que puede incurrir un hogar, captados con diferentes periodicidades (semanal, mensual, trimestral y anual). Entre sus objetivos específicos está el de obtener información para realizar actualizaciones del IPC, estimar líneas de indigencia y pobreza, y determinar la distribución del ingreso del hogar con respecto a características demográficas, educativas y económicas (DANE, 2018).

6. Costa Rica

La Encuesta Nacional de Hogares, llevada a cabo por el Instituto Nacional de Estadística y Censos (INEC) de forma anual, tiene como objetivo producir estimaciones del nivel de bienestar de la población. Se centra, en especial, en la conformación del ingreso de los hogares, su distribución y las características de los hogares y la población en situación de pobreza. El constructo principal y la motivación de esta encuesta se refieren a la pobreza multidimensional y la desigualdad, para lo cual se mide el ingreso promedio de los hogares por fuente y su distribución, su incidencia y gravedad, así como las brechas y perfiles. Esta encuesta permite obtener estas estimaciones a nivel de región y ha incluido algunos módulos especiales de victimización, gasto en los hogares y acceso a la salud (INEC, 2017).

La Encuesta Nacional de Ingresos y Gastos de los Hogares proporciona datos económicos de los hogares para conocer las diversas fuentes de ingresos que tienen y cómo distribuyen sus ingresos en la adquisición de los diferentes bienes y servicios. La encuesta suministra gran parte de la información necesaria para estimar la secuencia de cuentas del sector de los hogares, dentro del sistema de cuentas nacionales del país. También brinda los datos necesarios para actualizar la canasta de bienes y servicios que componen el IPC, entre otros estudios sobre la estructura de gastos de los hogares y la distribución del ingreso (INEC, 2018c).

7. Ecuador

El Instituto Nacional de Estadística y Censos cuenta con el Sistema Integrado de Encuestas a Hogares, con el cual se produce información sobre las características demográficas y económicas de los hogares y personas. Entre otras, el sistema realiza la Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU), con periodicidad mensual; la Encuesta Nacional de Ingresos y Gastos de los Hogares Urbanos y Rurales, con periodicidad quinquenal, y la Encuesta de Condiciones de Vida, con periodicidad cuatrienal. Estas encuestas cuentan con temas comunes, como condiciones de vivienda, caracterización de los hogares, educación de los miembros del hogar, actividad económica de los individuos, acceso a servicios públicos e ingreso (INEC, 2018e).

Por ejemplo, la Encuesta de Condiciones de Vida permite obtener indicadores sobre los niveles de vida y el bienestar de la población, en los que se relacionan varios factores, como la educación, la salud, la pobreza y la inequidad en la aplicación de las políticas públicas. En su edición de 2013-2014 incluyó temas nuevos, como hábitos, prácticas y uso del tiempo de los hogares, bienestar psicosocial, percepción del nivel de vida, capital social, seguridad ciudadana y retorno migratorio (INEC, 2018a).

8. El Salvador

La Encuesta de Hogares de Propósitos Múltiples es implementada cada año por la Dirección General de Estadística y Censos y su objetivo es generar información estadística relacionada con las condiciones económicas y demográficas de la población, con el fin de evaluar y orientar las políticas públicas del Gobierno para elevar el bienestar de la población. En esta encuesta se indaga sobre la información general de los miembros del hogar, su situación educacional (analfabetismo, escolaridad y asistencia), las características de la vivienda y la situación de ocupación de la población. Contiene, a su vez, un módulo de actividad del productor agropecuario, que recopila información acerca de la tenencia de la tierra, la superficie cultivada y la actividad agropecuaria del entrevistado. Por último, también pregunta acerca de variables de salud, dinámica de las remesas y gastos del hogar (DIGESTYC, 2018a).

La Encuesta de Ingresos y Gastos de los Hogares permite determinar la canasta de mercado para el desarrollo del IPC, el consumo privado en las cuentas nacionales y datos necesarios para los análisis de bienestar y pobreza. La encuesta también mide la educación, el empleo, las condiciones de la vivienda, la posesión de bienes durables, la construcción y los otros negocios y actividades agrícolas relacionados con el hogar (DIGESTYC, 2018b).

9. Guatemala

El Instituto Nacional de Estadística lleva a cabo de manera semestral la Encuesta Nacional de Empleo e Ingresos. Los objetivos de esta encuesta son dar seguimiento a un conjunto básico de variables e indicadores del mercado laboral y producir información que permita conocer el comportamiento y la evolución del empleo, el desempleo y las características, la composición, la estructura y el funcionamiento del mercado de trabajo. Además de investigar sobre aspectos generales del mercado laboral, en esta encuesta se indaga sobre las características de la informalidad, la ocupación y las formas de contratación. Tiene un componente de ingresos, e incluye también algunos aspectos demográficos y de educación de los hogares de Guatemala. Entre otros cambios recientes, se han incluido módulos de uso del tiempo y uso de las TIC (INE, 2018a).

La Encuesta Nacional de Condiciones de Vida tiene como principal objetivo conocer y evaluar las condiciones de vida de la población, así como determinar los niveles de pobreza existentes en Guatemala y los factores que los determinan. Con ese fin se caracteriza a la población pobre y no pobre del país y se brindan resultados a nivel nacional, regional y departamental (INE, 2016c).

10. Honduras

La Encuesta Permanente de Hogares de Propósitos Múltiples es una investigación semestral dirigida por el Instituto Nacional de Estadística de Honduras con el fin de recopilar información sobre las características generales de la población hondureña, en términos de vivienda, tasas de ocupación, desocupación y subempleo, ingreso de los hogares y acceso a las TIC (INE, 2018f).

La Encuesta de Condiciones de Vida de los Hogares es una investigación de carácter multipropósito que permite conocer los diferentes aspectos y dimensiones del bienestar de los hogares. Incluye, además de los ingresos y gastos de las unidades familiares, un conjunto de variables que describen los niveles de vida de los hogares. En ese sentido, esta publicación incorpora información sobre características de la vivienda, demografía, migración, educación, salud, antropometría, mercado laboral (género, personas con problemas laborales, trabajo infantil y juvenil), ingresos y gasto de los hogares, pobreza y otros temas de importancia (INE, 2004).

11. México

La Encuesta Nacional de Ocupación y Empleo (ENOE) es la principal fuente de datos sobre el mercado laboral de México. Proporciona información mensual y trimestral sobre diversos aspectos, como la fuerza de trabajo, ocupación, informalidad laboral, subocupación y desocupación. Es el proyecto estadístico continuo más extenso del país, que abarca cifras

a nivel nacional, en cuatro categorías de localidades, en cada una de las 32 entidades federativas, y en un total de 39 ciudades (INEGI, 2020).

La Encuesta Nacional de Ingresos y Gastos de los Hogares tiene como objetivo proporcionar un panorama estadístico del comportamiento de los ingresos y gastos de los hogares en cuanto a su monto, procedencia y distribución. Además, ofrece información sobre las características ocupacionales y sociodemográficas de los integrantes del hogar, así como las características de la infraestructura de la vivienda y el equipamiento del hogar (INEGI, 2016).

12. Nicaragua

Nicaragua lleva a cabo de forma trimestral la Encuesta de Medición de Nivel de Vida a través del Instituto Nacional de Información de Desarrollo. Su objetivo general es producir información continua sobre las características ocupacionales y demográficas de la población y la evolución de la pobreza. Esta encuesta indaga exhaustivamente acerca de las características demográficas de todos los miembros del hogar, además de la actividad económica y condición de los individuos en edad de trabajar y sus ingresos. Indaga también sobre el estado de la vivienda y sus características. Dispone de las variables necesarias para la producción de otras medidas de bienestar, como agregado de ingreso o necesidades básicas insatisfechas (INIDE, 2018).

13. Panamá

El Instituto Nacional de Estadística y Censo de Panamá realiza cada año la Encuesta de Propósitos Múltiples, cuyos objetivos están encaminados a la producción de estadísticas de empleo e ingresos y a la estimación de la situación del mercado laboral. Como la principal finalidad de la encuesta es la medición de los cambios del mercado laboral, se indaga sobre la actividad económica, ocupación, lugar de trabajo e ingresos. También se debe resaltar que la encuesta aborda, de manera no continua, algunos temas relacionados con el acceso a la tecnología, el interés y la colaboración con actividades de protección y conservación de los recursos naturales, la dinámica de turismo en los hogares, la identificación de recibo o envío de remesas y migración (desplazamiento interno y externo de la población durante un intervalo), así como el uso de servicios financieros (INEC, 2019).

La Encuesta de Ingresos y Gastos de los Hogares se realiza para actualizar la información de los ingresos de los hogares en Panamá y conocer cómo estos distribuyen los presupuestos destinados a diferentes bienes y servicios. La información recopilada tiene como principales objetivos obtener coeficientes de ponderación y canastas de consumo, que se utilizarán para el cálculo del IPC y la canasta básica familiar de alimentos (CBFA). Por otra parte, permite estructurar la demanda de los hogares en bienes de consumo privado (INEC, 2018b).

14. Paraguay

La Encuesta Permanente de Hogares, ejecutada cada año por la Dirección General de Estadística, Encuestas y Censos, tiene como objetivo la obtención de indicadores anuales sobre las principales características de las condiciones de vida de la población. Sus resultados se utilizan para elaborar las estimaciones de pobreza. Algunos de los constructos que se investigan en esta encuesta se relacionan con las características de las viviendas, la educación de los miembros del hogar, su salud, empleo e ingresos y su condición de ocupación, así como el acceso a programas sociales del Gobierno y remesas (DGEEC, 2018b).

La Encuesta de Ingresos y Gastos y de Condiciones de Vida tiene como principal objetivo actualizar la estructura de la Canasta Básica de Alimentos y la Canasta Total Familiar, cuyos valores definen las líneas de pobreza, además de caracterizar y analizar las condiciones de vida de la población del Paraguay. Esta información se recopila a través de cuestionarios que recogen datos acerca de temas de educación, salud, ingresos, actividades independientes no agropecuarias, perfiles de ingresos y de tipo productivo, entre otros (DGEEC, 2018a).

15. Perú

La Encuesta Nacional de Hogares sobre Condiciones de Vida y Pobreza es una investigación mensual que realiza el Instituto Nacional de Estadística e Informática, cuyo objetivo es la obtención de información estadística, social, demográfica y económica proveniente de los hogares para el cálculo de indicadores que se utilizarán en la medición de aspectos económicos y sociales. También sirve para conocer y explicar los determinantes o factores causales del comportamiento de dichos aspectos con miras al diseño, el monitoreo y la medición de resultados de las políticas públicas. Entre los módulos que aborda, se encuentran la caracterización de la vivienda y el hogar, la educación de los miembros del hogar, su estado de salud, la condición de actividad de empleo, los ingresos y gastos, el acceso a programas sociales y la participación ciudadana, así como la percepción de gobernabilidad y algunos temas relacionados con el fenómeno de la discriminación (INEI, 2016).

16. República Dominicana

La Oficina Nacional de Estadística cuenta con un Sistema Integrado de Encuestas de Hogares que agrupa, entre otras, la Encuesta Nacional de Hogares de Propósitos Múltiples, que se realiza con periodicidad anual, y la Encuesta Nacional de Fuerza de Trabajo, con periodicidad semestral. La primera es una encuesta orientada a recopilar periódicamente datos sobre diferentes temas sociales, económicos y ambientales. Por su parte, la segunda está orientada a obtener indicadores de la población en edad de trabajar y su ocupación. Algunos de los aspectos principales que evalúa el sistema de encuestas de hogares están relacionados con las condiciones y características de las viviendas y personas, así como

con la educación de los miembros del hogar, el acceso a las tecnologías de la información y algunas características de seguridad ciudadana y convivencia (ONE, 2018a).

Por ejemplo, la Encuesta Nacional de Ingresos y Gastos de los Hogares es un estudio estadístico que se realiza generalmente cada diez años con el fin de conocer la distribución del gasto en bienes y servicios de consumo de los hogares, así como los ingresos que estos obtienen de diferentes fuentes para financiar su consumo. Entre sus objetivos se encuentra el de obtener información para conocer el nivel y la estructura de los gastos de consumo de los hogares y su distribución en los rubros siguientes: i) alimentos y bebidas no alcohólicas; ii) bebidas alcohólicas, tabaco y estupefacientes; iii) prendas de vestir y calzado; iv) alojamiento, agua, electricidad, gas y otros combustibles; v) muebles, artículos para el hogar y para la conservación ordinaria del hogar; vi) salud, transporte, comunicaciones, recreación y cultura, y educación; vii) restaurantes y hoteles, y viii) bienes y servicios diversos (ONE, 2018b).

17. Uruguay

El Instituto Nacional de Estadística lleva a cabo mensualmente la Encuesta Continua de Hogares, mediante la que se obtienen los indicadores oficiales del mercado laboral (actividad, empleo y desempleo) y de ingresos de los hogares y las personas. Además, esta encuesta permite estimar cada año la proporción de hogares y personas que se encuentran por debajo de la línea de pobreza y de indigencia. El cuestionario de la encuesta indaga sobre las características de las viviendas y de los hogares, así como las características demográficas de los miembros del hogar y algunas variables de migración, acceso a la salud, educación, alimentación y uso de las TIC. Asimismo, se abordan de manera exhaustiva los constructos de actividad laboral, ingresos y egresos personales y del hogar (INE, 2016a).

La Encuesta de Gastos e Ingresos de los Hogares se realiza aproximadamente cada diez años y permite conocer la realidad económica y social del país. Con los resultados obtenidos, se elabora una canasta actualizada para el IPC y se determinan las líneas de indigencia y de pobreza nacionales. Se trata de una encuesta muy importante, ya que, a partir de los datos brindados, se puede obtener la información de base para elaborar indicadores que interesan a toda la sociedad, como los de inflación y pobreza (INE, 2016b).

18. República Bolivariana de Venezuela

La Encuesta de Hogares por Muestreo es una investigación que desarrolla el Instituto Nacional de Estadística con periodicidad semestral. Esta encuesta de propósitos múltiples brinda información sobre la estructura y evolución del mercado de trabajo, además de las características económicas y demográficas de la población. Algunos de las temas más relevantes de esta encuesta se centran en la actividad económica de los miembros del hogar y su situación de empleo, así como la caracterización de las viviendas y de los hogares y algunas variables educativas que dan origen a indicadores de analfabetismo. Por su parte,

la Encuesta Nacional de Presupuestos Familiares es una investigación por muestreo dirigida a los hogares, que tiene por objeto obtener información sobre sus ingresos, egresos, composición, características de la vivienda y otras variables económicas y sociales de sus miembros. Entre sus objetivos se encuentra el de conocer los cambios ocurridos en los patrones de consumo de los hogares y actualizar la canasta de bienes y servicios y las ponderaciones del IPC (INE, 2018e).

En los cuadros A1.1 y A1.2, se resumen las características de algunas de las encuestas recolectadas en la región.

■ Cuadro A1.1
Características de algunas encuestas repetidas en América Latina

País	Nombre de la encuesta	Tipo	Periodicidad	Rotación *(En porcentajes)*	Muestra de viviendas
Argentina	Encuesta Permanente de Hogares	Panel	Trimestral	50	25 000
Bolivia (Estado Plurinacional de)	Encuesta Continua de Hogares	Panel	Mensual	25	10 000
Brasil	Encuesta Nacional de Hogares	Repetida	Anual	-	11 5000
Brasil	Encuesta Nacional de Hogares Continua	Panel	Mensual	20	70 000
Chile	Encuesta de Caracterización Socioeconómica Nacional	Repetida	Bienal	-	84 000
Colombia	Gran Encuesta Integrada de Hogares	Repetida	Mensual	-	20 000
Costa Rica	Encuesta Nacional de Hogares	Repetida	Anual	-	13 000
Cuba	Encuesta Nacional de Ocupación	Panel	Anual	33	63 000
Ecuador	Encuesta de Empleo, Desempleo y Subempleo (ENEMDU)	Panel	Trimestral	50	16 000
El Salvador	Encuesta de Hogares de Propósitos Múltiples	Repetida	Anual	-	20 000
Guatemala	Encuesta Nacional de Empleo e Ingresos	Repetida	Semestral	-	6 000
Honduras	Encuesta Permanente de Hogares de Propósitos Múltiples	Repetida	Semestral	-	7 200
México	Encuesta Nacional de Ingresos y Gastos de los Hogares	Repetida	Bienal	-	20 000
Nicaragua	Encuesta de Medición de Nivel de Vida	Panel	Trimestral	20	7 500

País	Nombre de la encuesta	Tipo	Periodicidad	Rotación *(En porcentajes)*	Muestra de viviendas
Panamá	Encuesta de Propósitos Múltiples	Repetida	Anual	-	15 000
Paraguay	Encuesta Permanente de Hogares	Repetida	Anual	50	6 000
Perú	Encuesta Nacional de Hogares sobre Condiciones de Vida y Pobreza	Panel	Anual	20	32 000
República Dominicana	Encuesta Nacional de Hogares de Propósitos Múltiples	Repetida	Anual	-	34 000
República Dominicana	Encuesta Nacional de Fuerza de Trabajo	Panel	Semestral	25	10 000
Uruguay	Encuesta Continua de Hogares	Repetida	Mensual	-	53 000
Venezuela (República Bolivariana de)	Encuesta de Hogares por Muestreo	Repetida	Semestral	-	45 000

Fuente: Elaboración propia.

■ Cuadro A1.2
Características de algunas encuestas transversales en América Latina

País	Nombre de la encuesta	Año	Tamaño de la muestra
Argentina	Encuesta Nacional de Gastos de los Hogares	2017-2018	45 000
Bolivia (Estado Plurinacional de)	Encuesta de Hogares	2017	11 136
Brasil	Encuesta de Presupuestos familiares	2008-2009	53 154
Chile	VIII Encuesta de Presupuestos Familiares	2016-2017	15 239
Colombia	Encuesta Nacional de Presupuestos de los Hogares	2016-2017	87 201
Costa Rica	Encuesta Nacional de Ingresos y Gastos de los Hogares	2018-2019	9 828
Ecuador	Encuesta de Condiciones de Vida	2013-2014	29 052
El Salvador	Encuesta de Ingresos y Gastos de los Hogares	2005-2006	4 576
Guatemala	Encuesta Nacional de Condiciones de Vida	2014	11 536
Honduras	Encuesta de Condiciones de Vida de los Hogares	2004	8 155

País	Nombre de la encuesta	Año	Tamaño de la muestra
México	Encuesta Nacional de Ingresos y Gastos de los Hogares	2016	8 1515
Nicaragua	Encuesta de Medición de Nivel de Vida	2014	6 851
Panamá	Encuesta de Ingresos y Gastos de los Hogares	2007-2008	10 152
Paraguay	Encuesta de Ingresos y Gastos de Condiciones de Vida	2011-2012	6 000
Perú	Encuesta Nacional de Hogares sobre Condiciones de Vida y Pobreza	2017	36 996
República Dominicana	Encuesta Nacional de Ingresos y Gastos de los Hogares	2006-2007	8 358
Uruguay	Encuesta Nacional de Gastos e Ingresos de los Hogares	2016-2017	7 500
Venezuela (República Bolivariana de)	IV Encuesta Nacional de Presupuestos Familiares	2008-2009	45 768

Fuente: Elaboración propia.

Anexo 2

A. Programas informáticos para el análisis estadístico

En el diseño y análisis de la información proveniente de las encuestas de hogares, se debe contemplar el uso exhaustivo de las herramientas computacionales existentes. En este anexo se revisan con detalle las aproximaciones computacionales del *software* estadístico utilizado para realizar cada uno de los procesos que se necesitan para lograr el cometido de publicar cifras oficiales con altos niveles de precisión y confiabilidad, en particular, en lo referente a los siguientes procesos:

i) selección de muestras acorde al diseño de muestreo definido;

ii) generación de pesos de muestreo para cada individuo y hogar;

iii) modelación de la falta de respuesta e imputación estadística;

iv) calibración de los pesos de muestreo y ajustes por falta de respuesta;

v) estimación de los errores de muestreo para cada indicador de interés en los cuadros de producción estadística;

vi) análisis de las relaciones multivariantes entre las variables de la encuesta, y

vii) modelación de las estimaciones para la predicción del parámetro de interés en dominios pequeños.

En Naciones Unidas (2007, cap. XXII), se muestra la importancia de incluir la estructura del diseño de muestreo complejo en la inferencia que se realiza para la estimación de estadísticas oficiales a partir de encuestas de hogares. Mediante un ejemplo empírico, se advierte que, de no hacerlo, es posible que las estimaciones resultantes sean sesgadas y que los errores de muestreo se vean subestimados. A continuación se describen algunas de las características más importantes que incorporan los paquetes estadísticos computacionales en el manejo de datos que provienen de estructuras de muestreo complejas, como las de las encuestas de hogares. En Heeringa, West y Berglund (2010, apéndice A) puede encontrarse una revisión más exhaustiva y detallada, que incluye sintaxis y código computacional.

En general, estas herramientas computacionales están pensadas para hacer más eficiente el uso de las aproximaciones de varianza en muestras complejas, así como las técnicas de reproducción para obtener los estimativos de varianza determinados por el diseño de muestreo (Westat, 1997). Algunos de estos programas son de uso gratuito, aunque la mayoría corresponde a productos con licencia de pago. En general, estos productos, además de proporcionar estadísticas descriptivas (como medias, totales, proporciones, percentiles y razones), permiten ajustar modelos de regresión lineales y logísticos. Todas las estadísticas resultantes se basan en el diseño de muestreo de la encuesta.

1. *Software* R

R es un *software* gratuito cuyo uso es cada vez más frecuente en la investigación social, puesto que es muy probable encontrar programados en él los más recientes hallazgos científicos (R Core Team 2020). Al ser de uso libre, los investigadores pueden subir sus propias colecciones de funciones computacionales al repositorio oficial (CRAN) y ponerlas a disposición de la comunidad. El paquete *samplesize4surveys* (Gutiérrez, 2020) permite determinar el tamaño de muestra de individuos y hogares en encuestas de hogares repetidas, de tipo panel y con rotación. Los paquetes *sampling* (Tille y Matei, 2016) y *TeachingSampling* (Gutiérrez, 2015) permiten seleccionar muestras probabilísticas desde los marcos de muestreo con una gran variedad de diseños y algoritmos de muestreo. El paquete *survey* (Lumley, 2016), una vez que el diseño de muestreo se ha predefinido mediante la función *svydesign()*, permite analizar datos provenientes de encuestas de hogares y obtener estimaciones apropiadas de los errores estándar.

2. STATA

El entorno *svy* brinda un conjunto de herramientas para hacer una inferencia apropiada de las estadísticas oficiales provenientes de encuestas de hogares (STATA, 2013). El comando *svyset* permite especificar las variables que describen las características del diseño de muestreo de la encuesta, como los pesos de muestreo, los conglomerados y los estratos. El comando *svydescribe* proporciona tablas que describen los estratos y las unidades de muestra para una etapa determinada de la encuesta. Una vez cargadas las definiciones del diseño de muestreo, cualquier modelo puede ser estimado y sus estadísticos resultantes estarán basados en el diseño de muestreo de la encuesta. El entorno *svy* también permite la ejecución de algunos comandos predictivos.

3. SPSS

El módulo *complex samples* de SPSS (IBM, 2017) incorpora la selección de muestras complejas mediante la definición de un diseño de muestreo establecido por el usuario. Posteriormente, es necesario crear un plan de análisis mediante la asignación de variables de diseño, métodos de estimación y tamaños de las unidades de muestreo. Una vez definido el plan de muestreo, el módulo integra la posibilidad de estimar conteos, estadísticas descriptivas y celdas de tablas cruzadas. También es posible realizar estimaciones de razones y de coeficientes de regresión en modelos lineales, junto con las respectivas estadísticas de pruebas de hipótesis. Por último, el módulo permite estimar modelos no lineales, como regresiones logísticas, regresiones ordinales o la regresión de Cox.

4. SAS

Este *software* estadístico incluye un procedimiento para la selección de muestras probabilísticas llamado SURVEYSELECT, que integra los métodos de selección más comunes, como los de muestreo aleatorio simple, muestreo sistemático y muestreo con probabilidad proporcional al tamaño, además de algunas herramientas de afijación en los estratos. Para analizar los datos provenientes de muestras complejas, se han programado algunos procedimientos (SAS, 2010), como el de SURVEYMEANS, que permite estimar totales, medias, proporciones y percentiles, junto con sus respectivos errores estándar, límites de intervalos de confianza y pruebas de hipótesis. Por su parte, SURVEYFREQ sirve para estimar las estadísticas descriptivas (como totales y proporciones) de interés en tablas de una y dos vías, brindar las estimaciones del error de muestreo y realizar un análisis de la bondad del ajuste de las estimaciones, la independencia, los riesgos y las razones de momios (*odds ratios*). A su vez, SURVEYREG y SURVEYLOGISTIC permiten ajustar modelos de regresión lineal y logísticas, respectivamente. Con estos procedimientos, se estiman los coeficientes de regresión, con sus respectivos errores, y se adjunta un análisis exhaustivo de las propiedades de los modelos. Por último, SURVEYPHREG sirve para ajustar modelos de riesgos mediante la utilización de técnicas de máxima seudoverosimilitud.

Las encuestas de hogares brindan información valiosa para la toma de decisiones en materia de políticas públicas y el seguimiento de indicadores sociales, económicos, educativos y de salud, entre otros. Para garantizar resultados representativos, es necesario utilizar las herramientas apropiadas y considerar el diseño de muestreo en el estudio de estas operaciones estadísticas.

Esta publicación constituye una guía completa para la planificación y el análisis de las encuestas de hogares en América Latina. Abarca desde la definición de la población objetivo y la selección de la muestra hasta los métodos de análisis de datos más adecuados, incluidos la estimación de parámetros, la construcción de intervalos de confianza y las pruebas de hipótesis. También se aborda el análisis de la calidad de los datos y los posibles sesgos que pueden producirse en contextos de crisis, como durante la pandemia de enfermedad por coronavirus (COVID-19).

El documento, dirigido a estadísticos e investigadores de las oficinas nacionales de estadística, incluye ejemplos prácticos y aplicaciones reales que contribuyen a la comprensión de los conceptos teóricos, y representa una valiosa herramienta para la planificación, el diseño y el correcto análisis de este tipo de operaciones estadísticas en la región.

La colección *Metodologías de la CEPAL* se orienta a la divulgación de los fundamentos conceptuales, las especificaciones técnicas de elaboración y las aplicaciones de los instrumentos cuantitativos y cualitativos producidos y utilizados en el ámbito de la CEPAL. Su propósito central es contribuir mediante más y mejores instrumentos al diseño de políticas públicas basadas en datos empíricos que generen un desarrollo sostenible con igualdad.

CEPAL

Comisión Económica para América Latina y el Caribe (CEPAL)
Economic Commission for Latin America and the Caribbean (ECLAC)
www.cepal.org

ISBN 978-92-1-122127-5